Eliciting and Focusing Geographic Volunteer Work

Reid Priedhorsky, Mikhil Masli, Loren Terveen
GroupLens Research
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota, USA
{reid,masli,terveen}@cs.umn.edu

ABSTRACT

Open content communities such as wikis derive their value from the work done by users. However, a key challenge is to elicit work that is sufficient and focused where needed. We address this challenge in a geographic open content community, the Cyclopath bicycle route finding system. We devised two techniques to elicit and focus user work, one using familiarity to direct work opportunities and another visually highlighting them. We conducted a field experiment, finding that (a) the techniques succeeded in eliciting user work, (b) the distribution of work across users was highly unequal, and (c) user work benefitted the community (reducing the length of the average computed route by 1 kilometer).

Author Keywords

Wiki, geowiki, open content, geographic volunteer work, volunteered geographic information

ACM Classification Keywords

H.5.3 [*Group and Organization Interfaces*]: Collaborative computing, computer-supported cooperative work, web-based interaction.

INTRODUCTION

Two major revolutions in online content are occuring. One is *open content*, where users produce most or all of a site's value – Wikipedia has grown to nearly three million articles in English alone, Yahoo! Answers has an archive of over one billion answers, and Flickr users upload over 5,000 photos each minute. The other is *geographic content* – Google Maps and its peers make easy-to-use and high-quality maps available to anyone with a web browser, and their associated APIs support geographic "mashups" on a wide range of topics, from cab fare[1] to earthquakes[2] to "geogreetings"[3].

These revolutions are merging. Internet-based tools and communities are useful for communication even when people are physically present in the same city or neighborhood, and shared geography leads to shared geographic experiences and needs. You can't hire a plumber from another country to fix your toilet – ergo, Angie's List – nor can you go to another continent to pick up a free used piano – ergo Craigslist. Further, there are many neighborhood-based discussion forums, such as E-Democracy.Org.

Bicyclists are one particularly interesting community. While cyclists form an affinity group and share tips about bikes, gear, training, etc. with fellows around the world[4], there is also a strong geographically local component: bicycling is a physical activity, cyclists like to ride together, and they share knowledge about routes with other cyclists. Finding good routes is hard because the transportation network is mostly designed for another mode (i.e., motor vehicles), it changes over time at multiple time scales (e.g., daily weather, seasonal variation, road construction), individuals have differing purposes, attitudes, and abilities, and their preferences change over time [22]. Indeed, many local cycling communities already use online information-sharing resources[5].

This work analyzes Cyclopath[6], a web-based mapping application serving the route finding needs of bicyclists in the Minnesota cities of Minneapolis and St. Paul, an area of roughly 8000 square kilometers and 2.3 million people. While Cyclopath's interface resembles Google Maps and its peers, there is a critical difference – Cyclopath is a geographic wiki, or *geowiki*, unifying interactive web-based mapping with the open editing of wikis.

Open content systems like Cyclopath derive much value from the work of users. In the geographic setting, we call this *geographic volunteer work* (GVW), to emphasize the active role of end users and in contrast with the geographic community's term *volunteered geographic information* [11].

Getting users to participate in open content systems is an ongoing challenge: many online communities fail [4], and even those that succeed need to focus user work. For example, of the many techniques that Wikipedia uses to encourage par-

[1] http://yellowcabnyc.com/fareestimator
[2] http://earthquake.googlemashups.com
[3] http://geogreeting.com

[4] e.g., http://bikeforums.net
[5] e.g., http://mplsbikelove.com
[6] http://cyclopath.org

CSCW 2010, February 6–10, 2010, Savannah, Georgia, USA.

ticipation, one that serves to focus and motivate collective work is the featured article candidate process. An article is proposed as a candidate for appearing on Wikipedia's main page, and typically these articles receive a huge increase in editing as interested editors try to reach the goal. We believe that creating effective techniques to elicit collective work and focus it where needed is critical to the success of open content systems. This leads to our research questions.

RQ1. *What techniques lead to increased GVW contributions?* We found that (a) visually highlighting work opportunities resulted in more total work; (b) taking users to work opportunities in areas of the map they're familiar with leads them to do more work of certain types; and (c) users do significant "extra" work beyond what is visually highlighted.

RQ2. *How is GVW distributed across users?* We found that (a) work is distributed very unequally; (b) inequality varies for different types of work; (c) inequality may be reduced by issuing a "call to action" and introducing features to take people to areas that need work, but these effects go away if the features are removed; and (d) different editors are leaders in different work categories, which makes the system more robust.

RQ3. *Does GVW improve route finding?* We found that user work had a positive effect on route finding in Cyclopath – user input has decreased the length of the average route by 1 km since Cyclopath went live – and that this effect has a long-tail distribution.

