Precision CrowdSourcing: Closing the Loop to Turn Information Consumers into Information Contributors

Qian Zhao, Zihong Huang, F. Maxwell Harper, Loren Terveen, Joseph A. Konstan
GroupLens Research
University of Minnesota
{qian, zihong, harper, terveen, konstan}@cs.umn.edu

ABSTRACT
We introduce a theoretical framework called *precision crowdsourcing* whose goal is to help turn online information consumers into information contributors. The framework looks at the timing and nature of the requests made of users and the feedback provided to users with the goal of increasing long-term contribution and engagement in the site or system. We present the results of a field experiment in which almost 3000 users were asked to tag movies (plus a null control group) as we varied the selection of task (popular/obscure), timing of requests (immediate or varying delays), and relational rhetoric (neutral, system reciprocal, other users reciprocal) of the requests. We found that asking increases tags provided overall, though asking generally decreases the provision of unprompted tags. Users were more likely to comply with our request when we asked them to tag obscure movies and when we used reciprocal request rhetoric.

Author Keywords
Online communities; Crowdsourcing; Reciprocity

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces.

INTRODUCTION

User-contributed information is central to the rich and diverse set of online information resources on which we have all come to depend. Google Maps has rich information about places around the world thanks to user contributions through its Google Mapmaker website. Sites like TripAdvisor or Angie's List depend on reviews from their members or visitors. Information resources such as Wikipedia, genealogy databases, citizen science efforts, etc. depend on such volunteer-contributed information.

While the past success of such community-sourced information sites are substantial, the future is unclear. These online communities have orders of magnitude more information consumers than contributors. As Halfaker [8] observes, the number of contributors in some online communities is declining. And there are increasingly many requests for user's contributions of content and effort. Whether thought of as member contribution or crowd work, the future of community-sourced content requires careful nurturing. As Kittur [13] points out, "the future of crowd work requires that requesters and platform developers consider a broad set of motivations".

To illustrate the state of the art, consider just the first 100 messages of the inbox of one of the authors. Inside we find two messages from Amazon.com, one of them asking for feedback on a third-party vendor and a second asking for a rating of a recently-purchased product. We also find two messages from TripAdvisor, one of them congratulating the author on having reached an elevated level of reviewer and mentioning the number of times reviews have been read by others, and the second asking for help in answering another site user's question about a property the author had earlier reviewed (it turned out the question was already well-answered by others). Finally, there was a message from OpenTable asking the author "how dinner at <restaurant> was" and inviting submission of a rating and review to share with other diners.

These messages and requests illustrate the two points that inspire this work. First, that there are almost limitless degrees of freedom in how a site can request effort and information from its users. And second, that the field has little systematic understanding of the effectiveness of different types of request. To be clear, we understand that many individual site operators have substantial information on appeal effectiveness in their context, but these operators are not exposing that insight to the broader community or to being tested in different contexts.

To address this opportunity, we propose a basic framework we call *precision crowdsourcing* -- a systematic way of approaching the process of turning an information consumer into a long-term contributor through a series of requests, feedback, and interaction. The framework identifies five key decisions related to precision crowdsourcing.

CSCW'16, February 27-March 02, 2016, San Francisco, CA, USA
 ISBN 978-1-4503-3592-8/16/02 $15.00
DOI: http://dx.doi.org/10.1145/2818048.2819957

- **Who we ask.** the selection of users that we are requesting from, considering users' history of interactions with the system.
- **What we ask for.** the task type and contents, considering properties such as task effort, complexity, and relationship to the user's current context and history. Do we ask for immediate effort or a commitment for later effort?
- **When we ask.** how does the request relate in time to the user's activity, including any user interaction with the content related to the task? Do we interrupt, ask later, or even ask at a time triggered by other actions.
- **How we ask.** what is the rhetoric of the request? Do we frame contribution as self-benefiting? As a way to help others? As an obligation of membership?
- **What feedback follows.** Do we thank the users? How and when? Do we reference the impact of the contribution on others? Achievement of contribution milestones? Do we link past contributions to future requests?

Finally, the precision crowdsourcing framework defines a set of measurable outcomes. For any strategy of interaction, we can measure:

- **The immediate response to the request.** Does the user comply with the request? With what quantity and quality of contribution? If the response is a commitment for future effort, we measure both the commitment and the later delivery.
- **The long-term impact on contributions.** How does the quality and quantity of information contributed over the following weeks and months differ for those in different request conditions? Are we successfully creating contributors?
- **The impact on overall site engagement.** What is the effect of a precision crowdsourcing strategy on other measures of engagement (logins, activity, consumption or purchases, time on site). Are requests for contribution creating deeper engagement or scaring off consumers who want to avoid being asked?

Our long-term research goal is to complete a set of studies on different research platforms to build a theoretical understanding of precision crowdsourcing. The present study addresses three dimensions--what, when, and how--in the context of a non-commercial information system.

The core contributions of this paper are: (a) the precision crowdsourcing framework articulated above; (b) a set of findings on the immediate effects of three manipulations--specifically that time did not matter much but that reciprocal rhetoric and lower-popularity requests led to higher request completion; and (c) a set of findings related to long-term impact of these requests--specifically that engagement is unaffected, total contribution is higher, but unprompted contribution is lower as a result of the requests.

The remainder of this paper is structured as follows: in the next section, we'll review related work on motivation and inducements for contribution to online communities. Then we present the specific research questions for this study and describe the online experiment we conducted. We then present the results of the study, and conclude with a discussion of the implications of this work for both researchers and site designers, and review limitations of this work and directions for future work.

