Effects of User Similarity in Social Media

Ashton Anderson
Stanford University
ashton@cs.stanford.edu

Daniel Huttenlocher
Cornell University

Jon Kleinberg
Cornell University

{dph, kleinber}@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

There are many settings in which users of a social media application provide evaluations of one another. In a variety of domains, mechanisms for evaluation allow one user to say whether he or she trusts another user, or likes the content they produced, or wants to confer special levels of authority or responsibility on them. Earlier work has studied how the *relative status* between two users — that is, their comparative levels of status in the group — affects the types of evaluations that one user gives to another.

Here we study how *similarity* in the characteristics of two users can affect the evaluation one user provides of another. We analyze this issue under a range of natural similarity measures, showing how the interaction of similarity and status can produce strong effects. Among other consequences, we find that evaluations are less status-driven when users are more similar to each other; and we use effects based on similarity to provide a plausible mechanism for a complex phenomenon observed in studies of user evaluation, that evaluations are particularly low among users of roughly equal status.

Our work has natural applications to the prediction of evaluation outcomes based on user characteristics, and the use of similarity information makes possible a novel application that we introduce here — to estimate the chance of a favorable overall evaluation from a group knowing only the attributes of the group's members, but not their expressed opinions.

Categories and Subject Descriptors: H.2.8 **[Database Management]**: Database applications—*Data mining*

General Terms: Algorithms; Experimentation.

Keywords: User-to-User Evaluations, User Similarity, Status, Ballot-Blind Prediction.

1. INTRODUCTION

Many on-line social applications include mechanisms for users to express evaluations of one another, or of the content they create. These evaluations of users are defined in different ways, depending on the application; for example, one user can say that they trust another user's reviews on a product rating site such as Epinions [7, 9]; or that they like the answer another user provides to a question on a community discussion site such as Stack Overflow or Yahoo! Answers [1]; or that they are in favor of granting special levels of privilege to another user on a site built around large-scale collaboration such as Wikipedia or a massive open-source project [3, 14]. Such evaluations serve a crucial purpose in the functioning of these sites, directing users toward content that is highly favored, and — as in Wikipedia and other collaborative domains — enabling the formation of cohorts of highly trusted users who guide the operation of the site.

The synthesis of these types of evaluations is an important problem: when multiple people all provide evaluations of the same "target" person, how do we create a composite description of these evaluations that accurately reflects some type of cumulative opinion of the community? Several recent studies have approached this question through the observation that at an aggregate level, user-to-user evaluations seem to reflect overall levels of *status* in the community [7, 16] — that is, the extent of users' past contributions or achievement. Moreover, the process is further affected by the *relative status* between the user providing the evaluation and the user being evaluated [14]. At the same time, this framework has thus far provided only a first approximation to the factors affecting evaluations, and as we discuss below, there are a number of basic phenomena related to evaluation that it seems unable to explain.

In this paper, we show that the analysis of user-to-user evaluations can be significantly strengthened by taking into account the *similarity* in characteristics of users — such as the extent to which their contributions to the site have involved similar content, or have involved interactions with a common set of other users. We identify some fundamental principles that guide the ways in which similarity affects evaluations, and the way in which it acts in combination with relative status. Then we develop new methods demonstrating how community judgments can be more accurately extrapolated from a small set of evaluations when we have information about the similarity levels among users.

For our analysis, we focus on three primary domains that exemplify distinct forms of user evaluation: Wikipedia, Stack Overflow, and Epinions. On Wikipedia, we study the admin promotion process. When a user wants to acquire special adminship privileges on Wikipedia, she submits her credentials to a review and promotion process, involving public discussion followed by a public vote on the promotion [3]. On Stack Overflow, a question and answering site, a central mechanism allows users to vote for or against different user-contributed answers to a question, thereby collectively "up-voting" better answers and "down-voting" worse ones. On Epinions, users write reviews of various products and a similar mechanism to Stack Overflow's allows others to rate these reviews on a scale of 1 to 5.

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.

We find a range of common phenomena occurring across these domains, despite some differences in the way evaluations work across them. On Wikipedia, the target of the evaluation is the user herself, whereas on Stack Overflow and Epinions the target is a piece of content produced by the user, but in a setting where it is still clearly associated with the user's identity. We will thus speak of a user A evaluating a user B as shorthand either for a direct evaluation of B herself, or for an indirect evaluation via content that B has authored. Moreover, on Wikipedia, the evaluations we study are synthesized into a cumulative *outcome* — whether the user is promoted to adminship or not — whereas on Epinions the result of the evaluations is an unsynthesized composite of individual ratings.

We now provide a brief overview of some of our main findings about the effect of similarity, as well as our results on the use of similarity information to infer outcomes.

Overview of Results: The Effect of Similarity. As noted above, we measure similarity between users A and B using two different types of characteristics: similarity of interests using a distance metric capturing overlap in the types of content they produce; and similarity of social ties using a measure of the overlap in the sets of people they have evaluated. We measure status by the number of actions taken on the site, where "action" is defined in a way that fits the use of each site.

Across all the domains, we find an aggregate pattern in which users with more similar interests or social ties are more positive toward one another, and — more subtly — in which status differences matter less for evaluations among users who are more similar to each other. In other words, when people have more in common with each other, status seems to play a less important role in evaluation. This overall pattern varies in strength according to the specific site that we study and also the particular characteristics that are measured. For instance on Stack Overflow we observe only marginally more positive evaluations from more similar users when we measure similarity based on content, whereas evaluations are much more positive for highly similar users when using our social measure of similarity. Such distinctions can provide useful insights into the ways that different types of judgments — evaluating a user versus evaluating their content — affect the overall behavior we see on these sites.

Moreover, on Wikipedia, we see a further important effect: among the users who choose to provide evaluations of B, the high-status evaluators have overall more similarity to B than the low-status evaluators do. This suggests a selection effect where "elite" or high status individuals are more likely to participate in evaluations in their areas of interest or expertise.

The combination of these effects offers a novel explanation for the following intriguing observation from earlier work by three of us on admin promotion in Wikipedia [14] that we now also find in Stack Overflow and Epinions. A key finding in this earlier study was that the probability of A providing a positive evaluation of B has a pronounced local minimum when the status difference Δ between A and B is near zero — a striking effect where the aggregate evaluations are particularly negative among users of comparable status levels. It was an open question to explain the mechanism behind this effect. In Section 5 we illustrate how this effect is well explained by the combination of two underlying aggregate factors. The first is based on the role played by low-status users, who tend to be judged differently than those with moderate or high status. However, on Wikipedia the local minimum near $\Delta = 0$ persists (in a milder form) even when low-status users are not considered, and we argue that this can arise from the fact that users more similar to the target B tend to be more positive about B, and such high-similarity users are overrepresented among the high status evaluators. These findings suggest that much of the observed effect may be due to aggregate factors such as the set of people who show up to evaluate a given user, as well as the fact that status-consciousness decreases with greater similarity, rather than any individual effects of how users evaluate those of comparable status.

Overview of Results: Ballot-Blind Prediction. In the settings we study, users provide explicit evaluations (positive or negative). But there are many domains where users interact with a piece of content (or another user), and we know attributes of these users but not their evaluation of the content. Do our results have relevance for these types of domains?