The rest of our paper is organized as follows. We first present related work and discuss the nature of GVW in Cyclopath. We then describe the design of a field experiment to test our work elicitation techniques. Lastly, we present the results of this experiment, discuss their implications, and close with a brief summary.

RELATED WORK

Eliciting volunteer work. Social science offers insights into the problem of eliciting user work. Karau and Williams [15] developed the *collective effort model*, identifying key factors that can increase individual motivation to contribute to a collective activity, such as informing a person that they have *unique* knowledge or skill, increasing the *personal value* a person places on a group outcome, and *reducing the cost* of contributing. Other work has explored additional techniques that motivate participation, such as social comparisons [8, 9] and setting specific and challenging goals [19].

A stream of recent work has drawn on this body of knowledge to develop and evaluate techniques to elicit participation in open content communities. Ling et al. [18] describe several experiments that evaluated techniques built on factors like uniqueness and goal-setting, and Harper et al. [12] evaluated techniques based on social comparisons.

Most relevant to our work, Cosley et al. have developed automated techniques for eliciting work. In one, the researchers wanted users of the MovieLens movie recommendation site to edit information about movies [5]. In the other, they wanted Wikipedia editors to edit articles [6]. In both cases, techniques based on *familiarity* with the work users were asked to do were most effective. In MovieLens, this meant asking users to edit movies they had rated, and in Wikipedia, this meant asking users to edit articles that were related to articles they had edited. This use of familiarity followed the collective effort model. For example, rating a movie indicates familiarity with it, so editing familiar movies is easier (reduced *cost*). Also, users are more likely to like movies they have seen and rated, and thus to care enough to invest the effort of editing movies (increased *personal value*).

Distribution of volunteer work. A key goal of empirically-minded researchers in online communities is to illuminate the nature of participation. Many research methods are appropriate. For example, qualitative studies based on interviews have yielded interesting insights about issues such as how gender and power issues shape online interaction [23] and how Wikipedia editors change their behavior over time [3].

However, quantitative methods are more relevant to the current work. One central result is the highly unequal nature of participation in these communities. When the level of participation is visualized, it commonly looks like a "hockey stick" or exhibits a "long tail". Specifically, it often follows a power law [1]. Such relationships arise in many different kinds of online communities, including Usenet discussions [24], tagging [10], and Wikipedia edits [17, 22].

Recently, some work has applied the *gini coefficient* [2] to quantify distribution. If (say) wealth is distributed evenly, the gini coefficient is close to 0; if wealth is concentrated in a small set of individuals, the gini coefficient is close to 1.

The distribution of work is interesting for several reasons. While it is both inevitable and positive to have a core group of contributors, there are advantages to reducing the inequality. A greater number of active contributors can lead to quicker actions (e.g., answers to questions or fixes of malicious edits), robustness when active members leave the community, and perhaps increased diversity of perspectives.

Utility of volunteer work. The utility of user input in systems such as recommender systems is well established. For example, one of the seminal papers showed that movie recommendations based on a collaborative filtering algorithm outperformed recommendations from professional movie critics [13], and our prior work offers some suggestive evidence for the utility of geographic volunteer work [21]. Here, we quantify that utility.

Similar systems. We highlight two systems closely related to Cyclopath. Open Street Map[7] is an ambitious project building a global street map using the wiki model and starting from a blank map. While users can edit the transportation network, key wiki monitoring features like recent changes and watch lists are missing. Moreover, the focus

[7] http://openstreetmap.org

is solely on the transportation network, so users cannot enter tags, notes, ratings or other useful annotations. Another system, Google Map Maker[8], is very new at the time of this writing and does not allow editing in North America, Europe, and many other parts of the world. Users can edit the transportation network as well as points of interest and monitor the edits of others. However, it is not clear whether users can revert the edits of others or what the publication model is. Finally, neither system provides route finding for cyclists.

GEOGRAPHIC VOLUNTEER WORK IN CYCLOPATH

As a geowiki, Cyclopath gains much of its value from the information produced by the geographic volunteer work of its users, which affects two types of geographic objects. Users can enter and edit **points of interest** (bike shops, bike racks, restaurants, or other locations considered relevant) and descriptions thereof. Points can be used in routing and serve as landmarks for browsing the map. **Blocks** – i.e, the atomic segments of roads or trails which form the edges in the transportation network – support several types of work:

- Users can enter **bikeability ratings** of blocks on a scale from Excellent to Impassable. The routing engine's computations are enhanced by ratings; the more ratings users enter, the better the routes computed by the system.

- Users can add and edit text **notes** associated with blocks; these may point out traffic, hazards, or other properties, and they help users evaluate routes.