RELATED WORK

Motivating Contribution to Online Community. Ling et al. [18] studied the question of how to improve user response to requests for contributions of effort to an online community. Based on theories of individual and collective contribution, they tested framing uniqueness appeals and setting individual and group goals, finding that these approaches increased contributions. Cosley et al. [2] also found that uniqueness was an effective appeal.

Other researchers have more broadly used and extended theories of social behavior in an attempt to induce greater participation. Harper et al [9] conducted a field experiment to study the use of social comparisons to increase contributions in online communities. As the inequality aversion theory they used suggested, they found that comparisons were motivating, but users were more sensitive to gross contribution than to net benefit. Preece and Shneiderman [24] developed the *reader-to-leader framework* to help better explain the progression of users in technology-mediated social participation. Their illumination of the stages of reader, contributor, collaborator, and leader--together with an understanding of the reinforcements and conditions that support migration among these roles--can help guide communities trying to develop more engaged participants.

Nov et al. [23] found that tenure (length of membership) in an online photo-sharing community affects participation, but the effect depends on the type of participation activity, e.g. information-artifact sharing decreases for longer-tenure users, while meta-information sharing and social structures participation increases. Ma et al. [19] built on social psychology literature to understand how "IT-based features" are associated with contributions in online communities. They found that community IT artifacts have a positive effect on *perceived identity verification* which is related to user's satisfaction and contribution in online community. Nam et al. [22] found altruism, learning, and competency are frequent motivations for top contributors to participate, but that participation is often highly intermittent, in online communities of questions and answers.

We can support motivation by appeals explaining benefit to users themselves or to others and by displaying uniqueness or value of users' work. For example, Rashid et al. [26] found that displaying value of the contribution in the requests can increase users' contribution compared with not displaying value, and displaying the value of the

contribution to the user himself is more effective than to others. Zhu et al. [30] conducted a field experiment on Wikipedia to test the effects of different feedback types on members' contribution. They confirmed their hypothesis that negative feedback and direction increase people's efforts on focal tasks, positive feedback and social messages increase people's general motivation to work and the effects are stronger for newcomers.

Other-directed Motivation and Reciprocity. Motivation to contribute to online communities generally can be divided into self-directed and other-directed [6, 11, 17]. In the self-directed motivation, users are motivated by fun, perceived self-interest, and feelings of accomplishment or self-importance. One particularly promising approach to harnessing self-directed motivation is exemplified in gamification. Deterding et al. [3] used video game elements in non-gaming systems to engage more users.

By contrast, other-directed motivation includes altruism, which "aims to benefit others without intent to benefit self" [15] and reciprocity [2, 15], which involves the social norm of doing for others because they (either specifically or generally) have done or will do for you [7]. Falk et al [5] show that reciprocity does not always work intuitively as they present a formal theory of reciprocity which takes into account that "people evaluate the kindness of an action not only by its consequences but also by its underlying intention". The theory explains why outcomes tend to be fair in bilateral interactions whereas extremely unfair distributions may arise in competitive markets. Social exchange theory [4] by Ekeh generalizes from the fundamental idea of individuals exchanging goods or services. It defines generalized exchange, where recipients are unknown past or future parties, and group generalized exchange, where a "group" acts as a third party between individuals. This notion has been used to frame precisely the sort of open source, open content, and crowdsourcing systems we are investigating [1]. Building on the idea of both direct and generalized reciprocity, researchers have started to explore whether users experience reciprocity with systems as well as with other users. Larson et al. [16, 27] proposed the idea to request users of the recommender system for help when recommending items to users.

Relationship of the present work to prior work. Our work is inspired by this body of theory-driven and theory-developing research. We are also particularly inspired by Preece and Shneiderman's model [24]--we hope to eventually build a deep framework for the specific challenge of inducing desired levels of participation. Our goal in precision crowdsourcing is to experimentally build and validate models of the effect of different types of appeal, request, and feedback on the short- and long-term success of requests for information. By simultaneously looking at appeal and request, we also seek to identify which factors relate to users' fundamental judgment of a task vs. to the framing of the request. For example, Ling et al. [18] looked at uniqueness primarily as an appeal; in this study we look at it primarily as a fundamental judgment by offering more and less unique tasks without referencing the different in the request. Falk and Fischbacher's theory of reciprocity [5] leads us to consider the question of whether the user has already received benefit as a context that may be an important factor in the user's reciprocal response; we look at this issue as part of our exploration of when to make the request. Social exchange theory [4] and previous related work exploring the reciprocal relationship between systems and users [16, 27] motivate us to emphasize different types of generalized reciprocity when testing the effect of how to request users to help, specifically the question of reciprocity with system users collectively vs. with the system as an entity.

RESEARCH QUESTIONS

RQ1: What is the effect of asking for contributions?

In this research, we seek to deepen our understanding of several key questions related to the design of precision crowdsourcing requests. First of all, fundamentally we are interested in whether our requests are effective at eliciting contributions. The core idea of precision crowdsourcing is to turn information consumers into contributors through a series of interventions. We are motivated to study not only the immediate response to precision crowdsourcing requests, but also the impact on long-term behaviors. As shown by Masli et al. [20], "techniques that manipulate users into participating and contributing information may succeed in the short-term but might cause long-term harm, because users tend to recognize the manipulations and may consider them unfair". Looking at the positive side, previous research also shows that entry barriers and other opportunities for members to make community-specific investments can increase users' commitment to the system [14]. In this paper, we study two categories of long term behaviors: long-term commitment and long-term contribution. We are interested to know whether asking users to contribute information increases or decreases their usage of the system, or has no significant effect at all. Further, asking may achieve users' compliance as expected, however may affect users' voluntary contribution out of the requests when we stopping prompting.

