# How Oversight Improves Member-Maintained Communities

**Dan Cosley†, Dan Frankowski†, Sara Kiesler‡, Loren Terveen†, John Riedl†**
**CommunityLab***

†GroupLens Research
University of Minnesota
4-192 EE/CS Building, 200 Union St. SE
Minneapolis, MN 55455 USA

‡Human Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213 USA

{cosley, dfrankow, terveen, riedl}@cs.umn.edu, kiesler@cs.cmu.edu

### ABSTRACT
Online communities need regular maintenance activities such as moderation and data input, tasks that typically fall to community owners. Communities that allow all members to participate in maintenance tasks have the potential to be more robust and valuable. A key challenge in creating member-maintained communities is building interfaces, algorithms, and social structures that encourage people to provide high-quality contributions. We use Karau and Williams' collective effort model to predict how peer and expert editorial oversight affect members' contributions to a movie recommendation website and test these predictions in a field experiment with 87 contributors. Oversight increased both the quantity and quality of contributions while reducing antisocial behavior, and peers were as effective at oversight as experts. We draw design guidelines and suggest avenues for future work from our results.

### Author Keywords
online communities, participation, contribution, member-maintained, oversight, quality, collective effort model

### ACM Classification Keywords
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—Collaborative computing

### INTRODUCTION
Chad is the movie czar, the de facto dictator who determines which movies are included in MovieLens. MovieLens is an online recommender system that has thousands of active users and around 8,000 movies—almost all of which Chad entered. He is the guardian of quality, the finder of movie facts, the defender of decency, and the final authority on film. This control suits Chad well. He is pleased with the quality of his system's movie database.

*CommunityLab is a collaborative project of the University of Minnesota, University of Michigan, and Carnegie Mellon University. http://www.communitylab.org/


*CHI 2005*, April 2–7, 2005, Portland, Oregon, USA.


Not everyone agrees. Chad adds movies slowly, because MovieLens is one of his many activities[1]. Some members don't understand—or agree with—his movie inclusion criteria ("widespread U.S. theatrical release"). When members request movies, Chad rarely responds. Despite his pride in MovieLens, it is incomplete. Some recent movies are not yet in the database; some movies entered last year have since been released on DVD and need to be updated in MovieLens; some members want MPAA ratings, which MovieLens does not provide; and so on.

Many groups, online and off, have a Chad. They rely on key members who manage and maintain the community. These tasks include moderation, governance, welcoming new members, and building Frequently Asked Question (FAQ) lists. Such maintenance tasks are critical to the community but typically fall to the community's owners [2], those who created the site, bought the machines, maintain the software and monitor the community's health. By contrast, most members' contributions to an online community pertain to the day to day business of the community, what Preece calls its *purpose* [16]. Members post to discussion groups, rate movies, receive recommendations, and read each others' blogs. These contributions are visibly important but do not sustain the community's infrastructure.

### Member-maintained communities
Chad does not have to be the only person who adds movie information. MovieLens could allow everyone to add movie information as well as rate movies. We believe that such *member-maintained communities* can be more robust and valuable than owner-maintained communities. Member-maintained communities can reduce their reliance on key individuals, draw on the resources of all members to add value to the group, and scale as the community grows.

With the promise come challenges. Most online communities have no way for members to help. People offer to add movies, but MovieLens has no interface to let them. Even it it did, MovieLens would need to make contributors aware of the opportunity and show them how to perform the task. Even with help, some people will do a poor job, while others may deliberately sabotage the community. The interests of members may not align with those of owners—try adding

[1]As Putnam points out, people who are willing to contribute often have no shortage of groups willing to accept their help [17].

a link to a competitor in an Amazon book review. Though we know that people are willing to contribute some effort to maintenance [5], successful member-maintained communities will often need to motivate people to contribute more. As one MovieLens member wrote, "I don't want to add movies. I want them to be there for me."

Our long-term research question addresses all of these challenges: *How can we design mechanisms—interfaces, algorithms, economies, and social structures—that allow communities to maintain themselves and encourage members to provide valuable contributions?*

### Using editorial oversight to improve contributions

We created a mechanism that allows MovieLens members to contribute movie information, a task that formerly fell to Chad alone. In this paper, we explore how editorial *oversight*—using other people to review contributions—affects the quality and quantity of movie information MovieLens members contribute. Oversight is an important social mechanism employed by successful member-maintained communities including Slashdot, Amazon, and Wikipedia.

We conducted a field experiment asking MovieLens members to add information for movies other members had already suggested. We divided participants into three groups: (1) a no oversight group whose contributions went directly to the database; (2) a peer oversight group whose work was checked by another member; and (3) an expert oversight group, checked by a movie expert. We told half of the subjects in each group about the amount of oversight they would receive. We expected oversight to lead to a more valuable database, more total contributions, and less antisocial behavior. The experiment confirmed these expectations. Further, it showed that peer oversight worked about as well as expert.

## THEORY: WHY DO PEOPLE CONTRIBUTE?

*Social dilemmas* arise when a group would benefit if its members made a certain choice, but its members have an incentive to make the opposite choice [4]. Consider fishing in a lake. If everyone fishes without limit, the lake will soon empty, so the community benefits when people limit their catch. But for any one person, the rational choice is to fish fully. After all, one person's restraint will not save the lake, no matter what everyone else does. In the long run, the entire community loses. This is an example of the tragedy of the commons [8] social dilemma.

Contributions toward community maintenance can be modelled as a kind of social dilemma, the problem of providing *public goods* [9]. Public goods have two distinguishing characteristics. Once they are produced, everyone can consume them. Further, one person's consumption does not prevent others from enjoying the good as well. National defense and open source software are two examples of public goods. The products of community maintenance activities, such as moderations on posts and movie information added to a database, are public goods as well.