- Users can alter the geometry and topology of the transportation network itself by **editing blocks**, using visual tools modeled after standard drawing applications. Route finding becomes more effective as this network of roads, trails, and informal paths becomes more accurate and comprehensive.

This block editing is critical to the quality of routes that Cyclopath can generate (and we elaborate below on why this is so). However, we first note that unlike Open Street Map, Cyclopath did not begin with a blank map. Its database was initialized with the best existing geographic datasets available to us – road and bicycle facility datasets provided by the Minnesota Department of Transportation – but even these unified data have several undesirable properties:

1. **Incomplete.** Some important bicycle facilities, such as certain sidewalks, alleys, or dirt trails through parks, are informal, and thus will not appear in any official dataset; other facilities, while "official", were simply not mapped. Aerial photos show many facilities which are missing from the initial data.

2. **Unlinked.** The roads and bicycle facilities were two distinct datasets; thus, connections between the two types were recorded in neither, and automatic linking of the datasets was incomplete, leaving many missing links. Bicycle routes frequently involve riding on both roads and dedicated bicycle paths, but such routes can only be generated if the database has an accurate and comprehensive record of road-path links.

Work Type		Count
Revisions		8,622
Ratings		54,938
Point	additions	1,693
	edits	780
Block	additions	11,238
	edits	30,548
Note	additions	1,820
	edits	193

Figure 1. Number of user contributions of varying types.

3. **Static.** Conditions change; road construction and temporary closures are common. Further, seasonal factors such as the state of snow removal are also key to route choice.

These problems are common in geowikis, not exceptional, so most or all projects like Cyclopath will need to confront them. Therefore, the geographic volunteer work approach and our empirical results will be of general interest.

Current state of Cyclopath. The system went live for a limited group of testers in May 2008 and for the public in July 2008. As of fall 2009, over 1,500 users have registered accounts, and 15-30 registered users and 150 anonymous users visit the site each day. The key use of Cyclopath is generating routes; route requests average 150 per day in season and total over 35,000.

Figure 1 summarizes the GVW contributions of Cyclopath's users. As in Wikipedia or other wikis, a revision consists of a set of changes made by a user in one user-defined editing session. These may vary considerably in the amount of contained work (e.g., ranging from single edits like adding a note to a block to multiple, complex edits like adding some new blocks, connecting them to the existing network, and placing several notes and points); therefore, we prefer to report more specific units of work when possible.

Like other open content sites, Cyclopath has a small core of contributors who do much work, but most users do little or none (e.g., 423 logged-in users have saved a least one revision, but only 7 have saved more than 100). We wanted to investigate the extent to which users who had done little or no GVW could be nudged to do more: increasing the base of workers makes the system more robust, leads to better coverage (new workers may be familiar with areas that old ones are not), and may distribute work more evenly.

Furthermore, the Cyclopath database contains thousands of errors. Specifically, automated analysis of one class of errors revealed 7,000 *missing X nodes* – places where two blocks cross one another geometrically but no network node exists – and 6,300 *missing T nodes* – places where a dead-end block came within 20 meters of intersecting another block. While these potential nodes (i.e., street intersections) can be identified automatically, human judgment is required to deter-

[8] http://www.google.com/mapmaker

mine whether a node is actually appropriate; for example, a missing X node might consist of one road on a bridge over another, and a missing T might consist of two roads which come close but don't actually meet.

EXPERIMENT DESIGN

We carried out a field experiment to explore elicitation of this needed geographic volunteer work in Cyclopath. This section explains our experiment design and the interface techniques used to achieve it.

Three hypotheses drive our design:

H1. **Familiarity.** *Users will do more GVW in areas they are familiar with.* In MovieLens, the appropriate unit for computing familiarity is the movie, and in Wikipedia it is the article. In Cyclopath, the appropriate unit is a geographic area. We computed users' familiarity with an area based on how often they view it (a weak indicator of familiarity) and how much they rate or edit in the area (strong indicators of familiarity).

H2. **Visual Prompts.** *Highlighting specific work units will lead users to do more GVW.* The Cyclopath map is visually dense, perhaps even cluttered. Simply asking a user to enter ratings or edit blocks in an area is a fairly underspecified instruction. Thus, we used visual highlights to focus user attention on specific objects that may need work: blocks that need ratings are colored blue, and potentially missing nodes are indicated by maroon circles (see Figure 2).

H3. **Work Type-Familiarity.** *The familiarity effect of H1 will be stronger when users are asked to rate blocks than when they are asked to repair missed nodes.* As noted above, we knew we needed users to examine and repair potentially missed nodes. However, we decided to also ask users to enter ratings. The Cyclopath ratings database is very sparse, with about 43,000 ratings for 150,000 blocks (at the time of the experiment). More ratings would improve the accuracy of the route finding engine's block evaluations, thus enabling it to compute better routes.