RQ2 (What): What is the difference in effect between asking users to contribute information about more obscure content vs. more popular content?

There are many ways to select content for display in a precision crowdsourcing request. For instance, we might choose content that needs the most help, the newest content, or content that a user has recently acted upon. In this research, we are interested in varying the degree to which content is uniquely targeted to the subject. On the one hand, we might show a user content that is very popular and it may appear easy to act upon. On the other hand, we might show a user content that is rare but still familiar- as shown in previous research [25], asking people to perform tasks

that interest them and that they are able to perform increases contributions. In this case, the task is more uniquely targeted to the experiences of that user and may appear harder. This research question builds on the results of Ling et al. [18], who found that emphasizing users' ability to make unique contributions increased participation; in this work we do not explicitly call attention to uniqueness, to investigate this question from the perspective of content selection rather than framing.

RQ3 (When): What is the effect of requesting information immediately after a user logs in, as compared with waiting until later in the user's session?

It is possible that there are better and worse times to request information from a user. For instance, we could present a precision crowdsourcing request immediately after a user logs in, or we could wait until after the user has had a chance to enjoy some of the benefits of using the system. Based on the theory of reciprocity [5] by Falk et al., there may be a balance between consumption and contribution - and it may not generally be good practice to request before users actually consume some information from the system. On the other hand, asking later in a session might be quite disruptive to whatever task that user has chosen to pursue.

RQ4 (How): What is the effect of emphasizing the reciprocal relationship between the user and the community, and of emphasizing the reciprocal relationship with other user/contributors, as compared with a neutrally-framed appeal?

In online communities, users' contributions have benefits to other users and to the system itself. Thus, there is a reciprocal relationship that we can emphasize in framing our precision crowdsourcing requests. Previous research categorizes reciprocity into direct and indirect or generalized reciprocity [14]. In this research, we emphasize two different perspectives on indirect reciprocity, which we label *system-based reciprocity* and *user-based reciprocity*. *System-based reciprocity* emphasizes the reciprocal relationship between the user and the community, while *user-based reciprocity* emphasizes the reciprocal relationship with other users/contributors. We are interested in comparing the effectiveness of these reciprocity appeals and with neutrally framed appeals.

RQ5: What is the difference in terms of follow-up contributions between users who comply with the initial request and those who do not comply?

We are also interested in how well we can identify users who are generally willing to contribute or not by looking at their response to their first precision crowdsourcing request. The identification of willing contributors is one potentially useful outcome of a precision crowdsourcing intervention, since these users may be of higher value to the community overall.

EXPERIMENT DESIGN

Experiment Overview

We conduct an online field experiment that presents information requests to users of a movie recommender site *MovieLens* (https://movielens.org) to answer our research questions. The site allows users to browse and rate movies to get personalized movie recommendations. Users can also search and browse the movie database using tags - descriptive words or short phrases for movies that are provided by users.

In this experiment, we ask users to apply tags to movies. We choose tagging as the experimental task for several reasons. First, tagging movies creates content of value to the entire community of users, because tags link content together for improved browsing and searching, and because they add descriptive power. The tagging system improves as it receives more content, for both popular and rare movies. However, more than half (60.6%) of the site users have not applied any tags, and 38.7% of the site movies have zero tag applications. In contrast, rating movies as another primary feature of the system is more self-interest directed, because users rate primarily to improve their recommendations and to keep track of their movie watching history [10]. Second, tags are small contributions, typically requiring just a few seconds per application. Third, tags can be multiple-contributed, where subsequent applications continue to add value to the system by providing a information for ranking the "best" tags for each movie. Finally, we think tags are interesting to study because they are prevalent in many systems, including Tripadvisor, Flickr, and Youtube, and have been the topic of other recent crowdsourcing research (e.g, [28]).

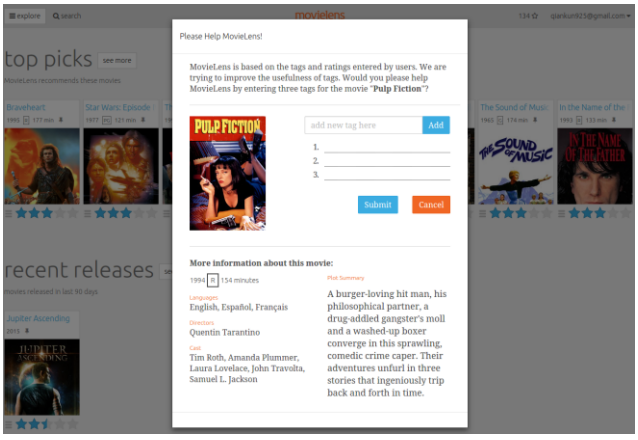

Figure 1. Screenshot of the experiment interface.

Our experimental interface (shown in Figure 1) is a pop-up request to apply three tags to a particular movie. We employ a pop-up interface because it is potentially shown in different parts of the system, depending on our experimental condition. The interface shows a variety of metadata about a movie, including the title, release date, genres, and a plot synopsis, but it does not show any tags.

The user is asked to either (a) apply exactly three distinct tags and press "Submit", or (b) press "Cancel". The tag input field supports free text entry, and also offers auto-completed suggestions from the database of all previously applied tags.