In fact, we find that they do: using the principles developed earlier in the paper, we show in Section 6 that knowing only the first few users $A_1, A_2, \ldots, A_k$ who evaluate another user B and their attributes (similarity to B and relative status) – but *not* the actual evaluations – we can make accurate predictions about the outcome of the election. The crucial point is that in this prediction task, we do not have access to the evaluations provided by $A_1, A_2, \ldots, A_k$, but only to their attributes as individuals relative to user B. As a result, we refer to this as the problem of *ballot-blind prediction*.

The results we obtain for ballot-blind prediction suggest that there is considerable power in the use of similarity information, and that the potential applicability of our framework for reasoning about the interplay of similarity and status extends beyond settings in which users provide explicit, observable evaluations.

2. RELATED WORK

The issue of evaluations in social media is a very general one, and it splits into several themes depending on the nature of what is being evaluated. A large amount of work has studied the evaluation of items such as products, movies, or Web sites; for example, the area of collaborative filtering and recommendation systems focuses on this issue (e.g., [8, 10, 17, 20]). In these settings, there is an inherent asymmetry between the people doing the evaluating and the items being evaluated, and hence the evaluation is best modeled in different ways — as a function of intrinsic properties of the item, or of some notion of the match of person to item. In our setting, on the other hand, the evaluators and the targets of evaluation are both people, and everyone can both evaluate and be evaluated. Thus in addition to user A judging user B purely based on B's characteristics, A can also take into account her own characteristics relative to B's ("Is B similar to me? Is B better than I am?"). This latter form is unique to our setting: when evaluating products such as a microwave oven, it doesn't make sense to compare the microwave to oneself – but in user-to-user evaluation, it is not only valid but natural to compare the subject of evaluation to oneself.

Other areas have also considered the evaluation of people in contexts different from ours. This includes the inference of opinion from natural-language text [19], as well as "higher-order" modeling of opinions about opinions [6]; it also includes the use of norms to control deviant behavior in on-line communities [5]. If we view evaluation as providing positive and negative signs on the edges of an underlying graph (denoting positive and negative evaluations between people), then we obtain a signed social network; recent work has considered the characteristic properties of such networks [2, 12, 16, 15, 21]. In contrast, our work here focuses on how attributes of the individuals comprising the nodes, and in particular their levels of similarity, affect the signs expressed on the links.

Finally, notions of similarity and status appear in a number of contexts throughout the sociology literature. The principle of *homophily* provides a principled basis for reasoning about the ways

Wikipedia language	N	$P_0(+)$	U
English	119,489	74.5%	10,558
German	78,647	67.7%	3,560
French	22,534	78.0%	1,552
Spanish	8,641	83.4%	917

Table 1: Wikipedia statistics. N = number of votes, $P_0(+)$ = baseline fraction of positive votes, U = number of users.

in which people tend to favor those who are similar to themselves [13]. Burt considers the ways in which differential status comparisons take place among people who view themselves as belonging to a set of peers — and how the fear of "falling behind" in comparison to this set can motivate people to succeed [4]. Work in network exchange theory has also produced experiments showing how perceived status differences between parties to an interaction can affect the balance of power in the interaction [22, 23].

3. DATASET DESCRIPTION

We use data from online social media applications where users explicitly evaluate other users, either directly (evaluating the users themselves) or indirectly (evaluating the content they produce).

Wikipedia is a collaboratively authored free encyclopedia. Active users can be nominated for promotion to admin status (admins have access to privileges that aid the maintenance of the site). Once nominated, a public deliberation process begins and other Wikipedia users cast either positive, negative, or neutral votes on the candidate. Votes are public, signed by the voter, and timestamped. After enough time has passed, a Wikipedia official reviews the votes and discussion and decides whether the election was successful or not. A public record of the election is archived online. We collected data from English Wikipedia on 3,422 elections that took place between September 17, 2004 and January 30, 2010. We extracted a total of 153K votes, but use a subset of 120K votes for which we could obtain similarity and status information. We collected similar data from French, German, and Spanish Wikipedias as well. Basic stats are shown in Table 1.

Stack Overflow is a popular question-answering site for programmers. Users post questions and answers, and can upvote or downvote other users' questions and answers. Heavily upvoted answers are prominently displayed on the question page, and heavily upvoted questions are publicized on the site. There is a visible reputation system that assigns a score to each user on the site. Users gain/lose reputation by receiving upvotes/downvotes on the content they produce. Users cannot downvote freely; it costs them a small amount of reputation to do so.

There are 1.1M questions, 3.2M answers, and 7.5M votes (1.8M on questions, 5.7M on answers) from the site's inception on July 31, 2008 to January 3, 2011. 93.4% of the votes are positive, and for each vote we know the identity of the voter and the time it occurred. Who voted on what is not publicly displayed on the site. Questions are annotated with tags describing relevant topics. There are 31K unique tags and 3.6M tag events.

Epinions is an online reviewing site where users can review products and rate each others' reviews. Our dataset has 132K users, 1.5M reviews, and 13.6M ratings of those reviews (on a scale of 1-5 stars) [18]. The ratings are overwhelming positive (78% are 5-star ratings), so we call a 5-star rating a "positive" evaluation and all others "negative".

In the results that follow, we have performed the analyses on all these datasets (four languages of Wikipedia, plus Stack Overflow and Epinions), except for analyses, where noted, that were specific to certain features of one of the sites. Due to the space limits, we will often report the results only for subsets of the datasets, designed to cover the space of different outcomes; the datasets where results are not reported have outcomes very similar to what is shown here. In particular, the results on Stack Overflow and Epinions are very similar; we only report Stack Overflow and comment on differences where there are any.

4. THE INTERPLAY BETWEEN SIMILARITY AND STATUS

Definitions of Similarity. To explore how user similarity affects evaluations, we first formalize a notion of similarity that measures how much two users' interests overlap.

In each dataset, we can think of users as being represented by the actions they take. By *action*, we will be referring to editing an article on Wikipedia, asking or answering a question on Stack Overflow, and rating a review on Epinions. Let user u's *binary action vector* be a vector of length M (where M is the number of all possible actions a user can take) with a 1 in the ith position if u took action i at least once, and 0 otherwise. The similarity between users u and v is then the cosine between their respective binary action vectors: $s(u,v) = \frac{u_a \cdot v_a}{||u_a|| \cdot ||v_a||}$ (where u_a is u's binary action vector). If either $||u_a|| = 0$ or $||v_a|| = 0$ (either u or v hasn't taken any actions), then we say $s(u,v)$ is undefined and exclude it from our calculations. Notice that a user's action vector changes over time as they make more actions. Whenever we compute the similarity $s(E,T)$ between an evaluator E and target T, we use the action vectors corresponding to the time at which E evaluated T.

On Wikipedia, we make one modification. The similarity metric defined above works well for most users, but breaks down for "typo-fixer" editors who make large numbers of minor edits, for example by writing automated bots that find and fix typos (since they edit a lot of articles, they are artificially similar to virtually everyone). To account for this, when computing user similarity for Wikipedia we restrict a user's binary action vector to only include the top k articles he edits the most. This eliminates the long tail of edits that typo-fixers make and limits our similarity measure to only take a user's main area of interest into account. After some experimentation we use $k = 100$.