We thought that the familiarity effect would be stronger for ratings because rating a block requires actual knowledge, while determining the disposition of a potentially missing node can frequently be done just by looking at the aerial photo.

Conditions. These hypotheses lead to three factors to test: Familiarity, Visual Prompts, and Work Type. Visual Prompts was a between-subjects factor: we believed it was a sufficiently compelling interface feature that once a subject saw it, he/she would be confused and perhaps unhappy if it was not present on their next trial. On the other hand, the other two factors were within-subjects: each time a user participated in the experiment, he or she was randomly assigned a Work Type (Ratings or Node Repair) and an Area Type (Familiar or Random).

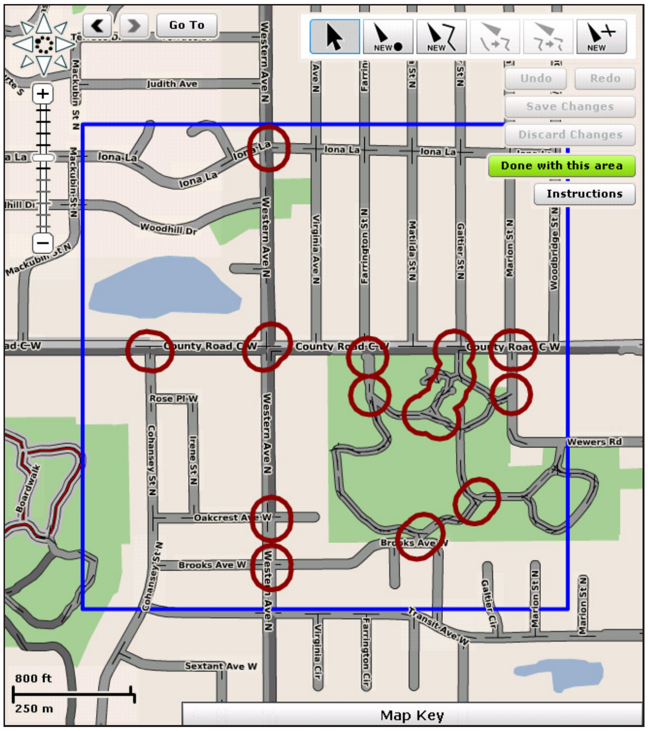

Figure 2. What a subject might see in the Visual Prompts / Node Repair condition. The blue square is the trial viewport – where the subject has been asked to do node repair work – and the maroon circles highlight potentially missed nodes within the viewport. In the No Visual Prompts condition, these visual highlights would be absent, but the blue square would still be present. In the Ratings condition, blocks needing rating would be highlighted in light blue and no node repair prompts would be present. There is no visible difference between the Familiar and Random Area Type conditions.

Partitioning subjects in Visual Prompts factor. Like most voluntary online activities (as noted above), Cyclopath activity is highly unequal. Therefore, we wanted to divide the most active users evenly between the Visual Prompts and No Visual Prompts conditions. We computed an overall participation score for users based on their viewing (a weak signal of commitment), rating, and editing (strong signals), sorted users by this score, and then stepped through this list, assigning users to Visual Prompts and No Visual Prompts alternately. The one subject who joined during the experiment was assigned a condition at random.

Computing familiarity. We used this participation score to estimate familiarity as well. We divided the Cyclopath map into a grid of 30,000 overlapping 1 km-square regions we call *viewports*. As all Cyclopath interaction is geographically grounded, we then computed a familiarity score for each (user, viewport) pair.

Soliciting participation. On March 26, 2009, we sent registered Cyclopath users an e-mail with the subject "Cyclopath needs your help!". The key passage was: *We have created a system which will automatically direct you to areas of the map that need work (more bikeability ratings entered or edits to the geography of the map itself).* The message also contained a link to take users directly into the experiment and provide instructions for participating. We also added a button to the interface called "How can I help?" that only logged-in users saw; clicking it also entered the experiment. On log-in, this button was highlighted with a popup window containing the text: *You Can Help Cyclopath. Cyclopath needs your help to improve the routes it computes for all users. Click "How can I help?" to begin.* The experiment was active for 10 days, and a total of 66 users participated.

Experiment procedure. The structure of an experimental trial (i.e., the user experience and experimental manipulations) is as follows.

1. *Begin trial.* A subject begins a trial by clicking on the "How Can I Help" button or following the e-mail link. To prevent a single subject from consuming all the work units, we limited subjects to 20 trials per day.