Our subjects include all site users who logged in between 12/21/2014 and 03/01/2015 and who had rated at least 15 movies at the time of login. We add the minimum rating requirement to ensure that we have enough content for display throughout the experiment, to ensure a minimum site commitment, and to filter out newly-registered users. We continued monitoring these users through 07/01/2015 to assess long-term effects. These subjects - excluding those in the control group, as discussed below - were shown pop-up requests to contribute tags. Subjects were possibly shown multiple requests (depending on whether they chose to return to the site), limited to no more than one request per week, and a maximum of four total requests. We limit the number or requests to avoid turning away site users. This experiment was reviewed by our institution's Institutional Review Board which approved a waiver of informed consent.

Experimental Manipulations

Our experiment tests three variables (described below) that vary the timing, content and appeal of the precision crowdsourcing request designed according to our research questions. We use a 2 x 4 x 3 between-subjects factorial design (*what* x *when* x *how*) for 24 experimental groups, plus one null control group that does not see any requests. We randomly assign subjects to a group. Once a subject has been assigned a group, that group membership does not change for the duration of the experiment.

What variable is operationalized by instrumenting different content selection algorithms for users in the two conditions.

- Subjects in the *easier/less targeted* group are shown informational requests referencing the most-often rated movies from their set of rated movies.
- Subjects in the *harder/more targeted* group are shown informational requests showing the least-often rated movies from their set of rated movies.

We validate that users can perceive difference in difficulty between these two conditions using a pretest, which was done from 7/25/2014 to 9/25/2014 with 167 subjects who are excluded from the main experiment. Using a within-subject design, we find that subjects take 81.0% (p=3.79e-09) more time applying tags to more targeted content, and rate the process to be more difficult (p=2e-16) as compared with less targeted content in a likert-scale of "*Very Easy, Easy, Neutral, Difficult, Very Difficult*".

When reflects the timing of the pop-up request and has four levels: *login, 1st/2nd/3rd movie detail page*. This variable controls whether we show the request immediately following the subject's login, or at the time of the subject's nth visit to a movie details page during that session (a page showing detailed information about a single movie in the site), where n = 1, 2, or 3. If a subject does not reach that number of page views, then we do not show a request.

How has three levels: *neutral, system-based reciprocity* and *user-based reciprocity*. To operationalize these levels, we frame the language based on the elements of composing reciprocal requests from [14]: priming norms of reciprocity by highlighting concepts that get people to think of their normative obligations, showing people what they have received from other users or the system, and highlighting opportunities to return favors to specific others. The language for the three conditions are as follows:

- Neutral: *Please Provide Three Tags For A Movie!*

  As a MovieLens user, you can not only rate movies, but also annotate them with descriptive tags. Would you please enter three tags for the movie "Pulp Fiction"?

- System-based reciprocity: *Please Help MovieLens!*

  MovieLens is based on the tags and ratings entered by users. We are trying to improve the usefulness of tags. Would you please help MovieLens by entering three tags for the movie "Pulp Fiction"?

- User-based reciprocity: *Please Help Other MovieLens Users!*

  In MovieLens, tags and ratings entered by other users help you get useful information about movies. Would you please help other users by entering three tags for the movie "Pulp Fiction"?

Measurements

Immediate Effect. Subjects' immediate response is measured using the *compliance rate* (percentage of fulfilled requests) of the requested tasks for different groups.

Effect on Long-term commitment. We measure the number of logins to the site in the four months following the user's assignment to a condition (triggered by a login during the experiment) as a proxy for long-term system commitment. (We did not find any significant difference in this measurement among different experimental conditions and hence will not report the data in the rest of the paper.)

Effect on Long-term contribution. We measure the number of tags in the four months following the user's assignment to a condition (triggered by a login during the experiment) as a proxy for subjects' long-term contributions. We break this period into separate two month intervals to investigate whether the effects of the manipulation persist or wear off. We measure both *voluntary tagging behavior* that is external to the experiment (i.e., part of the natural use of the system) as well as *total tagging behavior* which combines experimentally-induced tagging activity with voluntary activity when necessary to investigate aggregate effects.

RESULTS AND ANALYSIS

During our experimental period, we presented 2,978 subjects with at least one request, and made a total of 5,412 requests, with up to four per participant. In addition, we included 137 subjects in a null control group that received no requests. Collected data show that different groups of subjects through the randomization have no significant difference in terms of tenure, number of ratings and number of tags in the system. Table 1 lists the number of subjects according to how many requests they received throughout the experimental period.

number of requests	0	1	2	3	4
number of subjects	137	1856	360	212	550

Table 1. The number of subjects according to how many requests they received throughout the experimental period.

Throughout this section, we use three kinds of statistical tests for significance. For testing differences in proportions (e.g., percentage of fulfilled requests or percentage of users who tag) we use logistic regression. For testing count data, because of the over-dispersed nature of the data (i.e., many zeroes), we use zero-inflated negative binomial regression. Vuong tests [29] suggest that these zero-inflated models are significantly better than the standard negative binomial regression. The data we are dealing with is quite similar to the count data of Wikipedia edits in [30] which also uses the same kind of model. For testing count data leaving out zeros (e.g. number of tags added by tagging users – those who tagged at least once), we use zero-truncated negative binomial regression for the same reason of over-dispersion in the data. We build separate regression models for each pair-wise comparison. All these negative binomial models have significant non-zero dispersion parameters.

RQ1: Effect of asking

As Table 2 shows, 26.4% of the requests are fulfilled by users in the experimental group, giving 1.17 tags per user overall (i.e., including users who did not contribute any tags through the experiment), or alternatively, 4.75 tags per tagging user (i.e., users whose experimental contributions are non-zero). However, asking significantly decreases users' voluntary tagging behavior in the two months following the initial request, with 1.21 tags per subject, compared with 2.08 tags in the control group. This decrease stems from two effects in opposite directions: while more users from the experimental group tag one or more times (12.3% vs. 8.8%), these users apply many fewer tags on average as compared with users in the control group (9.9 vs. 23.6) . In the following two months, the two groups of subjects no longer exhibit significant differences.

