Spring 5-15-2016

Automatically Characterizing Product and Process Incentives in Collective Intelligence

Allen Brockhurst Lavoie
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science
Department of Computer Science and Engineering

Dissertation Examination Committee:
Sanmay Das, Chair
Yoram Bachrach
Roman Garnett
Roch Guérin
Yulia Nevskaya

Automatically Characterizing Product and Process Incentives in Collective Intelligence

by

Allen Brockhurst Lavoie

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2016
Saint Louis, Missouri

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my adviser, Professor Sanmay Das, for his guidance and mentorship. The other members of my dissertation committee, Doctor Yoram Bachrach and Professors Roman Garnett, Roch Guérin, and Yulia Nevskaya, have provided many useful suggestions and a lot of great advice, some of which had to do with this document.

Special thanks to Professor Malik Magdon-Ismail at RPI for many helpful discussions. Colleagues and friends at three universities (RPI, Virginia Tech, and Washington University) conspired to make the production of this dissertation both possible and enjoyable. I am grateful to friends and family for moral support along the way.

Much of this dissertation was written while I was supported with grants from the U.S. National Science Foundation, and would not have been possible otherwise. Specifically a CAREER award to Sanmay Das (IIS-0952918/1303350/1414452) and IIS-1124827.

Chapters 3, 4, 5, and 6 of this dissertation are based on previously published, co-authored work. Chapter 3 is based on work with Sanmay Das and Malik Magdon-Ismail published at CIKM 2013 [34]. Chapters 4 and 5 are based on work with Sanmay Das published at ICML 2014 [30] and AAMAS 2014 [31] respectively. Chapter 6 is based on work with a number of co-authors: Mithun Chakraborty, Sanmay Das, Malik Magdon-Ismail, and Yonatan Naamad. This work was published at AAAI 2013 [15]. Chapter 7 is based on unpublished work with Sanmay Das.

Allen Brockhurst Lavoie

Washington University in Saint Louis

May 2016

ABSTRACT OF THE DISSERTATION

Automatically Characterizing Product and Process Incentives

in Collective Intelligence

by

Allen Brockhurst Lavoie

Doctor of Philosophy in Computer Science

Washington University in St. Louis, May 2016

Professor Sanmay Das, Chair

Social media facilitate interaction and information dissemination among an unprecedented number of participants. Why do users contribute, and why do they contribute to a specific venue? Does the information they receive cover all relevant points of view, or is it biased? The substantial and increasing importance of online communication makes these questions more pressing, but also puts answers within reach of automated methods. I investigate scalable algorithms for understanding two classes of incentives which arise in collective intelligence processes. *Product incentives* exist when contributors have a stake in the information delivered to other users. I investigate product-relevant user behavior changes, algorithms for characterizing the topics and points of view presented in peer-produced content, and the results of a field experiment with a prediction market framework having associated product incentives. *Process incentives* exist when users find contributing to be intrinsically rewarding. Algorithms which are aware of process incentives predict the effect of feedback on where users will make contributions, and can learn about the structure of a conversation by observing when users choose to participate in it. Learning from large-scale social interactions allows us to monitor the quality of information and the health of venues, but also provides fresh insights into human behavior.

Chapter 1

Introduction

Online information aggregation venues grow their content out of the small contributions of a large number of individual participants. Even when participants are not compensated in any traditional sense, they collectively create products with significant economic and societal value. While there is often no explicit compensation, participants may receive value personally from the act of contributing, for example through social feedback, or from the later effects of a contribution, such as the comportment of a publicly displayed product with strongly held opinions. The goal of this dissertation is to identify such quantitative "microeconomic" principles underlying participation in collective intelligence.

Collective intelligence, the sometimes surprising ability of groups to aggregate and act on knowledge possessed by disparate individuals, is far from a new concept or phenomenon, but has become much more pervasive as a result of participatory mass communication. Prime examples include collectively edited knowledge bases ("wiki"s), with Wikipedia for example now containing millions of articles; rating websites on which users collectively describe and curate items (Yelp, Amazon, Netflix, etc.); prediction markets for quantitative forecasts (e.g. Betfair); question-and-answer websites such as Stackoverflow; link aggregators which create transient rankings of content (e.g. Reddit). All of these contain social components, from the detailed free-form discussions which pervade Wikipedia to the formal price-based interactions of prediction market participants,

leading to implicit incentive structures shared with each other and with social media more broadly, in addition to the unique incentives stemming from artifacts produced by each process.

Research into the principles behind collective intelligence has the potential for immediate practical impact. First, I describe and evaluate tools for determining the reliability of collectively produced content when strong external incentives (for example economic or political) influence participants. Once we can determine when the products of collective intelligence processes are representative of a broad consensus, we can begin to describe how venues should be organized to ensure that even contentious topics produce useful artifacts. In addition, learning why those who contribute do so brings us closer to designing venues from first principles to enable the efficient production and long-term maintenance of socially useful products, and to more effective monitoring of the health of existing collaborations.

This dissertation will explore two themes, product incentives and process incentives, in depth. *Product incentives* arise when a participant's utility depends on the final product of a collective intelligence process, for example on the point of view that is emphasized in a Wikipedia article. *Process incentives* encompass the utility that participants receive during the creation of an artifact, for example through social interaction or personal recognition. In exploring these incentives, this dissertation will employ models of large-scale collective intelligence processes, building on these models to produce empirically-grounded simulations of interactions between participants. Data will be sourced from the complete edit history of Wikipedia, from blogs, a prediction market experiment with human subjects, and from many Reddit communities. Models will be validated based on their ability to predict held-out data, and by a variety of side data depending on the venue being considered.

1.1 Product incentives

These incentives arise when users have a direct stake in the outcome of a collective intelligence process. Sometimes this is straightforward, such as a restaurant owner or product manufacturer hoping for a good rating (or a bad rating for their competitors), due to direct financial interest. Other times the incentives are less blatant, such as users with strong political views hoping to craft a collectively edited article to fit their preferred narrative. Some collective intelligence processes can create incentives to change the things they are trying to collect information about, such as a prediction market when participants help to control outcomes (e.g. small-town voting). I focus on three issues in this area. First, determining which points of view exist at scale by harnessing content disputes among users. Second, finding potential abuses of trust by elected leadership in collective intelligence processes, using Wikipedia as a case study. Finally, I analyze the results of a human-subject prediction market experiment where there are strong incentives for manipulation, finding that in some situations effective aggregation of information is nonetheless possible.

1.2 Process incentives

Tied to the creation of artifacts rather than the artifacts themselves, these incentives are either independent of the content produced by a collective intelligence process (as in the case of social interaction), or are tied to products only through community judgments of merit (as in the case of awards). This means that while process incentives do influence collective intelligence, they do not themselves create a conflict of interest between individuals and projects as a whole. As there is a well-developed literature on the effect of awards on user behavior (see for example Anderson et al. [3], Kriplean et al. [70]), I focus instead on the effects of social interactions. First, on what effect social interaction has on the choices users make in splitting their efforts between different

communities, both within and across different collective intelligence venues. This is important for reasoning about user retention and the long-term health of projects, and also touches on the problem of "bootstrapping" a new venue. Second, and tying in with point of view and manipulation issues brought about by product incentives, I explore the effects of deliberation on user retention, simultaneously modeling the opinions of users and their propensity to participate as a function of the opinions expressed by others.

1.3 Processes studied

This dissertation will examine a sample of socially important collective intelligence processes. While a comprehensive examination of all venues is not feasible (including wikis and blogs, there are at least hundreds of millions), I will use a diverse sample of venues to identify general principles.

1.3.1 Wikis

A wiki is a collaboratively edited collection of documents, often with a specific topical scope. Most publicly accessible wikis (as opposed to those used internally by organizations) are characterized by anonymous or pseudonymous editing by anyone at all, often with a significant fraction of content produced by a core set of users (see for example Kittur et al. [66]). Wikipedia is by far the highest profile wiki project, its users having created encyclopedias of varying size in many languages, with English being the largest by far.[1] With prominence comes an incentive for those with an agenda to use a wiki to push their message, i.e. product incentives. On the other hand, process incentives are increasingly relevant as Wikipedia in particular struggles with user retention

[1] https://en.wikipedia.org/wiki/List_of_Wikipedias

[50]. Non-public wikis face similar issues, with for example the U.S. intelligence community's Intellipedia struggling with participation [62].

1.3.2 Link aggregators

These social media have become popular venues for both news and culture. Common characteristics include submissions of links or text snippets by users, ranking of submitted content by both time (recent posts are heavily favored) and community voting, and comments on submissions which themselves are ranked by votes. Users often accrue points based on the voting scores of their submissions and comments. Popular examples include Reddit and Delicious, and formerly Digg. Reddit facilitates the creation of individual link aggregators, called subreddits, each with their own content and moderation but sharing user accounts. Hundreds of thousands of subreddits have been created to date.[2] This makes Reddit a useful data source for studying the flow of users between communities as a function of social process incentives. The visibility granted to highly ranked submissions can create strong product incentives, and additionally comment sections can be attractive places to shape discourse.

1.3.3 Blogs

While their content is quite varied, many blogs do function as collective intelligence venues. One notable example is the Polymath Project, a theorem-proving collaboration organized primarily through academic blogs [26]. Others feature more open-ended discussions, but nonetheless exhibit elements of consensus, either through voting or argumentation. Blog comments, as with those on link aggregators, create both product incentives and social process incentives.

[2] Via the unofficial `http://redditmetrics.com/`

1.3.4 Prediction markets

Inspired by the ability of financial markets to aggregate information about future prices and profits, prediction markets provide a similar profit motive for participants to reveal information about a much wider class of events. For example, someone with oceanfront property might create a market for securities which pay off $1 (likely inflation-adjusted) if sea levels rise more than 15cm in the next 20 years. The pricing of this security, if the market is efficient, then reflects a consensus probability estimate. Viewing the security's price over time as the product of this collective intelligence process, having better information than the market creates a process incentive to trade the security, and in doing so a participant moves the market price. Product incentives can arise when prediction markets are used to make decisions and participants want to influence those decisions, and process incentives may cause undesirable side-effects when participants influence the event a security is premised on (e.g. a market for the outcome of a small-town referendum creating an incentive for a group to collude on a surprising outcome). Automated market making algorithms can make prediction markets more liquid, increasing process incentives for traders who have new information.

Chapter 2

Background and Related Work

Part of emerging research in computational social science [71], work studying collective intelligence combines ideas from the social sciences—primarily economics, psychology, and sociology—with artificial intelligence, human computer interaction, and network science. This thesis has a focus on artificial intelligence, using techniques from machine learning and statistics to measure and simulate incentives in collective intelligence from a multi-agent perspective. This chapter covers related literature at a high level. Individual chapters introducing specific models cover application-specific background.

2.1 Wikipedia

Wikipedia attracts significant amounts of attention as one of the most socially useful collective intelligence processes. Of particular interest is controversy and the ability of crowds to coordinate on a single product [67]. One part of this coordination takes the form of process incentives for socializing new editors, including the feedback editors receive [141] and recognition they are given [70]. Wikipedia's governance, the enforcement of its rules and norms, falls to a relatively small group of users elected as administrators [12]. Even outside this group, Wikipedia is *de facto* quite reliant on the efforts of a relatively small number of power users [103, 66]. This centralization

and entrenchment has negative consequences for process incentives, making recruiting new users more difficult [50]. This dissertation is differentiated by its more formal approaches to process incentives (making quantitative forecasts of resulting user behavior), and its focus on quantifying product incentives which are otherwise embedded in complex social structures and text, difficult to access with traditional automated methods.

2.2 Social incentives

There is also significant interest in process incentives for retaining and motivating users in collective intelligence more broadly. The popular question and answer website Stack Overflow uses a system of badges, and user behavior seems to follow an economic model where users have utility for obtaining these badges and for maintaining a default mode of behavior, optimizing their total discounted utility by balancing these two goals [3]. Badges raise interesting mechanism design questions, including the choice of absolute or relative award criteria [36]. Social media exhibit a variety of interesting phenomena relating to their voting systems. When deciding whether to up-vote or down-vote a contribution, that contribution's current score is a strong consideration [92]. Users who receive negative feedback can be more motivated to continue contributing than are users who receive no feedback at all [20]. When they receive positive feedback, users are more likely to continue participating [138, 11], and to do so more frequently [49]. Social learning means that initial experiences can shape the long-term habits of users [13]. Given these results, it is interesting to note that votes in social media are often good reflections of community judgments [126], rather than the arbitrary results of a chaotic process. Users tend to adapt to the linguistic norms of communities they join, but reduce their adaptation over time as the norms themselves continue changing, and this second stage predicts a user leaving [27]. In contrast, this thesis will focus on simultaneously modeling the effects of multiple types of social feedback, and on making quantitative probabilistic predictions of user behavior based on these effects. These proper probabilistic

models allow us to perform multi-agent simulations of the interaction effects of many participants giving and receiving feedback.

Related to social incentives, the perceived popularity of opinions can influence willingness to express them, studied in the social science literature under the name "spiral of silence" [98]. The question of whether this effect applies online has attracted theoretical interest [115]. The effect can be observed experimentally in online forums [136], and seems to work similarly in online and offline settings [77]. This thesis proposes a generative model of both activity and long-term survival which is in part based on this effect.

2.3 Topic modeling

Bayesian topic modeling has a long history of useful applications in various domains, typically based on the popular Latent Dirichlet Allocation [7]. This type of modeling will be useful in this thesis when defining product incentives on Wikipedia, where users may have their own ideas about the proper way for the encyclopedia to present information *on a given topic*. Topic modeling will also be a useful component in a generative model of process incentives when users are choosing how to allocate their attention between communities, using in this case a non-parametric Bayesian model called the Hierarchical Dirichlet Process [128]. Extensions of topic modeling explain the word choices of authors [111], the word choices of authors when sending to a specific recipient [86] (part of a line of research making use of the corporate emails released after Enron's collapse), how word choices change over time [131], and how words vary between venues [78] or groups of users within venues [105]. Particularly relevant is a line of work modeling sentiment in text, including the discovery of "facets" or viewpoints [106], sentiment in movie reviews [76], and contrasting opinions from venues offering different perspectives [39] (focusing on major Chinese, U.S., and Indian news coverage of world events). Rather than modeling text, this thesis focuses on modeling

users, aggregating noisy signals of disagreement along with information about where users choose to allocate their time. This text-free approach enables modeling at unprecedented scale, including the entire edit history of English Wikipedia, and summarizes the otherwise labyrinthine product incentives which exist on the encyclopedia.

2.4 Network science

Networks, both social and otherwise, are an important tool for understanding collective intelligence. This thesis will analyze user behavior changes related to product incentives using metrics on a weighted network of artifacts. More generally, product incentives often involve implicit groups with mutually exclusive goals, and networks are a popular way to find and study such groups. While much has been said about methods for analyzing and discovering patterns in explicit social networks (e.g. Scott [116]), many collective intelligence processes take place without an explicit network among participants. Some interesting work focuses on using interactions to define and study dynamic implicit social networks [25], often defining a network which spans participants and the artifacts they create [65, 21]. These networks can help to identify the social roles of participants, and to explain the propagation of social roles to new users [132]. Relevant to this thesis's focus on point of view as a product incentive in wikis is community detection based on the use of topical themes [8, 113]. This thesis is differentiated by a novel generative model of adversarial actions as they relate to the points of view users attempt to promote within latent topical groupings.

2.5 Sentiment analysis

There is interest in opinion and sentiment online, especially in automatically recognizing its expression for use in product or brand advertising research. This thesis touches on sentiment analysis

with a text-free model of discussions which includes opinion as one factor determining participation. Mining text to determine sentiment automatically is a well-studied and active research area [104], but is in general tangential to the goals of this thesis. Of particular interest, however, is the integration of structural features of a conversation with traditional text-based sentiment analysis. Quotations, replies, and named references can help to identify the social roles of forum users [40]. Reply graphs provide useful guidance for aggregating the text-based opinion scores of comments [124, 125]. Augmenting sentiment analysis with reply structure can provide information about a user's role in a community [95]. Using only text, but with similar potential applications, is recent work on identifying contrasting sentences in document collections [140]. This thesis, on the other hand, will model the entrance, exit, and activity level of participants as a function of conversation and opinion dynamics, inferring opinion without any use of text.

2.6 Theoretical models of opinion

Opinion formation and aggregation has received a good deal of theoretical interest. Most relevant is work on game theoretic formulations of product incentives in idealized rating aggregators and wikis, finding that self-interested agents can collectively push venues close to a median of their beliefs [93]. Considering social networks, interest is in polarization of belief under different conditions, for example when participants can change their beliefs [84], in the presence of "contrarians" [85], or depending on the size of a user's communication network [121]. One focus of this thesis is on applying generative models in this spirit to observations from collective intelligence processes in order to understand product incentives and their effects on information aggregation.

2.7 Prediction markets

Prediction markets are interesting from both a theoretical perspective, creating complex incentives involving multiple agents and interactions with decision makers, but also from a behavioral perspective, with interesting questions surrounding their real-world efficacy and the interactions of humans with both process and product incentives created by deployed markets. One of the more visible controversies involving incentives in prediction markets concerned the Policy Analysis Markets, a DARPA-involved project which would have created markets for important world events. One criticism was that such a market could finance an event (e.g. an assassination or terrorist attack, the perpetrators being knowledgeable of market outcomes), despite financial markets creating the same problematic incentives with far more liquidity [81]. When one outcome is strongly preferred (e.g. a successful product launch in the context of corporate prediction markets), well-placed subsidies in a market can eliminate incentives for undesirable actions [118]. However, this type of subsidy is not appropriate when no outcome is preferred (e.g. a market for a referendum or opinion poll), one of the issues that this thesis will explore further. Another set of issues arise when prediction markets are used as input in a decision-making process [102], unless the preferences of traders over final decisions are known [19, 9]. On the empirical side, prediction markets have been studied in corporate [24] and academic [100] settings, and are interesting as aggregators of political information [5]. There has been some empirical work on manipulation, finding in controlled experiments that prices remain accurate even when some participants have an incentive to distort them [55]. Theoretically, manipulators can even make prices more accurate [54]. In contrast to previous work, the human-subject field experiment described in this thesis took place over a much longer period of time (several months) with many more participants and a goal of real-world prediction. The field experiment also compares market microstructures, with a goal similar to that of smaller controlled experiments by Brahma et al. [10].

Chapter 3

Product Incentives Among Wikipedia Administrators

3.1 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 [71], 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 [45], is another major target for manipulation.

Collective intelligence processes need some structure in order to effectively aggregate information. On Wikipedia, this structure takes the form of detailed rules of writing style and behavior, with an emphasis on consensus (and indeed, the rules themselves attempt to encode consensus-supported

norms). However, these rules are not self-enforcing. Rule enforcement ultimately falls to a group of users elected (via an open voting process) as administrators, with wide-ranging technical abilities to prevent editing on certain pages or by certain users. This centralized control makes an attractive target for individuals and groups with strong product incentives. For example, an advocacy group made plans to install an administrator by vote manipulation and the (pre-election) illusion of neutrality, with a goal of influencing coverage of the Israeli-Palestinian conflict on Wikipedia [14]. An administrator was recently banned from the project after accusations that he improperly promoted his own business school and denigrated competitors [119]. The common theme is misrepresentation: users attempting to conceal the influence that product incentives have on them in order to obtain a position of power, thereafter using the power to advance their interests.

The goal of this chapter will be to quantify these types of behavior changes, especially among newly-elected Wikipedia administrators (although the techniques apply to many other collective intelligence processes). Such an analysis is confounded by the natural and expected changes in behavior by Wikipedia users who become administrators, focusing more on rule enforcement and therefore dealing with more controversial content. To disambiguate the expected from the more unusual behavior changes that might indicate editing motivated by previously suppressed product incentives, I consider measures of both controversy and topical concentration. While users are expected to deal more with controversy after being elected as an administrator, doing so primarily on a specific topic which they had shown no interest in before is surprising.

Viewing a user's contributions to a collective intelligence process at a given time as a graph with "knowledge artifacts" (on Wikipedia, articles) as nodes and the similarities between these artifacts as edges, I use several metrics to quantify behavior. A version of the clustering coefficient for weighted graphs indicates how closely related contributions are. Associating with each artifact a measure of controversy, it is possible to quantify the overall level of controversy among the artifacts that a user chooses to contribute to. Both of these measures, clustering and controversy, are

interesting but not sufficient for examining administrator behavior on Wikipedia. We expect controversy to increase post-election (it does), and clustering alone is not suspicious. This motivates a measure of clustered controversy, quantifying the extent to which a user is focusing on a coherent controversial topic.

To validate these scores for use on Wikipedia, I evaluate their performance on a dataset of users who were blocked from editing the encyclopedia for manipulative behavior ("edit warring" to disruptively promote a specific version of an article, using multiple accounts to create the illusion of consensus, or violating rules regarding the biographies of living persons). Both overall controversy and clustered controversy are good indicators of manipulative behavior, effectively differentiating these users from similar users who were never blocked.

My analysis focuses on behavior changes in three populations. First, users who successfully become administrators, measuring behavior changes between their before-election behavior and their after-election behavior. Second, unsuccessful administrators provide a useful comparison, measuring behavior changes before and after their unsuccessful attempts to become administrators. Finally, I compare these two populations of users who do attempt to become administrators with similar users who never do, measuring behavior changes before and after a randomly selected edit. As expected, users who actually become administrators significantly increase their controversy scores. A subset of administrators, however, also significantly increase in measures of *clustered controversy*, which we would not expect in general. Users who never attempt to become administrators, and those who make unsuccessful attempts, tend to decrease their overall topical clustering and clustered controversy over time, corresponding perhaps to a broadening of interests. The subset of administrators with surprisingly large increases in clustered controversy tend to *increase* their topical clustering after being elected, focusing more tightly in an absolute sense and doing so on a more controversial topic. Even without the possibility of suppressed product incentives, these behavior changes undermine the transparency of Wikipedia's administrator selection process.

Can we identify administrators who go on to change their behavior, potentially misrepresenting themselves? The popular vote during an administrator's election is not helpful: Many who go on to change their behavior significantly receive near unanimous support. However, more sophisticated methods, taking the voting history of participants into account, show that information about who will change behavior is revealed in the voting process; see Figure 3.6. Alternative election processes might harness this collective intelligence more effectively, although more research is needed to find a mechanism that is both politically desirable and effective.

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 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.

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.

3.1.2 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-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 [43] (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.

3.1.3 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 [45] and very popular on the Web [122]. The dynamics of how articles become high quality and how information grows in collective media like Wikipedia have also garnered some attention [133, 33]. 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. [88].

Our work 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 [12]. 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.

3.2 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.

3.2.1 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:

$$\text{CScore} = \sum_k p_k c_k \tag{3.1}$$

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.

3.2.2 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 [63] to quantify how coherent an administrator's controversial edits are.

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) [7]—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.[3] 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.

[3] 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.

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 [63]. 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:

$$\iota(k) = c_k p_k \tag{3.2}$$

Then the clustering score of a page is:

$$\mathrm{clust}(k) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \iota(i)\iota(j)w_{ki}w_{kj}w_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} \iota(i)\iota(j)w_{ki}w_{kj}} \tag{3.3}$$

$\mathrm{clust}(k)$ 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[4]—that is, when a page is connected to a coherent and controversial topic which the user edits frequently. Note that $\mathrm{clust}(k)$ 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:

$$\mathrm{CCScore} = \sum_{k=1}^{N} \iota(k)\mathrm{clust}(k) \tag{3.4}$$

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

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

[4]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.

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 (3.4)—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 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 [75].

3.2.3 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.

3.2.4 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 can compare these measures to simply using the percentage of support votes a candidate receives during an RfA.

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. See Section 3.2.5 for details.

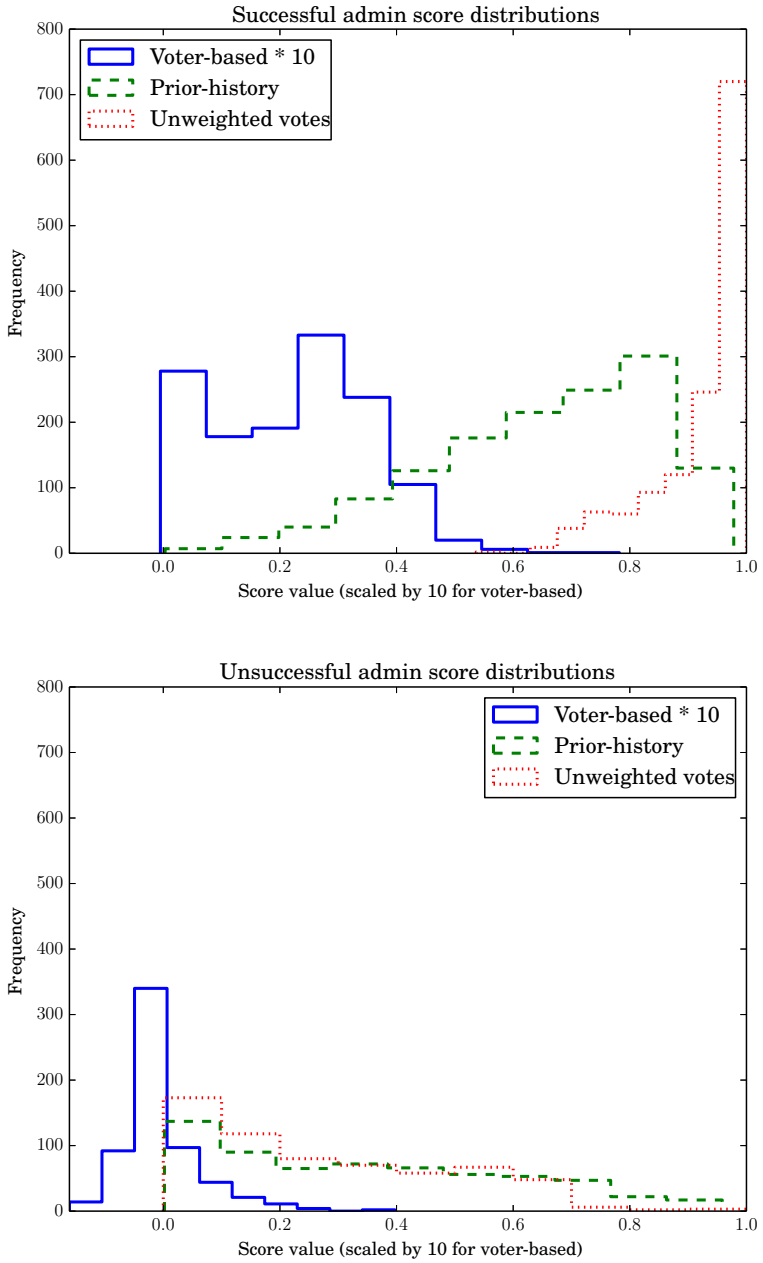

Figure 3.1: Distributions of the three RfA or pre-RfA scores for admin candidates. Successful candidates are shown on the left, unsuccessful on the right. The weighted-voter score is multiplied by a factor of 10 to show detail.

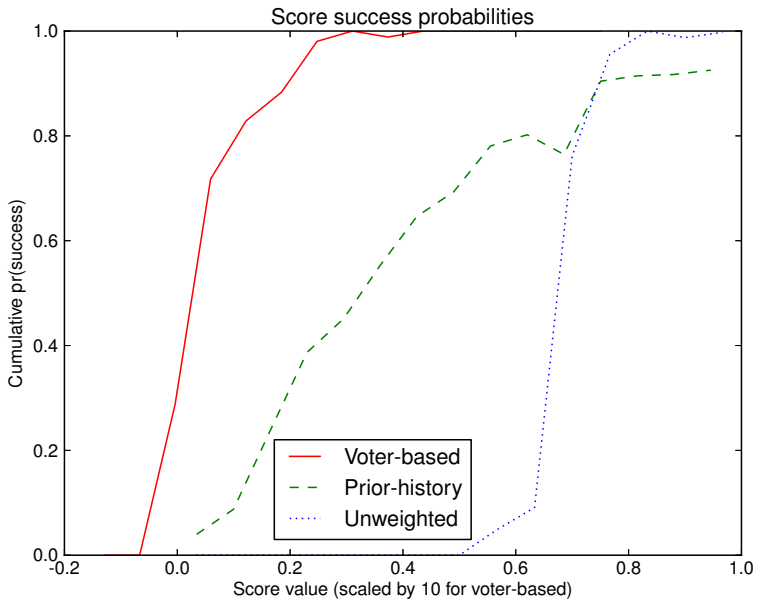

Figure 3.2: Probability of a successful RfA as a function of the weighted-voter score, the prior-history score, and the unweighted vote fraction. The weighted-voter score is multiplied by 10 to show detail.