On Stack Overflow, questions are annotated with tags that represent relevant topics. These can be used to provide a summary description of a user's interests by considering tags that annotate the questions a user evaluates as well as those annotating the parent questions of answers the user evaluates. We represent user u as a binary tag vector with a 1 in the ith position if tag i annotated a question u evaluated or a parent question to an answer u evaluated. "Tag similarity" refers to the cosine between binary tag vectors. Similarly, on Epinions, we measure similarity between users' binary action vectors (recall that on Epinions an action corresponds to rating a review).

Additionally, on all of our datasets, we also define another type of user-to-user similarity. We can characterize a user by a vector of other users that a user has evaluated and define similarity between two users as the cosine between their corresponding binary evaluation vectors. We call this *social similarity*.

We experimented with a number of standard similarity measures, including weighted cosine (where position i is assigned the number of times user u took action i), Jaccard, and others. For all of them, either the results are not qualitatively different, or if they are, it is due to an idiosyncrasy of the particular measure.

Similarity and Evaluations. Now that we have formally defined similarity, we can investigate how similarity affects the probability of a positive evaluation. Although it is natural to expect a relationship between status and similarity, it is not a priori obvious the

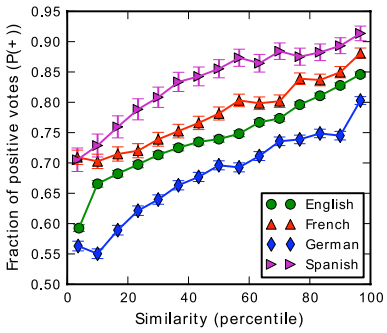


Figure 1: (Wikipedia) Probability of a positive evaluation ($P(+)$) as a function of the similarity (binary cosine) between the evaluator and target edit vectors ($s(e,t)$).

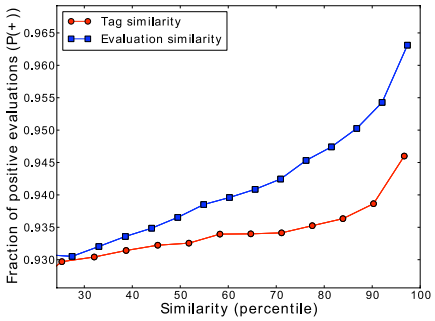


Figure 2: (Stack Overflow) Probability of E positively evaluating T as a function of similarity (note that errorbars are still plotted, but they are too small to be visible).

relationship is positive or negative. It could be that evaluators are more supportive of targets in their area of expertise. For instance, it's possible that evaluators on Wikipedia "cheerlead" their area's most promising editors into administrative positions. An alternative theory is that the more familiar the evaluator is with the target's work, the more likely the evaluator is to know about mistakes and weaknesses, or other negative things not uncovered by the type of superficial appraisal that evaluators from other domains can give, and are consequently harsher in their evaluations. Another reason evaluators could be harsher on similar targets is simple competition; an evaluator's power in their area can only diminish as the number of other admins around increases.

On Wikipedia we find strong evidence for the former theory – that the probability of a positive evaluation grows with the (content as well as social) similarity between the evaluator and the target. On average, overlapping interests with the target make it more likely that an evaluator will cast a positive vote. In Figure 1, we show the probability of a positive vote as a function of the cosine between the evaluator and target binary edit vectors (averaging over all values of Δ). (Note that throughout this paper, $P(+)$ is simply the fraction of positive evaluations in a given set of evaluations.) The monotonically-increasing, diminishing-returns relationship between them is clearly present. We plot results not only for English Wikipedia but also for French, German and Spanish as examples of how the results are similar for several datasets. The qualitative trend is the same across all of them. For ease of presentation, we will restrict Wikipedia plots to English Wikipedia for the remainder of the paper. The results are similar across all Wikipedia datasets, except where noted otherwise.

On Stack Overflow, the effect of similarity on evaluations is more nuanced. We still find that the more similar an evaluator-target pair is, the more likely the evaluation is positive. But in contrast to Wikipedia, the strength of this relationship depends on which notion of similarity we use. As discussed above, we can consider similarity on tags, which is a measure of content similarity, or we can consider similarity on evaluations, which is a measure of social similarity. These measure two fundamentally different things: similarity in the content two people are interested in and similarity in the users they have evaluated. In Figure 2, we plot the fraction of positive evaluations as a function of these two measures of similarity. Since the scale of the cosine values are different for the two measures, we normalize by their respective cumulative distribution functions so that we compare percentiles instead of raw values on the x-axis. We find that $P(+)$ increases with tag-similarity very gradually for most of the data (from the 25th to the 80th percentile, $P(+)$ only increases by 0.5%, a relative gain of 7%). But for evaluation-similarity, $P(+)$ is both everywhere higher than it is for tag-similarity (except for $s = 0$) and rises much more significantly (for the same range, $P(+)$ increases 1.5%, a relative gain of 21% – 3 times as much). This suggests that the social orbits users travel in influence evaluations much more than the topics they're interested in. Since social similarity is more informative on Stack Overflow, we will use it for the rest of the paper. As alluded to above, there is no strong difference on Wikipedia (and thus, we henceforth only consider action similarity on the Wikipedia datasets). We find similar effects for content similarity on Epinions. However, social similarity turns out to be too sparse to be meaningful in Epinions.

So far we have demonstrated that evaluator-target similarity is directly correlated with the probability of a positive evaluation. Next we examine how similarity and status jointly interact with user evaluations.

Similarity and status. We now consider some of the interactions between similarity and status; to do this, we first provide some definitions underpinning the measures of status. A user's status, or standing in the community, is an intricate function of the community's perception of his contributions. In this work, we follow previous research on notions of status in networks [3, 14] and use volumetric measures of contribution, or *activity level*, as simple yet reasonable approximations of status.[1] In particular, we consider user u's status at time t as the number of actions u has taken before time t (i.e. the number of non-zero entries in her corresponding binary action vector u_a at time t). Thus on Wikipedia, we use the total number of edits made by a user u before time t as u's status at time t; on Stack Overflow, we use the total number of questions asked and answers given[2]; and on Epinions, we use the number of ratings given. For the rest of the paper, "status" refers to the number of actions, and we use σ_A to refer to user A's status.

Previous research showed that evaluation behavior does not just vary with target status alone, but rather it varies with the evaluator

[1]This is consistent with sociological work showing how overt surface-level *markers* of status tend to serve as proxies in everyday interactions for more subtle forms of status [22]. Extending the simple formulations of status we use here to encompass the community's opinion and perception of one's status is an interesting direction for future work.

[2]A tempting definition of status on Stack Overflow would be the reputation score that is assigned to each user based on their activity on the site. However, since our main interest is in investigating the relationship between status and user evaluations, reputation is not an ideal measure. This is due to the fact that users directly gain reputation from positive evaluations and lose reputation from negative evaluations – and hence for this definition, status and user evaluations would be correlated in an uninteresting way.