Table 3 shows that asking has different effects on different users. We compare subjects in the experimental group in the *when=login* condition with the matched subjects in the control group. For those users who visit the site enough to

Group	Prompted Tags (Immediate Response)			Voluntary Tags First Two Months (Short Term)			Voluntary Tags Next Two Months (Long Term)		
	% of fulfilled requests	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users
Control	N.A.	N.A.	N.A.	8.82%	23.6	2.08	5.97%	5.75	0.343
Experiment	26.4%	4.75	1.17	12.3%	9.86 <ctrl, p~0***	1.21 <ctrl, p=0.0584.	4.91%	5.69	0.280

Table 2. Comparison of tagging behavior between the control and experimental groups. Prompted tags are those contributed directly through the experimental interface; voluntary tags are those contributed outside of the experiment. The two month intervals begin on each user's first treatment (or null treatment for the control group). Only significant comparisons are labeled with p-values.

Group	% Users With Enough Logins to Display 4 Requests	Voluntary Tags First Two Months (Short Term)			Voluntary Tags Next Two Months (Long Term)		
		% who tagged	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users
Control Who Would Have 4 Requests for *when=login*	52.5%	11.1%	28.0	3.10	11.4%	11.3	1.30
Experiment Subjects With 4 Requests *(when=login)*	32.1% <ctrl, p~0	29.5% >ctrl, p=0.002**	19.1	5.63 >ctrl, p=0.005**	17.5%	9.02	1.58

Table 3. Comparison of voluntary tagging behavior between subjects in the control group who would have received 4 requests if they were in the "*when=login*" condition (72 subjects) and subjects in the experimental group who actually receive 4 requests and are in the "*when=login*" condition (223 subjects). Only significant comparisons are labeled with p-values.

Group	Prompted Tags (Immediate Response)			Voluntary Tags First Two Months (Short Term)			Voluntary Tags Next Two Months (Long Term)		
	% of fulfilled requests	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users
Harder, More Targeted	30.2%	4.60	1.39	11.5%	10.8	1.24	5.64%	5.68	0.320
Easier,Less Targeted	22.2% <harder, p~0***	4.23	0.94	13.0%	9.00	1.17	4.16% <harder, p=0.064 .	5.71	0.237 <harder, p=0.07 .

Table 4. Comparison of prompted and voluntary tagging behavior between subjects in *what="harder, more targeted" and "easier, less targeted"* conditions. Only significant comparisons are labeled with p-values.

Group	Prompted Tags (Immediate Response)			Voluntary Tags First Two Months (Short Term)			Voluntary Tags Next Two Months (Long Term)		
	% of fulfilled requests	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users
Login	27.1%	4.94	1.79	14.7%	11.4	1.69	5.81%	5.69	0.331
1st Movie Detail Page	24.9%	4.8	1.20	12.9%	9.36	1.21	5.03%	4.56	0.229
2nd Movie Detail page	25.4%	4.71	0.926	11.3% <login, p=0.0505 .	8.92	1.01	4.64%	5.91	0.274
3rd Movie Detail Page	28.9%	4.32	0.774	10.3% <login, p=0.0107*	9.32	0.958 <login, p=0.061 .	4.20%	6.80	0.285

Table 5. Comparison of prompted and voluntary tagging behavior between subjects for different *when* conditions. Only significant comparisons are labeled with p-values.

see four experimental requests, users in the experimental group apply more tags than users in the control group during the first two months (5.63 vs. 3.10). In the following two months, the difference is no longer significant. However, asking when users log in significantly decreases the percentage of subjects who make it to the 4th request (32.1% experimental vs. 52.5% control). This points out a potential risk for scaring users off by asking them contribute multiple times before they actually start using the system. This is not general to all kinds of asking because we do not find a significant difference in users' logins in the first four months between the control and experimental groups.

Examining total tags (induced through the experiment or voluntarily added) over the four months following the initial treatment, we find that experimental subjects contributed more tag applications, on average, as compared with the control group (2.92 vs. 2.31, p~0).

RQ2: Effect of What to Ask

To examine the effect of asking for tags on harder/more targeted content versus easier/less targeted content, we restrict our analysis to the experimental group. As shown in the column *Prompted Tags* in Table 4, the harder/more targeted tasks have a significantly higher request compliance rate, as compared with the easier/less targeted tasks (30.2% vs. 22.2%). Subjects in the two groups are not significantly different in the first two months in their voluntary tagging behavior. However, in the following two months, subjects in easier/less targeted group provide fewer tags than subjects in harder/more targeted group (means: 0.24 vs. 0.32), although the difference is marginally significant with p-value=0.07.

RQ3: Effect of When to Ask

We compare the behavior of users in the different "*when*" conditions to examine the effect of asking for contributions at different points in a session. As Table 5 shows, there is no significant difference in compliance rate between asking immediately after a subject logs in versus later in a session. In the first two months, a significantly higher percentage of subjects in the *when="login"* condition provide tags voluntarily compared with *when="2nd or 3rd Movie Detail Page"*. The number of tags per tagging subject is not significantly different. Since many users have short sessions and do not actually view enough movie detail pages to see a treatment, we find that asking earlier leads to more voluntary tags overall (mean=1.69 for *"login"* versus mean=0.958 for *"3rd Movie Detail Page"*). We do not find a significant difference in the following two months in terms of voluntary tagging behavior or login behavior.