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 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 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. See Section 3.2.6 for details.

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 3.1 shows the distributions of all three scores for successful and unsuccessful candidates. Figure 3.2 compares the distribution of success probabilities associated with the weighted-voter score with that of the prior-history 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-history 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.

3.2.5 Activity-based RfA success prediction

Table 3.1 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 an area under the ROC curve (AUC) of 0.82.

3.2.6 Weighted-voter RfA scores

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

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 3.1: 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 (*** $p = 0.001$, ** $p = 0.01$, * $p = 0.05$).

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.

3.3 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 3.1, 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.

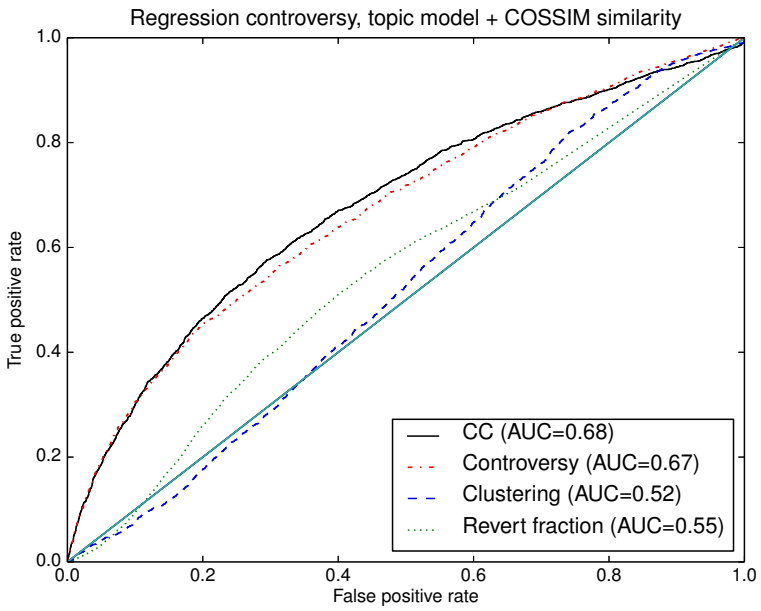

Figure 3.3: ROC curve for CC, Controversy, and Clustering Scores when differentiating between blocked and not-blocked users, based on 180 days of data. As a baseline, the fraction of a user's edits during this period which were reverts is also included. The CC and Controversy Scores effectively discriminate 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.

3.3.1 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 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 3.3 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.

3.3.2 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 concentrating on a single topic is useful here purely for validation. 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 [42] 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. At least with regards to U.S. politics, 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.

3.3.3 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 3.2, 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 3.1, an administrator who was banned from editing the encyclopedia for promoting the Indian Institute of Planning and Management (IIPM) and denigrating competitors [119]. Figure 3.4 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.

3.3.4 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 3.5 gives an overview of the (human-labeled) focus areas of administrators with very 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.

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%	Electronic voice phenomenon	0.4%
Vanessa Fox	2.1%	Web 2.0	0.4%
Scientology	1.6%	SS Edmund Fitzgerald	0.3%
Gamma-ray burst	0.8%	Collapse of the World Trade Center	0.2%
Search engine submission	0.8%		
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 3.2: 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.

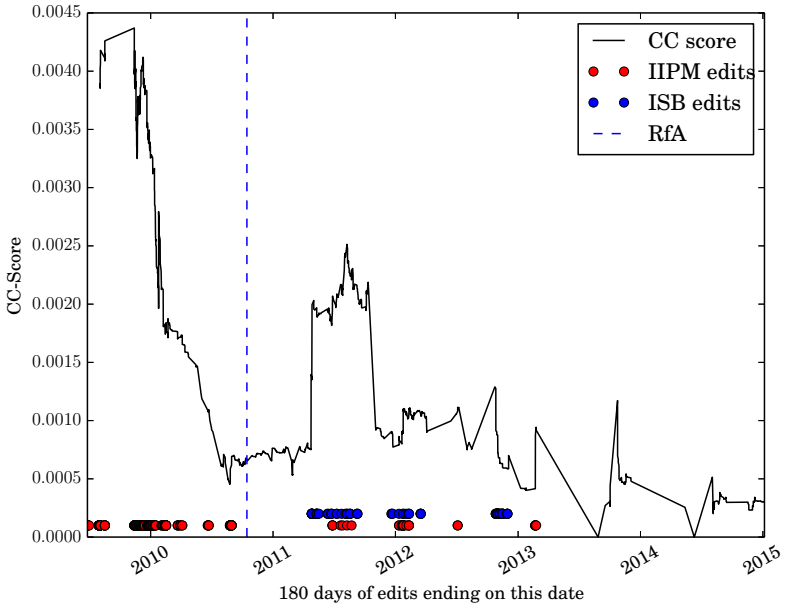

Figure 3.4: A plot of the CC-Score of one Wikipedia user, Wifione, over time. After joining Wikipedia in 2009, Wifione began heavily editing articles related to the Indian Institute of Planning and Management (IIPM), but significantly reduced this type of editing before making a successful request for 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) lead English Wikipedia's Arbitration Committee to ban Wifione in February 2015.

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.

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[5] 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. Group 3 above (users who have never participated in an RfA) serve as a baseline for what constitutes typical behavior changes over time.

[5]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.

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^{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.

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 3.6, 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.

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-history 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 3.7). 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 3.7).

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 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-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 model is (Figure 3.6, 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 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-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 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.

3.4 Alternative similarity and controversy

The CC score relies on two main components: page controversy and page similarity. We have defined the score in Section 3.2 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.

3.4.1 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.

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 [79] 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 [79]). 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:

$$v_f^{(i)} = \left(1 + \ln r_f^{(i)}\right) \ln \frac{D}{d_f} \tag{3.5}$$

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.

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.

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

$$\text{COSSIM}_{ij} = \frac{v^{(i)} \cdot v^{(j)}}{||v^{(i)}||||v^{(j)}||} \tag{3.6}$$

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:

$$\text{JSD}_{ij} = \frac{D_{\text{KL}}(u^{(i)}||M_{ij}) + D_{\text{KL}}(u^{(j)}||M_{ij})}{2} \tag{3.7}$$

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

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

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 3.8 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.

3.4.3 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 3.2, 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 lead 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 3.2, 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).

3.4.4 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

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.

Finding manipulative users

Controversy and the CC-Score, as defined in Section 3.2, differentiate users who are blocked for manipulative behavior from those who are never blocked (see Section 3.3.1). 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 3.9 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.

Administrator behavior changes

We find in Section 3.3, using the regression-based controversy and topic modeling with cosine similarity described in Section 3.2, 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 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 3.10 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 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 3.11 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.

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 3.11 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 3.11 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 3.11 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.

3.5 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 [135], 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 [51]). 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" [91]. 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 [110].

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. Here, 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.

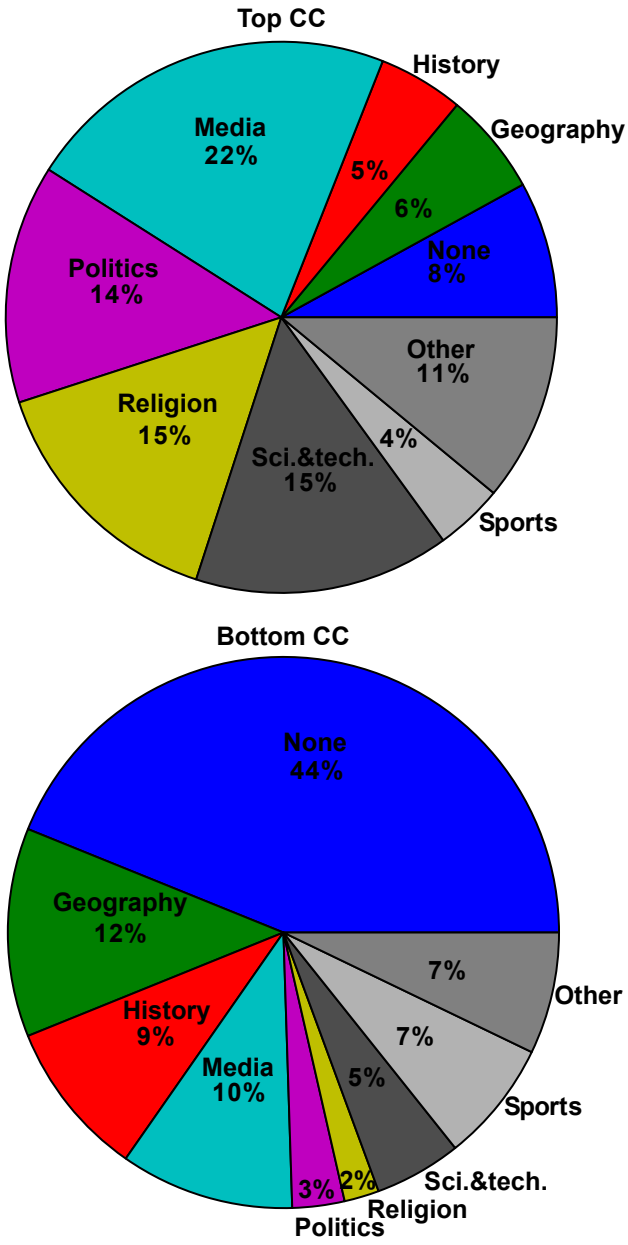

Figure 3.5: Blind human evaluation of the general category of edits (if any) for administrators directly after their RfA. The 100 highest and 100 lowest scoring administrators according to a previous version of the CC-Score are shown (using metadata page comparisons and a slightly different controversy measure). The charts illustrate the behaviors which the CC-Score selects for in administrators: controversial edits on a focused topic.

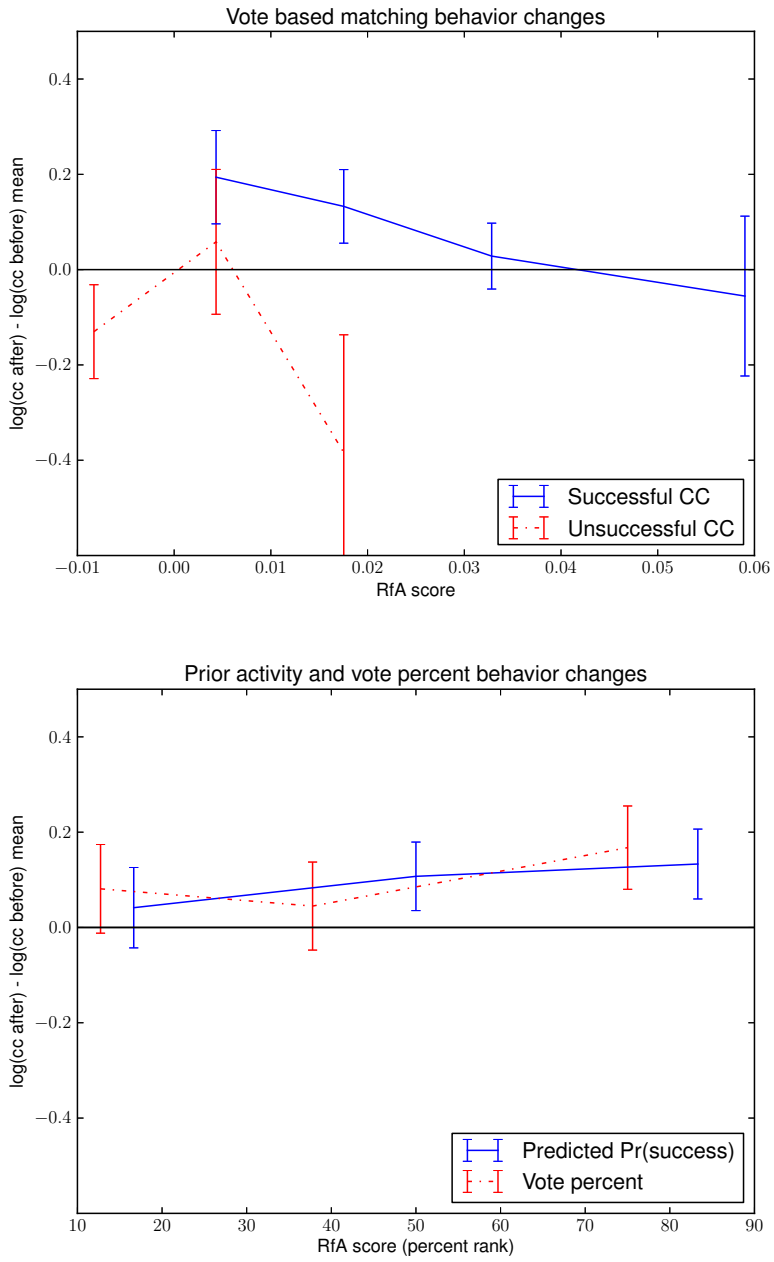

Figure 3.6: The vote-based score of a Request for Adminship (RfA) (left) discriminates between administrators who change their behavior significantly and those who do not; a small group with low vote-based scores skew the average for successful administrators. The activity-based score (right) does not filter out administrators who change their behavior; if anything, higher scoring administrators are more likely to change their behavior. Raw vote percentage performs similarly.

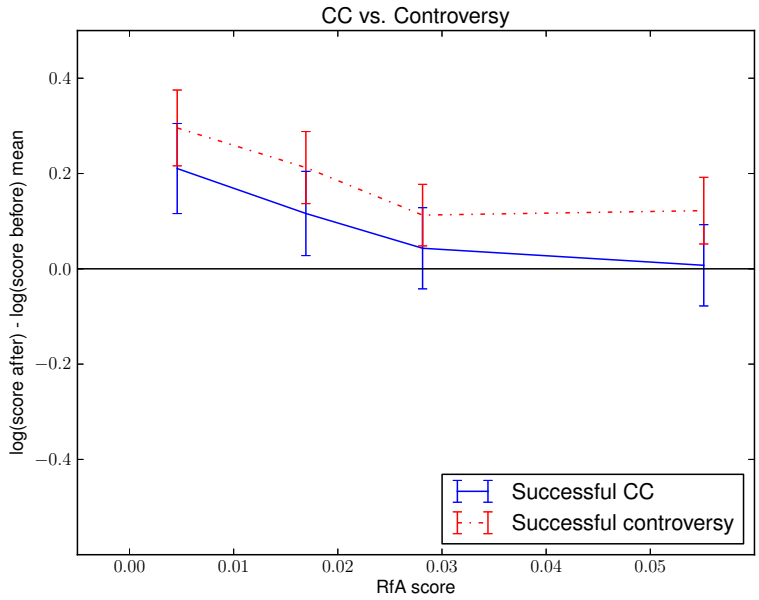

Figure 3.7: Behavior changes upon becoming an administrator, measured by the CC- and C-Scores for 180 days of edits before and after a successful Request for Adminship (RfA). The x axis is the vote-based RfA score, with a higher score implying a stronger consensus. The Controversy Score increases on average for both low and high scoring administrators, while only low scoring administrators increase their CC-Score.

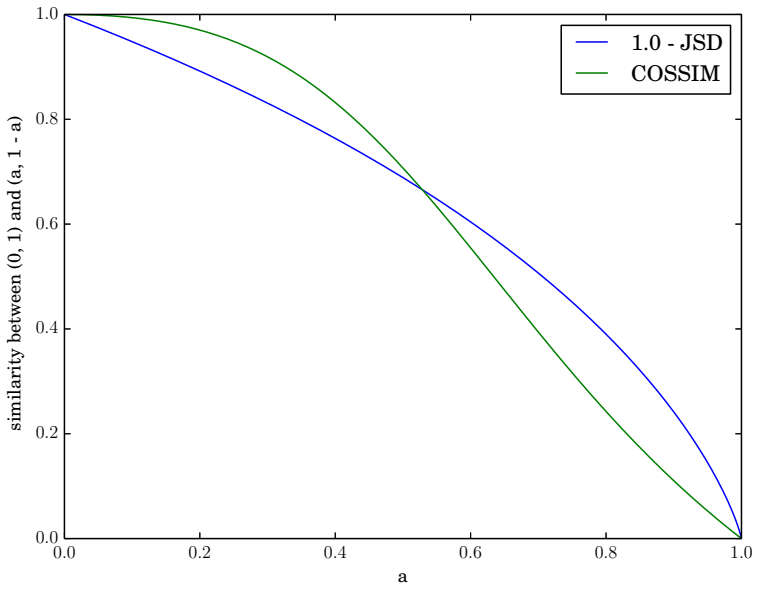

Figure 3.8: Cossine similarity and Jensen-Shannon divergence when computing the similarity between a fixed discrete distribution $(0, 1)$ and a family of distributions $(a, 1 - a)$ parameterized by a. Cossine similarity often takes more extreme values: closer to one than JSD when distributions are similar and closer to zero than JSD when distributions are dissimilar.

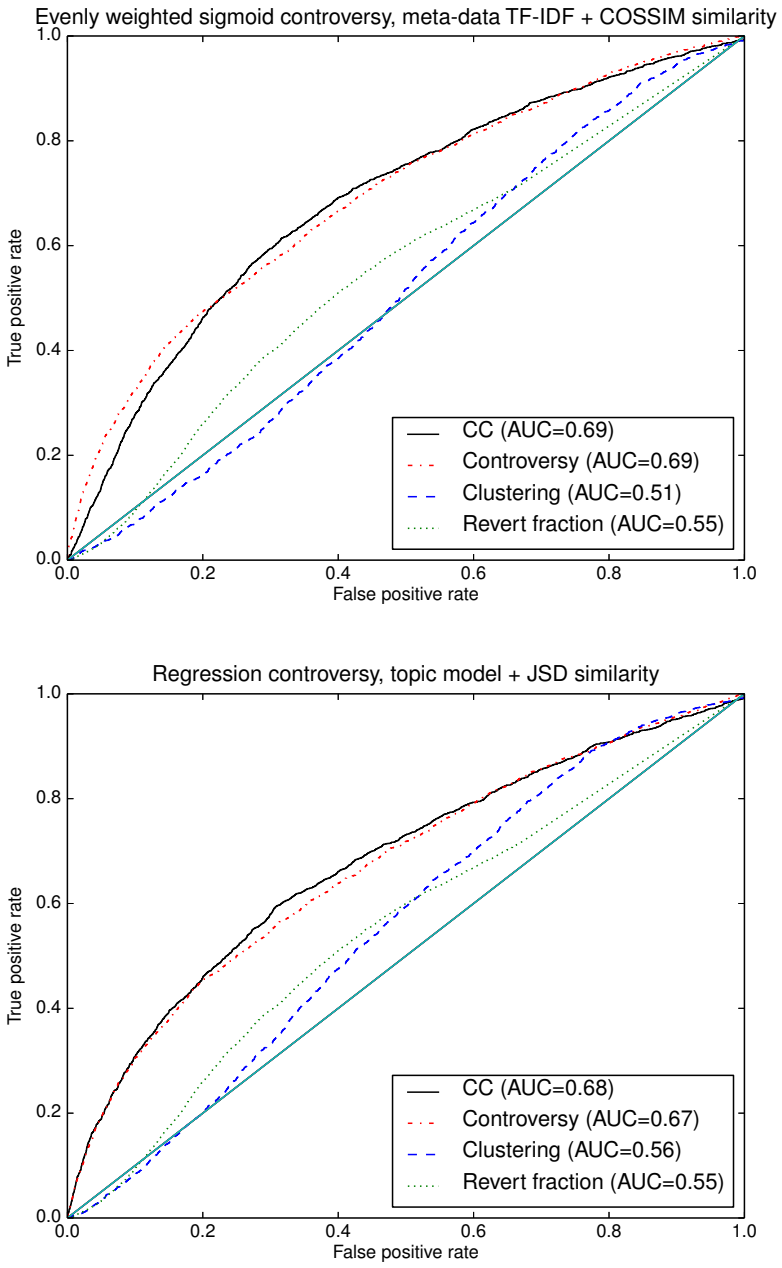

Figure 3.9: Orthogonal measures of controversy and similarity nonetheless produce consistent results when differentiating manipulative blocked users from users who were never blocked.

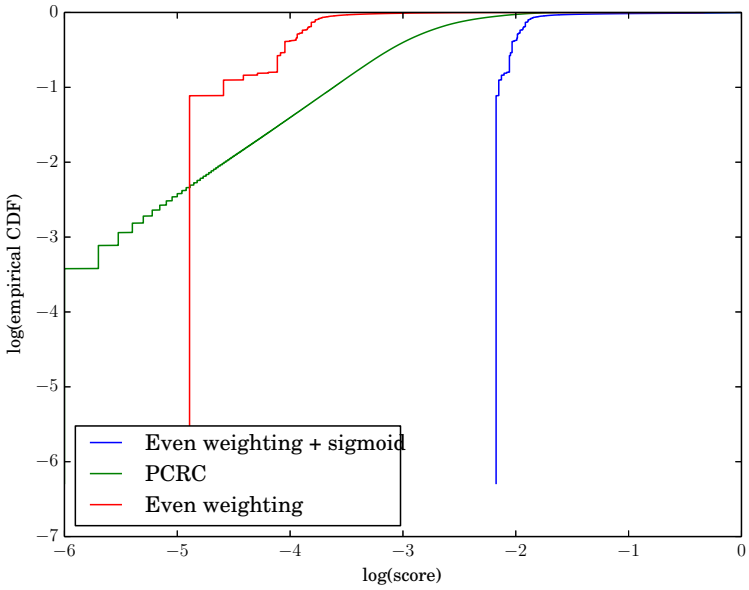

Figure 3.10: The CDFs of two different controversy scoring methodologies, an even weighting of four features and a regression-based measure, along with a sigmoid transform of the even weighting. The linearly-transformed scores assign high values to relatively few pages, with most pages getting very low controversy scores.

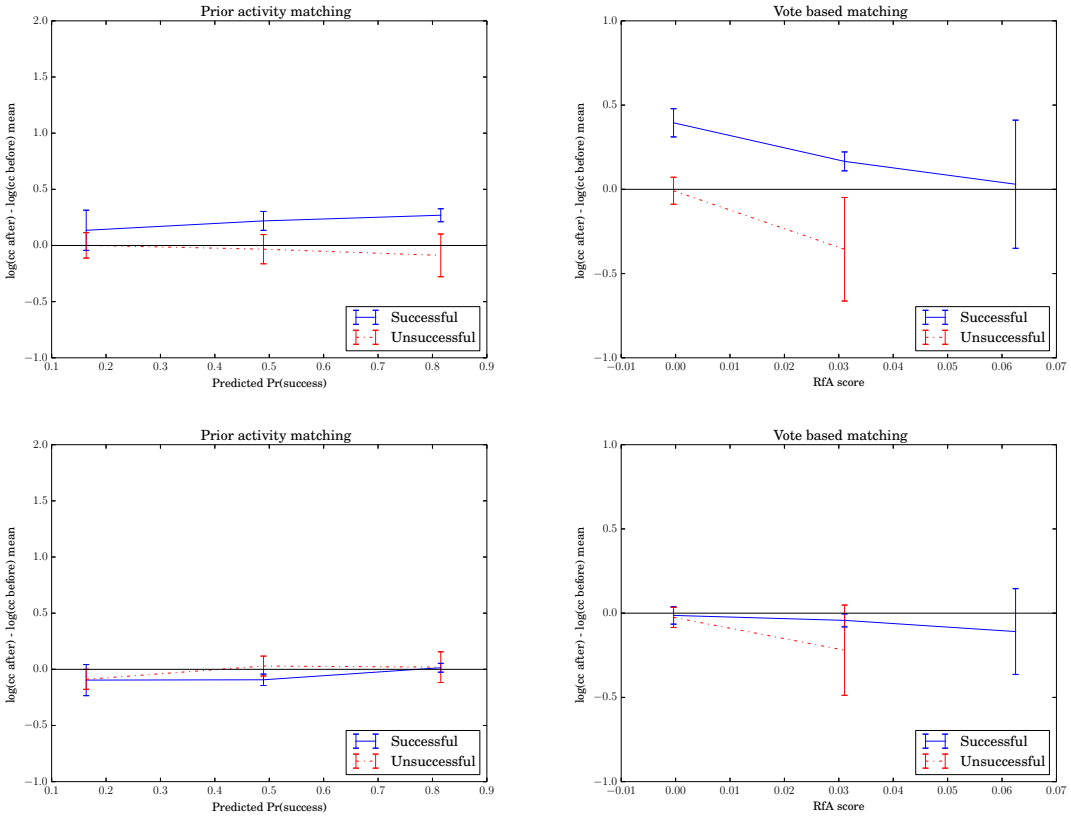

Figure 3.11: Evenly weighted controversy results (top, shown here with topic model page features and cossine similarity) echo our earlier administrator behavior change findings using the regression-based controversy score. The evenly-weighted score with a sigmoid transformation (bottom), which marks 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.

Chapter 4

Large-Scale Modeling of Product Incentives

4.1 Introduction

The Web has enabled an unprecedented democratization of information. We increasingly rely on decentralized sources such as blogs, social news, and wikis to stay informed. While this transition has many benefits, it also creates opportunities for individuals and groups to shape available information and thereby influence public perception. As a result, reliability has been a primary concern since the early days of knowledge-sharing platforms such as Wikipedia.

How can we define product incentives in a large collective intelligence project like Wikipedia? We have intuitive notions about incentives, such as those relating to political or religious conflicts, but manually defining conflicts across millions of articles is not feasible. Moreover, a useful model of product incentives would also estimate the positions that users are taking on an issue, and the proportion of users contributing a viewpoint to a given article. This chapter will develop an automated methodology for using disagreements between users to categorize disputes and their participants.

The model simultaneously identifies two properties of a dispute. First, the topical scope is defined by the articles a dispute is likely to come up on. This part of the generative Bayesian model is based on topic modeling, with users first choosing a dispute, then choosing an article related to

that dispute to contribute to. Second, having chosen a dispute and an article, the position within the dispute that a user has taken determines their reaction to other users. For example, two users interested in a dispute about anthropogenic global warming might both end up editing an article about coal. Under the generative model, their positions on the dispute would determine whether the user who arrives second disagrees with the first user by removing her contribution, a likely outcome if those positions are in opposition (note that disputes are not necessarily binary in the model). Inference involves learning distributions over articles for each dispute and simultaneously learning about the positions that each user takes within each dispute and the relationships between those positions.

I have trained this model on the full revision history of English Wikipedia, with 341 million total contributions by 32 million users across 10 million articles. The observed variables in the model are the sequence of users who have edited each article, and a (potentially quite noisy) binary signal for each edit which indicates whether the user disagreed with the previous user in the sequence. I use reverts, where one user removes the contribution of another, to provide this noisy signal of disagreement on Wikipedia. Inference on this scale is enabled by a highly parallel inference procedure based on collapsed Gibbs sampling. I compare the full model fit using this inference procedure with, among other baselines, a two-step procedure which first fits a standard topic model to users' edits (the articles selected taking on the role of word tokens in standard topic modeling, with each user being a document), finding that there is a significant benefit to using the noisy signals of disagreement to help define the topical scope of disputes.

In order to compare the performance of the approach with other possibilities on a prediction task relevant to product incentives, I have developed several datasets of users with known antagonistic relationships. The datasets are based on pairs of users where one has reported the other for violating Wikipedia's prohibitions on edit warring (specifically, for reverting edits on a single page more than three times within a 24 hour period). Understandably, these reports typically come from users who are opposed to the content promoted by the reported user, and so the pairs are very often going to

be composed of users on opposite sides of a specific debate. I compare the model and several other approaches based on their performance in differentiating these antagonistic pairs of users from pairs of users who are unlikely to have significant disagreements, and on several related prediction tasks. The model shows significant benefits compared to other approaches on these tasks.