2. *Assign within-subjects conditions.* The system randomly assigns the subject to either the Familiar or Random Area Type condition, and either the Ratings or Node Repair Work Type condition.

3. *Select viewport.* If the subject is in the Familiar condition, his or her most familiar viewport is selected; otherwise (Random condition), a viewport is selected at random. Viewports which (a) have already been visited by the subject, (b) do not contain sufficient work units (at least two potentially missed nodes, or at least 12 blocks or 75% of the blocks in the viewport not yet rated), or (c) intersect the current view are excluded from consideration.

4. *Display viewport.* The map is panned and zoomed to the selected viewport. If the subject is in the Visual Prompts condition, draw visual prompts for Ratings or Node Repair as appropriate (for work units within the viewport only). See Figure 2 for a sample viewport.

5. *Subject does work.* The subject now is free to use the system. We emphasize that subject activity within a trial is unconstrained: subjects may choose to do no work, prompted work (in the Visual Prompts condition), unprompted work (e.g., rate blocks that were not highlighted by a Visual Prompt), work of a different type (e.g., adding a note or a point), or even to pan and zoom to another part of the map. Figure 3 shows the work done in one trial.

6. *End trial.* The subject clicks "Done with this area". After completing a trial, subjects may return to normal Cyclopath use or do another trial immediately, in which case the process returns to Step 2.

RESULTS

We organize our results using our three research questions. Unless otherwise noted, for P-value computations we use the Welch two-sample t-test on the $\log(x+1)$-tranform of the data (in order to reduce non-normality somewhat and compensate for counts of zero). Throughout, we use the following significance codes: $\circ : P \leq 0.10$, $* : P \leq 0.05$, $** : P \leq 0.01$, and $*** : P \leq 0.001$.

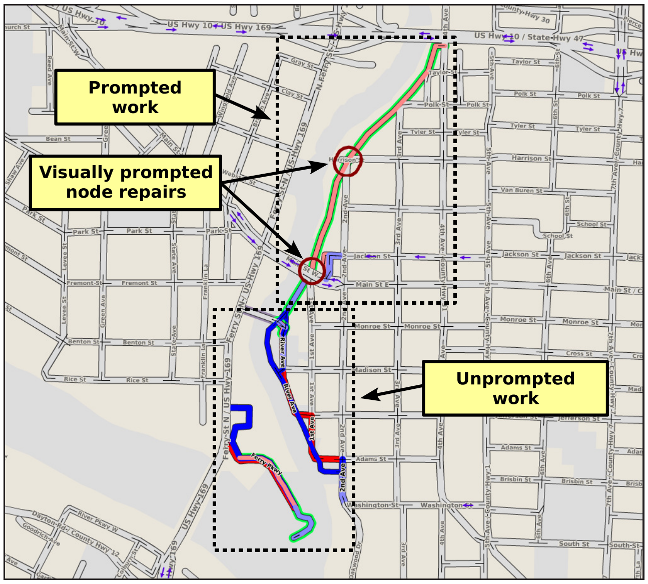

Figure 3. Work done in one trial; red and blue show the changed blocks. In this case, the subject has made block edits both related and unrelated to the prompted node repairs.

RQ1: Techniques to elicit and focus GVW

Our first research question was investigating the effect of techniques to elicit users to do specific geographic work. We structure the discussion of results around our three hypotheses: H1: Familiarity, H2: Visual Prompts, H3: Work Type-Familiarity. The experimental trial was the basic time interval for counting units of work. It also is useful to aggregate work, notably to the level of all the trials done within an experimental condition. In the analyses below, we typically report both work done per trial in a particular condition and total work done in that condition.

Metrics. We counted four different types of work. Clearly, we had to count *ratings* and *node repairs*, since those were the two types of work we solicited. We also counted *note edits*, an unsolicited type of work which subjects did anyway.

We also count *block edits*, a low-level construct whose mapping onto higher-level geographic editing actions like node repairs is variable (i.e., a node repair could involve 1, 3, 4, or more block edits). We would have preferred to count only higher-level actions, but defining these is a non-trivial task that may well require manual coding. Counting block edits was more tractable and still gave us a reasonable metric for quantifying work. Lastly, we wanted to distinguish block edits done in response to prompted node repair from those which were not; this too is difficult to determine in general. We approximated it by considering edited blocks passing within 80 meters of a prompted node repair to be prompted block edits. Figure 3 shows work done in one trial, highlighting prompted and unprompted block edits.