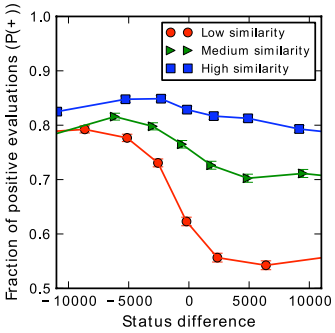
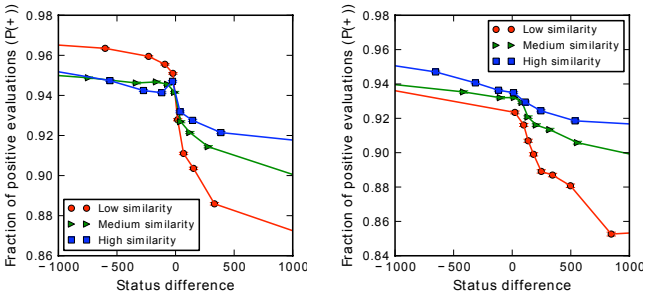


Figure 3: (English Wikipedia) Probability of E voting positively on T as a function of Δ for different levels of similarity

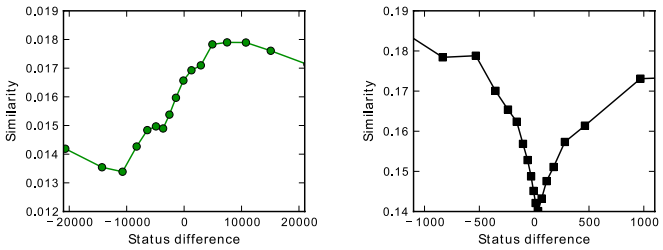


(a) All evaluators (b) No low-status evaluators

Figure 4: (Stack Overflow) Probability of E voting positively on T as a function of Δ for different levels of similarity (errorbars are plotted but are too small to be visible).

status and target status together [14]. We therefore need a way to compare the evaluator's status with the target's status. We use the *differential status* $\Delta = \sigma_E - \sigma_T$ to do this. Note that it is not a priori clear whether this is the right measure or not; for example, the ratio of their respective statuses could plausibly be a better measure. We show that the differential status is the appropriate way to compare statuses at the end of the section.

We now turn to a subtle effect that arises from the interplay of the similarity and differential status between two users. Previous work has found $P(+)$ depends on Δ, and that the relationship is natural: evaluators with higher status than the target tend to be harsher than the evaluators with lower status than the target. However, we find that this effect is significantly moderated by the similarity between the evaluator and the target. The more similar E and T are, the less their difference in status affects the vote. That is, if an evaluator and target have similar action profiles, the vote varies less with their status difference than if the evaluator and target operate in different domains. This suggests evaluators use status as a proxy for quality in the absence of more direct knowledge of the target, and depend less on status when they are more well-informed, i.e., more similar to the target. Evidence for this is shown in Figure 3, where $P(+)$ is plotted as a function of Δ for different levels of similarity in English Wikipedia. Notice that for low-similarity pairs status plays a major role, as can be seen by comparing the curves on left half of $\Delta = 0$ to the right half: for $\Delta < 0$, votes are about 80% positive, but for $\Delta > 0$, the fraction of positive votes plummets to below 60%. As similarity is increased however, the votes get more positive across all values of Δ and the curve becomes increasingly flat. This shows that status difference is less critical when the evaluator and target tend to edit the same articles.



(a) English Wikipedia (b) Stack Overflow

Figure 5: Similarity between E and T pairs as a function of Δ. On English Wikipedia we screen out low-status targets (those below 3000 edits) to show the trend doesn't depend on them, but the curve doesn't qualitatively change for any threshold on target status. On Stack Overflow, similarity on tags is shown, but the curve is robust to changes in the similarity and status metrics used.

The story on Stack Overflow is similar, but has an additional subtlety. The $P(+)$ vs. Δ curves for different levels of social similarity are shown in Figure 4a. When $\Delta > 0$, the picture is qualitatively the same as in Wikipedia: the higher the similarity, the higher $P(+)$ is. But for $\Delta < 0$, the situation is very different: the similarity curves are in the opposite order from before: evaluators with low similarity to the higher-status targets (since $\Delta < 0$) are more positive than evaluators with high similarity.

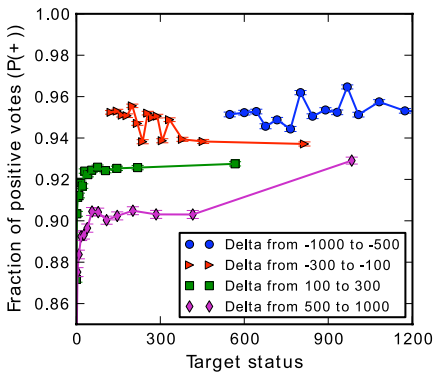
This is due to a particular property of Stack Overflow's reputation system, in which it costs a user a small amount of reputation to downvote a question or answer (issue a negative evaluation). This creates a disincentive to downvote which is most strongly felt by users with lower reputation scores (which correlates with our measure of status on Stack Overflow). When we remove low-status evaluators ($\sigma_E < 100$), this effect disappears and the overall picture looks the same as it does on Wikipedia (see Figure 4b).

We have just shown that similarity significantly controls the extent to which differential status influences evaluations. Now we highlight another key observation relating similarity and status.

In Wikipedia elections, we find that the evaluator's similarity to the target depends strongly on Δ (Figure 5a). In particular, the evaluators who have higher status than the target are significantly more similar to the target than are the evaluators who have lower status. This is an important *selection effect*: in aggregate, users with higher status than the target only tend to show up and vote on targets who are active in the same areas as they are. This means that voters who show up to evaluate a particular target are not necessarily representative of the voting population on Wikipedia. This effect is independent of the target's status – it holds and is quantitatively similar for both relatively high- and low-status targets.

This selection effect does not happen on Stack Overflow or Epinions (see Figure 5b). The shape of how s varies with Δ for evaluator-target pairs is almost identical to how it looks for pairs of random users, implying that there is no similar selection effect. The one difference between the two curves is that for random users the curve is symmetric across $\Delta = 0$, whereas for evaluator-targets pairs the similarity is a bit higher the left of $\Delta = 0$. Thus, there is a slight selection effect in the opposite direction on Stack Overflow: evaluators with lower status than the target tend to be more similar than when they have higher status. But we emphasize that this selection effect is slight in comparison to the one on Wikipedia, which completely overwhelms the baseline shape with a global minimum at $\Delta = 0$.