Public goods are a social dilemma because rational people might decide that if others care enough to provide the good,

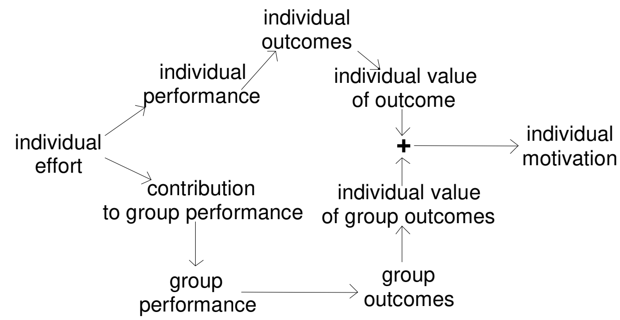


**Figure 1. A slightly simplified version of Karau and Williams' collective effort model.**

they need not expend the effort to help provide it themselves [12]. For instance, a small percentage of public radio listeners contribute money. The rest are free riders, enjoying the benefits without the cost. When people free ride, the public good is under-provided; the community would benefit if more of the good were produced. However, not everyone free rides. Many experiments have shown that people contribute to public goods under some conditions [5].

Social psychologists study the related problem of social loafing, the observation that people often expend less effort when working in a group than they would alone. Researchers have proposed a number of accounts of social loafing. For example, Kerr blames free riding [12]. Harkins suggests, more specifically, that people work less when it is hard for others to evaluate their individual contributions [10].

### The collective effort model

Karau and Williams integrate these and other accounts of social loafing into their collective effort model [11], a general model for understanding how people are motivated to contribute to groups. The model (see Figure 1) is based on Vroom's expectancy-value account of motivation [24]. This account suggests that one's motivation for a given effort depends on how well that effort translates into performance, what outcomes the expected level of performance is likely to lead to, and how much those outcomes are valued. The model posits that people consider these factors both from an individual and a group perspective. Further, in the case of groups, people consider whether their effort will make a contribution to the group's performance and whether that contribution will matter to the group's overall performance.

We believe the collective effort model is a rich tool for analyzing designs. Broadly, the model suggests that motivation increases when connections between elements in the model strengthen. For example, giving people a tool that reduces the effort required to add a FAQ entry should increase their motivation to do so, all other things being equal. By calling out these connections, the model can help designers reason about motivation and suggest strategies for increasing it.

Consider a professor deciding how to structure a group programming assignment. The model suggests students will be unmotivated if they think their efforts won't contribute to the group's performance, i.e., if they have no special skills to share. The professor might form groups where students have unique, required skills and, importantly, make students aware that their skills are unique so that they can see their contribution matters. Ludford et al. tried a similar approach in an online discussion group, telling people who had unique, relevant information to add to a conversation about their specialness. This tactic increased contributions [15].

We also could use the model to think about whether leader boards, like Amazon.com's list of top reviewers, encourage contributions. The model suggests leader boards might increase motivation by giving people a way to gain individual recognition, an outcome many people value. Leader boards also can demonstrate that one's effort visibly contributes to the group's performance. However, long-time participants have an advantage over newcomers, who may decide that they can never catch up and earn recognition. Newcomers might also reason that contributing a few reviews will mean nothing to the Amazon community, which has thousands of people who have written hundreds of reviews each. In other words, leader boards might actually reduce newcomers' motivation to contribute. A leader board showing both all-time and recent top contributors might prove to be a more effective design for increasing all members' motivation.

## OVERSIGHT IN MEMBER-MAINTAINED COMMUNITIES

We now examine how the use of oversight might improve member-maintained communities. We chose oversight because considering successful communities led us to realize that oversight can help by both discouraging low value contributions and motivating valuable contributors.

### Examples of communities that use oversight

Below, we discuss how we used the model to analyze how oversight would affect contributions to online communities. To ground the discussion, we introduce three representative online communities where member maintenance activities include oversight: Slashdot, Amazon, and Wikipedia. We also briefly describe peer oversight in MovieLens.

#### *Slashdot: moderating others' posts*

Slashdot lets technogeeks talk technospeak. Members read and comment on stories posted by editors and other members' comments. Members are sometimes allowed to moderate comments. Moderating a comment with a +1 or -1 rating and a descriptor such as Funny or Redundant takes a few seconds. Moderations are aggregated; readers can choose to see only highly rated comments or to see everything, a scheme Lampe and Resnick call *distributed moderation* [13].

#### *Amazon: rating review helpfulness*

Amazon encourages members to review books. These reviews are displayed on the book's web page on Amazon. Each review asks "Was this review helpful to you?" Readers may vote yes or no or report the review as inappropriate. Again, moderation is a low cost activity. The most helpful reviews for a book are designated as Spotlight Reviews and placed prominently on the page. Reader votes also are used to identify valuable reviewers, who get recognition both on a top reviewers list and with a special icon next to their names.

#### *Wikipedia: editing encyclopedia articles*

Wikipedia is building a collaborative encyclopedia using the *wiki* model of group editing [25]. Anyone can edit a wiki page, but every change is logged and can be easily repealed by later editors. Many regular wiki users use the "Recent Changes" button to see activity in the entire wiki, so they can reverse changes they view as harmful or improve new content. Users who care about a particular page may subscribe to all changes to the page so they can keep a watchful eye on modifications. Wikipedia uses these features together to enforce its neutral point of view philosophy and to quickly correct mistakes and deliberate vandalism [23].