When analyzing our data, we found an unanticipated correlation between Familiarity and work availability: as Figure 4 shows, the Familiar viewports we generated had more available work than Random viewports. This is because Cy-

Measure	fm	rd	*P*
Total blocks	112.0	53.0	***
Rateable blocks	102.7	52.8	***
Node repairs	2.89	2.32	**

Figure 4. Available work in an average trial viewport in the Familiar (*fm*) and Random (*rd*) conditions. For example, if a subject requests a trial and is assigned the Familiar condition, the average number of Node Repair work items in the trial viewport is 2.89. (T-test without log transformation; df=537, 531, and 723, respectively.)

Measure	Cd	Total	Per subject			*P*
			mean	Q3	max	
Ratings	rd	197	3.9	0	59	***
	fm	2676	47.0	55	537	
Node repairs	rd	219	4.4	4	47	
	fm	214	3.8	2	61	
Block edits	rd	1384	27.7	22.5	362	
	fm	1328	23.3	19	183	
Note edits	rd	7	0.14	0	3	○
	fm	48	0.84	0	26	

Figure 5. Total work completed in experiment trials for each type of work. We compare the Familiar (*fm*, n=57 subjects completing at least one Familiar trial) and Random (*rd*, n=50) conditions (df=67, 101, 101, and 70, respectively).

Measure	Cd	mean	Q2	Q3	max
Ratings	rd	0.55	0	0	45
	fm	7.19	0	5	119
Node repairs	rd	0.62	0	0	8
	fm	0.58	0	1	7
Block edits	rd	3.90	0	0	186
	fm	3.57	0	3	69
Note edits	rd	0.02	0	0	3
	fm	0.13	0	0	25

Figure 6. Work per trial, comparing Familiar (n=372 trials) and Random (n=355) conditions.

clopath users tend to live and work in (and be more familiar with) more densely populated areas, which have correspondingly denser roads and trails.

We performed our analyses using both raw counts (e.g., number of blocks rated) and counts normalized by the amount of available work (e.g., proportion of available blocks that were rated). The results were the same; thus, we report only raw counts for clarity, but we report the least favorable significance value of raw and normalized data.

H1: Familiarity, H3: Work Type-Familiarity. We hypothesized that users would do more work in Familiar viewports (H1), but that this effect would be stronger for ratings than for node repairs (H3). The results support H3, but only partially support H1. Figures 5 and 6 show that subjects entered an order of magnitude more ratings in the Familiar condition than in Random: 2676 total and 7.19 per trial vs. 197 total and 0.55 per trial. However, the amount of node repairs and block edits was virtually identical in the two conditions. And while many more notes were edited in the Familiar conditions, the total number were small (48 vs. 7), and the difference was only marginally significant.

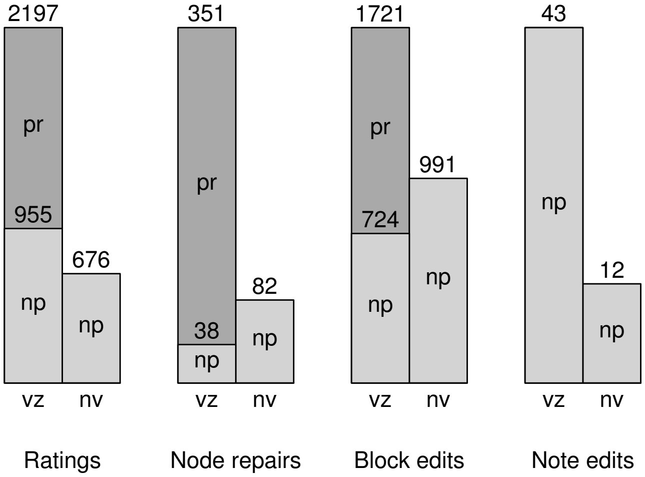

Figure 7. Total work completed in experimental trials, showing what was done in the Visual Prompts (*vz*) and No Visual Prompts (*nv*) conditions, divided into prompted (*pr*) and not prompted (*np*). For ratings and node repairs, prompted work included visually highlighted blocks rated and node repairs resolved; for block edits, prompted work includes all block edits related to a node repair prompts (as discussed above). Note that it is not meaningful to compare bar heights across work types, as maximal bar heights for each type have been equalized.

We speculate that very different effect of familiarity for the different work types is explained by the original rationale for H3: rating a block requires familiarity, but making a node repair does not. We suspect that familiarity giving no benefit for node repairs means that subjects essentially did not apply personal knowledge to this task; instead, they turned on aerial photos and looked to see whether a node was present. In contrast, in previous work that found familiarity to be the basis of effective work elicitation techniques [5, 6], the work elicited did require personal knowledge. It would be interesting to examine tasks in other domains that do not require personal knowledge. For example, Hoffman et al. [14] had subjects verify that certain information was present in a Wikipedia article. This task did not require subjects to know anything about the topic of the article. Would users be more motivated to do this task if they did know about the topic?