Having validated its performance, I explore several applications of the model in triaging product incentives. For example, the model allows us to find pages on controversial topics where almost all of the edits have been from a single point of view, pages which may benefit from the attention of editors with opposing views. Moreover, the model learns the views of editors as part of the inference process, enabling us to find editors with another perspective. Tagging the point of view of each revision also enables us to summarize changes to a page over time. Figure 8.4, for example, shows edits to the Wikipedia page on same-sex marriage, with an interesting shift in the popularity of one point of view early in Wikipedia's history. While many interesting applications exist on Wikipedia itself, transferring models of this type to less structured discourse in social media seems to be a promising direction. It is worth noting that disputes, even on Wikipedia, need not relate purely to product incentives. Removing content that one finds disagreeable or arguing with other users may also be a process incentive in some cases, and I explore this concept further with a form of survival modeling in Chapter 7.

While there has been some work on systematically identifying manipulation and bias in purely quantitative collective intelligence venues like rating systems [89], work on venues with free-form information (like Wikipedia) has been less thorough and more anecdotal. In order to move towards a principled, quantitative methodology for evaluating bias and trustworthiness in such venues, we introduce a generative model of users' points of view. Our method is based on Latent Dirichlet Allocation (LDA), a popular topic model [7]. By observing the pages a user chooses to edit and her interactions with other users on those pages, we are able to infer both topics and points of view simultaneously. While we focus on Wikipedia as a case study, our technique is general and can be used to study issues of bias and trust in many collective intelligence processes.

Two issues make identifying point of view a particularly hard—and, in my view, understudied—problem: subjectivity and scale. Point of view is notoriously difficult to quantify, even for humans considering single documents. There is little concrete information on which to base inferences about bias, and none of it is structured. The problem is exacerbated by an adversarial effect, whereby authors attempt to appear objective [34]. While the size and scale of the web are what makes sites like Wikipedia influential, this scale makes identifying point of view more difficult and human supervision problematic. The problem calls out for an efficient and accurate automated approach.

We pose the novel problem of identifying points of view in a large collective intelligence environment (e.g. Wikipedia). While point of view is fundamentally entangled with human communication, we posit that it also has observable characteristics in collective intelligence which do not involve natural language. In this chapter, we show how to use data on user interactions to quantitatively study bias and point of view both in aggregate and on the level of individual users and documents.

Contributions We introduce a generative model of topics and points of view based on user interactions, and give an efficient inference algorithm based on Gibbs sampling. We study the performance of the model in inferring topics and points of view from synthetic data, finding that both are recoverable from the model's observed variables. Using a complete Wikipedia dataset, we perform model selection and validate the approach by finding pairs of users with antagonistic relationships. Our approach, jointly modeling topics and points of view, significantly outperforms a social roles model, a model that fixes topics before considering points of view, and a non-Bayesian graph-based approach. Finally, we study the topics and points of view inferred from the entire history of English Wikipedia. This allows us to visualize shifts in point of view, revealing the evolution of the encyclopedia and its users over time, and provide insights into how Wikipedia functions. The model also provides a wealth of information about the process by which individual pages were

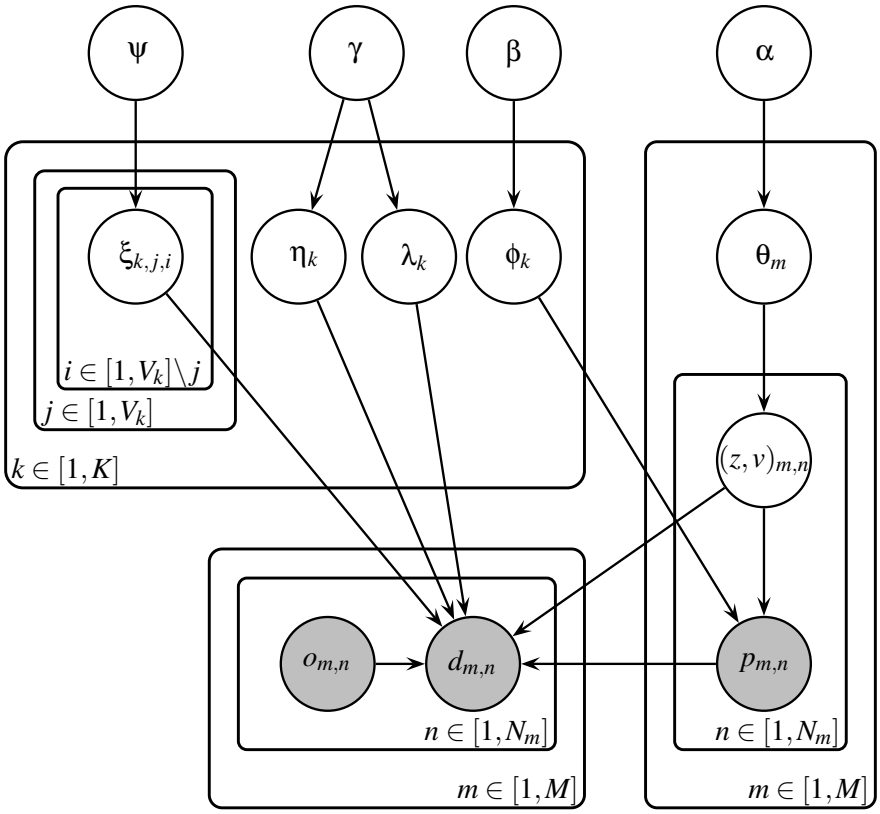

Figure 4.1: Graphical depiction of the model using plate notation, where plates (boxes) represent repeated variables. Nodes in the first row are beta or Dirichlet distributions, nodes in the second row are categorical or Bernoulli. The shaded nodes are observed.

created: as one example, we find pages on otherwise controversial topics which nonetheless are dominated by a single point of view.

4.1.1 Related work

Discovery of community structure in collective intelligence is a well studied problem. Kittur et al. (2007) use reverts to create a small-scale clustering of users while studying conflict and coordination on Wikipedia. Bogdanov et al. (2010) find communities on Wikipedia by comparing users based on multi-topic agreements and disagreements and then performing clustering, using LDA to inform the topic of text added or removed from a page. Pathak et al. (2008) propose a generative

	Beta, Bernoulli distributions
ψ, $\xi_{k,j,i}$	Revert probability between POVs $i \neq j$
γ, η_k, λ_k	Different topic (η_k), same POV (λ_k) revert
	Dirichlet, categorical distributions
β, ϕ_k	Pages associated with each topic
α, θ_m	Topic and POV preferences for a user
$(z, v)_{m,n}$	Topic and point of view (categorical)
	Observed variables
$p_{m,n}$	The page an edit is on (categorical)
$o_{m,n}$	Parent edit (ordering; not modeled)
$d_{m,n}$	Whether edit disagrees with parent
	Counts, parameters
M	Number of users
N_m	Number of edits by user m.
K	Number of topics
V_k	Number of points of view for topic k
P	Number of pages

Table 4.1: Notation: random variables, distributions, and model parameters.

model for community extraction, modeling communication content. Sachan et al. (2011) find communities based on user-to-user links in a social network using a generative model, also considering interaction types. We are the first to exploit an explicit synergy between user interests (topics) and interactions (governed by points of view within a topic) in community discovery.

Another line of literature uses topical structure to model human communications. Rosen-Zvi et al. (2004) model documents from multiple authors, where authors have a distribution over topics, and words in a document are generated by first choosing an author, then a topic from that author's distribution, and finally choosing a word from that topic. McCallum et al. (2007) model directed messages, where both the sender and recipient are significant. We show that points of view within a topic can be used to effectively model the sentiment of communications.

Several models have been proposed to extract points of view or related concepts from natural language. For example, Paul and Girju (2010) study a multi-faceted topic model which can be used to find viewpoints in text. Lin and He (2009) model words in movie reviews as having both sentiment and topical components. Fang et al. (2012) find contrasting opinions in collections of text written from different perspectives: press releases from U.S. politicians and articles from major Chinese, Indian, and U.S. news sources. Their model exploits the often disparate language used when framing an issue from different perspectives (e.g. "life" and "choice" when debating abortion). However, this line of work (1) relies on clean sources of ideologically-relevant material, and (2) operates on a much smaller scale than the large collective intelligence processes we target. We address the former issue in part by modeling point of view in a user-centric way, which provides the statistical strength necessary to differentiate between hundreds of points of view across different topics (where previous work assumes two or three total). The latter issue is addressed by using higher level observations: focusing on what users do rather than the specifics of what they say.

To summarize, we address three main challenges which prevent previous work on point of view modeling from applying to collective intelligence. (1) Very low signal to noise ratio: most Wikipedia revisions are not related to point of view. (2) Data on a completely different scale, with approximately four orders of magnitude more documents and words than previous work. (3) A lack of overarching ideologies across topics, which previous work takes advantage of in more specialized corpora. We present a novel model of point of view in collective intelligence based on user interactions, along with an efficient inference algorithm, which together address these challenges.

4.2 Model

We begin with a topic model much like Latent Dirichlet Allocation (LDA) [7]. LDA is often used to model a set of documents, where documents are assumed to be generated by first selecting a

topic from a document-specific distribution, and then selecting a single word from a topic-specific distribution, repeating this process for every word in each document. Instead of words we have pages, and instead of documents we have users: each user makes a collection of edits to different pages. As in LDA, each user (document) has a preference (probability distribution) over topics, and each topic has a distribution over pages (words). Each time a user makes an edit, they draw a topic from their personal topic distribution, and they then select a page to edit from that topic's distribution over pages. So far the only observable is the set of pages that each user chooses to edit, and edits are exchangeable within that set. This is exactly equivalent to LDA, where only the words in each document are observed.

Now suppose that each topic has a small number of points of view (POVs) a user editing on it can take, and each edit has one of these POVs associated with it in addition to its topic. Then each user has a distribution of preferences over (topic, POV) pairs rather than over topics alone. The chosen page for each edit still depends only on the topic of that edit, but POV determines interactions with other users on that page. We model this as a two-stage process: first every user chooses a topic, POV, and page for each of their edits, then interactions between users take place. These interactions are simple: for each edit, the user decides to disagree with its parent edit or not. Each edit's parent is observed but is not modeled: edits, except those at the beginning of a page, have an externally specified parent denoted $o_{m,n}$ in the model. There are three cases: (1) the parent edit is on a different topic; (2) the parent edit is on the same topic and the same POV; (3) the parent edit is on the same topic and a different POV. Disagreements in cases (1) and (2) might be mundane and unrelated to POV: style or formatting, for example. The probability of a disagreement in these cases is determined by the topic of the latter editor, and we would expect such disagreements to be unlikely. In case (3), however, the probability of a disagreement is determined by the relationship between the two POVs involved: some might have a very antagonistic relationship, others less so, but we would expect more disagreements here than in cases (1) and (2) as a fraction of the opportunities for disagreement. We refer to disagreements in case (3) as POV disagreements, and

for topic $k = 1 \to K$ **do**
  $\eta_k, \lambda_k \sim \text{Beta}(\gamma)$ // Non-POV revert probabilities
  $\phi_k \sim \text{Dirichlet}(\beta)$ // Page distribution
  for POV $j = 1 \to V_k$ **do**
    for POV $i = 1 \to V_k$ excluding j **do**
      $\xi_{k,j,i} \sim \text{Beta}(\psi)$ // Probability POV j reverts i
for user $m = 1 \to M$ **do**
  $\theta_m \sim \text{Dirichlet}(\alpha)$ // Topic and POV preferences
  for edit $n = 1 \to N_m$ **do**
    $(z, v)_{m,n} \sim \text{Categorical}(\theta_m)$ // Edit's topic and POV
    $p_{m,n} \sim \text{Categorical}(\phi_z)$ // Edit's page
for user $m = 1 \to M$ **do**
  for edit $n = 1 \to N_m$ **do**
    if $z_{o_{m,n}} = z_{m,n}$ **then**
      if $v_{o_{m,n}} = v_{m,n}$ **then**
        $d_{m,n} \sim \text{Bernoulli}(\lambda_k)$ // Disagree? Same POV
      else
        $d_{m,n} \sim \text{Bernoulli}(\xi_{k,v_{m,n},\text{POV}(o_{m,n})})$ // Diff. POV
    else
      $d_{m,n} \sim \text{Bernoulli}(\eta_k)$ // Different topic

Figure 4.2: Generative model pseudo-code. $x \sim D(y)$ indicates a random variable x drawn from distribution D parameterized by y.

to others as non-POV disagreements. The model does *not* interpret every Wikipedia revert as being a disagreement relevant to POV. Although we expect more reverts in case (3) as a fraction of opportunities for disagreement than we do in (1) or (2), the preponderance of cases (1) and (2) means that we would expect them to produce a significant fraction of all reverts.

Figure 4.1 depicts this model graphically, and Figure 4.1 summarizes the notation. Figure 4.2 provides a rigorous pseudo-code description of the model.

Simplifying assumptions We use symmetric Dirichlet distributions, with a single parameter (α and β for pages and (topic, POV) pairs, respectively). We set $\beta = 0.1$ and $\alpha = 5/(VK)$ where K is the number of topics and V is the number of POVs per topic: we expect users to be focused on a small number of (topic, POV) pairs (and as we add more topics or POVs, we expect them

to become increasingly specialized). These choices are similar to those of Griffiths and Steyvers (2004) for the equivalent LDA parameters. For the remainder of the chapter, $V = V_k$: every topic has the same number V of points of view. For the beta distributions, we set $\psi_\alpha = 0.8$, $\psi_\beta = 0.2$ and $\gamma_\alpha = 5$, $\gamma_\beta = 95$: for example, $\eta_k \sim \text{Beta}(\alpha = 5, \beta = 95)$. This encodes the belief that disagreement probabilities will be low for non-POV interactions, and may be higher for POV interactions.

For an edit B which has an opportunity to disagree with edit A, we refer to A as B's parent. If an edit C has an opportunity to disagree with B, then C is B's child. All references are between edits on the same page. We assume that each edit references (has the opportunity to disagree with) at most one other edit (its parent), and is itself referenced by at most one edit (its child). As in Kittur et al. (2007), an edit's parent is the immediately preceding edit, and a disagreement (if any) is only with that edit. This disregards complexities which can arise when an edit reverts multiple prior edits, or when a single edit makes a complex contribution and subsequent edits disagree with different parts of it, but simplifies while being correct in most situations.

4.2.1 Inference

Griffiths and Steyvers (2004) introduce a collapsed Gibbs sampler for inferring LDA's latent variables, integrating out the real-valued categorical distributions associated with documents and topics. A single Gibbs iteration samples each latent variable once according to its full conditional distribution (conditioning on the values of all other latent variables). For LDA, this means that each topic assignment is re-sampled taking into account the most recent topic assignments for all other words (edits in our case). The algorithm eventually converges to a stationary distribution, where topic assignments are drawn from their posterior distributions given the observed data.

We use a parallel approximation to collapsed Gibbs sampling for inferring the topic and POV of each edit. The collapsed Gibbs sampler is similar to that of LDA, with both topic and POV repeatedly re-sampled rather than the topic only. We use a tightly-coupled parallel approximation to Gibbs sampling—similar to the GPU inference of Yan et al. (2009) and Approximate Distributed LDA [96]—where each thread starting a (topic, POV) re-sample takes into account all previously recorded assignments, then is itself recorded before going on to the next edit (maintaining consistency). This tight coupling increases communication between threads (using shared memory extensively), but comes as close as possible to true Gibbs sampling (where sampling is serial).

Re-sampling is according to the full conditional probability of a (topic, POV) pair for a single edit, conditioning on the assignments of all other edits. As in LDA, this depends on the probability of the user selecting a given (topic, POV) pair, and the probability of selecting the observed page given that choice. Additionally, it depends on the probability of a disagreement between the edit and its parent (if any), and its child (if any). This full conditional distribution can be written as:

$$
\begin{aligned}
&p((z,v)_{m,n} = (k,j) \mid (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o}) \\
&\propto \left(n^{(m)}_{-n,(k,j)} + \alpha\right) \frac{n^{(p_{(m,n)})}_{-(m,n),(k,j)} + \beta}{n^{(\mathbf{p})}_{-(m,n),(k,j)} + P\beta} \\
&\quad p(d_{m,n} \mid (z,v)_{m,n} = (k,j), (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o}) \\
&\quad p(d_{\text{child}(m,n)} \mid (z,v)_{m,n} = (k,j), (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o})
\end{aligned}
$$

Where $n^{(A)}_{-B,C}$ denotes a count across object(s) A excluding the assignment of B on topic C. Bold variables denote a vector of all the associated values (see Figure 4.1), except $(\mathbf{z},\mathbf{v})_{-(m,n)}$, which excludes the n^{th} edit by user m (the edit currently being re-sampled). For edits lacking a parent or a child, the corresponding disagreement probability is omitted. We omit the normalizing constant on the factor $n^{(m)}_{-n,(k,j)} + \alpha$ (the probability of the user selecting this (topic, POV)), as the normalizing constant does not depend on the (topic, POV) being considered. The probability of observing a

disagreement depends on the topic and POV assignments of the edit r and its parent o_r:

$$p(d_r \mid (z,v)_r = (k,j), (z,v)_{o_r} = (k',j'), (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o})$$
$$= \begin{cases} k = k', j = j' & \frac{n_{-(m,n),k}^{(z=z',v=v',d)} + \gamma_\alpha}{n_{-(m,n),k}^{(z=z',v=v')} + \gamma_\alpha + \gamma_\beta} \\ k \neq k' & \frac{n_{-(m,n),k}^{(z\neq z',d)} + \gamma_\alpha}{n_{-(m,n),k}^{(z\neq z')} + \gamma_\alpha + \gamma_\beta} \\ k = k', j \neq j' & \frac{n_{-(m,n),k}^{(z=z',v=j,v'=j',d)} + \psi_\alpha}{n_{-(m,n),k}^{(z=z',v=j,v'=j')} + \psi_\alpha + \psi_\beta} \end{cases}$$

In the above counts, we ignore disagreements between the edit under consideration and its parent *and* child (as those depend on the edit's previous (topic, POV) assignment). The variable $n_{-B,C}^{(A)}$ again denotes easily computable counts based on the topic and POV assignments of other edits and whether those edits disagree or not.

Inference then consists of repeatedly drawing new (topic, POV) pairs with probability proportional to the full conditional distribution specified above. An iteration of Gibbs sampling consists of re-sampling the topic and POV of each edit once. To initialize, we randomize each assignment.

Computation We use 64 threads in parallel on a single machine (64 cores) for inference. Sampling with 200 topics and 4 POVs takes approximately 6000 CPU hours for 200 burn-in iterations and 400 additional samples (we save every fifth). However, this sampling need only be done once: the 80 saved assignments and one high-probability assignment of topics and POVs to revisions are all that is required to produce the results we present (aside from synthetic data and model selection), and these samples can be reused to compute new page and user statistics on the fly. The time complexity of each Gibbs sampling iteration is $O(KVN)$ where $N = \sum_m N_m$ is the total number of edits.

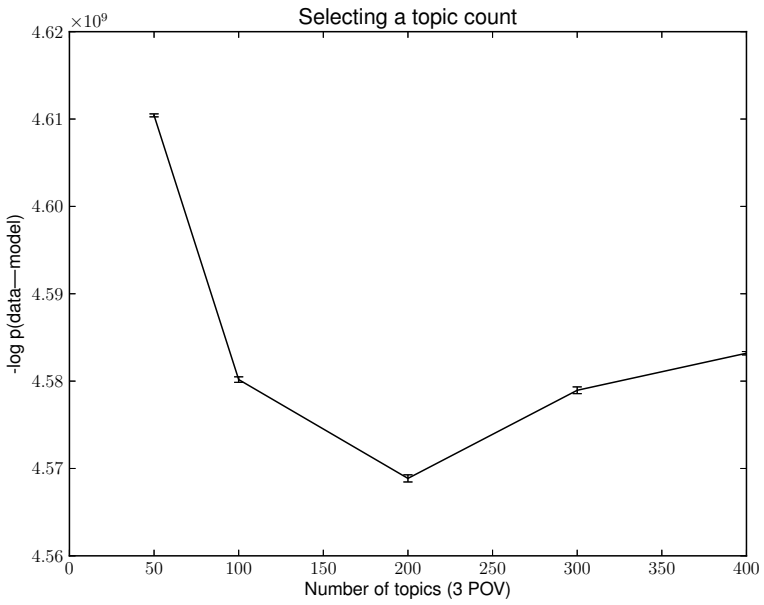

Figure 4.3: Negative log likelihood with the number of points of view fixed at 3. Error bars: twenty times standard error.

4.3 Data and model selection

Dataset We use the complete edit history of English Wikipedia as of November 2012, with 31583222 users, 9806233 pages, and 341026287 total edits. For anonymous users, we treat all edits from the same IP address as belonging to one user. Reverts are modeled as disagreements, either when the hash of a page matches the hash of a previous version of that page, or when "revert" or "rv" is mentioned in the edit comment. Wikipedia edits have a parent defined where applicable, which we honor except in rare cases where more than one edit has the same parent or the parent is on a different page (in the case of merged/split pages); in these cases, we treat the edit as not having a parent. We only use edits to pages in namespace 0 (the article namespace), ignoring talk and administrative pages.

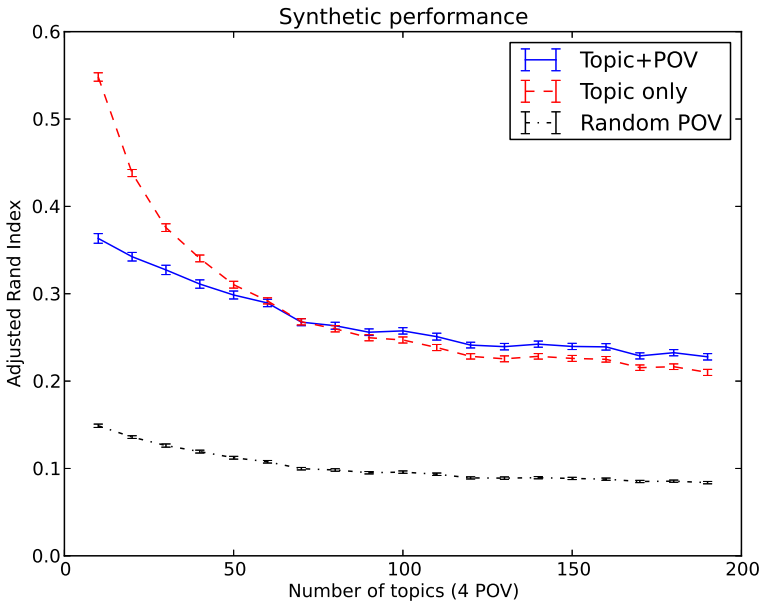

Figure 4.4: Synthetic performance, clustering edits according to their (topic, POV). Also shows performance when POV is ignored (Topic only), and when POV is randomized within a topic (Random POV). Error bars show the standard error of the mean.

Selecting a topic and POV count We perform model selection by estimating $p(\mathbf{p}, \mathbf{d}|K, V, \gamma, \psi, \alpha, \beta)$: the probability of the observed variables given the model, with the topic and POV assignments integrated out. We use an estimator due to Murray and Salakhutdinov (2009), which exploits forward and backward transition operators of the Markov chain and is easily implemented on top of a Gibbs sampler. For a comparison, see Wallach et al. (2009).

Fixing the number of POVs V at 3, we find that the data is most likely under a model with $K = 200$ topics; see Figure 4.3. The number of topics and the number of POVs per topic are not interchangeable: fixing the product $KV = 600$ and trying $V \in \{1, \ldots, 6\}$, $V = 3$ and $K = 200$ again assigns the highest probability to the data. We then optimize the number of POV V, fixing $K = 200$, and find that $V = 4$ maximizes the probability of the data, although the difference is smaller than for the number of topics. We now fix $V = 4$ and $K = 200$ unless otherwise noted.

While we do assume a fixed number of points of view, there is flexibility built into the model. Points of view are not necessarily antagonistic (the relevant prior, ψ, is very weak). If there are only two points of view A and B on a topic, the model is free to create points of view A_1, A_2, B_1, and B_2 such that A_1 and A_2 have a positive relationship to each other and a negative relationship to the B points of view. We can model any small number of "true" points of view in this way (below 5 if $V = 4$). We now turn to validation of the model assumptions, using real user relationships to test the inferred relationships between points of view.

4.4 Experimental validation

4.4.1 Synthetic experiments

Topic modeling can be viewed as a clustering problem, where words are assigned to topical clusters. In our case, we wish to assign edits to (topic, POV) clusters. In order to test the success of this method, we first generate the data directly *from* the model, and then perform inference and check the "correctness" of the inference. A perfectly "correct" clustering is unlikely – topics overlap, and there is limited additional information when there are multiple edits by the same user on the same page (although our model does leverage additional information in the form of disagreements). However, given the success of LDA in the past decade, we can compare our method with the baseline of simply using a topic model, ignoring points of view.

In order to do so, we generate data with 5000 users, 1547 pages (roughly keeping the ratio of users:pages the same as the Wikipedia data), and an upper truncated Pareto distribution for the number of edits per user [99] with an upper truncation of 147696 and a slope of 0.8, roughly matching the Wikipedia distribution. The parameters $\alpha = 5/(VK)$, $\beta = 0.1$, $\psi_\alpha = 0.8$, $\psi_\beta = 0.2$, $\gamma_\alpha = 5$, $\gamma_\beta = 95$, $V = 4$, and K are the same for generation and inference. We evaluate a

high probability assignment of topics and POVs found through iterative maximization after 100 iterations of Gibbs sampling, comparing the resulting clustering to the true assignments from the synthetic data.

Figure 4.4 varies the number of topics K, measuring clustering performance using the adjusted Rand index, which is corrected for chance (its values are between -1 and 1, with 0 being the expectation of a random assignment and 1 being a perfectly correct assignment). The algorithm is able to effectively infer points of view in addition to topics, on par with how well LDA infers topics. Small values of K make the topic assignment problem easier (consider a trivial example with a single topic), while a fixed number of POVs per topic keep the (topic, POV) assignment problem challenging even when considering very few topics.

4.4.2 Rule violation reports, reverts, and baselines

We turn now to validating our model on real data. We collect a dataset of *rule violation reports*, where one Wikipedia user reports that another has violated Wikipedia's Three Revert Rule (3RR): a user may not perform more than three reverts on a single page in a 24-hour period (subject to an administrator's interpretation and conditions on what constitutes a revert for 3RR purposes). The act of reporting another user implies a significant disagreement: reporting a user who shares your point of view, while a noble concept, is understandably unlikely in practice.

This gives us 7179 unique pairs of antagonistic users, where one has reported the other for a 3RR violation. Along with negative examples, they form comparisons:

RRP Randomly permuted reporting pairs provide negative examples, generating 7179 random pairings. Disputes are on specific topics, so random pairs with significant disagreements should be unlikely.

NR RRP with pairs of users who have reverted each other removed, leaving 986 reporting pairs.

WP With page information. Positive examples are where a reporting pair has edited consecutively on the same page, negative examples are from consecutive random edits by users who have never reverted each other (respectively 19683 positive and negative examples).

SP WP restricted to the set of pages that have both positive and negative examples, to eliminate any effects from choosing more controversial pages (4252 positive, 4536 negative examples).

For the datasets without page information (RRP and NR), we consider a thought experiment, placing edits by a pair of users next to each other on the same page, each edit serving as the parent with probability 0.5. We can then compute the expected probability of a POV disagreement over the possible assignments of topic and POV to the two edits, taking into account the topic and POV preferences of the users and the relationships between each POV. This POV disagreement probability measures the level of antagonism between users as inferred by the model. Viewing RRP as a ranking task, area under the ROC curve (AUC) is 0.85: a randomly selected true report pair will have a higher disagreement probability than a randomly selected non-report pair 85% of the time (0.5 is random guessing). Removing pairs who have reverted each other at least once (NR), the model is still quite discriminative, with an AUC of 0.72 on this more difficult task. How much of this performance is due to topical—rather than point of view—disparities between users in the permuted pairs? We address this question using the datasets WP and SP, which restrict examples to the same pages, and hence mostly to the same topic, and find that the model still performs well (Table 4.2). Computing the model's probability of any revert, rather than the probability of a POV revert specifically, yields significantly worse performance on all of these datasets: the model is not predicting reverts, it is predicting POV disagreements.