#### *MovieLens: verifying movie information*

MovieLens asks members to help build its movie database by entering movie information and checking information entered by others. After one member adds information for a movie, another member is asked to verify the information. MovieLens is explicitly anonymous, so there is no communication between members or recognition of valuable contributors. Verifying a movie's information takes several minutes and often requires seeking information such as video release dates on other websites.

Our example communities all employ oversight to help identify quality contributions. We now use the collective effort model to examine how oversight might affect the motivation of people to contribute to member-maintained communities.

### Discouraging low-quality contributions

Not all contributions to a community are valuable. Off-topic conversation, newbie questions, incorrect FAQ entries, flames and trolls, spam, and content-free posts like "just testing" all represent contributions that most members would like to avoid. Too many low-quality contributions can actually drive away valuable members who decide that the cost of participating is too high [22].

A key component in the motivation of spammers and trolls is that by posting advertisements or inflammatory messages (individual performance), they get responses (individual outcome). Distributed moderation schemes like Slashdot's can sever this connection by featuring valuable contributions and tucking low-quality contributions away into a dusty corner of the interface. The collective effort model predicts that reducing the link between posting a message and getting responses should reduce spammers' and trolls' motivation to make low-quality contributions.

Another common low-quality contribution is vandalism—deleting the contents of a wiki page one finds offensive, for instance. Graffiti is a real-world analogue. Gangs mark their territory, wanting everyone to feel their presence. Just as New York City dramatically reduced subway graffiti by quick removal [7], editing in online communities can erase the traces of vandals. In fact, page deletions in Wikipedia are corrected in an average of three minutes [23]. According to the model, tactics like quick editing that remove the link between an individual's contribution and the group's performance should reduce vandals' motivation.

### Motivating valuable contributors

Oversight may do more than just keep people in line. It also may encourage contributions. Distributed moderation leads more people to read valuable comments. The collective effort model predicts that high-quality posters will be more motivated if they know they will have more readers. Amazon also recognizes members who write valuable reviews by putting them on the list of top reviewers. This makes the connection between a member's efforts and their value to the group's performance more salient, which the model suggests will increase motivation to contribute.

The collective effort model also predicts that helping people achieve better performance for a given effort will motivate them to contribute. Oversight can do this in several ways. Wikipedia's change tracking help editors be more efficient by concentrating their effort on articles they most value. Mentoring and feedback also can improve contributors' ability to turn effort into performance. Both Slashdot and Amazon provide feedback in the form of quality ratings.

Collaborative editing can strengthen the link between effort and contribution to the group's performance, if people believe others will improve on their contributions. Wikipedia articles often start as *stubs*, tiny blurbs about a topic of interest that other members eventually improve. People post stubs hoping that more knowledgeable members will expand on their effort. Likewise, MovieLens members who are afraid of making mistakes might be more willing to contribute if they know someone will check their work.

Finally, the model predicts that oversight is important because it reassures high-quality contributors that their contributions matter. Reducing the impact of low-quality contributions makes high quality work matter more to the group's performance. Slashdot gets a sea of low-quality contributions, but good ones float to the top. Reducing low-quality contributions also leads to better group outcomes—a livelier discussion, a better article, a richer movie database—that Thorn and Connolly argue will keep high-quality members coming back to get more, and to give more as well [22].

## EXPERIMENT: ADDING MOVIES TO MOVIELENS

We tested several of these predictions in a field experiment using MovieLens. MovieLens is a movie recommender system that performs the dual roles of providing good movie recommendations as a service while serving as a platform for research in recommender systems (e.g., [3,18]) and more recently, online communities (e.g., [1,15]). It has about 80,000 registered users, thousands of whom log in regularly and 8,000 movies. On average, members rate about 120 movies.

When MovieLens started, the database was full of inaccurate information and duplicate movies because the task of adding movies fell to grad students in their spare time. Chad was an early MovieLens member who loved the service but hated the problems with the database. He volunteered to take over, cleaned up the mess, and has contributed information for about 6000 movies, adding new releases from the last five years as well as an eclectic selection of older movies.

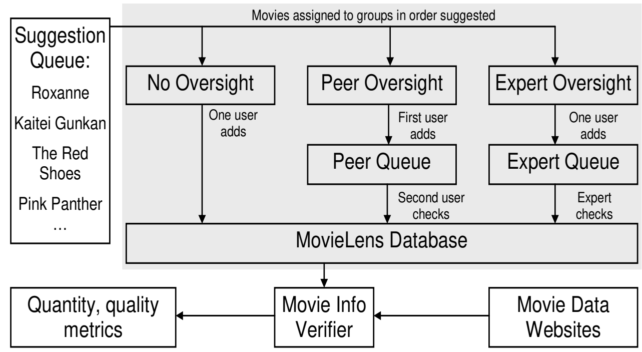


**Figure 2. A simplified overview of the experiment. Subjects were assigned to one of three oversight groups and assigned movies from the suggestion queue in the order received. At the end, we compared contributions against information from other movie data sites.**

Members sometimes want to rate movies that are not in the database. In August 2003 we added an interface for suggesting movies by entering their title and Internet Movie Database (IMDb, http://akas.imdb.com/) id. We decided not to ask people to add all of the information for a movie for two reasons. First, we wanted to keep members from wasting time entering information for movies that Chad would not accept. Second, Chad wanted to enter movie information himself to ensure high quality.

The suggestion feature was a hit, with 500 suggestions in the first month and 3000 in the first year. Members suggested so many movies that Chad could not keep up, creating a large backlog and making MovieLens members who felt they were being ignored unhappy. The backlog of suggested movies provided a reservoir of maintenance tasks we could ask members to perform in the experiment.