H2: Visual Prompts. We hypothesized that providing visual highlights to focus users on specific work opportunities would result in more work being done. Our results provide evidence for this. Figure 7 shows the total amount of work of each type done in the two conditions, and Figure 8 provides more detailed comparisons of work done in the Visual Prompts and No Visual Prompts conditions.

We were not surprised at the difference in the amount of work done: we had conjectured that when users are given something to focus on, they are likely to do more work than if given a visually complicated display for nothing stands out. While the data support our hypothesis, the reason for this is not exactly what we had supposed.

Measure	Cd	Total	Per subject			*P*
			mean	Q3	max	
Ratings	nv	676	21.1	21.8	224	○
	vz	2197	68.7	96	537	
	-np	955	38.8	46.8	416	
	-pr	1242	29.8	50.8	121	
Node repairs	nv	82	2.6	1.5	40	*
	vz	351	11.0	7.3	108	
	-np	38	1.2	0.3	11	
	-pr	313	12.0	12.3	102	
Block edits	nv	991	31.0	10.8	499	
	vz	1721	53.8	33.8	420	
	-np	724	22.6	15.3	226	
	-pr	997	31.2	25.5	272	
Note edits	nv	12	0.4	0	7	
	vz	43	1.3	1	26	

Figure 8. Total work completed in experiment trials for each type of work. We compare Visual Prompts (*nv*, n=32 subjects) to No Visual Prompts (*vz*, n=32 different subjects) (df=44, 48, 55, and 50, respectively); further, we show work done in Visual Prompts that was actually prompted (*pr*) and not prompted (*np*). *Total* gives the number of work units completed by all subjects during trials, while *Per subject* shows the the amount of work done by the mean, 75th percentile (*Q3*), and most prolific (*max*) subject. *P* gives the P-value code comparing Visual Prompts and No Visual Prompts.

Measure	Cd	n	mean	Q3	max
Ratings	rt,nv	80	7.03	5	105
	rt,vz	224	6.00	3	119
Node repairs	rp,nv	81	0.93	1	7
	rp,vz	268	1.30	2	8
Block edits	rp,nv	81	10.90	11	186
	rp,vz	268	5.90	7.3	80
Note edits	nv	161	0.075	0	3
	vz	566	0.076	0	25

Figure 9. Work units completed per trial in Visual Prompts (*vz*) and No Visual Prompts (*nv*) conditions, when asked to do Ratings (*rt*) or Node Repair (*rp*). Subjects were never asked to do Note Edits, so we include both for that measure. We report the number of trials in the condition (*n*) and the amount of work done in a mean, 75th percentile (*Q3*), and maximum trial.

Cd	Total	Per subject			*P*
		mean	Q3	max	
nv	161	5.03	5.3	25	*
vz	566	17.69	18	140	

Figure 10. Number of trials requested by subjects (df=49).

Specifically, the advantage of Visual Prompts was largely not because subjects did more work *per trial* (Figure 9) but because they requested over three times more trials (Figure 10). In other words, the *visual focus* vs. *visual clutter* distinction was not particularly supported. Subjects entered somewhat more ratings and block edits per trial in the No Visual Prompts condition, made somewhat more node repairs in the Visual Prompts condition, and made about the same number of note edits in the two conditions.

A second surprise was that in addition to subjects doing much work they were prompted to do, they also did much *unprompted* work. Figure 7 shows that subjects in the Visual Prompts condition entered more ratings of unprompted blocks than No Visual Prompts subjects did in total (though this result is not statistically significant).

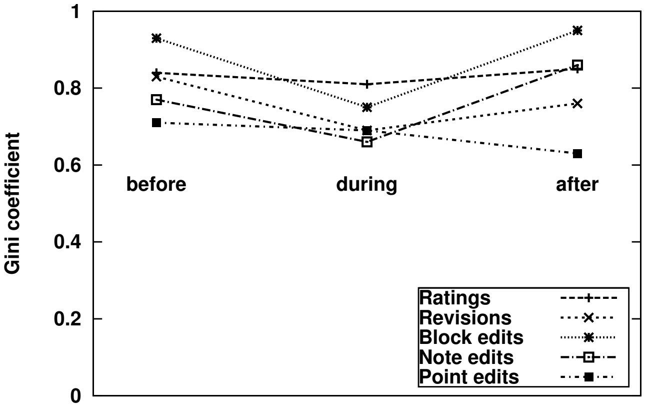

Figure 11. Distribution of work as measured by the gini coefficient. Work distributions are shown for diferent types of work during three time periods: before, during, and after the work elicitation experiment.