We hypothesize that the lack of a significant selection effect on



Stack Overflow

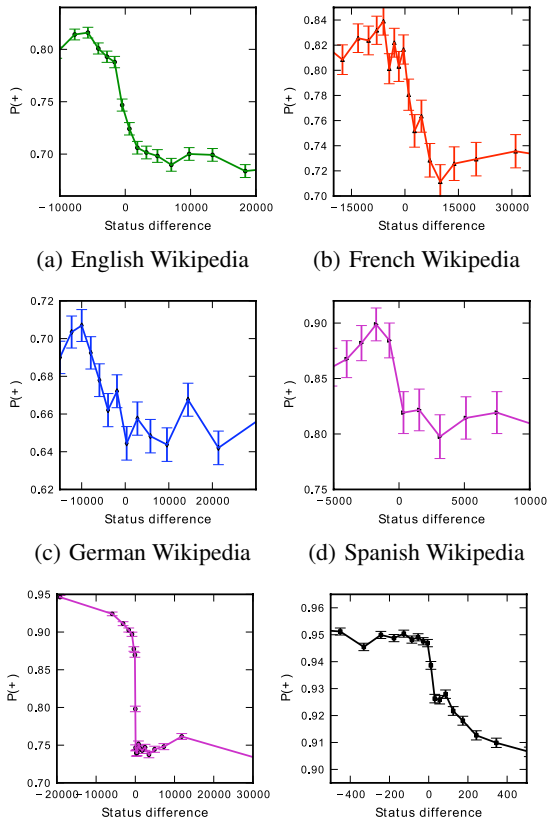
Figure 6: (Stack Overflow) Probability of E positively evaluating T versus σ_T for various fixed levels of Δ.

Stack Overflow similar to one we see on Wikipedia is due to the very different contexts that the evaluations occur in on Stack Overflow compared to Wikipedia. On Wikipedia, the vote is the central focus of the evaluator-target interaction we observe. An evaluator needs to care enough about the target's potential adminship in order to show up and vote – he must intend to vote. On Stack Overflow, evaluating questions and answers is a by-product of organic browsing behavior. Users mainly visit a question/answer page to learn something specific, and only then decide whether to offer an evaluation or not.

Absolute and Differential Status. Earlier we discussed the methodological question of how to compare an evaluator's status with a target's status, and we return to this issue now. For all of our datasets we find that status difference is the main factor determining evaluation outcome, as opposed to the absolute status or, for example, the ratio of the statuses. That is, the probability of evaluator E giving a positive assessment of target T conditioned on a status difference between them of $\Delta = \sigma_E - \sigma_T$ depends primarily on Δ, and not on their absolute status levels σ_E or σ_T.

This is illustrated in Figure 6, where we plot the fraction of positive evaluations $P(+)$ against the target's status σ_T within several narrow Δ ranges on Stack Overflow. If the status difference is really how users compare their status against others, then we would expect to see two things. First, that these curves are approximately flat (i.e. $P(+)$ doesn't depend on σ_T for a fixed Δ), because this would imply that for pairs separated by Δ, evaluation positivity does not depend on what their individual statuses are. Second, that the level of these constant curves depends on Δ, so that different Δ values result in different evaluation behavior. In Figure 6, this is exactly what we see. The fraction of positive votes does not significantly vary with absolute status level, meaning that even users of vastly different absolute status levels treat differences of Δ the same; and this constant level depends monotonically on Δ (the higher the Δ range, the lower $P(+)$ is). Neither of these observations hold if we consider narrow status ratio σ_E/σ_T ranges instead of status difference ranges. The last point on the $\Delta \in [500, 1000]$ curve is an outlier far from the rest of the curve and can be ignored.

Note that Δ does not control the result well for extremely low target status values, where evaluators are much more negative than just the difference in status would otherwise suggest. This is because the poor quality of targets with low status overwhelms the difference in status between evaluator and target. Thus absolute status, and not differential status, is the main criterion evaluators use to evaluate very low-status targets.



(a) English Wikipedia (b) French Wikipedia

(c) German Wikipedia (d) Spanish Wikipedia

(e) Epinions (f) Stack Overflow

Figure 7: Examples of the "dip" in various datasets.

To summarize, in this section we have analyzed how similarity affects user evaluation in various social media domains. The main insights we've discovered are: (1) positive evaluation is strongly positively correlated with similarity, (2) the more similar the evaluator is to the target, the less status-conscious he is, (3) there is a strong selection effect of high-similarity, higher-status evaluators showing up to vote on targets on Wikipedia (but not on Stack Overflow or Epinions), (4) methodologically, Δ is the correct measure for relative status, and (5) for low enough levels of target status, the target's low status overwhelms other factors.

5. MODELS OF AGGREGATE USER EVALUATION

We now turn our attention to designing a model that explains an interesting phenomenon first observed in our previous research [14]. As noted in Section 1, this work found that the probability of positive evaluation $P(+)$ is consistently non-monotonic in status difference Δ to the right of 0. We also observe this effect in our datasets as shown in Figure 7, with the "dip" emerging at small positive Δ. This phenomenon is intriguing because it applies across every user evaluation dataset we consider here, and it suggests that in aggregate, users are particularly harsh on each other when they have approximately the same status. Understanding why this happens, however, is challenging. In particular, the fact that users of comparable status are harsh on each other is intuitively at odds with our findings in the previous section that users who are similar in other dimensions tend to be more positive toward each other. Can we use our insights into how similarity affects evaluations to explain this dip and subsequent rebound in $P(+)$?

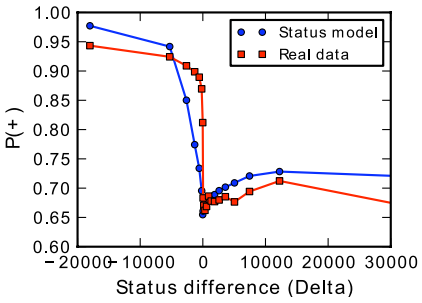


Figure 8: (Epinions) Fraction of positive votes versus status difference Δ.

In this section, we propose such an explanation. The explanation is two-pronged, designed to capture what appear to be two distinct aspects of this dip in $P(+)$. Recall first that there are two main regimes of user evaluations: (A) low-status targets tend to garner evaluations that depend mainly on the target's absolute status, whereas (B) evaluations of higher-status targets are functions of the differential status between the pair.

The distinction between (A) and (B) is reflected in two different behaviors for the dip in $P(+)$ across the datasets. For Stack Overflow and Epinions, we find that the dip disappears when low-status targets are filtered out, but for Wikipedia the dip persists even on the population consisting only of higher-status users. This provides an indication that different mechanisms may be contributing to the dip across different datasets. We now propose a pair of models that capture these different aspects.

Status model. First we present a model that is consistent with behavior on Stack Overflow and Epinions. When we examine $P(+)$ as a function of Δ, we are aggregating over the two different regimes of user evaluations. In Section 4, we showed that evaluation behavior is qualitatively different for low-status targets. Since most evaluators in the data have relatively low status, this implies that most evaluations in the absolute status evaluation regime will have small positive Δ values. And since these evaluations are much more negative than those made in the relative status regime, this can cause a dip slightly to the right of 0.

More concretely, we propose the following simple absolute-status-based mechanism that can generate the dip. We assume that evaluators do not discriminate among high-status targets, but for low-status targets they are increasingly negative as the target's status decreases. We capture this with a simple function: in the high-status target range ($\sigma_T > \sigma_{min}$) evaluators evaluate positively with a constant probability $P_E(+|T) = p^*$. In the low-status target range ($\sigma_T < \sigma_{min}$), evaluators evaluate a target positively with probability proportional to their status, $P_E(+|T) \propto \sigma_T$. We semi-simulate this process where we use real status values and $P_E(+|T) = (p^* - p_0)\sigma_T/\sigma_{min} + p_0$ where we set $p^* = 95\%$, $p_0 = 65\%$, $\sigma_{min} = 8000$. For each evaluation in our dataset, we use the real status values σ_E and σ_T (and hence real Δ) values, but replace the evaluation with a synthetic one sampled from the above $P_E(+|T)$. The resulting synthetic curve is overlayed with the real curve in Figure 8. Note that the model does not have a simple parametric form because the real σ_E and σ_T values were used.