Figure 2 gives an overview of the experiment. Subjects were assigned to one of three oversight mechanisms; half were told about about the mechanism. Each subject was assigned to only one condition. We asked subjects to add information for movies that other members had suggested or to check information that others subjects had added. At the end of the experiment, we used information from other movie websites to evaluate quality and counted useful and total contributions to measure quantity. We also conducted a survey open to all MovieLens members, asking them how they thought various oversight mechanisms would affect both their willingness to contribute and the quality of MovieLens.

### Hypotheses about oversight in MovieLens

The collective effort model suggests several hypotheses about how oversight would affect the quality and quantity of movie information contributed by MovieLens members.

Our first hypothesis is that oversight will reduce antisocial behavior. People can make low-quality contributions to MovieLens in several ways:

- Inaccurate or incomplete information: misspell a foreign title or fail to include video release dates.

- Deliberate sabotage: enter bogus or obscene information.
- "Hijack" a movie: instead of the movie assigned by MovieLens, add a different movie.

Deliberate sabotage is a low-quality, antisocial act. We also consider hijackings to be anti-social. To be fair to all, we wanted to add movies in the order they were suggested. People who add their own movies put their interests ahead of the community's. We believe subjects in oversight groups will expect hijackings and sabotage to be corrected, which according to the collective effort model should reduce their motivation to make these anti-social contributions:

**Hypothesis Less Antisocial Behavior.**
*More oversight leads to less antisocial behavior.*

We also wondered whether oversight would encourage each individual contributor to be more accurate when adding movie information: would knowing that someone would be checking their work motivate people to try harder? When we first discussed this question, we disagreed about whether oversight would decrease, increase, or have no effect on the accuracy of individual contributions. The collective effort model, alas, can support all three viewpoints.

- People will think their poor contribution will not affect the group's performance because someone else will fix it. Oversight reduces quality.
- High-quality contributors will decide their contributions matter more to the group's performance when low-quality contributions are suppressed. Oversight increases quality.
- People value the outcome of having more movies in MovieLens so much that they try their best no matter what. Oversight does not matter.

We decided that the combination of reducing low-quality contributions and motivating high-quality contributors would be the strongest effect, and thus that oversight would increase the accuracy of each individual contribution:

**Hypothesis Better Contributions.**
*More oversight leads to higher quality initial contributions.*

Our third hypothesis deals with the effect of oversight on the final quality of information that makes it into MovieLens. No matter what effect oversight has on individuals' accuracy when entering movies initially, we expect that a process where each contribution is checked will result in a higher final quality database than a no-oversight process that doesn't check contributions:

**Hypothesis Better Database.**
*More oversight leads to higher final quality information in the database.*

Our fourth hypothesis is that oversight will affect the quantity of contributions. The collective effort model predicts that oversight inhibits low-quality contributors and motivates high-quality contributors. We believe that most people are good and that a majority want to make high quality contributions, so we expect overall motivation—and quantity of contributions—to increase:

**Hypothesis More Contributions.**
*More oversight leads to more contributions overall.*

Finally, we hypothesize that telling people about oversight matters. In the model, motivation depends on strong links between effort, performance, and outcomes. Telling people about the existence of oversight will increase its effect because people will see these connections more clearly:

**Hypothesis Knowledge Is Power.**
*Knowing that oversight is present or absent will increase the effect of oversight.*

### Independent variables: oversight and visibility

We examined three oversight mechanisms, corresponding to the vertical paths in Figure 2.

- *No oversight*: one subject adds a movie, which goes straight to the database.
- *Peer oversight*: one subject adds a movie, another checks the information, which then goes to the database.
- *Expert oversight*: one subject adds a movie, a movie expert checks the information and adds it to the database.[2]

These mechanisms cover two fundamental design decisions related to oversight. First, is it needed at all? Second, can everyone provide oversight or must it be a chosen few? We considered other issues (e.g., should oversight only happen for new items or should it be continuous, as in a wiki?), but decided to limit the experiment to three mechanisms to reduce the number of subjects needed.

We also manipulated whether subjects knew about the level of oversight. We thought that, just as fishermen might be more likely to obey catch limits when they know game wardens are present, people might be more affected by oversight if they knew their work would be checked. Each subject was assigned to one of the three mechanisms. We told half of the subjects in each mechanism about the level of oversight their contribution would receive. We told the other half we were trying several different ways of obtaining accurate movie information from members, without details about oversight. We call the groups where subjects knew about the level of oversight visible groups (***NoneVis***, ***PeerVis***, and ***ExpertVis***), and the others not visible (***NoneNV***, ***PeerNV***, and ***ExpertNV***). This gave us a 3 by 2 design with oversight and visibility as our independent variables.

MovieLens members were invited to participate through a link on the main page asking them to beta test a feature for adding movie information. Only members who joined MovieLens before the experiment began were invited. Subjects who clicked the link were randomly assigned to one of the six groups and presented with instructions that contained our manipulation and help on adding movies to MovieLens.

[2] Chad was too busy to participate in the experiment. We simulated expertise by having a MovieLens developer check movie information against other movie websites.

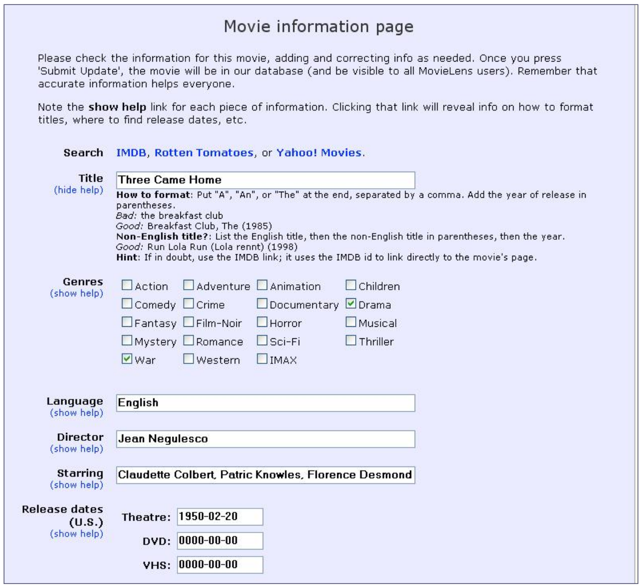


**Figure 3. The movie information interface.**

After reading the instructions, subjects could proceed to the movie information interface, shown in Figure 3.