RQ4: Effect of How to Ask

We compare the behavior of users in the different *"how"* conditions to examine the effect of two different reciprocity

Group	Prompted Tags (Immediate Response)			Voluntary Tags First Two Months (Short Term)			Voluntary Tags Next Two Months (Long Term)		
	% of fulfilled requests	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users
Neutral	21.5%	4.14	0.89	11.0%	11.5	1.27	5.64%	6.05	0.341
System-based Reciprocity	31.1% >neut, p~0***	4.50	1.40	13.1%	7.10 <neut, p=0.008**	0.938 <neut, p=0.026*	4.74%	5.60	0.266
User-based Reciprocity	26.3% >neut, p=0.0315* <sys, p=0.0283*	4.45	1.17	12.5%	11.23 >sys, p=0.0114*	1.41 >sys, p=0.0117*	4.35%	5.32	0.231

Table 6. Comparison of prompted and voluntary tagging behavior between subjects for different *how* conditions. Only significant comparisons are labeled with p-values.

framings. See Table 6 for a summary of the results. As measured by compliance, requests including a reciprocal appeal increases compliance as compared with a neutrally-framed appeal, and system-based reciprocity outperforms user-based reciprocity (system-reciprocity=31.1% > user-reciprocity=26.3% > neutral=21.5%). However, during the first two months, the direction of the effect on subjects' voluntary behavior is reversed: asking with system-based reciprocity decreases subjects' voluntary tagging behavior per all users and per tagging user. Subjects in the system-based reciprocity group also voluntarily contribute significantly fewer tags as compared with the user-based reciprocity group. During the following two months, these differences vanish.

The contradicting effect of reciprocity-based appeals leads us to analyze total tag contributions (including both prompted and voluntary tags) in the four month period. Table 7 shows that both reciprocity appeals outperform the neutral appeal on this metric, though the difference between system-reciprocity and neutral is only marginally significant (p=0.09).

Group	# Total Tags Per All Users, First Fourth Months
Neutral	2.50
System-based Reciprocity	2.60 >neut,p=0.09 .
User-based Reciprocity	2.81 >neut,p=0.06 .

Table 7. Comparison of total tags (prompted + voluntary) per subject in the first four months for different *how* conditions. Only significant comparisons are labeled with p-values.

RQ5: Difference between subjects who comply or not in the first tagging request.

Subjects who comply with the first request also comply more often in the following request, as compared with subjects that do not comply with the first request (36.4% vs. 13.0%, p~0). There is also a dramatic difference between these two groups of users in their long-term behaviors. See Table 8. Subjects who comply with the first request are more likely to provide voluntary tags, and provide more tags overall. This demonstrates that the response to the first request is very helpful in identifying subjects who are generally willing to contribute or not.

DISCUSSION

Our analysis shows that asking users to contribute effort (in this case tags) has a mixed effect. Looking across a four month period, users receiving the experimental treatment contributed slightly more tags, on average. On the other hand, asking to contribute appears to diminish the level of voluntary contributions outside the experiment in the subsequent two months (see Table 2). While it is not surprising that asking users to contribute raises overall participation, we do not have a clear understanding of why it diminishes voluntary contributions. We have several hypotheses. Prompting people to do work may change the framing from a "fun activity I found and do for myself" to "work I was asked to do." This framing may quash the intrinsic motivation of those who would have contributed anyway. Another possibility is that asking may set up the expectation that many others will be contributing as well, setting up a "why should I contribute more when there are lots of other people giving" situation. Or it may simply reframe contribution as something that is not needed except when asked (if you needed more tags, you'd ask me). As yet we cannot distinguish these possible causes (though future work may probe into this question), but we leave system designers with the caution that they should recognize that "prompted contribution" is not a once-and-learned behavior. If that is your strategy, you need to plan for repeated prompts.

We also find evidence that there is a real risk of driving users away by making requests of them in the context of their system use. Users who received an experimental request immediately after login were less likely to log in enough times to receive a fourth request over the course of the experiment, as compared with the control group (see Table 3). This finding echoes the results of a recent analysis from Google that led their Google+ product team to stop

Group	Voluntary Tags First Two Months (Short Term)			Voluntary Tags Next Two Months (Long Term)		
	% who tagged	# tags per tagging user	# tags per all users	% who tagged	# tags per tagging user	# tags per all users
Comply with the 1st Request, When=login	26.3%	13.4	3.52	10.2%	4.40	0.45
Do Not Comply with the 1st Request, When=login	10.3% <comply, p~0***	10.1	1.04 <comply, p~0***	4.27% <comply, p=0.003 **	6.28	0.26 <comply, p=0.002**

Table 8. Comparison of voluntary tagging behavior between subjects who comply with the 1st request and who do not in *when=login* condition. Only significant comparisons are labeled with p-values.

showing "interstitial" ads asking users to install the mobile app version of the site [21]. They cite evidence that the interstitial ads dramatically increased the number of users who "abandoned" their product.

In our manipulation of *what* we asked users to do, we found that asking subjects to tag less popular movies was associated with a significantly higher compliance rate compared with asking them to tag more popular ones. Given that both movies were ones they have seen before, and that our pre-test showed that the less popular ones took more time to tag, we have two different possible explanations for this result. First, users might recognize that they are more uniquely suited to provide tags for the obscure movies they've seen. Under Karau and William's collective effort model [12], this uniqueness would increase willingness to commit effort to a shared goal because they recognize that their effort is more irreplaceable and therefore the collective (implicit) goal may fail without them. In other words, don't ask me to tag "Star Wars" or "Titanic" because I know everyone else can do that. Similarly, users may not explicitly think of uniqueness but just simple value -- a tag for a popular movie isn't that valuable because there are probably already many of them. Alternatively, it may be that the extra challenge itself is motivating. This cause would be consistent with Ling et al.'s result that challenging (but attainable) goals increase contribution [18]. Rating popular movies may be "too easy" to be a motivating goal.