Similarity model. The previous status-based model aligns well with the behavior of the dip in $P(+)$ on Stack Overflow and Epinions. On Wikipedia, however, the dip persists even when we eliminate targets of low status. Here we propose a mechanism that takes similarity into account in order to fully explain the phenomenon.

In Section 4 we observed that high similarity between evaluator and target yields more positive votes, and that Wikipedia elections have a persistent selection effect: higher-status evaluators are more similar to the target than lower-status evaluators are. In the following we show that these two effects alone are sufficient to cause the non-monotonicity we observe. Note that the function we are seeking to model is $\Pr[+ \mid \Delta]$, and that this aggregates over all levels of similarity. We also know that for small positive Δ values, the mix of similarities shifts more towards high-similarity pairs than it does elsewhere. Finally, we know that higher similarity correlates well with positive evaluations. The combination of these two observations leads us to the following model where we aim to recreate a dip at $\Pr[+ \mid \Delta = 0]$: We write:

$$\Pr[+ \mid \Delta] = \sum_s \Pr[+ \mid \Delta, s] \cdot \Pr[s \mid \Delta]$$

where the summation ranges over the domain of similarity s. By Bayes' Rule the right-hand side can be written as

$$\Pr[+ \mid \Delta] = \sum_s \frac{\Pr[+, \Delta, s]}{\Pr[\Delta, s]} \cdot \frac{\Pr[\Delta, s]}{\Pr[\Delta]}$$

Next we show how to model $\Pr[+ \mid \Delta, s]$. Figure 3 showed that the change in $P(+)$ as a function of status difference is more drastic for low values of similarity. This suggests that we can model $\Pr[+ \mid \Delta, s]$ with a function of the form

$$\Pr[+ \mid \Delta, s] = (a - bh(s)\Delta),$$

for constants $a, b > 0$ and a function $h(\cdot)$ that is monotone decreasing in s. In other words, for each fixed s, the probability $\Pr[+ \mid \Delta, s]$ is approximately linearly decreasing in Δ, with a slope controlled by s in such a way that larger values of s produce shallower downward slopes. Since the definition of similarity is arbitrary, we can re-parametrize the definition of similarity so that this monotone function $h(\cdot)$ becomes $h(s) = 1 - s$. We use this functional form and now we have:

$$\begin{aligned}
\Pr[+ \mid \Delta] &= \sum_s \Pr[+ \mid \Delta, s] \cdot \Pr[s \mid \Delta] \\
&= \sum_s (a - b(1-s)\Delta) \cdot \Pr[s \mid \Delta] \\
&= \sum_s (a - b\Delta + b\Delta s) \cdot \Pr[s \mid \Delta] \\
&= (a - b\Delta) \sum_s \Pr[s \mid \Delta] + b\Delta \sum_s s \cdot \Pr[s \mid \Delta] \\
&= (a - b\Delta) + b\Delta E[s \mid \Delta].
\end{aligned}$$

The advantage of working with this last expression is that one does not need to work with the full set of conditional probabilities $\Pr[s \mid \Delta]$, but instead only with the conditional expectations $E[s \mid \Delta]$ as a function of Δ. The conditional expectation $E[s \mid \Delta]$ is simply the mean similarity among evaluator-target pairs (A, B) for which the status difference is Δ. We write $f(\Delta) = E[s \mid \Delta]$.

Function $f(\cdot)$ can be approximated from the data. Based on our findings in Figure 5 we know that $f(\cdot)$ has a roughly inverted unimodal shape with a unique local minimum near 0, and that it rises higher to the right of 0 than to the left. It is also natural to assume that the similarity eventually saturates as Δ moves away from 0 in either direction — in other words, that

$$\lim_{x \to -\infty} f'(x) = \lim_{x \to \infty} f'(x) = 0.$$

A natural function that has a unique local minimum at $x = 0$ and

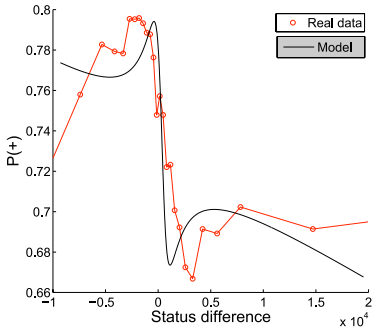


Figure 9: (English Wikipedia) The dip in real data versus the differential status model.

the saturation phenomenon is

$$f(x) = \frac{c(\alpha x)^2}{1 + (\alpha x)^2}$$

for positive constants c and α.

We show results for this model in Figure 9 and compare it to the real Wikipedia "dip". For $f(x)$ we set parameters $c = 0.98$ and $\alpha = 40$. Notice that the behavior qualitatively well matches the real data. Moreover, we also note that in our model for any fixed level of similarity, the probability of positive evaluation decreases monotonically with Δ. The more similar the evaluation-pair, the less steeply $P(+)$ drops off. This is consistent with our previous observations in Figure 3.

Thus although any individual $P(+)$-vs.-Δ curve for a particular similarity level is monotonically decreasing (as in Figure 3), the persistent dip on Wikipedia can be obtained by a weighted average over all these curves with the selection effect observed in Section 4. This highlights the importance of the selection effect and explains why the dip persists after thresholding out low-status targets on Wikipedia but not on Stack Overflow or Epinions (because the selection effect only occurs on Wikipedia).

In this section, we applied the insights we explained in Section 4 to theoretically model the mechanism behind the interesting "dip" phenomenon first observed in [14]. Now we turn our attention to showing that Section 4's discoveries are also useful in a practical learning setting.

6. BALLOT-BLIND PREDICTION

Here we aim to show that evaluator-target similarity has predictive power. The task we focus on is predicting administrator promotion on Wikipedia from the early voters, without looking at the sign of their votes. Just by looking at properties of the first few voters who show up (and importantly, their similarity with the candidate up for election), we are able to predict whether the election succeeds or not with accuracy that significantly improves on a number of natural baselines. Since we're not allowed to look at the actual votes to make our prediction, we call this task *ballot-blind prediction*.

Inferring information from a small prefix of users can be useful for any application with implicit user feedback. Our results raise the possibility of being able to infer an audience's approval or disapproval of a piece of content purely from the makeup of the audience (e.g., readers of a news article, viewers of a video, etc.).

We are interested in predicting a final outcome about a particular user from the first few evaluations we see. For this task we focus on Wikipedia because it has the most natural outcome variable that is a direct result of the evaluations we observe (election success).

Experimental setup. The Wikipedia data consists of public elections where users vote whether candidates should be granted administrative status or not. Formally, an election (T, Ω, R) is specified by a target (or *candidate*) T, a set of ordered votes Ω, and a result $R \in \{+1, -1\}$ denoting whether T was promoted or not. Each vote $v_i \in \Omega$ is a tuple (E_i, s_i, t_i) where E_i is the evaluator, $s_i \in \{+1, -1\}$ denotes whether the vote is positive or negative (we disregard neutral votes), and t_i is the time when E_i cast the vote. The votes are sorted by time, so $t_1 < t_2 < \ldots < t_m$, where m is the number of votes cast in the election. The task then is to predict R from the first k votes $v_1, \ldots, v_k$ (without looking at the sign s_i of each vote) using the similarity and status of the evaluators relative to the target. Our key finding is that knowledge of the similarity and status of the early voters alone (and not what they actually voted) provides enough useful information to successfully predict the final result.