Adding a movie to MovieLens is not easy. Movies have alternate titles; films often have multiple release dates, both in theatres and on video; MovieLens asks for five representative actors. The interface provides copious formatting help, links to the assigned movie on several popular movie sites, and accepts several date formats. It still took subjects an average of six minutes to add or verify a movie's information.

Whenever a subject visited the movie information interface, the experiment assigned them either the next movie from the suggestion queue or, if movies needed checking in ***PeerVis*** or ***PeerNV***, a movie that needed checking. The interface displayed the assigned movie and asked subjects to add or verify the movie's information. The interface also reinforced the oversight manipulation. It reminded subjects in visible groups that they were adding or checking information and told them whether the information would go directly to MovieLens or who it would be checked by (see the top paragraph of Figure 3). After making a contribution, a thank you page again reminded subjects about the manipulation. Subjects in the non-visible groups saw more generic instructions asking them to enter the movie's information and telling them their contribution would appear within a few days.

### Dependent variables: quality and quantity

We needed metrics for the quantity and quality of contributions. Quantity has an obvious metric: count the number of movies added or checked. We also counted live movies (movies successfully added to MovieLens) and hijackings (where someone entered information for a different movie than they were assigned). We removed hijacked movies as they came to our attention. This normally took about a day, because members mailed ten times during the experiment to let us know about hijacked movies. This surprised us because to know a movie is hijacked requires extra work. We realized that some people were monitoring our "Newest Additions" link and checking newly added movies for us, much as wiki users can monitor recently changed pages.

Choosing quality metrics was more complicated. Movies have several fields of information: title, genres, release dates (theatre, DVD, and VHS), actors, director, and language. We created an automatic checking program to evaluate contribution quality by comparing contributed information to movie information from the same websites our movie expert used. High-quality contributions contain complete information that matches the automatic checker.

Note that although automatic checking was useful for our experiment, it is not a general answer for either measuring quality or for helping people perform maintenance tasks. It is not a good general strategy for measuring quality because many tasks, like checking Wikipedia articles, do not lend themselves to automation. Also, harvesting data is not necessarily useful for helping people do maintenance. Though it was tempting to use the data our checker harvested to help people enter movie information, doing so would have run afoul of IMDb's terms of use (and would have made it harder to detect quality differences in the experiment).

We tried several metrics for scoring a contribution's quality. A simple version counts the number of correct fields, giving partial credit in cases where it was reasonable to do so. For example, if a subject misspelled one actor and spelled four correctly, they got 0.8 credit for the actor field. We tried versions of the counting metric that weighed each field by how often users search on it (title 50 times as often as language, for instance) and by how important users rated each field on a 1 to 5 scale (title is 3 times as important as VHS release date, for example). Weighted versions of the metric correlate strongly with the simple count ($r^2 > 0.80$) so we only report results using the counting metric. Hijacks were compared to the assigned movie's information just like any other contribution, typically earning scores near zero.

| Group | Total Subjects | Contributors | Total Work | Live Movies | Hijacks |
|---|---|---|---|---|---|
| ***NoneVis*** | 31 | 10 | 61 | 20 | 41 |
| ***NoneNV*** | 33 | 14 | 88 | 67 | 21 |
| ***PeerVis*** | 38 | 19 | 130 | 63 | 11 |
| ***PeerNV*** | 35 | 13 | 64 | 38 | 2 |
| ***ExpertVis*** | 32 | 17 | 140 | 109 | 11 |
| ***ExpertNV*** | 35 | 14 | 92 | 83 | 7 |

**Table 1. A summary of participation in the experiment.**

## RESULTS

The experiment lasted six weeks. A total of 204 users followed the invitation link, with 87 making at least one contribution. Table 1 shows how many subjects were assigned to each group, how many contributed, and the total number of contributions and movies added by each group.

Contributions in MovieLens such as ratings and posts often follow an exponential distribution—most people contribute

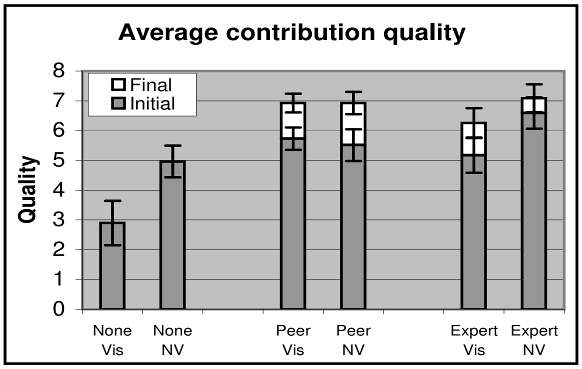


**Figure 4. Quality of the initial and final contributions for a movie. For *NoneVis* and *PeerVis* the initial and final contributions were the same. No oversight performs worse on final quality, especially for the *NoneVis* group.**