Why does the model work well, and are there alternative, simpler models that may be as powerful? We consider two alternatives. One hypothesis is that there are roughly four social roles on Wikipedia, and that users can be described just as well by these four groups as by many groups split across

Table 4.2: Model comparisons (AUC). Pairwise differences within each dataset are significant ($p < 0.01$) except for starred* pairs, computed via empirical overlap across 10^4 bootstrapped datasets. Bootstrapped 95% confidence intervals are $\pm.01$ for RRP, $\pm.02$ for NR, $\pm.01$ for WP, and $\pm.02$ for SP and Reverts.

Model	RRP	NR	WP	SP	Reverts
Social roles	.57	.48	.69	.75	**.88**
Hierarchical	.80	.68	.71	.72	.80
Simultaneous	.85*	**.72**	**.82**	**.80**	.85
ApproxMaxCut	.85*	.57	.59	.62	.51*
RevedSamePage	**.88**	.62	-	-	.50*

topics (i.e. disputes are not topical). In order to test this hypothesis, we consider a baseline model with a single topic and four points of view, using the same methodology as for the full model when ranking pairs of users. This model nets an AUC of 0.57 on the full permuted user interactions dataset, but loses its discriminative power when we remove pairs who reverted each other (AUC 0.48). It does much better when non-reporting pairs are chosen to have edited consecutively on the same page (WP, 0.69), and better still when the reporting and non-reporting pairs are on the same set of pages (SP, 0.75). Social roles seem to play a role in animosity, but fail in cross-topic comparisons.

A second idea is that disputes are *entirely* topical. Is there any benefit to having a full model of topics and points of view over first determining topics and then clustering within topics to find points of view? We evaluate a two-level model which does the latter: first fixing 200 topics (standard LDA, but a single high probability assignment), then clustering revisions within those topics into four points of view. This hierarchical baseline consistently performs significantly worse than the simultaneous model. Table 4.2 summarizes these model comparisons.

Table 4.2 includes results for revert prediction: given held-out pairs of users with page information (794 reverts in 6835 edits, 654 pages), determine whether the latter editor reverts the former. The social roles model does very well on this task. This is likely because most reverts are not related to

POV disputes, but instead are typical "maintenance" tasks on Wikipedia; simple revert prediction is not our goal. The value of the simultaneous model is in domain adaptation: trained on reverts, it not only predicts those, but also more reliable indicators of user relationships, as demonstrated by the other datasets. Even though the majority of reverts are maintenance tasks, other reverts do contain information about deeper topical disputes, which can be harnessed by considering topics and points of view.

Also in Table 4.2 are results for two simple non-Bayesian baselines. ApproxMaxCut first builds graphs of users, with an edge if either user has reverted the other at least once on a given page. It then partitions users on each page into two groups via an approximate maximum cut, computed by selecting the best of 50 greedily optimized random partitions. For the datasets with page information (WP and SP), it predicts a positive label if the users are on opposite sides of that page's cut and negative otherwise. For the datasets without page information (RRP and NR), its predicted scores are a zero-one average across pages the users have edited in common (zero if they are on the same side of a cut, one otherwise). The performance is generally poor, with the exception of RRP. For RRP, memorizing pairs of users who have any relationship at all is profitable, since the permuted pairs are unlikely to have any pages in common while the reporting pairs almost certainly do. To illustrate this, RevedSamePage predicts a positive label when two users have reverts on at least one page in common, and a negative label otherwise. The Bayesian models do not memorize, instead summarizing relationships with a small number of topics and points of view, yet still excel on RRP.

4.5 Discussion

As we become increasingly reliant on collective social processes to aggregate information, understanding these processes is critical. In the presence of incentives for manipulation, having information sources wear their biases "on their sleeves" has enormous value both for users who must evaluate information from multiple sources, and for information sources themselves attempting to maintain credibility. We propose a scalable model which takes a first step toward uncovering bias in collective intelligence processes, and which can help aggregation venues police themselves.

This kind of modeling applies to a wide variety of collective intelligence and aggregation venues, and has the potential to make online information sharing more transparent. By augmenting human judgment with machine inferences from large datasets, we can ease the transition from traditional centralized information aggregation models, allowing more reliable and more useful information sharing.

Chapter 5

Modeling Social Process Incentives

5.1 Introduction

Collective intelligence processes are driven by large groups of people, and so crowd dynamics are critical for understanding how venues can attract and maintain participation over time. Social process incentives motivate individuals to contribute to a collective intelligence process, and that participation itself creates social process incentives for others. Figure 5.1 provides a concrete and model-free view of one type of social process incentive. Along with other types of process incentives, including for example inherent interest or altruism, social process incentives can determine which venues receive enough participation to create socially or economically useful products. Given that each individual has a finite capacity to participate in collective intelligence processes, there is an element of competition between venues for the social capital they need to maintain collective intelligence.

While the consequences of these incentives are most apparent in aggregate at the venue level, they act by influencing the choices of many individual users. We can think of users as playing a large-scale sequential coordination game, where they get higher utility for some actions (for example via inherent interest) but also receive utility for choosing actions which other users take later on (via social process incentives). As we are interested in a model of which actions users

actually take, behavioral game theory is a natural tool. Based in part on a model developed in that literature [37], I evaluate a Bayesian model of participation as a function of social process incentives and inherent interests.

Whereas the model I adapt from behavioral game theory studies situations where incentives are known quantities, typically because an experimenter is explicitly providing them, incentives in collective intelligence are much less straightforward. There are multiple types of incentives which could individually be more or less motivating; in this study I focus on voting scores and comment replies. It is then important to learn the relative importance of each type of social feedback. The approach I take is related to inverse reinforcement learning: Observing the actions that users take, and assuming a functional form, learn how a user's utility varies with observed social feedback. The learning algorithm attempts to find utility functions which, under the action model from behavioral game theory, give a high probability of each user's observed actions.

The social media website Reddit is an interesting testbed for this kind of modeling. Reddit is partitioned into communities called subreddits with a wide range of topics and participation levels. As accounts are shared between subreddits, participation data from Reddit provides the kind of longitudinal observations of users which allow us to draw inferences about how social feedback influences the effort allocation of individual users between different communities. I use the majority of this longitudinal data for learning the model parameters, but reserve a set of observations chronologically after this training set on which to evaluate the model. The model's predictions are probabilistic, and so I estimate the divergence function of a proper scoring rule in order to make principled performance comparisons between models [16]. I compare the full inverse reinforcement learning model to a variety of empirical and model-based baselines, finding that it is important to model both inherent interests and the tendency of interests to shift over time, and further that learning a translation from social feedback to motivation provides better predictions of these shifts in effort allocation.

Reddit is an interesting testbed because of easily accessible longitudinal data, but this type of research is especially interesting because collective intelligence venues are competing with each other and with other online venues for the attention of users. Process incentives often seem to be an unmitigated good for collective intelligence venues, attracting users who bring information and contribute effort without the likelihood of manipulation created by users who contribute because of product incentives. Moving from training and evaluation to simulations, I make an analogy between communities competing on Reddit and this broader competition between online venues.

Given that social feedback is important for growing or maintaining participation in a collective intelligence process, one natural question is how new collective intelligence processes can attract participants. Without participants, new users receive very little social feedback and are therefore less likely to continue participating, a Catch-22. I augment the probabilistic community choice model referenced above with a simple model of how users provide feedback to others once they have selected a community. Simulations show a strong herding effect: Communities with more participants provide more social feedback to their users, making them more attractive to new users and creating the expected rich-get-richer effect. However, the efforts of a relatively small number of agents who provide consistent feedback to participants can overcome this effect, mirroring the founding story of Reddit itself, where the founders initially contributed heavily under pseudonyms until there were enough users to form a self-sustaining venue.

Effects on user behavior of social-psychological feedback have been documented recently in social media on YouTube and Digg [138], and on Wikipedia [141]. While social news is often viewed as a way for participants to influence public awareness and opinion, the act of sharing and its associated social feedback have a much more direct effect on those doing the sharing. What are the implications of millions of users providing and receiving feedback, influencing and being influenced? We liken social media to a game, where a user's strategy helps to determine the social feedback received by others, and the choices made by other users influence a user's own utility. As users learn and adapt their strategies, they create and abandon groups, communities, and whole venues.

Understanding the complex social dynamics governing the evolution of these communities is a key challenge for those who study multi-agent systems and collective intelligence. In this chapter, we investigate how our interests are determined, individually and collectively, by feedback from others.

We introduce a model of behavior in response to social-psychological feedback in social media. This model builds on work in human game playing with matrix games [37], in the behavioral/ reinforcement learning tradition. Rather than attempting to find an optimal strategy, players make updates to mixed strategies in response to the feedback they receive. We combine this learning model with a more sophisticated model of initial preferences (based on the Hierarchical Dirichlet Process [128]), and create an inference algorithm which discovers the dynamics of the learning process itself along with factors which constitute behavior-altering feedback.

We test the algorithm on real and synthetic social media data, where users choose between thousands of communities based on their initial preferences and the feedback they receive. On synthetic data, the inference algorithm is able to recover a user's true distribution of community preferences—a mixed strategy under the game analogy—with near perfect accuracy. We then apply the learning model to real data from the social news site Reddit, which at any one time is composed of thousands of active communities. The learning model outperforms a plethora of static and adaptive baselines on this probabilistic prediction task. Moreover, the model provides easily interpretable explanations. It allows for a form of inverse reinforcement learning where probabilistic human behavior changes are viewed as the product of a set of features, whose relative importance can then be determined by the inference algorithm.

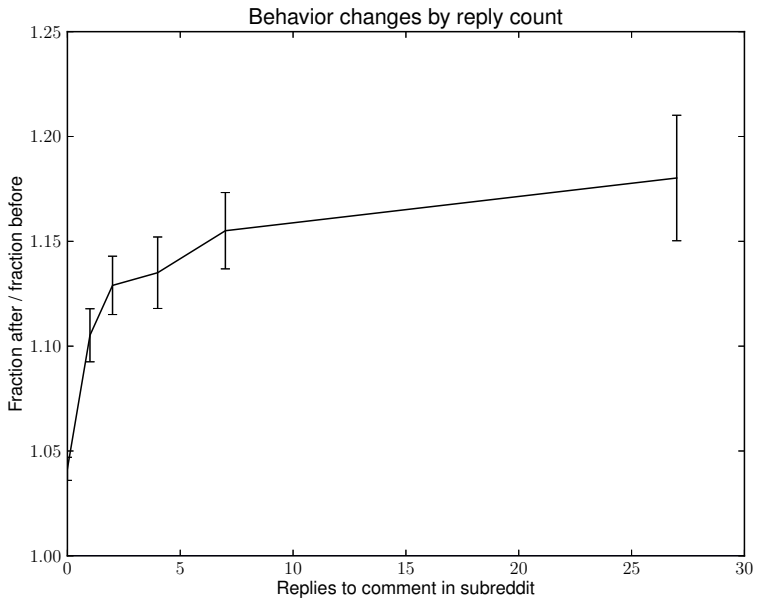

Figure 5.1: The ratio of the fraction of time spent on a subreddit after a comment to the fraction of time spent on that subreddit before the comment (excluding the contribution itself in both cases), as a function of the number of replies to that contribution. Receiving more responses makes a participant more likely to allocate more of their effort to that subreddit in the future, consistent with a learning effect in response to social feedback.

5.1.1 Related work

There has been significant interest in the population-level dynamics of collective intelligence. Wu and Huberman [137] study the relationship between a novelty factor of content and its popularity. Szabo and Huberman [127] predict the long-term popularity of content given the initial reaction to it. Networks play an important part in social media dynamics. Lerman and Hogg [73, 72] introduce stochastic models of popularity which distinguish network effects and visibility effects resulting from the content venue when predicting and explaining content popularity. Ahn et al. [2] study the relationship between participation and contribution on user generated content platforms. Although it is not considered in this chapter, user heterogeneity (e.g. [64]) is an interesting and

relevant direction for this line of research. Lerman *et al* [74] show that social proximity predicts sharing behavior in social media. Heaukulani and Ghahramani [56] present a Bayesian model for reasoning about unobserved interactions in social networks. We begin with a more basic model of individuals, where the existence of interactions are more important than the identities of the other participants in those interactions.

Another line of research considers the effects of social interactions and social biases in social media. Wu *et al* [138] find in social media that users who contribute have an increased propensity to contribute in the future, and that this effect along with social interactions explain the distribution of the number of contributions per user. Cobot [61] was developed to track and respond to social feedback in a virtual world. On Wikipedia, Zhu *et al* [141] test the effects of positive and negative feedback on users' propensities to contribute in the future. Huberman *et al* [60] show that attention predicts future contributions on YouTube, and that a lack of attention is demotivating. Muchnik *et al* [92] study the effects of social influence bias by conducting a randomized experiment where vote counts are seeded with a small positive or negative value. They find asymmetric herding effects as a result of these initial signals. Hsieh *et al* [59] study predictors of volunteer socialization on Reddit. Gilbert [44] finds that users are unlikely to participate in evaluating new content on Reddit, citing a lack of social interaction as one potential reason. To the best of our knowledge, we are the first to quantify the effects of social feedback on individual user behavior by explicitly modeling complex future decisions.

We make an analogy between social feedback and the rewards received in a game-theoretic setting. There is related work studying social media in game-theoretic terms, and studying human behavior in game playing. Munie and Shoham [93] put wikis, ratings and similar collaborative venues into a game theoretic framework. They show that users taking actions under a simple myopic rule converge to an equilibrium. Genter *et al* [41] investigate ways to control groups of agents exhibiting flocking behavior, reminiscent of our community seeding experiments. Erev and Roth [37] study a model of human game playing, explaining the changes in strategies resulting from the

rewards received in repeated matrix games. They show that simple learning models outperform equilibrium predictions which assume rational behavior. We adapt this model to explain and predict the effects of social feedback in social media.

Inverse reinforcement learning [97] and apprenticeship learning [1] recover the utility functions of actors based on their actions, with the hope producing similar (often optimal) policies in unseen situations. Knox and Stone [68, 69] add manual signals in a reinforcement learning setting, allowing humans to affect policies and speed learning. Chernova and Veloso [22] deal with the stochasticity of human demonstrations of policies in an inverse reinforcement learning setting using Gaussian mixture models. We are agnostic as to whether the behavior we are learning about is truly optimal: we wish to learn how behavior changes in response to the results of prior actions, by learning utility functions and update rules from observed behavior that are predictive of future behavior.

5.2 Model

Our goal is to model the behavior of users on social media sites such as Reddit. One feature of interest on these websites is strong community structure. On Reddit, for example, users choose to belong to communities called subreddits, which are user-run and organized around a theme: anything from New York City to cat pictures. What drives community selection? We make an analogy to games: users choose to post in a specific community, analogous to picking a pure strategy or action. Based on the action, they get a reward in the form of social feedback from other users who have also chosen that community. Users play this repeated community selection game, giving and receiving social feedback. Rather than explicitly specifying a goal, we focus on jointly learning how players adapt to rewards and what form those rewards take.

5.2.1 Background

We begin with some background from previous work, then present a model for players in repeated community selection games.

Human game playing

We are interested in how humans play community selection games. Our model builds on the three-parameter model of Erev and Roth [37] for humans playing mixed strategies in relatively simple matrix games with small, fixed numbers of strategies. A player begins with equal propensities for playing each pure strategy k:

$$q_k(1) = Z$$

Propensities are updated with a non-negative reward (which can be achieved in matrix games by subtracting the minimum reward). For a reward R after playing strategy k, we have:

$$q_k(t+1) = q_k(t) + R$$

Propensities for other strategies $j \neq k$ remain unchanged ($q_j(t+1) = q_j(t)$) under the one-parameter model. When picking a strategy, a player selects from her normalized propensities:

$$p_k(t) = q_k(t) / \sum_k q_k(t)$$

The three parameter model adds recency and exploration parameters ϕ and ϵ to the initial propensity strength parameter Z. For a reward R after taking action k:

$$q_k(t+1) = (1-\phi)q_k(t) + (1-\epsilon)R$$

Other actions are also updated under the three-parameter model. For a strategy j which was not taken:

$$q_j(t+1) = (1-\phi)q_j(t) + R\epsilon/(M-1)$$

Where M is the number of available strategies.

Hierarchical Dirichlet Process

The strategy space of this game, the space of all communities, has several interesting properties. First, it is not finite: users are free to start their own communities at any time. It is also quite large, with thousands of active communities at any one time. Finally, users start the game with strong prior preferences over strategies: given that a user is from New York City, she has a good chance of remaining active in that community. Likewise for hobbies, interests, organizations, and so on. Further, some communities are much more popular than others across all users.

Initial propensities are not of great importance in matrix games with small and finite strategy spaces. However, in games of the kind we are considering, with infinite strategy spaces over which users may have strong prior preferences, the representation of the initial propensities becomes critical. The model must imply a proper probability distribution over this infinite strategy space, and should also be a natural model for preferences. To this end, we adapt a nonparametric Bayesian model used in machine learning and clustering, the Hierarchical Dirichlet Process [128], as a model for initial propensities; we present only the necessary special case here. This model allows global preferences, meaning that some strategies may be more popular overall across all users, and also models a user's personal prior preferences. First, an infinite discrete distribution β—representing global preferences over strategies—is drawn from an infinite-dimensional Dirichlet distribution with concentration parameter γ:

$$\beta \mid \gamma \sim \lim_{L\to\infty} \text{Dirichlet}(\gamma/L, \ldots, \gamma/L)$$

$q^0 \sim \text{Dirichlet}(\alpha_0\beta)$ // Initial propensities from HDP
$q \leftarrow q^0$
for $i \in C_u$ **do**
  // For each of this user's actions (in order)
  $s_i \sim \text{Categorical}(q/\sum_j q_j)$ // Strategy picking
  $q \leftarrow q(1-\phi)$ // Forgetting
  $q_{s_i} \leftarrow (1-\epsilon)R(r_i) + q_{s_i}$ // Direct reward
  $q \leftarrow q + \epsilon R(r_i) q^0$ // Exploration

Figure 5.2: Pseudo-code for the generative model of a single user's behavior. The symbol "$\leftarrow$" denotes assignment, and "$\sim$" indicates a draw from a probability distribution.

γ, α_0, β	Hierarchical Dirichlet Process[128] parameters
ϕ	Forgetting (beta prior)
ϵ	Exploration (beta prior)
q^0	Initial propensities for a user
R	Reward function (same for all users)
C_u	Sequence of actions by user u
r_i	Feedback/reward features for action i
s_i	Strategy of action i (observed)

Table 5.1: Summary of notation.

Next, second-level distributions are drawn according to the base measure β and concentration parameter α_0:

$$\pi_j \mid \alpha_0, \beta \sim \text{Dirichlet}(\alpha_0\beta)$$

Each π_j is again an infinite discrete distribution, sharing the same "atoms" as β. This distribution π_j is our model of a user's initial propensities.

5.2.2 Model description

We use these models of human reinforcement learning and of initial preferences to create a generative model of human behavior in response to social-psychological feedback in large social processes. Under this generative model, we first draw global preferences $\beta \sim \lim_{L\to\infty} \text{Dirichlet}(\gamma/L, \ldots, \gamma/L)$,

where $\gamma \sim \mathrm{Gamma}(\gamma_\alpha, \gamma_\beta)$ is the first-level concentration parameter. We also draw a second-level concentration parameter $\alpha_0 \sim \mathrm{Gamma}(\alpha_{0\alpha}, \alpha_{0\beta})$, used for generating user-specific initial preferences.

Instead of fixing the global parameters of the learning model, we also sample these from their respective prior distributions: exploration $\epsilon \sim \mathrm{Beta}(\epsilon_\alpha, \epsilon_\beta)$ and forgetting $\phi \sim \mathrm{Beta}(\phi_\alpha, \phi_\beta)$. These parameters are analogous to those in the three-parameter model of Erev and Roth[6]. We assume that the reward function R is linear in a set of non-negative "reward features" with non-negative weights, the weights having independent Gamma priors.

Finally, users play the game-analog: repeatedly picking actions, receiving rewards, and updating their mixed strategy based on the reward received and the global learning model. Algorithm 5.2 describes this process formally, and Table 5.1 summarizes the notation. Users draw strategies from their current propensities. For each decision made, a user "forgets," scaling down his weights by $1 - \phi$ and creating a recency effect. Users receive a reward $R(r_i)$, depending on the reward features r_i. An exploration parameter ϵ distributes some fraction of this reward to the user's initial propensities q^0, with the rest going to the propensity of the chosen strategy. Each user repeats this process, choosing strategies and learning based on the resulting reward.

We do not explicitly model the reward features r_i received in response to an action, aside from the aforementioned assumption of non-negativity. These features will typically rely on the actions of other players, as they do in our experiments. This implies an additional mechanism which translates from all user strategies to reward features for each user; we do not model it. In the setting of social media, this mechanism specifies who replies to whom given where users choose to make their comments. To simplify notation, we assume all rewards accrue before the user takes a new action.

[6]The third parameter, strength of initial propensities, is redundant in our model with the coefficients of the reward features, which can be scaled to simulate any initial propensity strength.

5.3 Inference

Having specified a generative model for learning in response to social-psychological feedback, our goal is to reverse this process, making inferences about user learning from observed data. The main idea will be to separate inferences about users' initial preferences from inferences about the learning process[7]. Having done this, we use Gibbs sampling for approximate Bayesian inference.

To implement this separation, we begin with a series of binary latent variables, one associated with each action by a user. These variables are sampled according to their full conditional distributions given the values of all other latent variables (i.e. Gibbs sampling). That is:

$$p(\iota_i = \text{Initial} \mid s_i, \cdot) = \frac{p(s_i \mid \iota_i = \text{Initial}, \cdot)p(\iota_i = \text{Initial} \mid \cdot)}{\sum_{\iota_i'} p(s_i \mid \iota_i = \iota_i', \cdot)p(\iota_i = \iota_i' \mid \cdot)} \tag{5.1}$$

Where the sum in the denominator is over the two possible assignments of ι_i: initial or reinforcement. We use $\cdot$ as shorthand for conditioning on the values of all of the latent and observed variables except those pertaining to i (s_i, ι_i, and several we have not yet introduced). This application of Bayes' rule allows us to condition on both these latent variables and the observed action s_i.

To compute the "prior" probabilities (those not relying on s_i) in Equation (5.1), we refer to the simulation in Algorithm 5.3. Having run this simulation, which relies on the values of the latent learning parameters, we can explicitly compute those probabilities as follows:

$$p(\iota_i = \text{Initial} \mid \cdot) = \frac{q^i_{\text{init}}}{q^i_{\text{init}} + \sum_j q^i_j} \tag{5.2}$$

[7]Note that this model, where contributions come from a mixture distribution over initial and reinforcement distributions, is exactly equivalent to the more standard one in which each contribution comes from the "reinforced" version of the initial distribution.

$q_{\text{init}} \leftarrow 1$
$q \leftarrow \vec{0}$
for $i \in C_u$ **do**
  // For each of this user's actions (in order)
  $q^i, q^i_{\text{init}} \leftarrow q, q_{\text{init}}$ // Record current weights
  $q \leftarrow q(1-\phi)$ // Forgetting
  $q_{\text{init}} \leftarrow q_{\text{init}}(1-\phi)$
  $q_{s_i} \leftarrow (1-\epsilon)R(r_i) + q_{s_i}$ // Direct reward
  $q_{\text{init}} \leftarrow q_{\text{init}} + \epsilon R(r_i)$ // Exploration

Figure 5.3: Pseudo-code for the reinforcement learning simulation, which forms part of the inference algorithm.

$p(\iota_i = \text{Reinforcement} \mid \cdot)$ is simply $1 - p(\iota_i = \text{Initial} \mid \cdot)$. This leaves the probability of observing strategy s_i given the assignment of ι_i. In the case that $\iota_i = \text{Initial}$, this is the probability of drawing s_i from the Hierarchical Dirichlet Process conditioned on all of the other initial strategy observations (but not any of those where $\iota_j = \text{Reinforcement}$). We reproduce this probability here, but see Teh *et al* [128] for background and details:

$$p(s_i \mid \iota_i = \text{Initial}, \cdot) = \frac{\alpha_0 \beta_{s_i} + n^{-i}_{u_i,s_i}}{\alpha_0 + n^{-i}_{u_i}} \tag{5.3}$$

$$n^{-i}_{u_i,s_i} = \sum_{j \in C_{u_i} \setminus i} I(\iota_j = \text{Initial and } s_j = s_i)$$

$$n^{-i}_{u_i} = \sum_{j \in C_{u_i} \setminus i} I(\iota_j = \text{Initial})$$

Here, I is an indicator function which is 1 if its argument is true, and 0 otherwise. u_i is the user associated with action i. The global propensities β and second-level concentration parameter α_0 are part of the HDP, and this probability corresponds to the direct sampling scheme in Teh *et al* [128].

The equivalent probability for the case when $\iota_i = \text{Reinforcement}$ depends only on q^i:

$$p(s_i \mid \iota_i = \text{Reinforcement}, \cdot) = q^i_{s_i} / \sum_j q^i_j \tag{5.4}$$

In the case that $\sum_j q^i_j$ is 0, $\iota_i = \text{Initial}$ deterministically.

This concludes the sampling scheme for each ι_i: evaluate Equation (5.1) using (5.2), (5.3), (5.4), and Algorithm 5.3, then draw a Bernoulli random variable according to that probability. This forms the bulk of the inference procedure. However, we have neglected the global latent variables (learning parameters, HDP parameters) to this point.

Teh *et al*[128] include or give reference to sampling schemes for β, α_0, and γ (the latter being related to our inferences through β), which we use. We do not reproduce them here; see Teh *et al* and Escobar and West [38] for details.

This leaves the global learning parameters ϕ and ϵ, and the feature weights of the reward function R. We sample these parameters using Metropolis-Hastings: proposals are generated from a proposal distribution (we use a Gaussian), then accepted or rejected based on the probability of the proposed parameter and the proposal distribution. This allows us to indirectly sample from the full conditional distributions of these parameters. For example, consider the forgetting parameter ϕ:

$$p(\phi \mid s, \iota, \cdot) = \frac{p(s, \iota \mid \phi, \cdot) p(\phi)}{\int p(s, \iota \mid \phi', \cdot) p(\phi') d\phi'} \tag{5.5}$$

Where ι and s are the vectors of action-specific indicators and action types respectively. $p(s, \iota \mid \phi, \cdot)$ is easy to compute with Algorithm 5.3 and Equations (5.2) and (5.4), but the resulting distribution is difficult to sample from directly. Instead, we generate a proposal ϕ', and accept that proposal ($\phi \leftarrow \phi'$) with probability:

$$\min\left(1, \frac{p(\phi' \mid s, \iota, \cdot) p(\phi' \to \phi)}{p(\phi \mid s, \iota, \cdot) p(\phi \to \phi')}\right)$$

$p(\phi \to \phi')$ is the probability of moving from ϕ to ϕ' using the proposal distribution. The integral in the denominator of Equation (5.5) cancels, so sampling is as easy as rerunning Algorithm 5.3 for each user. The remaining learning parameters can be sampled likewise.

The overall inference procedure is then Gibbs sampling: pick initial values for the latent variables, then sequentially sample from each full conditional distribution. Repeating this sequential sampling, the procedure draws from the posterior distribution given our observations in the limit, and in practice is a good approximation (using a finite number of samples) after a burn-in period.

5.4 Experiments

With a model and inference algorithm, we turn to empirical questions: how much data is required to recover the parameters? Is the model useful for describing real data? What can we learn from it?

5.4.1 Data

We collected a set of 174783 submissions and comments on submissions by 1696 users from the social media website Reddit, along with 2024160 related comments and submissions from other users which we use to compute reply counts. The comments and submissions are across 7037 "subreddits", communities of varying sizes which compose Reddit. These subreddits are the actions in our model: users select between them when choosing to post on a *new* submission. We group together comments by the same user on the same submission, and this grouping is reflected in the count of 174783 comments and submissions. Users were selected through a crawling process, excluding self-identified bots.