On the *when* dimension we did not find significant differences in compliance with our requests. There are many dimensions to the timing of requests, including whether we are interrupting and how much benefit users have experience, and we expect that more careful isolation may be needed to find any effects. The fact that long-term contribution is higher for those who were invited upon login is interesting. It could be that users pay more attention to messages at login (even when declining to participate) than later. Validating this hypothesis would require a separate study of whether users were familiar with the tagging feature.

On the *how* dimension we have strong evidence supporting the hypothesis that requesting users to do tasks in a reciprocal way can achieve a higher compliance rate as compared with requesting in a neutral way. We experimented with two different forms of reciprocity: system-based reciprocity, which emphasizes the reciprocal relationship between the requested users, and user-based reciprocity, which emphasizes the reciprocal relationship between them and other users. Interestingly, while system-based reciprocity led to the highest compliance rates, it appears to actually diminish voluntary tagging in the subsequent two months as compared with either the neural appeal or the user-based reciprocity appeal (see Table 6). Possibly, this framing exaggerates the overall effects of asking for contributions, whose reasons we speculate above. Or, this effect could be specific to our experimental platform, because the site is mainly a movie recommendation and information service provider that lacks social features that connect users together closely. Exploring this effect in other platforms with strong social features is something we would like to do in future work.

Finally, we find users who comply to the initial request not only comply more often in the later request but also voluntarily contribute more than those who do not. This suggests that users' response to the initial request demonstrates a lot about their general willingness to contribute in the long term. We might need to consider different treatments on those two groups of users, based on their initial response. For example, by changing the type of request, or by requesting follow-up feedback to elicit their thoughts on the perceived value and difficulty of the task.

LIMITATIONS AND FUTURE WORK

This work has several limitations that we hope to address in future studies. First and foremost, the study is limited to a single movie recommendation system. To provide generalizable theory will require replication in other environments. Second, the number of requests a user would receive, and the opportunities a user would have to voluntarily contribute, depend heavily on the frequency of usage of that user. Many users logged in only once during the experimental period, and we cannot assume that the motivations or activity of those users are the same as those who visit weekly or more often. We will continue to follow user behavior to look for longer-term effects, but some factors will always be confounded with individual differences in usage behavior.

We are interested in studying the quality differences between voluntary and prompted contribution. However,

we do not have a gold standard for measuring the quality of the contributed tags; longer-term analysis of tag usage is a possible solution, but remains future work.

At the start of this paper we outlined a theoretical framework for precision crowdsourcing consisting of five dimensions: *Who, What, When, How* and *Feedback*. In this study, we investigate specific manipulations along three of these dimensions*: What, When, How*. Based on these initial results, we see three lines of future research:

- Conducting parallel studies in other online communities. The essence of generalizable theory is the confidence that a predicted effect will work across a range of domains. We are working with a variety of other online community sites to identify areas where manipulations around selecting *what* to ask and *how* to ask it--specifically the issues of popularity/uniqueness and reciprocal rhetoric--can be confirmed or better understood.
- Conducting further investigations along these dimensions. We are particularly interested in the longer-delay *when* question that explores the tradeoffs between asking for effort during the usage of a site vs. asking afterwards (e.g., through a follow-up e-mail). We're also interested in exploring requests for immediate effort vs. delayed commitments (e.g., we could ask someone going to see a movie if they're willing to review it for us afterwards).
- Exploring the *who* and *feedback* dimensions. Our results strongly suggest that one of the more important challenges for a site is to identify which users are actually likely to contribute (either when asked, or on their own) and perhaps to individualize the interaction to address them. We are also interested in a more systematic exploration of gratitude--particularly distinguishing between instance (and content-free) gratitude and messages that reflect the accumulated impact of a contribution.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation with awards IIS-0808692, 0964695, 0968483, 1017697 and 1319382 and Google Social Computing Focused Research program. We thank users of MovieLens who participated this study and anonymous reviewers for providing valuable feedback to our paper. We also acknowledge the thoughtful discussions and helpful feedback from Brent Hecht, Shuo Chang, and other members of GroupLens Research.

REFERENCES

1. Coye Cheshire, and Judd Antin. "The social psychological effects of feedback on the production of Internet information pools." *Journal of Computer Mediated Communication* 13.3 (2008): 705-727. http://dx.doi.org/10.1111/j.1083-6101.2008.00416.x
2. Dan Cosley. "Mining Social Theory to Build Member-Maintained Communities." *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors.* 2005.
3. Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. "Gamification. using game-design elements in non-gaming contexts." In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 2425-2428. ACM, 2011. http://dx.doi.org/10.1145/1979742.1979575
4. Peter P. Ekeh. Social exchange theory: The two traditions. *London: Heinemann*, 1974. http://dx.doi.org/10.1525/aa.1977.79.1.02a00230
5. Armin Falk, and Urs Fischbacher. "A theory of reciprocity." *Games and Economic Behavior* 54.2 (2006): 293-315. http://dx.doi.org/10.1016/j.geb.2005.03.001
6. Paul Fugelstad, Patrick Dwyer, Jennifer Filson Moses, John Kim, Cleila Anna Mannino, Loren Terveen, and Mark Snyder. 2012. What makes users rate (share, tag, edit...)?: predicting patterns of participation in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 969-978. http://dx.doi.org/10.1145/2145204.2145349
7. Alvin W. Gouldner. "The norm of reciprocity: A preliminary statement." *American sociological review*(1960): 161-178. http://dx.doi.org/10.2307/2092623
8. Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. "The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline." *American Behavioral Scientist* (2012). http://dx.doi.org/10.1177/0002764212469365
9. Maxwell F. Harper, Yan Chen, Joseph Konstan, and Sherry Xin Li. "Social comparisons and contributions to online communities: A field experiment on movielens." *The American economic review* (2010): 1358-1398. http://dx.doi.org/10.1257/aer.100.4.1358
10. Maxwell F. Harper, Xin Li, Yan Chen, and Joseph A. Konstan. "An economic model of user rating in an online recommender system." In *User Modeling* 2005, pp. 307-316. Springer Berlin Heidelberg, 2005. http://dx.doi.org/10.1007/11527886_40
11. Alicia Iriberri and Gondy Leroy. A life-cycle perspective on online community success. *ACM Comput. Surv.* 41, 2, Article 11 (February 2009), 29 pages. http://doi.acm.org/10.1145/1459352.1459356
12. Steven J. Karau, and Kipling D. Williams. "Social loafing: A meta-analytic review and theoretical integration." *Journal of personality and social psychology* 65.4 (1993): 681. http://dx.doi.org/10.1037/0022-3514.65.4.681