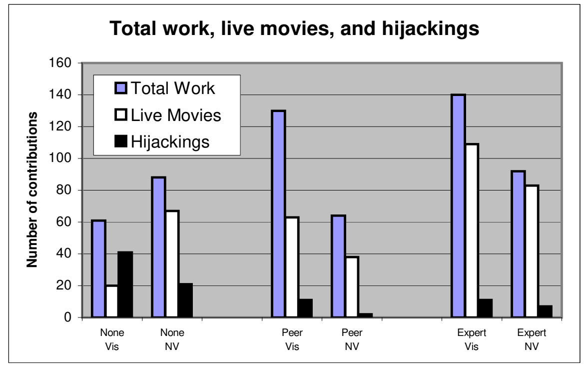


**Figure 5. Total contributions, movies successfully added, and hijackings. Peer oversight groups required two contributions per movie while other groups required only one. Groups with no oversight hijacked more movies.**

a little, a few contribute a lot. Movie information contributions were no exception. A few early subjects entered dozens of movies, so we capped the number at contributions per subject at 20 for fear of running out of suggestions to add, a fear that was well-founded:

> *I also didn't like how you capped the number of additions I was able to make. In all honesty I would have added several hundred more had I been allowed.*

For the early users who made over 20 contributions, we used only their first 20 in our analysis.

### How oversight affected quality

Figure 4 shows the average quality of contributions made by each group at two stages:

- *Initial quality*: The quality of the initial contribution for a given movie, when it was first added by a member. We use this to test our **Better Contributions** hypothesis.

- *Final quality*: The quality of the a given movie's information when it was inserted into the database. This we use to test our **Better Database** hypothesis.

We used repeated measures ANOVAs to analyze these data, with oversight and visibility as between-subjects fixed effects and subject id as a random effect (an unbalanced repeated measure, since each subject added up to 20 movies).

On initial quality, all groups performed about the same except for ***NoneVis***, which was worse than the others. ***ExpertNV*** appears to perform slightly better. Neither oversight nor visibility, nor their combination, has a statistically significant effect, as it turns out that almost all of the variation is explained by differences between users. This analysis is based on 460 movie entries contributed by 68 subjects.

Final quality was influenced by oversight. Here, ***NoneNV*** and especially ***NoneVis*** lag behind all of the oversight groups. A repeated measures ANOVA, based on 478 movie entries from 60 subjects, shows that oversight has a significant effect on quality, $F(2, 97) = 10.58, p < 0.01$. It appears that there was little difference between peer and expert oversight, but people who were told or who discovered their work would not be checked felt free to hijack movies and make low-quality contributions. It also suggests that the extra quality added by an editor makes a difference.

### How oversight affected quantity

For each group, Figure 5 shows how many contributions it made, the number of movies successfully added, and the number of hijacked movies. Note that peer oversight groups needed two contributions to add a movie while other groups only needed one. The difference between total work and live movies is wasted work: contributions toward hijacked movies or movies our expert rejected as too obscure.

Among the visible groups, ***PeerVis*** and ***ExpertVis*** did the most work, the three non-visible groups much less, and ***NoneVis*** the least work at all. The differences are visually striking, but an ANOVA comparing the average amount of work, using all 204 subjects, showed that the differences were not statistically significant, $F(5, 198) = 1.23, p = 0.30$. We can combine groups based on how much oversight the group knew it would receive. ***PeerVis*** and ***ExpertVis*** knew there would be oversight. ***NoneVis*** knew there would be none. The three non visible groups did not know either way. An ANOVA that compares groups that knew about oversight, those that knew about no oversight, and those that didn't know either way shows that quantity tends to increase with oversight, $F(2, 201) = 2.56, p = 0.08$.

***NoneNV*** and ***NoneVis*** commit significantly more hijackings than other groups, $\chi^2(5, 495) = 119, p < 0.01$. ***PeerVis*** also had a relatively large proportion of hijacks, all of them perpetrated by one checker who replaced all of the information entered by other subjects. Surprisingly, no one attempted to enter bogus movie information during the experiment.

### How people think oversight affects contributions

Our post-experiment survey of people's attitudes toward oversight supports these results. We invited all MovieLens users to evaluate five oversight mechanisms: no oversight,

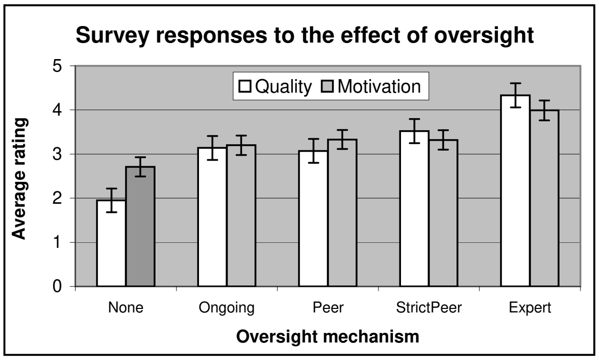


**Figure 6. Users' estimates of how well various oversight mechanisms would work. Users preferred experts to peers and peers to no oversight. Estimated quality and motivation to participate correlated strongly.**

peer oversight, expert oversight, *strict peer* oversight, and *ongoing* oversight. Strict peer oversight is like peer oversight, except a movie is checked until someone does not change any information. Ongoing oversight, like no oversight, lets members add movies directly to the database, but then any member can edit a movie at any time. Several people noted ongoing oversight is similar to wiki-style editing.

Respondents evaluated the mechanisms in two ways:

- *Estimated quality*: How well would each oversight mechanism work for getting accurate information?
- *Estimated motivation*: How likely would they be to add movie information if MovieLens used that mechanism?

Figure 6 shows the results. Low quality and motivation correspond to ratings of 1, while high quality and motivation correspond to ratings of 5. People prefer expert oversight and shun no oversight compared to the peer mechanisms. An ANOVA showed significant differences (quality, $F(4, 404) = 11.19, p < 0.01$ on 82 responses; motivation, $F(4, 400) = 59.03, p < 0.01$ on 81 responses). Willingness to participate and estimated quality correlated strongly, $r^2 > 0.52$ for all five systems.