5.4.2 Features

One important consideration is feature extraction: what is the exact form of the reinforcement function R? The main social-psychological reward features we use are relative reply counts and relative "karma", the latter being the result of other users voting a comment or submission up or down. These features are first transformed into quantiles (between 0 and 1) of the karma and reply counts respectively across all the data we collected.[8] When grouping contributions on the same submission, we pick the contribution with the most prominent "level", breaking ties by picking the user's first comment on that submission. Submissions are the first level, followed by top-level comments to submissions, replies to those comments, and so on. Huberman *et al* [60] find that more experienced users eventually measure feedback relative to their own previous contributions rather than contributions by other users; this is an interesting direction for more complex future models.

We also include three binary features which indicate the prominence of the contribution. These binary features are 1 respectively if (1) the contribution is a submission, (2) a reply to a submission, or (3) a reply to another comment, and 0 otherwise. These features serve as intercepts for the regression (they are mutually exclusive). To summarize, we fit the following reinforcement utility function:

$$\begin{aligned} R(r) =& A \cdot r_{\text{karma}} + B \cdot r_{\text{replies}} + C \cdot I(r_{\text{type}} = \text{Submission}) \\ &+ D \cdot I(r_{\text{type}} = \text{Top}) + E \cdot I(r_{\text{type}} = \text{Reply}) \end{aligned}$$

With observed feature vector r and regression coefficients A-E.

[8] We considered using quantiles within a subreddit and contribution level, but analogs of Figure 5.1 indicate that the data does not support this grouping.

5.4.3 Priors and parameters

We perform inference on all of the model parameters rather than fixing their values, but must first specify prior distributions for each. Using shape and rate parameters for Gamma distributions, we have weak priors $\mathrm{Gamma}(1, 0.1)$ and $\mathrm{Gamma}(1, 1)$ on the HDP parameters γ and α_0 respectively as in Teh *et al* [128]. For both learning parameters ϕ and ϵ, we use the prior $\mathrm{Beta}(1, 9)$. The reward feature coefficients and intercepts have priors $\mathrm{Gamma}(1.5, 4)$.

After a burn-in period of 3000 samples, we average predicted distributions across 250 samples, skipping 30 iterations between each to avoid storing and processing correlated samples. For Metropolis-Hastings samples, we use Gaussian proposals with standard deviation $\sigma = 0.01$, except for ϕ and ϵ, for which we use $\sigma = 0.005$ to avoid excessive rejections.

5.4.4 Scoring probabilistic predictions

What is the right way to score predictions about a user's next subreddit choice? There is only one "correct" choice, corresponding to what the user actually does, but this choice is dependent on many unobserved external factors (perhaps unobservable under some conceptions of free will). Rather than attempting to make a single prediction, we focus on quantifying our uncertainty.

Given a prediction and an event's true outcome, a scoring rule quantifies the performance of that prediction. For one well known class of scoring rules, *strictly proper* scoring rules [114], an agent maximizes her score in expectation by truthfully revealing her beliefs. However, we are interested in comparing the performance of several different models. Strictly proper scoring rules have an associated *divergence function* [47], which measures the divergence of a predicted distribution from an unknown true distribution.

We assume that there is a distribution S over states of the world, in our case encompassing a user's entire history and current state of mind. For a given state of the world $\sigma \sim S$ there is a true distribution over observations f_σ and a predicted distribution g_σ, i.e. a model's probabilistic prediction. Based on a set $X = \{x_1, \ldots, x_N\}$ of observations (i.e. subreddits) drawn from the mixture distribution $\sigma_i \sim S$, $x_i \sim f_{\sigma_i}$, we want to estimate an expected divergence $E_{\sigma \sim S}[d(f_\sigma || g_\sigma)]$. In other words, how well does the model predict the true distribution of observations? If d is the divergence function of strictly proper scoring rule Q, $\frac{1}{N} \sum_{i=1}^{N} Q_{x_i}(g_{\sigma_i})$ approximates this expectation up to a constant (a generalized entropy term depending only on the true distribution). We use the quadratic scoring rule, which has divergence function $d(f||g) = ||f - g||_2^2$. Comparing models empirically with the quadratic scoring rule compares their predictions' expected squared Euclidean distance from the unknown true distribution of observations.

The quadratic scoring rule is computed as:

$$Q_i(p) = 2g_i - \sum_j g_j^2 \tag{5.6}$$

Where i is the true outcome, and g a vector of predicted probabilities. The score ranges from 1 when all probability mass is placed on the true outcome, to -1 when all probability mass is placed on an incorrect outcome.

5.4.5 Models and baselines

Reinforcement is the full model described previously, where users update their propensities in response to social feedback.

UserAll predicts that a user posts in a subreddit proportional to the number of times he or she has done so previously.

UserKMax predicts a subreddit will be chosen next proportional to the number of times it was chosen by the user in that user's past K contributions, maximizing over K. On the real data, this is achieved at approximately $K = 20$; we omit this baseline for synthetic data to simplify presentation.

Global predicts that users pick a subreddit proportional to the number of times it has been picked globally (across all users).

ErevRoth removes the learned initial propensity model, assuming instead that q^0 (Algorithm 5.2) is uniform over communities.

Initial removes the learning aspects of the reinforcement model (setting the reinforcement function R to 0 deterministically), leaving only the initial propensities. This model smooths a user's local preferences by incorporating global popularity.

InitKMax like UserKMax, trains the initial propensity model on the past K comments from each user, maximizing over K ($K = 25$ on the real data).

True For synthetic data, subreddits are drawn from a true distribution, which serves as an omniscient baseline for that data (but is unfortunately unavailable for real data).

5.4.6 Performance

Figure 5.4 shows the performance of the reinforcement model and a variety of baselines on held-out real and synthetic data. The parameters used to generate the synthetic data were chosen to approximately match those inferred from the real data. The performance of the inference algorithm is almost exactly the same as that of the true distribution of held-out subreddit choices on synthetic data. This indicates that the inference algorithm is effective, and that there is enough data available to make accurate inferences (the synthetic and real datasets are approximately the same size).

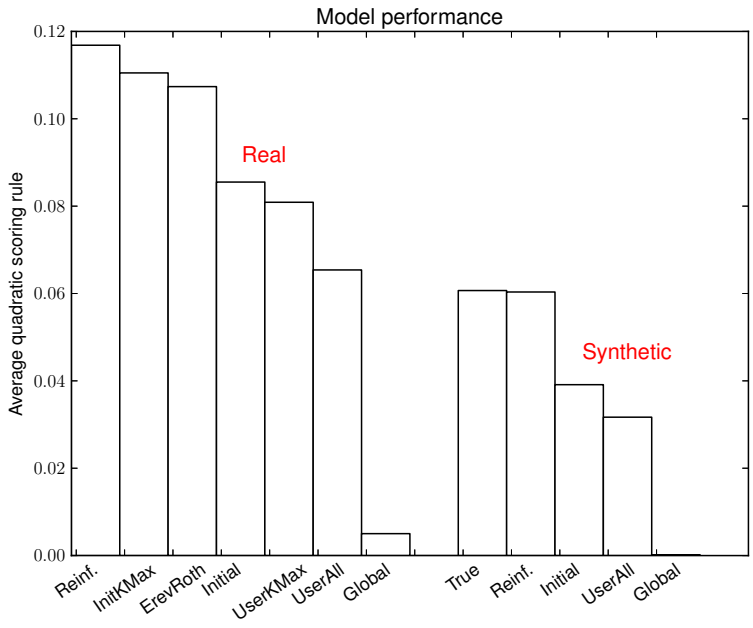

Figure 5.4: Performance of the models on real and synthetic data. Performance is averaged over the last 7 comments from each user, with each being respectively held out (along with all following comments) and its distribution predicted, among users with at least 8 comments (most having significantly more).

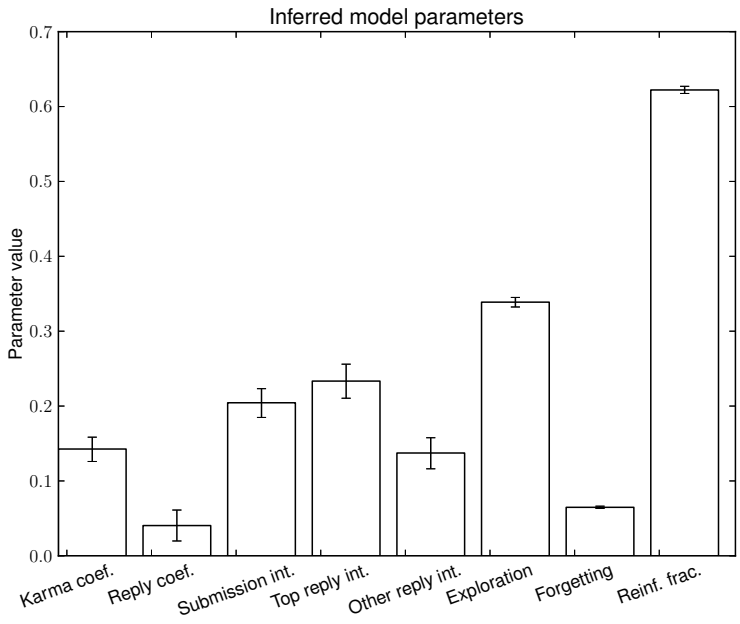

Figure 5.5: Inferred parameter values using the reinforcement model on real data. Reinforcement function coefficients, intercepts, learning parameters, and the fraction of contributions which were inferred to be the result of reinforcement (ι indicators) are shown. Error bars show empirical 95% credible intervals (contiguous about the mean).

Turning to real data, we see very similar relative performance. The reinforcement model comes closest to the true distribution of subreddit choices. Simple baselines such as UserAll perform fairly well considering that they are static models, not allowing for behavior changes over time. The UserKMax baseline attempts to compensate for this by throwing out older data, and does offer a significant improvement over the static UserAll baseline. While predictions based on global subreddit popularity are not very accurate alone, smoothing the user baseline by including global subreddit popularity turns out to be quite effective, as evidenced by the initial propensity baseline. As with the user baseline, we can attempt to adapt the static method to this dynamic setting. InitKMax is the best performer of the non-learning baselines, but the performance loss compared to the reinforcement model is statistically significant ($p = 2.8 \times 10^{-31}$ using a paired t-test).

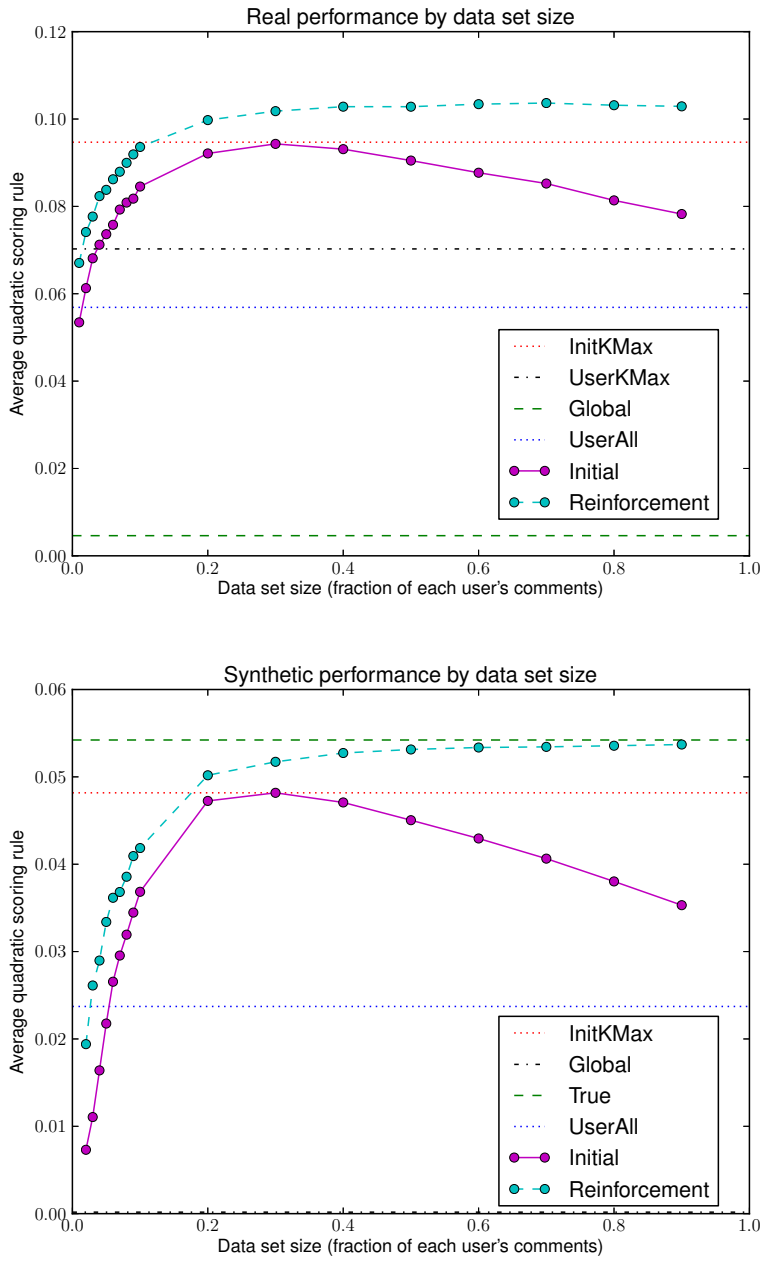

Figure 5.6: Performance of the reinforcement and initial propensity models as the amount of data varies, for both real and synthetic data. Performance is measured on held-out test data, which is always each user's last contribution. In order to truncate the data, we remove earlier contributions by each user first, leaving a contiguous training set directly before the test data. There are approximately 150000 contributions in the training set for the synthetic data, and about 170000 for the real data.

Figure 5.6 shows the performance of the reinforcement and initial propensity models as a function of the amount of data they are trained on, on both synthetic and real data. On synthetic data, the reinforcement model quickly approaches the true distribution, while the initial propensity baseline peaks and then declines as agents change their behavior in response to (simulated) feedback. The performance of these two models on real data is strikingly similar to their performance on the synthetic data: the reinforcement model quickly climbs and then stabilizes, while the initial propensity model peaks and then declines as it is "weighed down" by older data. Maintaining a moving window (as in InitKMax) removes the decline, partially compensating for unmodeled user learning.

While sub-optimal, this sliding window in combination with our initial propensity model does quite well despite ignoring a learning dynamic unquestionably present in the synthetic data and, considering Figure 5.1, also in the real data. However, the sliding window model does not offer a generative model of behavior or any explanatory power; by contrast, the reinforcement model actually explains the process by which changes in behavior take place, rather than just playing catch-up with observed behavior changes. We explore the value of this explanatory power in the next two sections.

5.4.7 Inferred parameters

Now that we have established that the reinforcement learning model is useful for describing real data, what can it tell us about human behavior? Figure 5.5 shows the inferred parameters.

Exploration is fairly high, at about 34%. This is sensible for an environment like social media where diversity of interests is critical. Even if rewards are heavily concentrated in a small number of subreddits, users will still pursue a broad range of interests. At the same time, the majority of contributions (about 62%) are modeled as previous experiences rather than initial propensities.

Perhaps the most interesting insights come from the reinforcement function R: what motivates people in social media? One surprise is the relative prominence of the intercept, applied regardless of the social-psychological feedback received in response to a contribution. Interpreting the intercept presents a puzzle: does the contribution simply provide information about a behavior change which has already happened, or does the user's behavior change as a result of having made that contribution? This is an extremely difficult question to answer in general, but we can provide some related insights which hint toward the latter explanation.

We split the intercept in the reinforcement model into three contribution levels: (1) submissions, (2) comments which are top-level replies to submissions, and (3) replies to other comments. Submissions and their replies (levels (1) and (2)) are far more visible: anyone visiting the subreddit will see contributions in level (1), and anyone who looks at the comment thread for a submission will see comments in level (2). Replies to other comments (level (3)) are much less prominent, and are sometimes not displayed on the main comment page at all without additional user interaction. The inferred intercept in case (3) is significantly lower than those inferred for cases (1) and (2). Under a causal interpretation, this might correspond to a higher level of social commitment associated with cases (1) and (2), and therefore a correspondingly more pronounced behavior change. Under the non-causal interpretation, an alternative hypothesis is that users who have already committed more heavily to a specific subreddit are more likely to make prominent contributions to that subreddit.

In either case, social feedback in the form of voting and replies is of greater relative importance when the user's contribution is a reply to a comment than it is for submissions and top-level comments. Surprisingly, the voting score of a contribution is more significant than the number of replies that contribution receives. One potential explanation is that replies are not always positive, while a high voting score is a clear indicator of community approval.

The reinforcement model enables a form of regression for user behavior changes. Here we have included two social feedback features (voting and reply counts), but other features are possible. Not only can we predict changes in user behavior, but we are also able to articulate specific reasons for those behavior changes.

5.5 Conclusions

We have shown that a simple model of learning can capture complex behavior changes in social media. Users spend more time in communities where they have received social-psychological feedback, and in communities where they have previously invested more time. While behavior is stochastic, an analogy to humans playing mixed strategies in matrix games provides a simple and effective learning model in this setting. Our quantitative model gives insight into individual user behavior in social media, and provides a solid foundation for studying the dynamics of communities of agents with mutual feedback and complex collective learning.

Chapter 6

Field Experiment on the Effects of Prediction Market Process Incentives

6.1 Introduction

The Instructor Rating Markets (IRMs) [15] were a field experiment, conducted by a collaboration (myself included) at Rensselaer Polytechnic Institute, to test the ability of prediction markets to provide real-time feedback. Markets were run on two-week intervals, the security associated with each market liquidating at the end of each period based on an opinion poll of students in a class about the performance of their instructor. Ten courses, each with their own security, had markets during each of five two-week periods, for a total of 50 markets over the course of the Fall 2010 semester.

The IRMs were designed to provide *real-time feedback* to instructors on the progress of their courses. Premised on periodic opinion polls, the markets provided an incentive for anyone with information *about* the next opinion poll to place a trade immediately. A student seeing their roommate struggling with a poorly worded assignment for a course could sell that course's security, profiting if the next opinion poll were below the sell price. Incentives in the IRMs were based on a virtual currency, with prizes raffled off periodically based on performance in terms of this currency.

Students in a participating class were able, along with every other person participating in the field experiment, to trade that course's security on the IRMs. An opinion poll of the students in a course determined the security's liquidation value. Students in a course were asked to rate their professor high (100) or low (0), and the corresponding security's value was an average of these ratings. For smaller courses, this was at times just a handful of people. The resulting incentives, including incentives for manipulation, are an interesting aspect of this field experiment, explored further in Section 6.1.1.

Figure 6.1 shows that market prices in the IRMs were good predictors of opinion polls. That is, the markets succeeded in aggregating information about future instructor ratings. Given the potential for manipulation, it would be reasonable to wonder whether the IRM opinion polls agree with official institute ratings (where there was no incentive to manipulate). There is in fact strong agreement, with a correlation coefficient of 0.86, indicating that responses to opinion polls were largely truthful.

The IRMs are a rich source of data from users who both traded in the markets and provided ratings. Information generally came from insiders, i.e. students who were taking the course they traded; insiders as a group provided fresh information, whereas outsiders in aggregate tended to trade based on previously revealed information (this does not rule out *some* outsiders bringing new information to the market). The pattern of insider and outsider behavior is consistent with the overall goal of incentivising the immediate revelation of information which affects a course evaluation, information which we would expect to lie primarily with the students actually taking a course. There was surprisingly little successful group manipulation (students in a course profiting by coordinating their ratings), with only one likely instance identified based on an enumeration of groups who provided the same rating during a period and held corresponding positions in the market.

Another purpose of the IRMs was to compare different market making algorithms in a longer-running prediction market setting. These algorithms provide liquidity to markets which would

otherwise be very sparse, allowing participants to place small corrective trades which are executed immediately, rather than placing limit orders on a nearly empty order book. In other words, market makers ensure that there are adequate process incentives for participating in a market. Sometimes this takes the form of an explicit subsidy, with the market maker expected to lose money, but market makers which break even can nonetheless increase process incentives, especially in new markets which would otherwise have a cold-start problem.

6.1.1 Product incentives in the IRMs

Having purchased shares of her course's security, *homo economicus*[9] would then unfailingly rate her professor high regardless of her own personal opinion. Moreover, the students in a course could decide to coordinate their votes, for example collectively deciding to rate their professor high, giving each of them strong knowledge of the security's liquidation value.

These incentives arise because the Instructor Rating Markets are a compound collective intelligence process with two components: opinion polls and prediction markets. The markets are premised on opinion polls of students taking participating courses, and these polls alone have only the incentives common to all opinion polls, for example product incentives to have the poll results reflect one's own opinion. However, the introduction of prediction markets changes this. Having traded in the markets, a participant has a new external product incentive to make the opinion poll match not her own opinion but her market position. For example, believing that the opinion poll results will be above the current market price and so buying the corresponding security, a participant would then have an incentive to rate her instructor high despite believing that the class was going poorly (as she stands to do better in the markets the higher the security liquidates). Here, process

[9] A likely fictitious perfectly rational and self-interested participant, intent in this case on maximizing prize winnings.

incentives in the prediction markets (i.e. the potential to profit from buying a security which liquidates high or selling a security which turns out to be worthless) are creating product incentives for the collective intelligence processes they are designed to track.

6.1.2 Related Work

In recent years, prediction markets have gone from minor novelties to serious platforms that can impact policy and decision-making [134]. There has been a concomitant rise in interest in prediction markets across academia, policy makers, and the private sector [135, 5, 117, 4, 17]. There has been some research on the design and deployment of live prediction markets [6, 100, 24]. There have been small experiments to test the impact of insiders on small, short-running, experimental prediction markets [53, 54]. The IRM is more of a "field experiment" than these controlled studies. It is significantly longer in duration and larger in the number of participants. It is also unique in its goal of explicitly providing dynamic feedback to instructors that can be correlated with real measures of performance.

A second motivation of this work is to provide a framework for comparing prediction market structures. There has been little systematic work in this area. While much of the literature on liquidity provision discusses the pitfalls and advantages of different algorithms [107, 101, 18, 108], only recently have there been attempts to simultaneously compare market microstructures in controlled experimental designs involving human traders (such as the work of Brahma et al. (2012)). However, Brahma et al. make these comparisons using short, ten-minute experiments. We open the door to studies of such issues in longer-horizon markets.

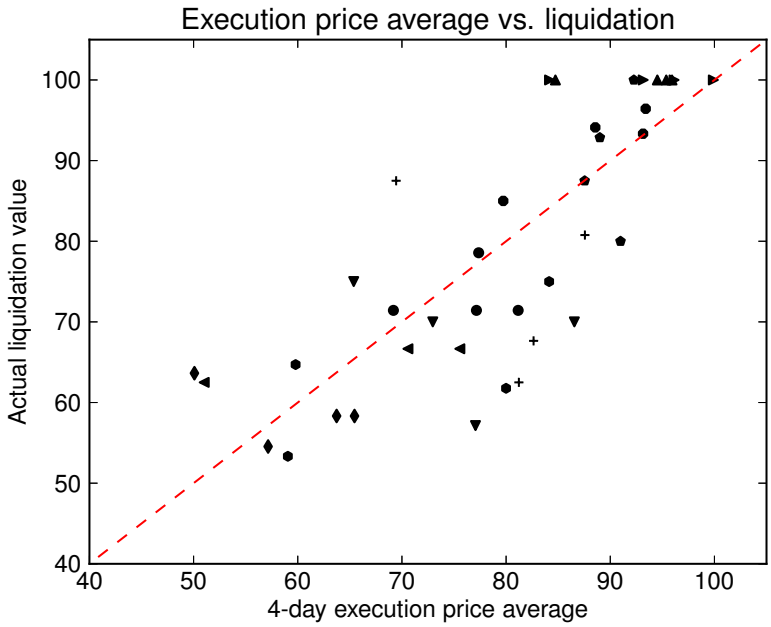

Figure 6.1: Traded prices predict liquidation values well.

6.2 Description of the Markets

We ran virtual cash markets for 10 different classes during the fall semester of 2010. The experiment was divided into five periods of approximately two weeks each, with one period extended due to a holiday break. Each instructor-course pair was one security (and a fresh security was instantiated for each instructor-course pair at the beginning of each of the five periods). Each security could be traded by anyone at the institute. At the end of each period, students enrolled in a particular class rated their instructor. The payoff (liquidating dividend) of the security for that instructor-course pair for that period was the average rating given by the students.

6.2.1 Ratings

Students taking one of the ten subject classes were given keys at the beginning of the semester which enabled them to rate their instructor at the end of each trading period. Each student who registered a key was sent a reminder email for each trading period. For the first period, rating was done through the same website as trading; for the remaining four periods, students could also rate directly from their reminder email. Initially, rating could be either thumbs-up (100%) or thumbs-down (0%), but a neutral option (50%) was added beginning with the third period. The initial limitation reflected the idea that only 0 and 100 were rational choices for traders seeking to maximize their wealth; we relaxed this limitation in response to feedback from students who did not want to rate their instructor either positive or negative. The liquidation value $\in [0, 100]$ of a market was the average of all ratings cast for the associated trading period.

6.2.2 Incentives

After each liquidation at the end of a trading period, a trader's account value was equal to their cash balance plus the liquidation value of any shares they held. All trader accounts were then re-initialized for the next trading period (there was no carryover).

Prizes were awarded twice: once after the second period of trading, and once after the fifth period of trading. Six prizes were raffled off each time, based on a trader's rank and account value in each period. The top 3, 5, 10, and 20 accounts in each period were eligible for the 1st, 2nd, 3rd and 4th prizes respectively.

If an account featured in the top 3 accounts of a period, it was eligible for all the prizes; if an account featured in the top 5 (but not top 3), it was eligible for all prizes but the top prize, etc. The fifth prize was a participation prize awarded uniformly at random to one of the top 50% of traders

Periods	Prizes					
	1st	2nd	3rd	4th	5th	6th
1-2	$69	$49	$40	$30	$20	$20
3-5	$150	$100	$60	$40	$20	$20

Table 6.1: The value of awarded prizes.

in each period. The sixth prize encouraged participants to rate their professors and was drawn with probability proportional to the number of times a trader provided ratings. The prizes were awarded from 1st to 6th, with the restriction that once an account was awarded a prize, it became ineligible for any subsequent prize. Prize values are summarized in Table 6.1.

From a theoretical standpoint, these incentives create complex utility functions. We could instead have awarded prizes with probability proportional to a trader's total account value. However, making such a scheme sufficiently rewarding was not practical given reasonable constraints on the value of awarded prizes; rank order incentives such as those used in the IRM can be significantly more effective than proportional payments [83] due to decreased risk aversion in traders. Simply paying participants based on their performance without vastly increasing the total amount awarded would likely have been demotivating [46]. While linear rewards for participation would seem to at least yield simple incentives, even this does not occur in practice, as other motivations are found to have a significant impact on experiment participants [80].

6.2.3 Microstructure

Traders interacted with the markets by placing market orders through a Web interface. Traders were presented with a full history of traded prices and liquidation values for each security, along with links to the associated course website. They were also shown the current (spot) price of the security, and could place a market buy or sell order for a desired quantity – they would then receive a price quote for their entire order, and were asked to confirm. For the first two periods,

users started with 50000 units of virtual currency and 50 shares of each market. For the final three periods, users started with 100 shares of each class and the same amount of currency.

Price quotes were generated using two different market making algorithms (only one algorithm was used for any given market during a particular trading period). We used an implementation of Hanson's logarithmic market scoring rule (LMSR) [52] with a b parameter of 125 (restricting loss to 8664.34 in any given period)[10], and an implementation of the Bayesian Market Maker (BMM) described by Brahma et al. (2012). Both market makers are initialized at the beginning of a trading period so that the quoted price in each market is the same as the close of the previous trading period.