We evaluate performance using accuracy on leave-one-out cross-validation, where we train on the entire dataset minus one example and test on the example for every example in the dataset.

In English Wikipedia, we have approximately 120,000 votes made by approximately 7,600 unique voters across 3,422 elections on 2,953 distinct candidates. (The number of elections exceeds the number of distinct candidates since some candidates go up for election more than once). We use $k = 5$ in our experiments, so that we're only looking at the first 5 voters. This ballot-blind prediction is difficult for two reasons: first, we are only using information from the first 5 voters (the average election length on English Wikipedia is 44 votes, so we only see the first 11% of the votes), and second, we aren't allowed to look at the actual sign of the vote.

Features for learning. We work with the following information. For each vote v_i, we know the identity of the evaluator E_i, her similarity $s(E_i, T)$ with the target, her status σ_{E_i}, and the status difference $\Delta_i = \sigma_{E_i} - \sigma_T$ between her and the target.

Note that in the previous sections our analysis and models have focused on aggregate behavior. Since we're now predicting on a per-instance basis, it makes sense to use per-instance features. We use each voter's historical fraction of positive votes P_i (excluding the current vote, since we aren't allowed to look at it), which we call their *positivity*. If they have no other votes in the dataset, we define their positivity to be the global positivity across the entire dataset (the overall fraction of positive votes across all voters).

We use two classes of features:

- Simple Summary Statistics (**S**) of the status and similarity of the target and the evaluators: $\log(\sigma_T)$, mean similarity $\bar{s} = \sum_{i=1}^{k} s(v_i, T)/k$, and $\bar{\Delta} = \sum_{i=1}^{k} (\sigma_{E_i} - \sigma_T)/k$.
- Δ-s quadrants: We partition the Δ-s space into into four quadrants (shown in Figure 10), divided by $\Delta = 0$ and a similarity value that roughly splits the evaluations into two equally-sized buckets (for example, $s = 0.025$ in English Wikipedia). This gives us four features where each simply counts the number of voters coming from a particular Δ-s quadrant.

Baselines. Before developing our final approach in detail, we describe three baselines we compare against.

The first baseline is a logistic regression classifier that uses all of the features we described above: 4 features of Δ-s quadrants and the **S** statistics. We consider this a very strong baseline as it uses a wide range of basic features that describe the voter population in a given election. We refer to this model as **B1**.

The second baseline models the probability of voter E_i voting positively (without considering his relation to the candidate). We estimate the probability of user E_i voting positively $P(E_i = 1)$ as

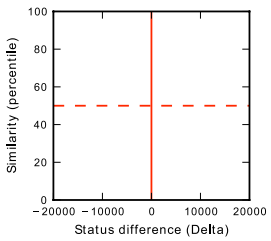


Figure 10: The Delta-similarity half-plane. Votes in each quadrant are treated as a group.

the empirical fraction E_i's votes that are positive. The estimated fraction of positive votes in a election is then simply the average $\frac{1}{k}\sum_{i=1}^{k} P_i$ of first k voters. We then learn the optimal threshold to predict election outcome. We call this baseline **B2**.

We also define a "gold-standard" (**GS**) that represents the best possible performance we can hope to achieve. It "cheats" by examining the values of actual votes themselves (s_i, for $i \leq k$), computes the empirical fraction of positive votes and learns the optimal threshold to predict the election outcome.

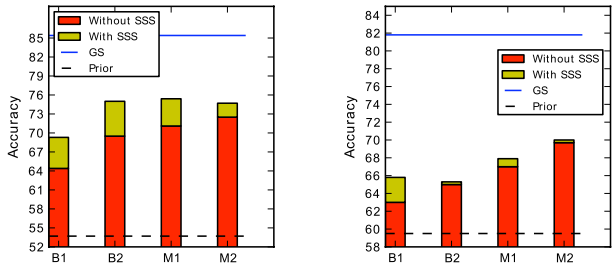
Proposed methods. Now we develop our main methods. The main idea is to directly incorporate how similarity and relative status drive deviations from default voting behavior. Thus, our models have two components: (1) P_i, which captures the overall positivity of user i, and (2) a deviation component that quantifies the deviation in the positivity of user i as a function of her similarity s_i and status difference Δ_i with the target.

In the first method (**M1**), we model the probability that E_i votes positively as $P(E_i = 1) = P_i + d(\Delta_i, s_i)$, where P_i is E_i's positivity, and $d(\Delta_i, s_i)$ is the average deviation of the fraction of positive votes in the Δ_i, s_i bucket compared to the overall fraction of positive votes across the entire dataset. This way we are using the evaluator's positivity as an "absolute" level and then adjusting it, based on how s and Δ affect the probability of positive votes across the *entire* population. We compute the average $P(E_i = 1)$ for $i = 1 \ldots k$ and then threshold it to make a prediction. **M1** thus uses 4 deviation parameters (one for each Δ-s quadrant), the threshold, and P_i for each user i.

In the second method (**M2**), we extend the model to factor in how s and Δ affect a voter's behavior at the *individual* level. We model $P(E_i = 1)$ as $P(E_i = 1) = \alpha \cdot P_i(\Delta_i, s_i) + (1 - \alpha) \cdot d(\Delta_i, s_i)$, where d is the same as above, and $P_i(\Delta_i, s_i)$ is E_i's positivity in the Δ, s_i bucket. This can be thought of as establishing an absolute level of the evaluator's positivity (as above), but then making relative adjustments based on personal as well as global deviations based on s and Δ. Now we are directly modeling how s and Δ affect E_i's likelihood of voting positively. We use $\alpha = 0.6$ in the results below, but the results are essentially the same with α anywhere between 0.6 and 0.9. **M2**'s parameters are: 4 deviation parameters for each user, P_i of each user, α, and the threshold.

Discussion of results. We find that the performance of the different approaches varies across our datasets, reflecting the different attitudes towards voting displayed by the various Wikipedia populations we're studying. Here we examine both the results on English Wikipedia as well as German Wikipedia to illustrate the differences (French and Spanish Wikipedia results were very similar to the German results).

Figure 11 plots the classification accuracy of our models. In English Wikipedia, where the prior on elections (i.e., the accuracy of random guessing) is around 54%, **B1** trained on Δ-s features alone give us 64% accuracy (and using the simple summary statistics **S** boosts its performance close to 70%). This shows how surprisingly informative similarity and status are, since we're not utilizing voter identities here.



(a) English Wikipedia (b) German Wikipedia

Figure 11: Ballot-blind prediction results.

Baseline **B2**, which looks at the identities of the individual evaluators and uses their positivity, gives an additional significant boost. On English Wikipedia, using just the positivities alone (**B2** without the **S** features) does as well as **B1** with **S** (70%). Adding the **S** features to **B2** gives an unusually large boost to 75%, which we only observed on English Wikipedia. On German Wikipedia, **B2** performs about as well as **B1** and hardly benefits from **S**.