13. Aniket Kittur et al. "The future of crowd work." *Proceedings of the 2013 conference on Computer supported cooperative work.* ACM, 2013. http://dx.doi.org/10.1145/2441776.2441923
14. Robert E. Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. Building successful online communities: Evidence-based social design. *Mit Press*, 2012.
15. Stacey Kuznetsov. "Motivations of contributors to Wikipedia." ACM *SIGCAS computers and society* 36.2 (2006): 1. http://dx.doi.org/10.1145/1215942.1215943
16. Martha Larson, Paolo Cremonesi, Alan Said, Domonkos Tikk, Yue Shi, and Alexandros Karatzoglou. "Activating the crowd: exploiting user-item reciprocity for recommendation." In the first workshop on Crowdsourcing and human computation for recommender systems, *ACM Conference Series on Recommender Systems, ACM RECSYS. 2013.*
17. Min Kyung Lee, Tawanna Dillahunt, Bryan Pendleton, Robert Kraut, and Sara Kiesler. 2009. Tailoring websites to increase contributions to online communities. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '09*). ACM, New York, NY, USA, 4003-4008. http://doi.acm.org/10.1145/1520340.1520608
18. Kimberly Ling, Gerard Beenen, Pamela Ludford, Xiaoqing Wang, Klarissa Chang, Xin Li, Dan Cosley et al. "Using social psychology to motivate contributions to online communities." *Journal of Computer Mediated Communication* 10, no. 4 (2005): 00-00. http://dx.doi.org/10.1145/1031607.1031642
19. Meng Ma, and Ritu Agarwal. "Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities." *Information systems research* 18.1 (2007): 42-67. http://dx.doi.org/10.1287/isre.1070.0113
20. Mikhil Masli, and Loren Terveen. "Evaluating compliance-without-pressure techniques for increasing participation in online communities." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2915-2924. ACM, 2012. http://dx.doi.org/10.1145/2207676.2208698
21. David Morell. Google+: A case study on App Download Interstitials. Retrieved July 27, 2015 from *Google Webmaster Central Blog*: http://googlewebmastercentral.blogspot.fr/2015/07/google-case-study-on-app-download-interstitials.html.
22. Kevin Kyung Nam, Mark S. Ackerman, and Lada A. Adamic. "Questions in, knowledge in?: a study of naver's question answering community." *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009. http://dx.doi.org/10.1145/1518701.1518821
23. Oded Nov, Mor Naaman, and Chen Ye. "Analysis of participation in an online photo‐sharing community: A multidimensional perspective." *Journal of the American Society for Information Science and Technology* 61.3 (2010): 555-566. http://dx.doi.org/10.1002/asi.v61:3
24. Jennifer Preece, and Ben Shneiderman. "The reader-to-leader framework: Motivating technology-mediated social participation." *AIS Transactions on Human-Computer Interaction* 1.1 (2009): 13-32.
25. D. Raban, M. Harper, Motivations for Answering Questions Online. Book chapter in New Media and Innovative Technologies (Caspi, D., Azran, T. eds.), 2007. http://dx.doi.org/10.1111/j.1460-2466.2009.01429.x
26. Al M. Rashid, Kimberly Ling, Regina D. Tassone, Paul Resnick, Robert Kraut, and John Riedl. "Motivating participation by displaying the value of contribution." In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 955-958. ACM, 2006. http://dx.doi.org/10.1145/1124772.1124915
27. Alan Said, Martha Larson, Domonkos Tikk, Paolo Cremonesi, Alexandros Karatzoglou, Frank Hopfgartner, Roberto Turrin, and Joost Geurts. "User-Item Reciprocity in Recommender Systems: Incentivizing the Crowd." In *UMAP Project Synergy(UMAP ProS) Workshop* "A forum for UMAP related projects to exchange ideas and practices. 2014.
28. Luis Von Ahn, and Laura Dabbish. "Labeling images with a computer game." *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004. http://dx.doi.org/10.1145/985692.985733
29. Quang H. Vuong. "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica: Journal of the Econometric Society*(1989): 307-333. http://dx.doi.org/10.2307/1912557
30. Haiyi Zhu, Amy Zhang, Jiping He, Robert E. Kraut, and Aniket Kittur. "Effects of peer feedback on contribution: a field experiment in Wikipedia." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2253-2262. ACM, 2013. http://dx.doi.org/10.1145/2470654.2481311