6.2.4 Market Participation

Overall there were 226 registered users, with registration limited to current RPI students, faculty, and staff. Of these, 198 users made at least one trade during the experiment. Participation declined as the experiment progressed, with 117 active traders in the first period and only 33 in the fifth period. Rating was more steady, peaking at 93 raters during the second period, but never dropping below 70 raters during any period. The backgrounds of participants were mixed: from undergraduates studying physics to faculty in computer science.

6.3 Information Content of Prices

Prediction markets attempt to aggregate information and to incentivize the dissemination of information that is otherwise difficult to obtain. One question is whether traded prices in the IRM

[10]From a market making perspective, a real-valued dividend $\in [0, 100]$ is equivalent to the more typical 0-1 dividend, modulo a constant factor; the price computation for LMSR is exactly the same, and the loss bound is determined by the extreme values of the dividend.

provide any new information about *future* instructor ratings. If traders simply provide a noisy realization of the previous rating (dividend), for example, then the prices themselves do not provide useful dynamic instructor feedback. Do the markets have predictive power?

6.3.1 Predictivity of prices

Figure 6.1 answers this question in the affirmative. In the figure is a scatter plot of the upcoming liquidation value versus the average traded price. Different markets are referenced with different symbols. Also shown is the ideal outcome (the line $y = x$). Modulo noise in the data, there is good agreement between the data and the ideal line. We use an average traded price because prices are noisy and averaging can provide a better proxy for the market value than the price of any single executed trade.[11] Such smoothed prices were significantly more predictive than previous liquidation values, with a four-day average price yielding an R^2 value of 0.58, while previous liquidations produced an R^2 of 0.48. This finding is robust to different averaging windows for prices and different aggregates for previous liquidations.

To further validate that market prices are a better predictor of future liquidation values than prior prices, we ran a regression using both the previous liquidation and the market price as independent variables in the following model:

$$\mathrm{Liq}_{s,\rho} = \beta_1 \mathrm{Liq}_{s,\rho-1} + \beta_2 \mathrm{Price}_{s,\rho} + \alpha \tag{6.1}$$

where $\mathrm{Liq}_{s,\rho}$ is the liquidation value of market s in period ρ, and $\mathrm{Price}_{s,\rho}$ is the 4-day average market price before liquidation. The sample size for this regression is 40, since we have no previous liquidation value for securities in the first period.

[11] Collecting information from prices in this pilot deployment would have been difficult to do in real time because of volatility, especially when using LMSR as the market maker. One could combat this by increasing the loss tolerance of LMSR, effectively performing smoothing with the market maker itself.

The significance of the previous liquidation value at the $p = 0.05$ level disappears when price is included in the linear model above, showing that previous liquidation value provides no additional information beyond price in this regression. This result is robust with respect to the choice of how price is smoothed. For $\mathrm{Price}_{s,\rho}$ equal to the 4-day average price, we find that β_2 is the only statistically significant coefficient (at $p < 0.05$). The results are qualitatively unchanged when adding random effects controls for per-period and per-stock variations.[12]

6.3.2 Insider trading/sources of information

Having shown that prices are predictive, we would like to know where the new information is coming from. While this is sometimes done by looking at the trade prices of different types of traders, that methodology is more appropriate for markets with limit orders. In a market-maker mediated market, it makes more sense to look at the directions of trades. Consider a single trade on the IRM: either this trade moves a price toward the corresponding instructor's future rating, or away from it. By examining the set of all IRM trades in this manner, we can get an idea of the information revealed by groups of traders. We would expect that in-class traders, since they determine instructor ratings, would provide more information than out-of-class traders. Indeed, in-class traders traded toward the future liquidation 53.9% of the time (95% confidence interval 53.0% to 54.8%), while out-of-class traders traded toward the future liquidation only 52.5% of the time (95% confidence interval 52.3% to 52.8%). The difference is statistically significant ($p = 0.015$). This tells us that in-class traders brought more information to the IRM. However, we know that previous liquidations are a good predictor of future liquidations; how many of these trades are simply based on old information?

[12] We added α_s and α_ρ as random effects, assumed to be normally distributed with mean zero, representing random per-stock and per-period variations respectively.

To determine which traders bring *new* information to the IRM, we can examine trades that occur at prices between the previous liquidation price and the future liquidation price (see Figure 6.3). In such situations, if insiders are truly the sources of fresh information, we would expect them to trade more in the direction of the future liquidation, while others trade more in the direction of the last liquidation. Examining the data confirms this hypothesis. In situations where the execution price was in between the last liquidation value and the next liquidation value, in-class traders traded toward the future liquidation 53.5% of the time (95% confidence interval 51.7% to 55.2%, so also significantly more than 50% of the time). Out-of-class traders favored the previous liquidation, trading toward the future liquidation only 47.7% of the time (95% confidence interval 47.0% to 48.4%, so significantly less than 50% of the time). The difference is, of course, statistically significant ($p = 7.1 \times 10^{-7}$). This is compelling evidence that out-of-class traders were mostly trading on old information, and the markets serve to disseminate the inside information of in-class traders to the world, and provide feedback to instructors in doing so.

6.3.3 Qualitative features of prices

Figure 6.2 shows the traded prices and liquidation values for a selection of markets (traders saw this information in a similar format, although they were not aware of which market maker was used in which period). The figures highlight certain interesting qualitative features of the price processes. First is the effect of volatility, which may make the instantaneous price a less useful piece of information for the instructor at any point in time than a smoothed version of the price, as discussed earlier. An alternative would be to use a less volatile market making scheme (different parameters or a different algorithm). In fact, volatility does appear to be significantly less for markets using BMM [10], which we discuss later; this lower volatility does not come with any loss in predictive ability of the resulting prices. Second, prices often move towards the previous liquidation right after that value is revealed, without moving all the way there. We see this behavior clearly in

Course 1, especially during periods four and five. Two of the IRM classes always liquidated at a value of 100, and in these classes the security prices slowly converged to 100; the slow rate of convergence is probably because the incentive to buy a security near 100 even given a sure liquidation at 100 is very small.

Summarizing the evidence: the markets are useful and predictive, providing information on future ratings that instructors will receive. We find strong evidence that most of the useful new information is added by in-class traders. Meanwhile it appears that out-of-class traders help in providing market stability by trading toward previous liquidation values, offsetting large noise trades.

6.4 Trading and Rating Behavior

One of the unique benefits of the IRM is that we have data on both the trading and rating behavior of the participants. This allows us to explore issues in market manipulation and trader behavior in ways that were previously not possible. For example, we present evidence not only that the IRM succeeded in its primary goal of providing dynamic predictive information on how a professor is doing, but also that this information was mostly provided by students enrolled in the class. Here we look more deeply into the behavior of users.

6.4.1 Insiders, Manipulation, and Collusion

Traders who had rating credentials in a market (in-class traders, or insiders) could both trade in the market and affect the dividend through their rating. Therefore, not only did they have better information on the professor being traded than other participants, but they had the opportunity to explicitly choose how to rate the professor based on their position in the stock. We define "manipulation" as situations in which students provide a rating they do not truly believe in order to

maximize their profits from the IRM. There were plenty of opportunities for manipulation: several classes had only 3-5 raters, and information on how many ratings contributed to a particular liquidation value was made easily accessible on the trading interface (along with the prior liquidation values), allowing raters to estimate their impact on a market's liquidation. Of course, knowing if manipulation actually occurred is difficult, but we provide several pieces of data that make the case that there was little manipulation.

6.4.2 IRM Ratings Were Not Manipulated

Do IRM student ratings correspond well with what they actually thought of the class? Since seven of the ten classes were in the Computer Science Department, we were able to measure the correlation of IRM ratings with the official end-of-semester student evaluations.[13] We averaged the ratings and prices of periods 3-5 in the IRMS. The coefficient of correlation of the IRM ratings to the official ratings for these 7 classes was 0.86, and the coefficient of correlation of the IRM prices averaged over these periods with the official ratings was 0.75. The strong correlation between IRM ratings and official ratings validates the usefulness of our markets in terms of a real benchmark that is "outside the system," and also indicates that students were rating honestly in the IRMs, and that we do not need to worry about experiment-wide misbehavior.

6.4.3 Little Evidence For Manipulation in IRM Prices

We considered any group of raters who both gave the same rating and made a significant amount of money (1000 virtual currency each) trading the associated security during a given period as candidates for having colluded. We observed collusive behavior in course 2 during period 4. A group

[13]To protect instructor confidentiality, we gave the IRM ratings and prices to the Department Head, who ran the correlations against end-of-semester student evaluations.

of 3 raters together made about 9000 in virtual currency by selling course 2's security and rating the course low. These 3 students controlled 20% of the liquidation value; since most liquidations were between 60 and 100, this was enough for the manipulators to reduce the security's price significantly below the market's expectation. This liquidation was Course 2's lowest, although it is not apparent from the liquidations alone that manipulation was involved (see Figure 6.2). Pairs of raters made somewhat smaller amounts of virtual currency in several other markets, but it is not clear if intentional manipulation was involved.

More surprising than the observed manipulation in the IRM was its relative scarcity. Most markets did not see any successful collusion based on the criteria that raters both made money and rated together during a given period. Perhaps students did not understand the opportunities for manipulation, or perhaps giving accurate feedback was more important than winning prizes for some raters.

We note that the potential for manipulation was not limited to groups or to simple rating manipulation. Examining the trading records of raters who made more than 1000 virtual currency trading in a given security during a given period, however, seems to indicate that such opportunities where not successfully exploited; we do not observe significant shifts in trading activity by these raters. Manipulation by non-raters seems significantly less likely given the relative lack of information and influence.

6.4.4 Trading Strategies and Profits

Traders varied wildly in their activity levels, strategies, and apparent rationality. While some amassed large quantities of virtual currency by frequently monitoring for mispriced securities, others seemed eager to cause as much havoc as possible while divesting themselves of their entire initial capital. Figure 6.4 shows the number of trades and the number of shares traded per user and

per period, grouped by the user's account value at the end of that period. We see that the defining feature of the most successful traders was activity; while they did trade more shares overall, they did so in almost twice as many transactions as the less successful traders. The worst traders also stood out, making a moderate number of massive trades.

6.5 Effects of Microstructure

The IRM is a powerful platform for testing the effects of different microstructures on price dynamics. We tested two different market-making algorithms. Brahma et al. (2012) develop a Bayesian Market Maker (BMM) (building on [32, 29, 28]), and compare with Hanson's Logarithmic Scoring Rule (LMSR) market maker [52]. In short, 10 minute trading sessions, they find that BMM can offer comparatively higher price stability and smaller spreads than LMSR without suffering losses in expectation. On the flip side, LMSR comes with a strong loss bound, while BMM may occasionally take high losses. We provide additional evidence for these conclusions in a more realistic long-running experiment, with ample opportunities for strategizing and manipulation.

6.5.1 Description of LMSR and BMM

LMSR is a purely inventory-based market maker. For a single security with payoff in $[0, 1]$ (as noted above, the cost/price function and loss-bound is exactly equivalent to the case of binary payoffs), the spot price at an inventory level q_t is given by $p(q_t) = e^{q_t/b}/(1+e^{q_t/b})$, where b is a positive parameter, and the cost for a change Q in the inventory is $C(Q; q_t) = b \ln \left[(1 + e^{(q_t+Q)/b})/(1 + e^{q_t/b})\right]$. Thus, for a buy or sell order of size Q at an inventory level q_t, the market maker quotes a volume weighted average price (VWAP) $|C(Q; q_t)/Q|$ where Q is positive for buys and negative for sells. The inventory is updated to $(q_t + Q)$ only if the trade is accepted, and the market maker waits for

the next order. Note that, in our implementation, all these quantities are multiplied by 100 to keep the prices in the range $[0, 100]$.

BMM, an information-based market maker, maintains a Gaussian belief distribution $N(\mu_t, \sigma_t^2)$ for the value of the market; the spot price is equal to the mean belief μ_t. The underlying assumption is that trader valuations are normally distributed around the true value V. A fixed trade size parameter (α) determines quoted prices: every buy/sell order of size Q is imagined to be a sequence of $k = \lceil Q/\alpha \rceil$ independent mini-orders of sizes $\{\alpha_i\}_{i=1}^{k}$ which are all α except possibly the last one. The market maker then quotes a VWAP and updates its state depending on the trader's decision (acceptance/cancellation); the precise updates are non-trivial, but efficient (see [10] for details). Even though the Bayesian belief updates converge, BMM can adapt to market shocks, where the market's value changes dramatically. To do so, BMM maintains a "consistency index" that quantifies how consistent the trades in a window of size W are with the current belief. When trades are inconsistent with the belief, the belief variance rapidly increases, allowing quick adaptation.

LMSR is simple and loss bounded: the loss is at most $b \ln 2$. Moreover, being inventory-based, it is difficult to manipulate; and, assuming rational traders who learn consistently from prices, an LMSR-mediated market converges to a rational expectations equilibrium. Though the loss is bounded, LMSR does typically run at non-zero loss. One drawback is that a single parameter b controls various aspects of the market such as the loss-bound, liquidity, and adaptivity; therefore, achieving a trade-off can be difficult. Moreover, Brahma et al. find (and we confirm here) that if the beliefs of the trading population do not converge, prices can be very unstable. BMM, on the other hand, is not loss-bounded but makes much less loss in expectation while providing an equally liquid market. Moreover, in the absence of market shocks, BMM's belief (and hence the spot price) converges owing to a monotonically decreasing variance, even if the traders maintain heterogeneous valuations.

	Periods	Avg profit	Max loss	Std	Liq dev
LMSR	35	1341.67	-5298.58	8.6	16.9
BMM	15	8273.13	-13763.40	3.0	9.6

Table 6.2: Overview of statistics for LMSR and BMM, showing average profit, max loss, the standard deviation of prices, and deviation of prices from the market's liquidation value.

6.5.2 Exploiting BMM

The variance of BMM's belief distribution determines its spread. A simple implementation can be manipulated by artificially tightening the spread, with a sequence of alternating small orders followed by a large order to exploit the low spread. To avoid this, we perform inference on BMM's variance parameter only once for each trader unless an intervening trader also places an order. This idea can be easily extended to pairs of colluding traders, but could suffer from Sybil attacks. Such manipulation strategies are highly non-obvious, and, further, we limit traders to a single account by requiring an institute email address for authentication. Ultimately, exploitation of BMM did not become an issue.

6.5.3 Comparison of Market Makers

We confirm the major findings of Brahma et al.'s previous comparison of BMM to LMSR. In essence, BMM offers more stable prices (see Figure 6.2 and Table 6.2), while making higher profits and maintaining lower spreads (see Table 6.2). We set LMSR's b parameter to 125; by increasing b one can get lower spreads and more stability, but at the expense of other tradeoffs. For example, the b parameter of LMSR is an explicit market subsidy, increasing not only the loss bound but the expected loss of the market maker in reaching a given equilibrium price. Since LMSR actually *made* money on average[14], this could be an acceptable tradeoff. BMM already made more money

[14]This does not contradict our previous results concerning the information content of prices. While LMSR only makes money when the first market price is more informative than the last, we find that an average of the last few

on average in the IRM, however, and so comparing volatility is quite reasonable. It is interesting to note that the median trader made money when trading with LMSR, although the mean was below 0, whereas both the mean and median traders lost money with BMM. The volatility of prices and the deviation from the future liquidation value suggest that not only was the BMM price more stable than that of LMSR, it also provided a better estimate of the liquidation value. These results are robust and significant when regressing with per-security random effects.

6.6 Discussion

The Instructor Rating Markets are a field experiment in the space of agent-mediated prediction markets that incentivize humans to truthfully reveal their information, and, in doing so, provide useful dynamic feedback (in this case to instructors). The IRMs are a platform for studying the behavior of insiders and potentially manipulative participants in unprecedented depth. Many of the questions we study here would not be amenable to either short, intensely controlled lab experiments, or to study based on the data from prediction markets deployed "in the wild." Perhaps the most fruitful questions to pursue in similar, medium-sized experiments in the future revolve around manipulation and the role of market design (including the design of automated market mediators) in achieving good information dissemination properties.

prices is more informative still. Setting the last price to this average and computing the hypothetical profit, LMSR loses money on average and in most cases.

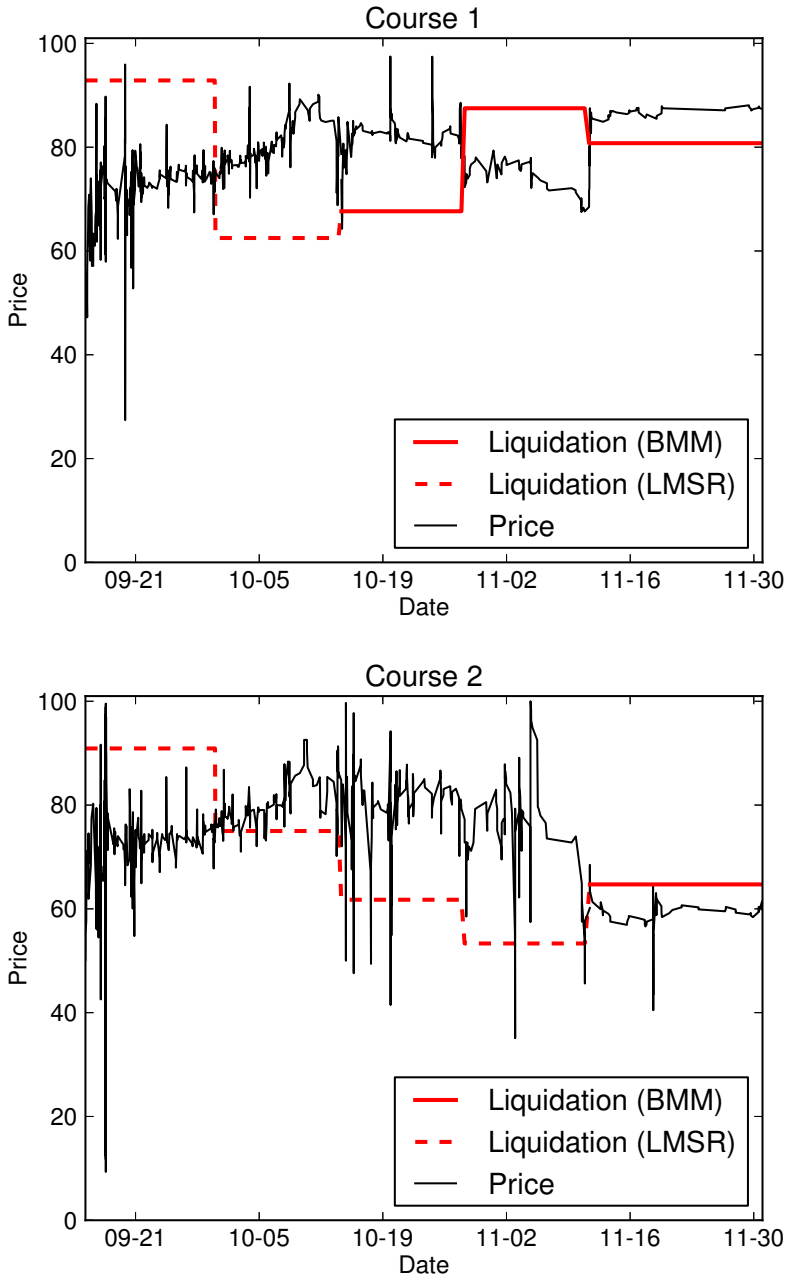

Figure 6.2: Price charts and liquidation values for selected markets, with line style indicating the market making algorithm. Each trade is plotted according to its transacted price with no smoothing.

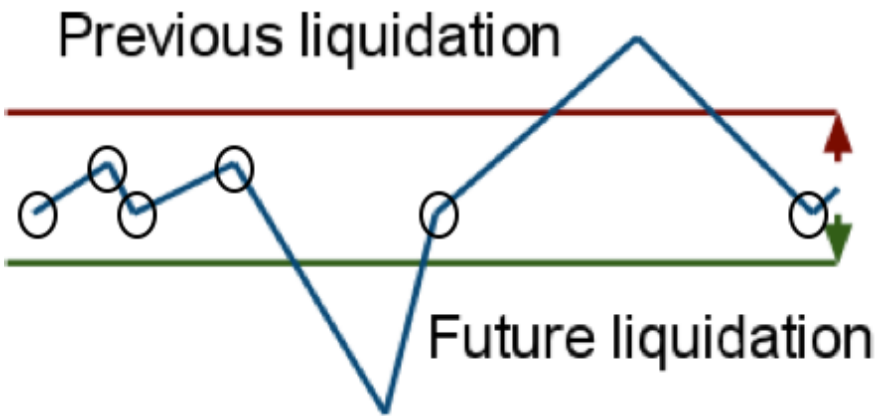

Figure 6.3: Methodology for determining who brings new information to the market. Trades that occur between the previous and future liquidation prices are circled, and move the price either in the direction of the future liquidation (new information) or the past liquidation (old information).

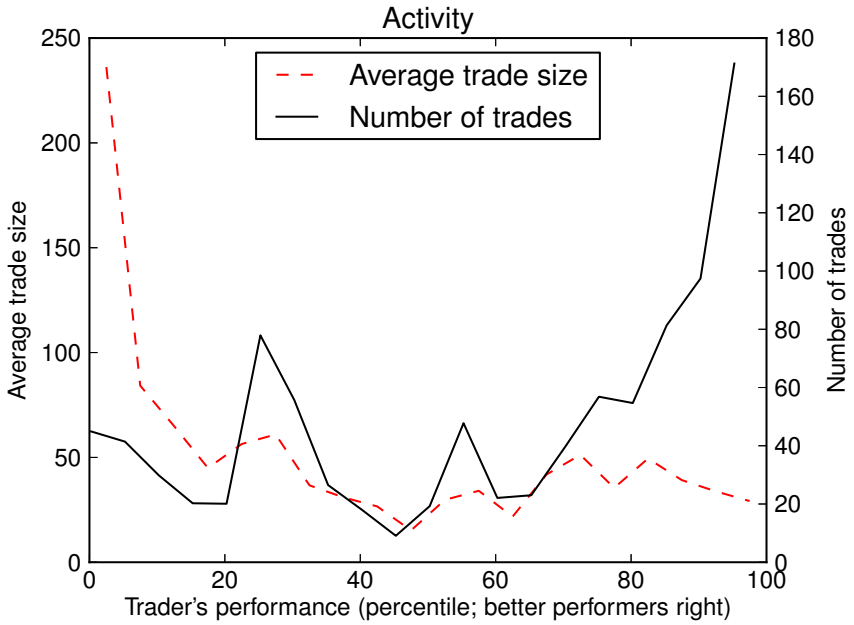

Figure 6.4: Successful traders made many smaller trades.

Chapter 7

Inferring Incentives from Participation

7.1 Introduction

Collective intelligence requires the aggregation of information and opinion from a diverse population. The process of consensus building relies not only on eloquent rhetoric and argumentation to convince, but also on the dynamics of participation. Online collective intelligence processes rely on voluntary participation, and the choice of where to exert effort or whether to participate at all can be influenced by existing consensus, real or illusory. How does this voluntary participation affect the outcome of collective decision making? Our goal is to characterize the dynamics of individual decision making and its impact on consensus formation.

In order to do this, we develop a novel model for participation in collective intelligence processes. The model is driven by two latent variables, one a feature of the conversation as a whole and the other a feature of each individual user. The first, which we call the *response parameter*, characterizes a conversation on a spectrum from cooperative to adversarial. Consider a forum focused on baseball. An example of a cooperative conversation would be one where "like attracts like" – for example, bursts of activity from users interested in the history of the game interspersed with bursts of activity from those interested in statistics and sabermetrics. An adversarial conversation would

be one where participants went back-and-forth, arguing different sides of an issue (for example, proponents of sabermetrics arguing with those who prefer traditional measures of performance).

The second critical latent variable, which we dub *interpinion* (a portmanteau of interest and opinion), quantifies a user's preferences with respect to a conversation. We intentionally coin a new term to escape the connotations associated with words like "opinion", "interest", or "tone" in common language. Interpinion can take on different flavors of these terms in different contexts. Interpinion interacts with the response parameter as follows: In cooperative conversations, having the same interpinion as the conversation increases the probability that a user will participate (in this case, the latent variable can be interpreted more as "interest"), while in adversarial conversations, having a different interpinion from the current state of the conversation increases the probability that a user will participate (and the latent variable can be interpreted more as "opinion").[15] Figure 7.1 depicts this relationship graphically.

We have described two modes of activity for online forums: cooperative, where like attracts like, and adversarial, where opposites attract. These can be interspersed and nested in real discussions, but here we are interested only in the *principal* mode. To build intuition for how it is possible to infer interpinion from only participation in a discussion, consider a simple example where discussions are either entirely cooperative or entirely adversarial, and users likewise have binary interpinions. We observe the sequence "ABAC" of contributors to a discussion (Alice, Bob, Alice, Carol). If the discussion is adversarial, we expect Alice to participate next, since based on their participation times Carol likely disagrees with Alice and agrees with Bob. If the discussion is cooperative, the users likely all have the same interpinion, and therefore we would predict that either Alice or Bob will participate, each with probability 0.5. Observing "ABACA" is then a weak signal that the discussion is adversarial, while "ABACB" indicates a cooperative discussion. In either case, knowing something about the discussion type tells us something about the interpinions

[15] We would like to stress that these are merely linguistic interpretations of a variable; the model's mathematical semantics and empirical quality are, of course, independent of our interpretation.

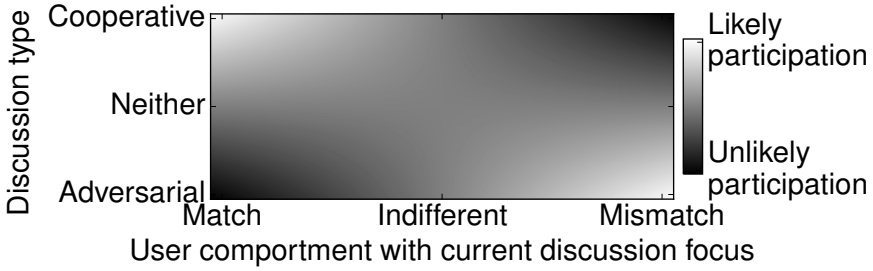

Figure 7.1: Under the generative model, users are more likely to participate if their interpinion matches that of a cooperative discussion, or differs from an adversarial discussion's. Short- and long-term effects are considered separately.

of participants. Over many observations, and with a probabilistic model which tolerates deviation from strict cooperative or adversarial discussion, we can learn about users while only observing *when they choose to participate*. We can fold this into a probabilistic model and simultaneously infer the adversarial/cooperative nature of the discussion and the interpinions of participants, and in this chapter we develop and validate precisely such a model. Our validation is on two datasets: Wikipedia talk pages and the anti-vaccine blog Vactruth.com. In both cases we (probabilistically) predict the next user to participate given the history of a conversation, and demonstrate that our model outperforms several baselines.

This analysis touches on both product and process incentives. Discussing the content of an article can function as a process incentive for participation, i.e. through social interaction. On the other hand, users who have pre-existing product incentives relating to the content of an article also have an incentive to steer discussions toward their beliefs. Incentives may depend on the conversation type, with purely cooperative discussions likely driven by process incentives, and conversations with an adversarial component experiencing both product and process incentives.

7.1.1 Related work

There is continuing academic interest in the birth, death, and health of online communities. Danescu-Niculescu-Mizil et al. [27] study linguistic predictors of user retention, finding that users adapt early on but eventually are left behind by the evolution of community norms. Dror et al. [35] study predictors of churn, finding that users on question and answer websites who answer more questions and who receive more positive social feedback are likely to remain active. Choi et al. [23] find that the type of feedback given to new participants on Wikipedia has a large impact on their retention. Das and Lavoie [31] study reinforcement effects of social feedback on Reddit. Rowe [112] studies churn prediction, identifying "lifecycle trajectories" and finding that users diverge from community norms before leaving. Kittur et al. [67] study controversy and explicit disagreements (reverts) on Wikipedia. Using these explicit disagreements, Das and Lavoie [30] identify the topics and points of view expressed on Wikipedia.