## DISCUSSION

Our theoretical analysis and results raise a number of issues for discussion. We begin by analyzing the evidence for and against the hypotheses:

- **Less Antisocial Behavior** was supported. Groups with no oversight committed many more hijackings.
- **Better Contributions** was not supported. Groups with no oversight made lower-quality initial contributions but the difference was not statistically significant.
- **Better Database** was supported. The quality of movies that reached the MovieLens database was significantly lower for groups with no oversight.
- **More Contributions** had some support. Quantity tended to increase with oversight, and survey respondents were more willing to contribute with more oversight.
- **Knowledge Is Power** was supported. Knowing about oversight affected both quantity and database quality.

Oversight improved both the quantity and overall quality of information added to MovieLens with no negative effects. Below, we discuss other issues raised by our results including the relationship between peers and experts, between owners and members, and between theory and design in online communities.

### From peers to experts

Although when surveyed, people clearly preferred experts, in the experiment peer oversight did about as well as expert oversight on both the quantity and quality of contributions. This is important. For every contribution a subject in ***ExpertVis*** and ***ExpertNV*** made, our expert had to make a matching contribution. This does not scale. Since peer oversight works about as well as expert, designers of member-maintained communities should consider letting members provide oversight as well as content.

Teasing out when peer and expert oversight work best will be important. Peer and expert oversight achieved roughly the same quality in a domain of structured, factual information. In Wikipedia the content has less structure, contains arguments as well as facts, and sometimes requires specialized knowledge. How do peers compare to experts in domains like Wikipedia? Other factors, such as the size of a community and how its members regard authority, are also likely to impact how a community responds to oversight.

The collective effort model suggests that mechanisms for improving user performance such as training and mentoring may be important. User variability accounted for large differences in contribution quality. Reducing this variability through improving everyone's competence may amplify the effect of mechanisms such as oversight. The model also predicts that increasing individuals' competence should increase their motivation to contribute—though this effect may be countered by the fact that going through training increases the cost of participating. Finally, high quality contributors should feel reassured—and more motivated—knowing that the average contribution will be of higher quality.

One way to approach training and mentoring is to structure maintenance tasks around roles that require more or less expertise. The idea of communities of practice [14] suggests that role structures encourage the growth of members in the community. Having multiple roles also allows members to make the kinds of contributions they prefer and allows designers to use status and power to reward and recognize contributors, both of which are likely to increase motivation according to the collective effort model.

### From owners to members

Moving from owner to member maintenance poses risks. The move can threaten key members like Chad:

> *I am proud of the work I have done...this is like an architect having to watch his building being torn down. I fear that any more than a few days of this and the cleanup will become incomprehensible...you have to re-*

*member how IMDb keeps their stuff clean!! It's certainly not by way of the users!!*

Losing Chad would be costly; his contributions have great value. How can groups keep key contributors content, especially when their interests are at odds with those of other members? "What movies should be added to MovieLens?" has a simple answer when Chad does all the work: whatever Chad adds. A natural way to further include members in maintaining MovieLens would be to let them help decide what gets added. But they want to add lots of movies: obscure movies, TV movies, and non-U.S. releases. This is at odds with Chad's vision of MovieLens, while a large influx of movies might be bad for our recommendation algorithm. Designing a mechanism for adding movies that balances Chad's wishes, the needs of the algorithm, our research goals and the desires of members will be a challenge.

### A broader look at quality

In this paper, we called a newbie question a low-value contribution. But for the asker, for lurkers with the same question, and for members who want to demonstrate their knowledge, the contribution has high value. Even the quality of movie information is not as clear-cut as we might like. Some members have very high standards:

> *Inaccuracies make [IMDb] a first stop only...MovieLens would have to find people who have at minimum access to a set of reference books beginning with basic volumes...and who specialize further in a given genre or form (e.g. post-war American underground cinema..)*

On the other hand, 25 movies were added with no information except the title. No one complained about these movies and many were rated often. People did mention them in survey comments but most said the increased volume of movies made up for the lapses in quality.

This reminds us that quality is in the eye of the beholder. Distributed moderation does a good job of giving the oversight task to everyone, but it makes only aggregate quality judgments. It might be better to use people's ratings as input to a collaborative filtering system [19], allowing each member of the community to give more weight to the moderations of people they agree with.

### From hijacker to freedom fighter?

In the experiment we asked people to add information for movies other members suggested. Many suggestions were hijacked: members chose to enter data for a movie they cared about rather than for the movie they were asked to add. We considered hijackings to be low-quality because they violated our goal of fair processing of the suggestion queue. Some MovieLens researchers argue that hijackings were not necessarily low-quality contributions. Maybe the hijackers were freedom fighters, looking out for the community by adding movies they thought everyone would want instead of obscure movies that one person had suggested.[3]

Real member-maintained communities should allow members to "scratch their own itch". The collective effort model is clear that aligning individual and group outcomes will increase motivation to contribute. We expect that designs like Wikipedia that allow members to choose tasks are more likely to succeed. In retrospect, it was a mistake to seek fairness in processing the suggestion queue because it placed individual motivation in opposition to group motivation.

### From theory to design, and back

Using social science theory to drive and critique design is promising, but not foolproof. For instance, the collective effort model has several weaknesses. It does not directly account for the cost of an effort. Cost can be factored into valuing outcomes but should be explicitly considered. The model also does not address opportunity cost and how people decide between courses of action. Factors that affect motivation such as duty and morality do not fit neatly into the model. Finally, motivation is hard to quantify, and as we saw, the model sometimes makes ambiguous predictions.