Our best results come from the personal approaches that explicitly model deviations from voters' positivities (models **M1**, **M2**). In German Wikipedia, both models improve significantly on the baselines, with the **M2** approach improving on both **B1** and **B2** almost as much as the baselines improve on the prior (70% versus 64% and 66%, where the prior is just below 60%). In English Wikipedia, both models **M1** and **M2** significantly outperform **B1**, and marginally out-perform **B2** (the gain is smaller because of **B2**'s significant, idiosyncratic gain from the summary statistics).

Furthermore, notice that on German Wikipedia **M2** without the informative **S** features beats both baselines with access to **S**. Adding **S** to our second model gives us performance that is halfway to even the gold standard. This means that by only using s, Δ, and P_i values of the early voters – and ignoring the votes they casted – we can still recover half of the information contained in their votes.

We also ran experiments to measure how much predictive power similarity and status each contribute. We created two versions of our best model (**M2**): one which only uses similarity and one which only uses status. Thus instead of splitting the Δ-s space into quadrants, the first just splits along the similarity dimension (creating two buckets) and the second splits along the Δ dimension. In English Wikipedia, the original **M2** without **S** achieves 72.7% accuracy. Using only status, it achieves 70.7%; using only similarity, it scores 68.1%. Similarity's 2% contribution is the majority of **M2**'s gain over **B1**. On German Wikipedia, the full **M2** model scores 70%, whereas it scores 65.3% using only status and 68.5% using only similarity. In both cases, similarity contributes a significant amount of predictive power, and in German Wikipedia it is even more predictive than status. This experiment shows that similarity is a major component in the success of our models.

All of our approaches give us a quantitative picture of how evaluations depend on where they fall in the Δ-s space. For example, in our first model **M1**, where we model $P(E_i = 1) = P_i + d(\Delta_i, s_i)$, the $d(\Delta_i, s_i)$ function imposes an ordering over quadrants indicating which quadrants have the highest fraction of positive evaluations. Baseline **B1**, which learns weights corresponding to each Δ-s quadrant, also learns coefficients that impose a similar ordering. In all the experiments we ran on all of the models we introduced in this section, the relative importance of the quadrants remained

	$\Delta < 0$	$\Delta > 0$
High s	4	3
Low s	2	1

Table 2: Relative importance of quadrants in Δ-s space (1 = highest weight, 4 = lowest weight). This ranking is robust to changes in the election prefix length, similarity metric, etc.

unchanged. This ranking, shown in Table 2, agrees with all of the observations we made in Section 4: votes from evaluators with low similarity to the target are more predictive of election success than those with high similarity, and this split between high and low similarity is more important than the differential status between evaluator and target — a further indication of the important role that similarity is playing in this analysis. Also, votes from higher-status evaluators are more predictive of election success than those from lower-status evaluators.

These results demonstrate that without even looking at the actual votes, it is possible to derive a lot of information about the outcome of the election from a small prefix of the evaluators. As we mentioned at the start of this section, our results suggest that very informative implicit feedback could be gleaned from a small sampling of the audience consuming the content in question, especially if previous evaluation behavior by the audience members is known.

7. CONCLUSION

We have investigated a number of fundamental ways in which similarity affects how users evaluate each other on-line. This has enabled us both to explain open theoretical questions — such as the low aggregate evaluations given by users to others of comparable status [14] — and also to provide new methods for predicting the outcome of group evaluations from observing only the attributes of the evaluators and not their individual evaluations, which we term ballot-blind prediction.

Our work suggests several promising directions for further research. One direction is based on the fact that we can predict outcomes simply from the statuses and similarities of the users who show up to provide evaluations, without ever seeing the positive/negative values of the evaluations themselves. This suggests intriguing potential applications in which the composition of an audience can tell us something about the audience's reaction; it also may have potential synergies with recent techniques for modeling users in collaborative filtering contexts [11]. Another direction is to explore the spectrum of evaluation ranging from judgments about the content produced by individuals to overt judgments about the individuals themselves. We have seen that many of the underlying phenomena are remarkably similar across domains (Wikipedia and Stack Overflow) that occupy different points along this spectrum. Nonetheless different kinds of evaluation bring contrasts into play as well, and developing an understanding of these contrasts can help shed further insight into the ways in which users form collective judgments in social applications.

Acknowledgements. We thank Stack Overflow for providing us with their data. This research has been supported in part by a Google Research Grant, a Yahoo Research Alliance Grant, NSF grants IIS-0910664, CCF-0910940, IIS-1016099, IIS-1016909 and CNS-1010921, the Albert Yu & Mary Bechmann Foundation, IBM, Lightspeed, Microsoft and Yahoo.

8. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proc. WWW*, pages 665–674, 2008.

[2] M. J. Brzozowski, T. Hogg, and G. Szabó. Friends and foes: ideological social networking. In *Proc. CHI*, pages 817–820, 2008.

[3] M. Burke and R. Kraut. Mopping up: Modeling wikipedia promotion decisions. In *Proc. CSCW*, pages 27–36, 2008.

[4] R. S. Burt. *Neighbor Networks: Competitive Advantage Local and Personal*. Oxford University Press, 2009.

[5] D. Cosley, D. Frankowski, S. B. Kiesler, L. G. Terveen, and J. Riedl. How oversight improves member-maintained communities. In *Proc. CHI*, pages 11–20, 2005.

[6] C. Danescu-Niculescu-Mizil, G. Kossinets, J. M. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proc. WWW*, pages 141–150, 2009.

[7] R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proc. WWW*, 2004.

[8] W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In *Proc. CHI*, pages 194–201, 1995.

[9] S. Kamvar, M. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proc. WWW*, pages 640–651. ACM, 2003.

[10] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to usenet news. *Commun. ACM*, 40(3):77–87, 1997.

[11] Y. Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. *TKDD*, 4(1), 2010.

[12] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The Slashdot Zoo: Mining a social network with negative edges. In *Proc. WWW*, pages 741–750, 2009.

[13] P. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, and C. H. Page, editors, *Freedom and Control in Modern Society*, pages 18–66. Nostrand, 1954.

[14] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Governance in social media: A case study of the Wikipedia promotion process. In *Proc. ICWSM*, 2010.

[15] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in on-line social networks. In *Proc. WWW*, pages 641–650, 2010.

[16] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proc. CHI*, pages 1361–1370, 2010.

[17] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[18] P. Massa and P. Avesani. Trust metrics in recommender systems. *Computing with social trust*, pages 259–285, 2009.

[19] B. Pang and L. Lee. *Opinion Mining and Sentinment Analysis*. Number 2(1-2) in Foundations and Trends in Information Retrieval. Now Publishers, 2008.

[20] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proc. CHI*, volume 1, pages 210–217, 1995.

[21] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci.*, 107(31):13636–13641, 2010.

[22] S. Thye, D. Willer, and B. Markovsky. From status to power: New models at the intersection of two theories. *Social Forces*, 84:1471–1495, 2006.

[23] D. Willer (editor). *Network Exchange Theory*. Praeger, 1999.