Much of the theoretical interest in discussions and bias is in statistical physics. However, Munie and Shoham [93] present a simple but interesting game theoretic formulation of users taking turns influencing the product of a collective intelligence process (wikis, ratings, etc.). Our interest is primarily empirical, and we are interested in a more nuanced description of user behavior which is unlikely to be as theoretically tractable. From the statistical physics literature, Martins [84] studies continuous opinions and discrete actions, finding that allowing participants to change their opinions significantly reduces the incidence of extremism. Martins and Kuba [85] show that adding "contrarians" can limit the formation of extremism in theoretical models of opinion formation. The main departure of this chapter from this literature is the empirical focus, both the evaluation in terms of predictivity on real data and in the characterization of observed discussions.

Finally, there is an important line of literature from psychology and sociology which focuses on the "spiral of silence" [98]. This is the idea that individuals may choose not to speak out because of

a perceived risk of ostracism. Our work is in this spirit, attempting to operationalize and quantify effects which amplify or degrade a user's propensity to express opinions online. Recent work of particular interest is Woong Yun and Park [136], experimentally testing the willingness of subjects to post messages where their views would be in the minority.

7.2 Model

We model user activity and turnover in a discussion as a function of individual *interpinion* and the discussion's *response function*. A discussion is a stream of contributions from users, each with their own static interpinion. When a user participates, she contributes her interpinion to the discussion. This contribution then affects the activity and retention of other users. New users join over time, and existing users eventually become inactive. We consider both short-term effects of the previous contribution on who the next contributor will be, and also long-term effects of average interpinion on user retention.

In a generative Bayesian description, at step s we have a number U_s of users, each with weight $w_{u,s} \in [0, 1]$ (the weight is used to model survival, with lower weights being associated with a higher chance that the user has left the discussion). Each user $u \in \{1, \ldots, U_s\}$, has interpinion $t_u \in [-1, 1]$. After each post, all users draw a candidate inter-arrival time from an exponential distribution:

$$b_{u,s} \sim \text{exponential}\left(w_{u,s}(R(t_u, t_{a_{s-1}}) + \gamma)\right) \tag{7.1}$$

Here, $t_{a_{s-1}}$ is the interpinion of the previous user who participated (having arrived at time $s-1$), R is a response function that depends on the characteristics of the conversation, and γ is a term that is constant across all users that can be thought of as a participation probability bias.[16] The interpretation here is that the parameters of the exponential distribution in Equation 7.1 tell us the

[16] We estimate γ in the Bayesian framework as well, with prior $\gamma \sim \text{gamma}(1, 1)$.

probabilities of each user being the next to participate (in a joint distribution with the timing of when that participation will happen). One of these participation events (the one with the smallest $b_{u,s}$) is actually realized, it becomes the next contribution to the conversation, and then the parameters for who will participate next get reset depending on the interpinion of that participant and the response function of the conversation. Thus we have a heterogeneous (the distribution of inter-arrival times varies) and non-anonymous Poisson process (not simply counting events, but also modeling the type of event).

Response function A user's inter-arrival time distribution is determined by the function R, which depends on the user's interpinion and that of the previously active user. This function describes how users respond to interpinion: are they more active when their interpinion disagrees with the previously active user, or are they instead more active when they agree? This is determined by a response parameter r:

$$r \sim \text{beta}(0.5, 0.5) \tag{7.2}$$

$$R(t, \tau) := t \cdot \tau \cdot (2r - 1) + 1 \tag{7.3}$$

If $r > 0.5$, agreement ($t \cdot \tau > 0$) increases the response function. If $r < 0.5$, agreement instead decreases it. We refer to the situation where $r > 0.5$ as *cooperative*, with users who agree more likely to reply, and to situations where $r < 0.5$ as *adversarial*, with users being more active if they *disagree* with the previous interpinion. When $r = 0.5$, interpinion does not affect user activity.

$R(t, \tau)$ ranges from 0 to 2. The scale of this function is arbitrary: multiplying the response function and bias γ by a constant factor has no effect on the distribution of the lowest inter-arrival time a_s. Larger values of $R(t_u, t_{a_{s-1}})$ result in a higher parameter for the exponential function, lower inter-arrival times $b_{u,s}$ and a higher probability of arriving first.

Survival Weights $w_{u,s} \in [0,1]$ model users being less likely to come back after periods of inactivity. Decay depends on a longer-term average of the interpinion of a discussion.[17] Specifically, we introduce an averaging parameter $\alpha \sim \text{beta}(1.5, 10.5)$. Averaged interpinion at the beginning of round s, θ_s, is:

$$\theta_s := (1-\alpha)\theta_{s-1} + \alpha \cdot t_{a_{s-1}} \tag{7.4}$$

Higher values of alpha correspond to more reactive discussions, with recent interpinion having more weight. Decay depends on θ_s, the user's interpinion t_u, a second response parameter ρ, and a bias δ:

$$\rho \sim \text{beta}(0.5, 0.5)$$

$$\delta \sim \text{gamma}(2, 4)$$

$$P(t, \tau) := \frac{t \cdot \tau \cdot (2\rho - 1) + 1 + \delta}{2 + \delta} \tag{7.5}$$

$P(t,\tau)$ ranges from 0 to 1, with a fraction $2/(2+\delta)$ depending on interpinion and the survival response parameter ρ. We can now define user weight decay. Users do not post twice in a row: $w_{u,s} := w'_{u,s} I(u \neq a_{s-1})$

$$w'_{u,s} := [P(t_u, \theta_{s-1})]^{b_{a_{s-1},s-1}} \cdot \begin{cases} 1 & u = a_{s-1} \\ w'_{u,s-1} & \text{else} \end{cases} \tag{7.6}$$

The active user a_s has their weight "reset" to 1 at the beginning of round s, while other users continue their weight decay from the previous round's weight.

P defines the decay *rate*, while the value of the lowest inter-arrival time determines how much "time" passes between rounds. This captures the intuition that a third user should be no more likely to leave a discussion if she misses a rapid exchange of messages between two other users.

[17] Activity could also be modeled this way. We fit a model of this kind to data, but the posterior nearly always indicated that only the previous user's interpinion was important.

New users We model new users joining a discussion by having a new user draw an inter-arrival time during every round. If this draw is lower than any other inter-arrival time, then the user is added to the active set and is thereafter treated like every other. If an existing user draws a lower inter-arrival time, the user leaves, but other candidate new users draw inter-arrival times in subsequent rounds. New users have their own interpinion, again drawn uniformly: $t_{N,s} \sim \text{uniform}(-1, 1)$. New user inter-arrival draws have a separate bias η (i.e. there is some variability in the rate of new user arrivals across conversations):

$$\eta \sim \text{gamma}(1, 1) \tag{7.7}$$

$$b_{N,s} \sim \text{exponential}(R(t_{N,s}, t_{a_{s-1}}) + \eta) \tag{7.8}$$

If $a_s = N$, $U_{s+1} = U_s + 1$, $t_{U_s+1} = t_{N,s}$, and $w'_{U_s+1,s} = 1$. In this case we will abuse notation slightly by using a_s to refer to user $U_s + 1$. Otherwise, if an existing user draws the lowest inter-arrival time, $U_{s+1} = U_s$ and the interpinion $t_{N,s}$ has no further effect.

Initial conditions The generative model begins with $U_1 = 0$. With no other inter-arrival times drawn, a new user joins during round 1. Unable to participate again until the third round, a second new user joins during round 2. Participation is thereafter probabilistic. Initial page interpinion is set by the first user: $\theta_2 := t_1$ and $\theta_1 := 0$.

7.2.1 Inference

Our goal is to infer the posterior distribution of the model parameters given observed data. We consider the use of real time as an observable, then develop an efficient description of the model.

Real time vs. model time

Observed data unquestionably include the sequence of users who have participated (user 1, 2, 1, 3, ...); that is, the values of a_s for each observed round s are known. However, there is a question of the role of the physical time elapsed between contributions to a discussion. On one hand, we could make maximal use of observed timings, designating the value of the lowest inter-arrival time $b_{a_s,s}$ as a second observed variable. However, this would necessitate significant normalization to account for cycles in activity (day vs. night in the more active time zones, weekends, holidays, etc.). Without normalization or significantly more parameters, the minimum of several exponentially distributed random variables is a poor model for this heterogeneous activity, so we leave inter-arrival times as latent.

On the other hand, we define a *rate* of decay in (7.5), which can much more comfortably operate on physical time: intuitively, the more physical time users spend away from a discussion, the less likely they are to return. Thus we replace—when performing inference on real data—the exponent $b_{a_{s-1},s-1}$ in (7.6) with a scaled version of real time, introducing a scale factor $\lambda \sim \mathrm{gamma}(1,1)$ which divides physical time in order to produce the exponent in (7.6). When performing simulations based on the inferred model parameters below, we maintain the original model ((7.6) as written) but scale the exponent to approximately match average time elapsed per contribution on real data.

Model likelihood

We perform inference using the No U-Turn Sampler (NUTS) [58] implementation provided by Stan [123]. Stan performs automatic differentiation on a log-likelihood specification, developed below, providing gradient information to NUTS.

Since we have substituted real time for the value of the lowest inter-arrival time in (7.6) for the purposes of inference on real data, we can integrate out the values of all inter-arrival times $b_{\cdot,\cdot}$:

$$W(u,s) := \begin{cases} w_{u,s}(R(t_u, t_{a_{s-1}}) + \gamma) & u \leq U_s \\ R(t_{N,s}, t_{a_{s-1}}) + \eta & u = U_s + 1 \end{cases} \tag{7.9}$$

$$p(a_s = u | \ldots) = \frac{W(u,s)}{\sum_{u' \in \{1,\ldots,U_s+1\}} W(u',s)} \tag{7.10}$$

This follows from a well-known property of the exponential distribution, that for a set of random variables $z_i \sim \text{exponential}(Z_i)$, $p(j = \text{argmin}_i(z_i)) = Z_j / \sum_i Z_i$. While this reduces the number of parameters which must be sampled significantly, we are still left with one nuisance parameter every time an existing user participates: $t_{N,s}$ for each round s where $a_s \neq N$. We integrate it out:

$$\begin{aligned} p(a_s = u, a_s \neq N | \ldots) &= \frac{1}{2} \int_{-1}^{1} \frac{W(u,s)}{R(t_{N,s}, t_{a_{s-1}}) + \eta + \sum_{u' \in \{1,\ldots,U_s\}} W(u',s)} dt_{N,s} \\ &= \frac{W(u,s) \log\left(\frac{t_{a_{s-1}} \cdot (2r-1) + 1 + \eta + \sum_{u' \in \{1,\ldots,U_s\}} W(u',s)}{-t_{a_{s-1}} \cdot (2r-1) + 1 + \eta + \sum_{u' \in \{1,\ldots,U_s\}} W(u',s)} \right)}{2 t_{a_{s-1}} \cdot (2r-1)} \end{aligned}$$

If a new user joins, we do need to reason about their interpinion, as it then affects other users' activity and retention. In that case, we fall back to (7.10). In combination with with the likelihoods from the priors, this allows us to sample from the posterior. Sampling is in terms of an interpinion for each user, $t_1, t_2, \ldots$, and the global parameters (activity biases η and γ, averaging parameter α, response parameters r and ρ, survival bias δ). We draw 100 samples using NUTS, after 100 burn-in samples.

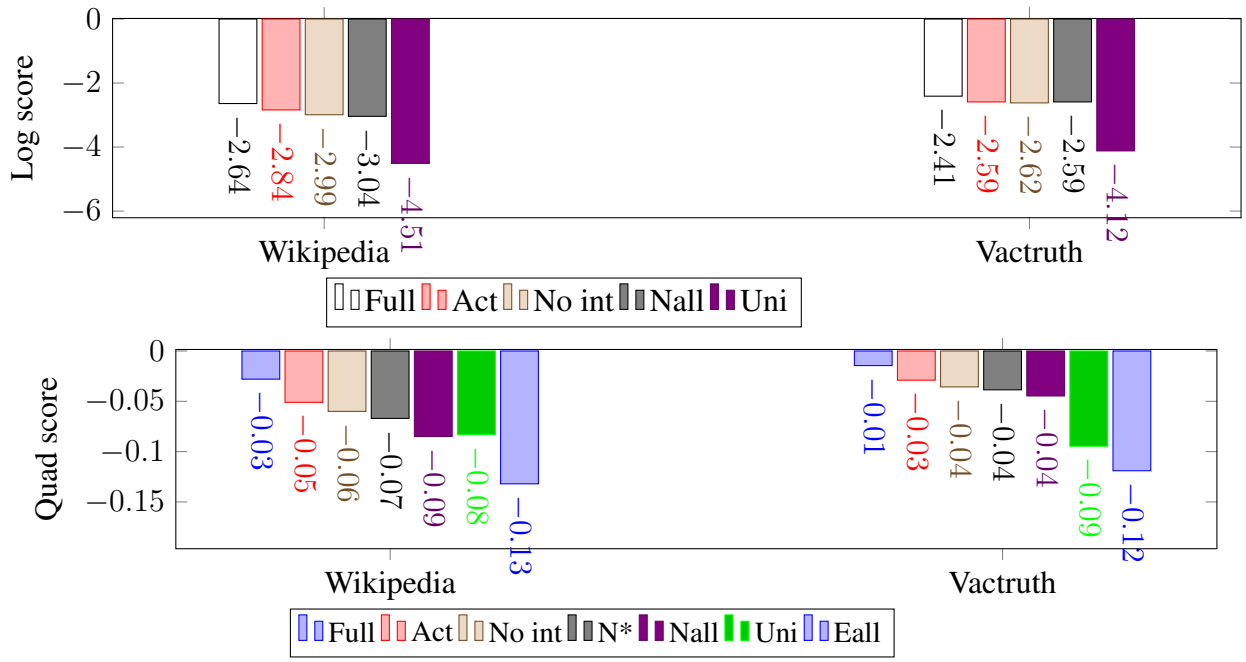

Figure 7.2: Results of model comparisons. Evaluation is in terms of logarithmic and quadratic scoring rules on a held-out user prediction task. Most differences are statistically significant, with Wilcoxon signed-rank between adjacent methods (when sorted by score) yielding $p < 0.01$ for Wikipedia data (exception: N* and No int quadratic, $p = 0.03$), and $p < 0.05$ for Vactruth data (exceptions: Act and Nall logarithmic are indistinguishable; N* and No int quadratic are indistinguishable, as are N* and Nall; N* and Act quadratic are indistinguishable, but comparing N* and Full yields $p = 0.02$). We omit error bars, as they are inappropriate for related samples (via the consistent test set).

7.3 Evaluation

We evaluate our model on a specific, *a priori* difficult task: For a given conversation, predict the next user to participate given the amount of time elapsed before that participation (e.g. the models are told that a user will participate in 5 minutes and asked to predict which one). As this is a probabilistic task, we evaluate performance in terms of the distance between predicted and true distributions.

7.3.1 Data

Wikipedia We collect participation and timing information from the 4000 most popular English Wikipedia talk pages, from a February 2012 database dump. We replace sequential contributions from the same user with a single contribution, using the timestamp of the first contribution in the sequence. We also remove edits marked as minor. After pre-processing, we split conversations into blocks of no more than 1000 contributions, resulting in about 4700 extracts ranging in size from 100 to 1000 posts, the mean being 610. For each of the extracts, we hold out the last contribution and use the models and baselines to predict a distribution over the contributing user (models are trained 4700 times, each giving a performance signal from a held-out example).

Vactruth This dataset consists of 30000 comments on blog posts from Vactruth.com, an anti-vaccine blog. Treating the comments on each blog post as a separate conversation yields an average new-user rate of 72% per post, with about 1.6 contributions per user, not enough to test a user modeling algorithm (the Wikipedia extracts have about 7 contributions per user). Instead, we group blog comments by time in blocks of 200 (ordered within blocks by post time), yielding 150 conversations with about 2.5 posts per user on average. Comments from different blog posts are interspersed, the justification being that users are likely to visit multiple recent posts, and are likely to hold similar opinions on each. We again remove adjacent contributions by the same user, and hold out the last contribution in each block for testing. The small number of posts per user makes this second task far more challenging, as models can not rely on *confident* inferences about any particular user. Nonetheless, reasoning about interpinion turns out to be beneficial even on this dataset.

7.3.2 Models and baselines

Uniform distribution (Uni) A zero-intelligence reference point, predicting uniform probabilities.

Empirical user distributions (Eall, Nall, N*) A simple approach is to predict the empirical distribution of past users (Eall). Because we have pre-processed data such that no user contributes twice in a row, we re-normalize the predicted distribution so that it does not place probability on impossibilities. This ignores new users, so we have also considered a baseline which, upon seeing a new user, increments a count not only for that user (as in Eall) but also a "new user" count. Since there is an expectation of user turnover, we have also considered "past N" user activity models, which predict the empirical distribution of behavior over some limited number of previous contributions. We report only the best-performing model (40 back on Wikipedia, 60 on Vactruth), dubbed N*. N* and Eall, sometimes predicting probability 0, are omitted from logarithmic evaluations.

Exponential activity decay (No interpinion / No int) The remaining models are fit using the model and inference described previously, simply fixing certain parameters to achieve simplifications of the full model. The first such baseline does not include interpinion: the interpinion response parameters r and ρ are set to 0.5, meaning that interpinion has no effect on activity or retention. This results in a simple baseline, with users' activity decaying exponentially in terms of real time but being reset each time they make a post (with a fixed new user arrival rate).

Interpinion affecting activity only (Act. only / Act) Again setting the "survival" response parameter ρ to 0.5 deterministically, we allow interpinion to affect only activity.

Full model / Full Finally, with no parameters fixed, the model applies interpinion not only to immediate activity predictions, but also to the rate at which users' weights decay.

7.3.3 Evaluation methodology

As we have models making probabilistic predictions, we estimate the distance between their predicted probability distributions and the true distribution of held-out examples. We estimate the expected logarithmic scoring rule, or negative cross-entropy, for each model on this task. Cross-entropy is an additive constant away from KL-divergence. However, it heavily penalizes some baselines, which occasionally assign probability zero to users who are still active. To accommodate this, we also evaluate the methods in terms of average quadratic scoring rule, an additive constant from negative squared Euclidean distance. See a detailed justification in Chakraborty et al. [16]. The logarithmic scoring rule is in $(-\infty, 0]$, quadratic $[-1, 1]$.

7.3.4 Results

Figure 7.2 shows the results of performance comparisons on the conversation extracts (predicting the held-out next user to participate). The full model, with interpinion affecting both activity and survival, scores the highest according to both the logarithmic scoring rule and the quadratic scoring rule, indicating a lower KL-divergence and squared Euclidean distance to the unobserved true distribution. The full model decreases KL-divergence by at least 13% on Wikipedia data over the empirical baseline[18], with significant contributions from reasoning about interpinion in terms of survival and activity. This result, as with most other comparisons (see the caption of Figure 7.2), is statistically significant, with $p < 0.01$ (Wikipedia) or $p < 0.05$ (Vactruth) using the Wilcoxon

[18]Cross-entropy is reduced to 2.64 from 3.04. "At least" due to the nature of probabilistic scoring, which cannot tell us whether the nature of our data are probabilistic, or if so what the entropy of the true distribution is. Therefore we make a worst-case improvement claim, assuming that the true distribution is deterministic.

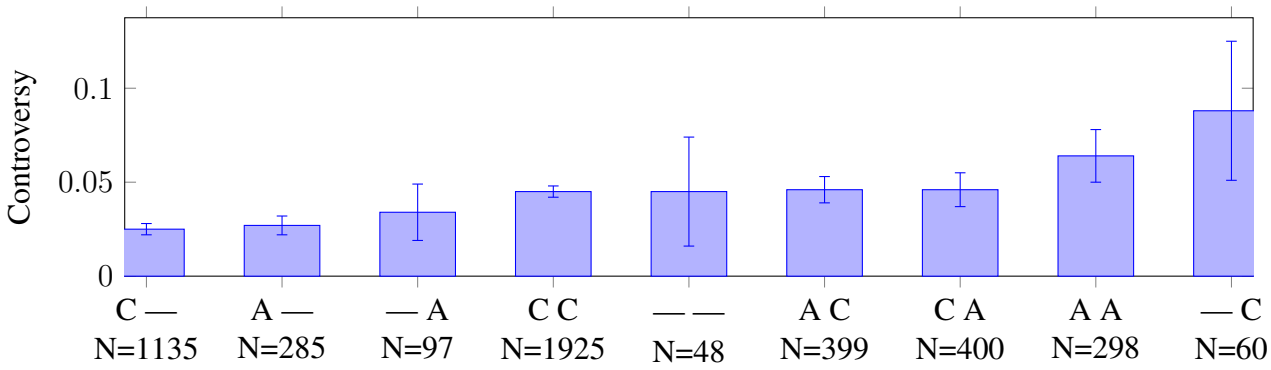

Figure 7.3: Mean controversy scores for Wikipedia talk pages. Discussions are cooperative (C), adversarial (A), or neither (—) in terms of the long-term survival parameter ρ (first letter) and short-term activity parameter r (second letter). Double-adversarial pages have higher controversy scores (excepting small-N "— C"). Error bars: 95% confidence.

signed-rank test. Especially on Vactruth, this performance is due in part to predicting when new users will arrive. However, we note that this in turn relies on determining when old users are no longer participating. The "No interpinion" ablated baseline has the same model of user decay and new user entry as the full model, and so improvements over this baseline are due to inferences about interpinion which have predictive power (i.e. improve performance on the test set). That this works even when there are very few per-user data is a strength of the model.

While attractive due to its lack of parametric assumptions, Wilcoxon tests the median rather than the mean of the samples. As the relationship between scoring rules and divergence functions involves an *expectation* [47], we also computed a paired t-test (testing the mean, but also assuming normality). Many of the logarithmic scoring rule t-test comparisons involving the Bayesian models on Wikipedia data are not significant, due to the presence of three extremely low scores (< -100). This highlights an interesting weakness of the model: the exponential weight decay is too aggressive, leading the Bayesian models to put inappropriately small probabilities on the re-appearance of users after significant periods of inactivity. Treating these three points as outliers puts the t-test results in line with the Wilcoxon tests on the logarithmic scores; the quadratic scores, being

bounded, are not sensitive to these points. The one notable exception to the correspondence between Wilcoxon and t-test results is the logarithmic Vactruth test between the full model and Nall, with $p = 0.15$ (vs. 10^{-7} from Wilcoxon), the difference likely due again to a non-Gaussian distribution of logarithmic scores. The corresponding comparison using quadratic scores is significant.

7.4 Identifying controversy

Having established the utility of our model for prediction, we now ask what it can tell us about the nature of online discussions, by classifying Wikipedia talk pages as cooperative or adversarial in terms of activity and survival, and relating these features to an independent measure of how controversial the Wikipedia page associated with that talk page is.

We first classify Wikipedia talk page extracts in terms of inferred model parameters. Pages with a high ρ, having fewer than 2.5% of their posterior samples below 0.5, we call cooperative in terms of survival. This means that users are more likely to remain part of the discussion if their interpinion agrees with the page's average interpinion θ. If fewer than 2.5% of posterior samples are above 0.5, we call survival adversarial, users being more likely to remain if they disagree with the current page average. Similarly, discussions can be cooperative or adversarial in terms of activity (r parameter). Pages matching these criteria typically have average parameter values very close to 0 or 1 (e.g. in cooperative survival discussions the mean inferred ρ is greater than 0.99).

We then analyze the mean *controversy score* of each page type, using a method based on the predicted controversial revision count of Kittur et al. [67] and normalized to be between 0 and 1. Figure 7.3 shows that conversations with both adversarial survival and adversarial activity are significantly more controversial than the other sufficiently sized categories (testing A A vs. C A with Mann-Whitney U, $p = 10^{-5}$).

7.5 Conclusions

How do discussions work to aggregate information and opinion in collective intelligence? We introduce a new model for participation in online communities that simultaneously reasons about a user-specific latent variable (measuring some notion of opinion or interest) and a conversation-specific one (measuring its adversarial or cooperative nature) and demonstrate its utility in predicting participation. Our model can be efficiently applied to real discussions, and the fact that it is language agnostic makes it broadly applicable. We also use the inferred values of latent variables to characterize.

Chapter 8

Simulations, Applications, and Conclusions

Having explained and validated models of process and product incentives by testing their predictivity on real data, this chapter revisits those models and asks what we can learn from their inferred parameters about social processes. Given a generative model with good predictivity, one approach is to simulate user actions in couterfactual situations. That is the approach I take in Section 8.1, where I explore the problem of starting a new community from scratch, given that there is initially a negative feedback loop of no users providing no social feedback and the lack of social feedback failing to retain users. However, one advantage of using interpretable models is that the parameters often summarize data in useful ways. In Section 8.2 I use the parameters learned from the topic and point of view model on Wikipedia to find potentially biased pages and explore the evolution of conflict on the encyclopedia. Section 8.3 briefly concludes with some thoughts on future directions for this area of research.

8.1 Seeding communities

How do you start a social news site from scratch? If no one is participating, new users will be turned off by a lack of activity and content. Reddit's founders faced this problem, and solved it by posting content from fake accounts for the first few weeks of Reddit's existence [90]. Social

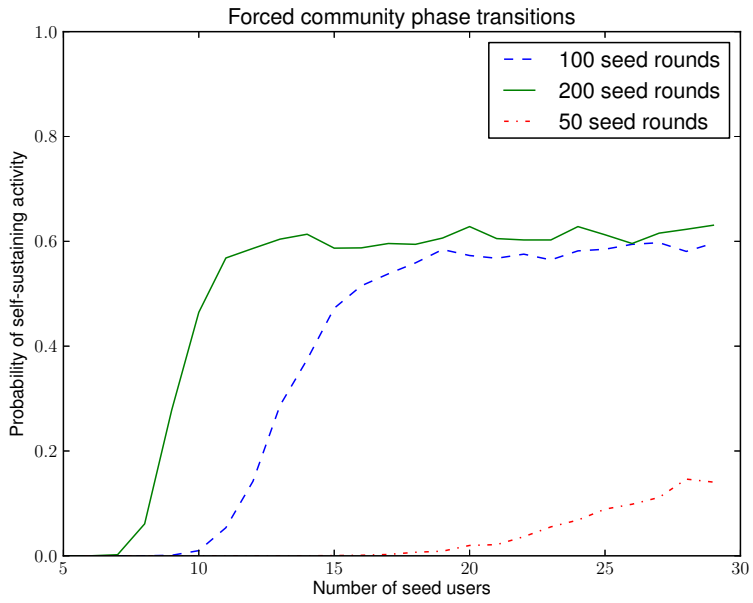

Figure 8.1: Seeding success probability depends heavily on the number of users doing the seeding, and on the time spent.

feedback exhibits a similar issue: without existing users to provide the feedback, new users will not receive enough interaction to keep them interested. Given that users go where the feedback is, how do you start a community from scratch? One answer is providing feedback through an initial set of "seed" participants, who may be sybils or paid participants. We simulate the effects of such seed participants on a group of agents, using the generative model of user behavior learned with the reinforcement model of Chapter 5 to better understand the dynamics.

Consider four communities A, B, C, and D with a common user base of 100 users. For simplicity, each user has identical initial propensities $(0.3, 0.3, 0.3, 0.1)$, participating in community D with probability 0.1. Users take turns, selecting communities according to our reinforcement model using the same parameters inferred from real data above. Having selected a community, they reply to a random *new* comment (since their last visit). In that same community, they provide positive feedback via voting to each new comment independently with probability 0.3. Each reply or vote increases the associated feedback feature (element of r_i) by 1; since users receive feedback in every

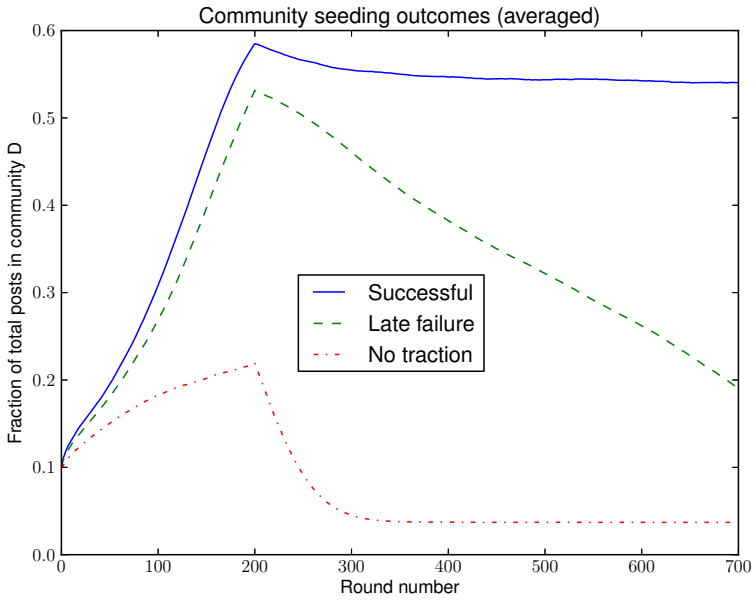

Figure 8.2: Three types of observed outcomes in synthetic community seeding experiments, with 200 seed rounds and 9 seed users. Sequences were grouped based first on the fraction of interest in community D at round 200: no traction (≤ 0.4) or some early traction. Of those with early traction, there are late failures (≤ 0.5 at 700) and successfully seeded communities. Curves are averaged within each group.

community, the magnitude of this feedback quickly becomes irrelevant, only its relative frequency affecting participation probabilities.