Theory must be applied carefully. We originally speculated that oversight would reduce contributions. We focused on the collective effort model's predictions about inhibiting low-quality contributions, and only later realized that oversight will reassure—and motivate—high-quality contributors as well. Social scientists who can identify relevant theories and ease the friction when theories collide are a valuable asset. Theory properly used can help designers avoid mistakes and suggest approaches they otherwise might not consider. Applying social science theory to design is a promising, but under-explored approach [1].

We also believe in using design to help probe theory. The collective effort model was primarily derived from carefully controlled lab experiments on somewhat unrealistic tasks such as shouting in a group. The results of the field experiment support using the model in real online communities, where both "real" and "online" extend the reach of the model beyond lab experiments. When theory informs design, both stand to benefit.

## CONCLUSION

Based on our findings, we offer a number of guidelines for online community designers who wish to use oversight in helping members maintain their communities.

- Oversight improves outcomes and increases contributions. Use oversight mechanisms to improve quality, reduce antisocial behavior, and help reduce the risks of member-maintained communities.
- We found no differences between peer and expert oversight in quality or quantity of contributions. Take the burden off of community owners and share it with the members. Some of them really want to help.
- Major differences in quality can be attributed to individuals. Increase the quality of contributions by selecting for

[3]To test this theory we added twenty hijacked movies and the movies they replaced to MovieLens. We then watched ratings behavior over a period of four weeks, and there was no difference in ratings behavior between the original and the replacement movies. This suggests that the hijacked movies were *not* more popular than the movies they replaced.

the best contributors, and by improving the capabilities of individual users, e.g., through training.

- Telling people about oversight may increase their motivation to contribute. Tell them about oversight to encourage good contributors and discourage bad ones. (We do not recommend lying about oversight. Users will find out.)
- A number of users surveyed said they did not see our invitation link. Make opportunities to contribute obvious. Do not assume that ignoring an offer is intentional.

We also encourage designers and researchers in online community to incorporate theory into their design practice. Researchers have developed a number of tools such as reputation systems (e.g., [20]), social proxies (e.g., [6]), and mechanisms for making contributions visible (e.g., [21]) that might be useful in helping motivate members to maintain their communities. We believe that using theory to critique and drive design of such features, along with focused experimentation to validate the design choices, is an effective way to understand how and when to apply these mechanisms to support member-maintained communities. We hope that our exploration of the value of oversight helps other designers and researchers in their efforts to build and deploy systems that help people discover—and build—community online.


## ACKNOWLEDGEMENTS
Special thanks to Venkateswaran Udayasankar, who implemented much of the code required to conduct the experiment. This work was supported by grants 0325837 and 0324851 from the National Science Foundation.


## REFERENCES

1. G. Beenen, K. Ling, X. Wang, K. Chang, D. Frankowski, P. Resnick, and R. E. Kraut. Using social psychology to motivate contributions to online communities. In *Proc. CSCW2004*, Chicago, IL, 2004.
2. B. Butler, L. Sproull, S. Kiesler, and R. Kraut. *Community Building in Online Communities: Who Does the Work and Why?* Leadership at a Distance. Lawrence Erlbaum Publishers, Inc., Mahwah, NJ, 2005.
3. D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of SIGCHI*, pages 585–592, Ft. Lauderdale, 2003.
4. R. M. Dawes. Social dilemmas. *Annual Review of Psychology*, 31:169–193, 1980.
5. R. M. Dawes and R. H. Thaler. Anomalies: Cooperation. *The Journal of Economic Perspectives*, 2(3):187–197, 1988.
6. T. Erickson et al. Socially translucent systems: social proxies, persistent conversation, and the design of babble. In *Proc. SIGCHI*, pages 72–79, 1999.
7. A. Graycar. Graffiti: Implications for law enforcement, local government and the community. In *Graffiti and Disorder: Local Government, Law Enforcement and Community Responses*, Brisbane, August 2003.
8. G. Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.
9. R. Hardin. *Collective Action*. Johns Hopkins, 1982.
10. S. G. Harkins. Social loafing and social facilitation. *Journal of Experimental Social Psych.*, 23:1–18, 1987.
11. S. J. Karau and K. D. Williams. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4):681–706, 1993.
12. N. L. Kerr. Motivation losses in small groups: a social dilemma analysis. *Journal of Personality and Social Psychology*, 45:819–828, 1983.
13. C. Lampe and P. Resnick. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of SIGCHI*, pages 543–550, Vienna, Austria, 2004. ACM Press.
14. J. Lave. *Situating Learning in Communities of Practice*. Perspectives on Socially Shared Cognition. APA, 1993.
15. P. J. Ludford, D. Cosley, D. Frankowski, and L. Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of SIGCHI*, pages 631–638, Vienna, Austria, 2004. ACM Press.
16. J. Preece. *Online Communities: Designing Usability, Supporting Sociability*. John Wiley & Sons, Ltd, 2000.
17. R. Putnam. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, 2000.
18. A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan, and J. Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proc. IUI*, pages 127–134, 2002.
19. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of CSCW*, pages 175–186, Chapel Hill, NC, 1994.
20. P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.
21. M. Smith. Tools for navigating large social cyberspaces. *Communications of the ACM*, 45(4):51–55, 2002.
22. B. K. Thorn and T. Connolly. Discretionary data bases: A theory and some experimental findings. *Communication Research*, 14:512–528, 1987.
23. F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of SIGCHI*, pages 575–582, Vienna, Austria, 2004. ACM Press.
24. V. H. Vroom. *Work and Motivation*. Wiley, N.Y., 1964.
25. Wikipedia. Wikipedia:about - wikipedia. http://en.wikipedia.org/wiki/Wikipedia:About, 2004.