Under this model, participation in community D *decays* from its initial proportion of 0.1 to about 0.04: users get more reinforcement in other communities, and so only visit D because a third of the feedback they receive goes to their initial propensities ($\epsilon = 0.33$). This result is quite stable: D has almost no chance of growing. Is it possible to seed the community?

We add K seed users who participate only in community D, providing 1 voting feedback to every new comment during their turn. Their purpose is to make D a *self-sustaining* active community by providing extra social feedback during an initial seed period. The game proceeds in rounds, with every user taking one turn during each round in a consistent order. The seed users participate for the first S rounds. We consider D to be self-sustaining and active if the average non-seed user spends 50% of their time in community D after an additional 500 rounds without any seed users.

Figure 8.1 shows the probability of successfully seeding community D as a function of the number of seed users K and the number of seed rounds S. Even a large number of users has a small chance of seeding a community in a short time, but relatively small numbers of users over a long period of time can force phase transitions. Figure 8.2 shows example dynamics of three common outcomes of the seeding process: no traction, late failure, and successful seeding of a self-sustaining community.

8.2 Uncovering Bias

What can points of view tell us about Wikipedia? Having validated such a model in Chapter 4, we now explore some of its insights.

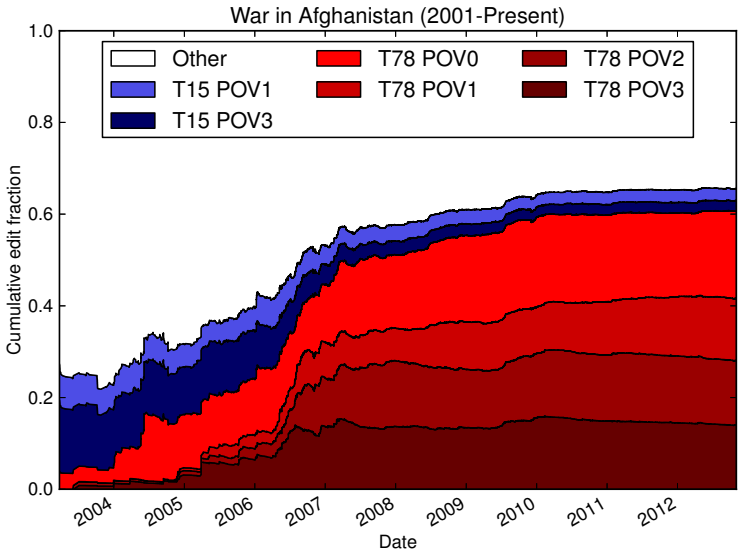

Figure 8.3: Cumulative fraction of edits on the top 6 topics and POVs on the article on the War in Afghanistan, showing specialization over time. Topic 15 encompasses many disputes—terrorism, politics, and articles about Wikipedia itself—and is used to explain many early edit conflicts. As Wikipedia matured, users specialized more: topic 78 can be described as "contemporary wars", and better explains later conflicts on this page. Topic 78, POV 0 is composed of casual editors (17 on-POV edits/user), while POV 3 consists of "power editors" (269 edits/user).

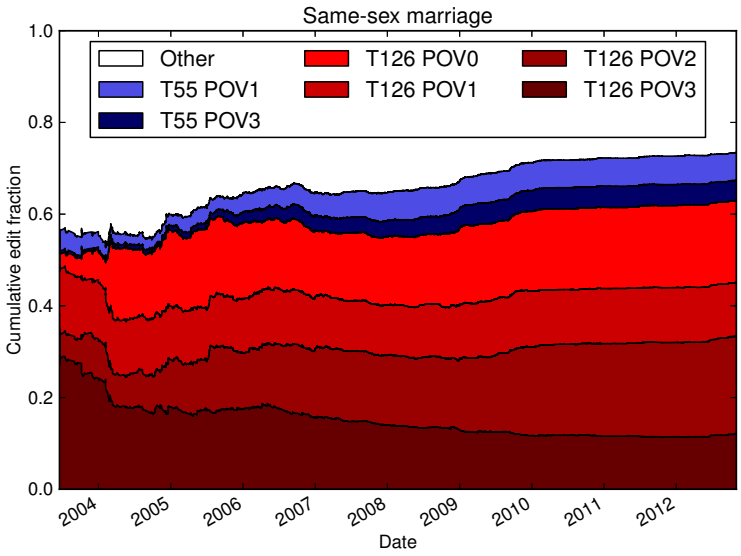

Figure 8.4: Cumulative fraction of edits on the top 6 topics and POVs on the article on Same-sex Marriage. Topic 126 covers issues related to human gender and sexuality, with POV 0 generally taking a more socially conservative stance. POV proportions on this page are relatively stable, after an initial increase in opposition (POV 0) as the encyclopedia became more notable. Topic 55 explains the interactions between vandals and those who remove vandalism, and shows up on many popular pages.

Table 8.1: Selected topics, with the top pages by number of edits on that topic (ignoring POV). From a high probability assignment.

Topic 61	Topic 68	Topic 23
Killer whale	Anarchism	2006 Lebanon War
Tiger	Race and IQ	Muhammad
Lion	Capitalism	Gaza War
White shark	Libertarianism	Islam
Cougar	Iraq War	Israel
Giraffe	Socialism	Lebanon

Table 8.2: Active pages (more than 100 editors) which—as of November 2012—had more than 60% of their edits on a single, controversial POV of a controversial topic.

Page title	POV%
Private finance initiative	64%
World War II casualties	67%
John Prendergast (activist)	65%
1948 Palestinian exodus from Lydda and Ramle	70%
Chilean presidential election, 2005–2006	69%

Figure 8.5: Word cloud showing additions from POV 0 (more conservative, shown in red) and a more liberal point of view. Size is based on a simple TF-IDF weighting among all four points of view.

Unlike in traditional topic modeling, where documents are mixtures of topics, we model *users* as such mixtures. This leads to topics where pages are grouped semantically—as we would expect from a traditional topic model on documents and words—only if user behavior is well explained by those topics. Table 8.1 includes examples of such topics that contain pages which are deeply semantically similar. However, our model also reveals topics which have more to do with user behavior than with the subject of edited pages: for example, one topic deals exclusively with vandalism and those who remove it from the encyclopedia. A user may then be a mixture of not only several topics but also several kinds of topics (e.g. animals and anti-vandalism).

Changes over time Labeling each revision with a point of view allows us to visualize page dynamics. Has the nature of a conflict changed over time? Were the current points of view always well represented? Figure 8.3 shows a shift in the topics used to explain editing and edit conflicts: early Wikipedians were often—by necessity, considering the number of editors—generalists. With growth, editors became increasingly specialized. This shift is reflected in the topics represented on the page, and in the points of view used to explain the changing conflict.

Figure 8.4 shows a traditional topic—dealing with the page's subject matter—coexisting with a behavioral topic explaining the interactions between vandals and anti-vandals. Points of view show a similar duality: POVs in Figure 8.3 deal as much with types of users (casual vs. heavy editors) as with page content, whereas those in Figure 8.4 are more focused on subject matter disputes (see Figure 8.5 for an informal comparison of two points of view).

As an aside, some point of view disputes are not apparent from natural language, e.g. the "modern wars" topic includes a dispute over WWII casualty numbers. Many disputes over figures have this property, and vandalism is another case where actions are more informative than words.

Page and user statistics Modeling POV provides a rich source of information about pages and users. Consider the problem of finding pages which could benefit from contributions by editors with a different POV: the model allows us to not only find these pages, but also to find users on different POVs who could be interested in the topic. For example, Table 8.2 shows the five most controversial pages that had more than 60% of their edits come from a single, controversial POV of a controversial topic. Here we define controversial topics as those with rare same-POV reverts ($< 3\%$) and more common different-topic reverts ($\geq 6\%$), and controversial POVs as those that have a high probability of reverting or being reverted by a different POV on the same topic ($\geq 30\%$). The model provides flexibility in querying for specific patterns over topics and POVs.

8.3 Conclusions

This dissertation presents several views on two general classes of incentives in collective intelligence. Product incentives, where users have an incentive to change the final output of a collective intelligence process, and process incentives, which provide users with more innocuous reasons to contribute. The models and analyses presented for these classes of incentives are not in any sense exhaustive, but serve to show that this qualitative framework is a powerful way to think about issues of manipulation and participation. With the exception of the Instructor Rating Markets, this work has been purely observational in nature, which is both exciting because of the wealth of social data now available and also limiting in the sense that there are alternative interpretations of any observational model.

There is a strong potential for this form of modeling to enable more useful and healthful venues for collective intelligence processes. The major work towards this goal is in further enumerating incentives with quantitative models of individual behavior, in validating incentives with experiments, and finally in synthesizing and optimizing to form a concrete picture of better venue design. In

the meantime many intermediate products, such as those touched on in this chapter, can help us to understand and nurture collective intelligence.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML*, page 1, 2004.

[2] Dae-Yong Ahn, Jason A Duan, and Carl F Mela. Managing user generated content: A dynamic rational expectations equilibrium approach. 2014. Working paper.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Steering user behavior with badges. In *Proceedings of the twenty-second conference on the World Wide Web*, WWW '13, pages 95–106, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2035-1. URL `http://dl.acm.org/citation.cfm?id=2488388.2488398`.

[4] K.J. Arrow et al. Statement on prediction markets. AEI-Brookings Publication No. 07-11, 2007.

[5] J. Berg and T. Rietz. Prediction markets as decision support systems. *Inf. Sys. Front.*, 5(1): 79–93, 2003.

[6] J. Berg, R. Forsythe, F. Nelson, and T. Rietz. Results from a dozen years of election futures markets research. *Handbook of Experimental Economics Results*, 1:742–751, 2008.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[8] Petko Bogdanov, Nicholas D. Larusso, and Ambuj Singh. Towards community discovery in signed collaborative interaction networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 288–295, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4257-7. doi: 10.1109/ICDMW.2010.174.

[9] Craig Boutilier. Eliciting forecasts from self-interested experts: Scoring rules for decision makers. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '12, pages 737–744, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0-9817381-2-5, 978-0-9817381-2-3. URL `http://dl.acm.org/citation.cfm?id=2343776.2343802`.

[10] Aseem Brahma, Mithun Chakraborty, Sanmay Das, Allen Lavoie, and Malik Magdon-Ismail. A Bayesian market maker. In *Proceedings of the Thirteenth ACM Conference on Electronic Commerce*, pages 215–232, 2012. ISBN 978-1-4503-1415-2. doi: 10.1145/2229012.2229031. URL `http://doi.acm.org/10.1145/2229012.2229031`.

[11] Michael J Brzozowski, Thomas Sandholm, and Tad Hogg. Effects of feedback and peer pressure on contributions to enterprise social media. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 61–70. ACM, 2009.

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

[13] Moira Burke, Cameron Marlow, and Thomas Lento. Feed me: Motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 945–954, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518847. URL `http://doi.acm.org/10.1145/1518701.1518847`.

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

[15] Mithun Chakraborty, Sanmay Das, Allen Lavoie, Malik Magdon-Ismail, and Yonatan Naamad. Instructor rating markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 159–165, 2013.

[16] Mithun Chakraborty, Sanmay Das, and Allen Lavoie. How to show a probabilistic model is better. *ArXiv e-prints*, abs/1502.03491, 2015.

[17] Y. Chen and D. Pennock. A utility framework for bounded-loss market makers. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 49–56, 2007.

[18] Y. Chen, S. Dimitrov, R. Sami, D.M. Reeves, D.M. Pennock, R.D. Hanson, L. Fortnow, and R. Gonen. Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica*, pages 1–40, 2009.

[19] Yiling Chen, Xi Gao, Rick David Goldstein, and Ian Kash. Market manipulation with outside incentives. American Association for Artificial Intelligence, 2011.

[20] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[21] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238. IEEE, 2008.

[22] Sonia Chernova and Manuela Veloso. Confidence-based policy learning from demonstration using Gaussian mixture models. In *Proc. AAMAS*, pages 233:1–233:8, 2007. ISBN 978-81-904262-7-5. doi: 10.1145/1329125.1329407. URL `http://doi.acm.org/10.1145/1329125.1329407`.

[23] Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. Socialization tactics in Wikipedia and their effects. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pages 107–116, New York, NY, USA, 2010. ACM.

[24] B. Cowgill, J. Wolfers, and E. Zitzewitz. Using prediction markets to track information flows: Evidence from Google. Working paper, 2010.

[25] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 160–168, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401914. URL `http://doi.acm.org/10.1145/1401890.1401914`.

[26] Justin Cranshaw and Aniket Kittur. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1865–1874. ACM, 2011.

[27] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the twenty-second conference on the World Wide Web*, pages 307–318. International World Wide Web Conferences Steering Committee, 2013.

[28] S. Das. A learning market-maker in the Glosten-Milgrom model. *Quantitative Finance*, 5 (2):169–180, 2005.

[29] S. Das. The effects of market-making on price dynamics. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '08, pages 887–894, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-0-9817381-1-6. URL `http://dl.acm.org/citation.cfm?id=1402298.1402347`.

[30] Sanmay Das and Allen Lavoie. Automated inference of point of view from user interactions in collective intelligence venues. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 82–90, 2014.

[31] Sanmay Das and Allen Lavoie. The effects of feedback on human behavior in social media: An inverse reinforcement learning model. In *Proceedings of the Thirteenth International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '14, 2014.

[32] Sanmay Das and Malik Magdon-Ismail. Adapting to a market shock: Optimal sequential market-making. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 361–368. 2009.

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

[34] Sanmay Das, Allen Lavoie, and Malik Magdon-Ismail. 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, pages 1097–1106, 2013.

[35] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn prediction in new users of yahoo! answers. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 829–834. ACM, 2012.

[36] David Easley and Arpita Ghosh. Incentives, gamification, and game theory: An economic approach to badge design. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pages 359–376, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1962-1. doi: 10.1145/2482540.2482571. URL `http://doi.acm.org/10.1145/2482540.2482571`.

[37] Ido Erev and Alvin E Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Amer. Econ. Rev.*, 88(4): 848–81, September 1998.

[38] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *J. ASA*, 90(430):577–588, 1995.

[39] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, pages 63–72. ACM, 2012.

[40] Mathilde Forestier, Julien Velcin, and Djamel Zighed. Extracting social networks to understand interaction. In *Advances in Social Networks Analysis and Mining (ASONAM)*, pages 213–219. IEEE, 2011.

[41] Katie Genter, Noa Agmon, and Peter Stone. Ad hoc teamwork for leading a flock. In *Proc. AAMAS*, pages 531–538, 2013. ISBN 978-1-4503-1993-5. URL `http://dl.acm.org/citation.cfm?id=2484920.2485005`.

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

[43] Arpita Ghosh, Satyen Kale, and Preston 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`.

[44] Eric Gilbert. Widespread underprovision on Reddit. In *Proc. CSCW*, pages 803–808, 2013. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441866. URL `http://doi.acm.org/10.1145/2441776.2441866`.

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

[46] Uri Gneezy and Aldo Rustichini. Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810, 2000.

[47] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[48] Thomas L. Griffiths and Mark 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.

[49] Alexander Halavais. Do dugg diggers digg diligently? feedback as motivation in collaborative moderation systems. *Information, Communication & Society*, 12(3):444–459, 2009.

[50] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, page 0002764212469365, 2012.

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

[52] R. Hanson. Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets*, 1(1):3–15, 2007.

[53] R. Hanson. Insider trading and prediction markets. *Journal of Law, Economics, and Policy*, 4:449–463, 2008.

[54] R. Hanson and R. Oprea. A manipulator can aid prediction market accuracy. *Economica*, 76(302):304–314, 2009.

[55] R. Hanson, Ryan Oprea, and David Porter. Information aggregation and manipulation in an experimental market. *J. Econ. Beh. and Org.*, 60:449–459, 2006.

[56] Creighton Heaukulani and Zoubin Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. In *Proc. ICML*, pages 275–283, 2013.

[57] M. Hindman, K. Tsioutsiouliklis, and J.A. 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.

[58] Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(Apr): 1593–1623, 2014.

[59] Gary Hsieh, Youyang Hou, Ian Chen, and Khai N. Truong. "Welcome!": social and psychological predictors of volunteer socializers in online communities. In *Proc. CSCW*, pages 827–838, 2013. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441870. URL http://doi.acm.org/10.1145/2441776.2441870.

[60] Bernardo A Huberman, Daniel M Romero, and Fang Wu. Crowdsourcing, attention and productivity. *J. Inf. Sci.*, 35(6):758–765, 2009.

[61] Jr. Isbell, CharlesLee, Michael Kearns, Satinder Singh, ChristianR. Shelton, Peter Stone, and Dave Kormann. Cobot in LambdaMOO: An adaptive social statistics agent. *AAMAS*, 13(3):327–354, 2006. ISSN 1387-2532. doi: 10.1007/s10458-006-0005-z. URL http://dx.doi.org/10.1007/s10458-006-0005-z.

[62] Joab Jackson. Intellipedia suffers midlife crisis. *Government Computer News*, 18, 2009.

[63] Gabriela Kalna and Desmond 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.

[64] Wagner A Kamakura and Gary Russell. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26:379–390, 1989.

[65] Gerald C Kane. It's a network, not an encyclopedia: A social network perspective on Wikipedia collaboration. In *Academy of Management Proceedings*, volume 2009, pages 1–6. Academy of Management, 2009.

[66] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19, 2007.

[67] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 453–462, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240698.

[68] W. Bradley Knox and Peter Stone. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *Proc. AAMAS*, pages 5–12, 2010. ISBN 978-0-9826571-1-9. URL http://dl.acm.org/citation.cfm?id=1838206.1838208.

[69] W. Bradley Knox and Peter Stone. Reinforcement learning from simultaneous human and MDP reward. In *Proc. AAMAS*, pages 475–482, 2012. ISBN 0-9817381-1-7, 978-0-9817381-1-6. URL http://dl.acm.org/citation.cfm?id=2343576.2343644.

[70] Travis Kriplean, Ivan Beschastnikh, and David W. McDonald. Articulations of wikiwork: Uncovering valued work in Wikipedia through barnstars. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 47–56, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-007-4. doi: 10.1145/1460563.1460573. URL http://doi.acm.org/10.1145/1460563.1460573.

[71] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[72] Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. WWW*, pages 621–630, 2010. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772754. URL http://doi.acm.org/10.1145/1772690.1772754.

[73] Kristina Lerman and Tad Hogg. Using stochastic models to describe and predict social dynamics of web users. *ACM Trans. Intell. Syst. Technol.*, 3(4):62:1–62:33, September 2012. ISSN 2157-6904. doi: 10.1145/2337542.2337547. URL http://doi.acm.org/10.1145/2337542.2337547.

[74] Kristina Lerman, Suradej Intagorn, Jeon-Hyung Kang, and Rumi Ghosh. Using proximity to predict activity in social networks. In *Proc. WWW*, pages 555–556, 2012. ISBN 978-1-4503-1230-1. doi: 10.1145/2187980.2188124. URL http://doi.acm.org/10.1145/2187980.2188124.

[75] Chenliang Li, Anwitaman Datta, and Aixin 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.

[76] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646003.

[77] Xudong Liu and Shahira Fahmy. Exploring the spiral of silence in the virtual world: lndividuals willingness to express personal opinions in online versus offline settings. *Journal of Media and Communication Studies*, 3(2):45–57, 2011.

[78] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link LDA: Joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 665–672, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553460. URL http://doi.acm.org/10.1145/1553374.1553460.

[79] Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong 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.

[80] George Loewenstein. Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, 109(453):25–34, 1999.

[81] Robert E Looney. Darpa's policy analysis market for intelligence: Outside the box or off the wall? *International Journal of Intelligence and Counterintelligence*, 17(3):405–419, 2004.

[82] Rui Lopes and Luis 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.

[83] Stefan Luckner and Christof Weinhardt. How to pay traders in information markets? Results from a field experiment. *Journal of Prediction Markets*, 1(2):147–156, 2007.

[84] André C. R. Martins. Mobility and social network effects on extremist opinions. *Physical Review E*, 78(3):036104, 2008.

[85] André CR Martins and Cleber D Kuba. The importance of disagreeing: Contrarians and extremism in the CODA model. *Advances in Complex Systems*, 13(05):621–634, 2010.

[86] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email. 2005.

[87] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, October 2007. ISSN 1076-9757.

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

[89] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology*, 7(4):23, 2007.

[90] Kevin Morris. How Reddit's cofounders built Reddit with an army of fake accounts. *The Daily Dot*, June 2012. URL `http://www.dailydot.com/business/steve-huffman-built-reddit-fake-accounts/`.

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

[92] Lev Muchnik, Sinan Aral, and Sean J. Taylor. Social influence bias: A randomized experiment. *Sci.*, 341(6146):647–651, 2013. doi: 10.1126/science.1240466. URL `http://www.sciencemag.org/content/341/6146/647.abstract`.

[93] Michael Munie and Yoav Shoham. Joint process games: from ratings to wikis. In *Proc. AAMAS*, pages 847–854, 2010.

[94] Iain Murray and Ruslan Salakhutdinov. Evaluating probabilities under high-dimensional latent variable models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1137–1144. 2009.

[95] Claudiu-Cristian Musat, Boi Faltings, and Philippe Rousille. Direct negative opinions in online discussions. In *Social Computing (SocialCom), 2013 International Conference on*, pages 142–147. IEEE, 2013.

[96] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed inference for latent Dirichlet allocation. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1081–1088. MIT Press, Cambridge, MA, 2008.

[97] Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 663–670, 2000.

[98] Elisabeth Noelle-Neumann. The spiral of silence a theory of public opinion. *Journal of communication*, 24(2):43–51, 1974.

[99] Felipe Ortega. Wikipedia: A quantitative analysis. Doctoral thesis, Universidad Rey Juan Carlos, Móstoles, Spain, April 2009.

[100] A. Othman and T. Sandholm. Automated market-making in the large: the Gates-Hillman prediction market. In *Proc. EC*, pages 367–376, 2010.

[101] A. Othman, T. Sandholm, D. Pennock, and D. Reeves. A practical liquidity-sensitive automated market maker. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, pages 377–386, 2010.

[102] Abraham Othman and Tuomas Sandholm. Decision rules and decision markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, pages 625–632, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-0-9826571-1-9. URL `http://dl.acm.org/citation.cfm?id=1838206.1838288`.

[103] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: A study of power editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, GROUP '09, pages 51–60, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-500-0. doi: 10.1145/1531674.1531682. URL `http://doi.acm.org/10.1145/1531674.1531682`.

[104] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[105] Nishith Pathak, Colin Delong, Arindam Banerjee, and Kendrick Erickson. Social Topic Models for Community Extraction. In *The 2nd SNA-KDD Workshop '08 (SNA-KDD'08)*, August 2008.

[106] Michael Paul and Roxana Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. *AAAI Conference on Artificial Intelligence*, 2010.

[107] D. Pennock and R. Sami. Computational aspects of prediction markets. In N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, editors, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[108] D.M. Pennock. A dynamic pari-mutuel market for hedging, wagering, and information aggregation. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 170–179, 2004.

[109] Martin Potthast, Benno Stein, and Robert 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.

[110] 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.

[111] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press. ISBN 0-9749039-0-6.

[112] Matthew Rowe. Mining user lifecycles from online community platforms and their application to churn prediction. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 637–646. IEEE, 2013.

[113] Mrinmaya Sachan, Danish Contractor, Tanveer Faruquie, and Venkata Subramaniam. Probabilistic model for discovering topic based communities in social networks. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, CIKM '11, pages 2349–2352, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063963.

[114] Leonard J. Savage. Elicitation of personal probabilities and expectations. *J. ASA*, 66 (336):pp. 783–801, 1971. ISSN 01621459. URL http://www.jstor.org/stable/2284229.

[115] Anne Schulz and Patrick Roessler. The spiral of silence and the internet: Selection of online content and the perception of the public opinion climate in computer-mediated communication environments. *International Journal of Public Opinion Research*, 24(3):346–367, 2012.

[116] John Scott. *Social network analysis*. Sage, 2012.

[117] E. Servan-Schreiber, J. Wolfers, D.M. Pennock, and B. Galebach. Prediction markets: Does money matter? *Electronic Markets*, 14(3):243–251, 2004.

[118] Peng Shi, Vincent Conitzer, and Mingyu Guo. Prediction mechanisms that do not incentivize undesirable actions. In *Internet and Network Economics*, pages 89–100. Springer, 2009.

[119] Alastair Sloan. Manipulating Wikipedia to promote a bogus business school. *Newsweek*, March 24 2015.

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

[121] Dongyoung Sohn and Nick Geidner. Collective dynamics of the spiral of silence: the role of ego-network size. *International Journal of Public Opinion Research*, page edv005, 2015.

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

[123] Stan Development Team. Stan: A C++ library for probability and sampling, version 2.5.0, 2014. URL `http://mc-stan.org/`.

[124] Anna Stavrianou, Julien Velcin, and Jean-Hugues Chauchat. Definition and measures of an opinion model for mining forums. In *International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, pages 188–193. IEEE, 2009.

[125] Anna Stavrianou, Julien Velcin, and Jean-Hugues Chauchat. Prog: A complementary model to the social networks for mining forums. *From Sociology to Computing in Social Networks: Theory, Foundations and Applications*, page 59, 2010.

[126] Greg Stoddard. Popularity dynamics and intrinsic quality in Reddit and Hacker News. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[127] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Comm. ACM*, 53(8):80–88, 2010.

[128] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *J. ASA*, 101(476), 2006.

[129] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, and Hady Wirawan 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.

[130] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553515.

[131] Xuerui Wang and Andrew McCallum. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150450. URL `http://doi.acm.org/10.1145/1150402.1150450`.

[132] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc 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`.

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

[134] J. Wolfers and E. Zitzewitz. Five open questions about prediction markets. Stanford GSB Working Paper, 2004.

[135] Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

[136] Gi Woong Yun and Sung-Yeon Park. Selective posting: Willingness to post a message online. *Journal of Computer-Mediated Communication*, 16(2):201–227, 2011.

[137] Fang Wu and Bernardo A. Huberman. Novelty and collective attention. *PNAS*, 104(45): 17599–17601, 2007. doi: 10.1073/pnas.0704916104. URL `http://www.pnas.org/content/104/45/17599.abstract`.

[138] Fang Wu, Dennis M Wilkinson, Bernardo Huberman, et al. Feedback loops of attention in peer production. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 409–415. IEEE, 2009.

[139] Feng Yan, Ningyi Xu, and Yuan Qi. Parallel inference for latent Dirichlet allocation on graphics processing units. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2134–2142. 2009.

[140] Maarten de Rijke Zhaochun Ren. Summarizing contrastive themes via hierarchical non-parametric processes. In *Proceedings of the ACM SIGIR conference on research and development in information retrieval*, 2015.

[141] Haiyi Zhu, Amy Zhang, Jiping He, Robert E Kraut, and Aniket Kittur. Effects of peer feedback on contribution: a field experiment in Wikipedia. In *Proc. CHI*, pages 2253–2262, 2013.