RQ2: Distribution of Geographic Volunteer Work

We were interested in several specific questions about the distribution of GVW in Cyclopath. What is the general distribution of work? Do different types of work (e.g., block edits and ratings) differ in their distributions between users? Are different individuals leaders in different types of work? Finally, did our work elicitation techniques *change* the distribution of work?

Metrics. We computed two metrics to answer these questions. We used the gini coefficient to quantify overall distribution of work across users and *unique-at-N* to quantify whether the same or different individuals were the leaders in different types of work. For a given level of N (say 5), we found the set of top-N contributors for each work type, then took the union of these sets, i.e., the set of individuals who were among the top-N producers of any kind of work. We did this for ratings, block edits, note edits, and point edits.

We computed our metrics for all work done before, during, and after the experiment. Note that during the time period of the experiment, we included work done outside of experiment trials, including by users who did not participate in the experiment at all. We thought this was a more realistic situation: any online community that conducts a work campaign cannot compel users to participate, only entice them. Therefore, our results give a more accurate picture of the global effects on work that maintainers of an online community could produce with such a campaign.

Distribution of work. Figure 11 shows the gini coefficients for the different types of work over the three different time periods. The results tell some interesting stories.

First, as expected, work of all types is very unequally distributed, meaning a small proportion of users do most of the work. Values of the gini coefficient range from nearly 0.95

(for blocks edited after the experiment) to a low of 0.63 (for points edited after the experiment).

Ortega et al. [20] analyzed distribution of editing work across 10 different language Wikipedias, reporting gini coefficients ranging from 0.92 to 0.97; thus, general editing activity in Wikipedia is highly concentrated, as much or more so as editing blocks in Cyclopath. Similarly, Kittur and Kraut [16] analyzed concentration of edits to a *single* Wikipedia article. Using a sample of over 23,000 articles, they found that work at this level was distributed much more evenly, with a mean gini coefficient of 0.26 ($\sigma = 0.18$), much more equitable than Cyclopath.

Second, the distribution of different types of work varies somewhat. Ratings and block edits are most concentrated, while point and note edits are less concentrated. These differences need more investigation. However, we can speculate on a few factors that may contribute to more or less concentration of work. Rating the bikeability of a block requires a cyclist to recall his or her impression of a past experience on that block. Thus, it requires both experience and memory of the experience at a fairly detailed level. Editing blocks is difficult at the interface level (clearly the most so of any of the four work types) and requires either personal experience or reference to correct aerial photos. On the other hand, editing notes and points is technically easy, and this type of information might be more memorable: for example, it might be easier to remember the location of a cafe or bike shop you frequent or a bike rack you use than to remember how much you liked riding a particular block.

Third, our work elicitation campaign affected the distribution of work, with the distribution of each type of work becoming less unequal. The work type where the decrease is largest is block edits: the gini coefficient decreased from 0.93 to 0.75. There are several possible explanations for this. First, the visual prompts for missed intersections made a particular type of block editing task more prominent and technically easier. Further, the instructions suggested turning on aerial photos. If users had not been aware of this before, just learning about it would make the task more accessible. Second, the 20-trials-per-day limit played a role: three of the ten most prolific block editors reached this limit at least once.

Fourth, once the work campaign ended, distributions of work (with the potential exception of points edited) largely returned to the pre-campaign status quo. This is an important finding: a work campaign can have two key benefits: accomplishing work and creating workers. For example, Drenner et al. [7] found that changes to the MovieLens new user process resulted in those users participating at higher rates even after they had completed the entry process. However, it appears that our campaign did the first, but not the second. We think this is largely due to how we ended the work campaign: we took away the "How can I help?" button and the visual prompts – essentially, we made block editing harder again! We did this because, while we intend to incorporate these features into the Cyclopath interface permanently, some changes were required before we could do so. Thus, perhaps the right lesson to draw from this is not so much about the effect of work campaigns on the distribution of work, but rather about the effect of interface features.

		Time Period		
		before	during	after
Rank Limit (N)	5	11	13	10
	10	23	23	26
	20	43	44	53

Figure 12. *Unique-at-N* metric for N = (5, 10, 20) and three time periods: before, during, and after the experiment. The minimum value of the metric is N and the maximum 4N.

Different leaders for different types of work. Figure 12 presents the results of the *unique-at-N* metric; these results show consistency in overlap between the top contributors of different work types. At level 10, for example, the number of unique contributors was 23 before and during the experiment, and 26 after. Further, the work campaign had little impact on diversity of leadership.

RQ3: How GVW affects route finding

Cyclopath's primary service is generating routes. Therefore, we measured the effect of volunteer work on the quality of routes Cyclopath generates. We randomly selected 800 of the of the 6,700 unique routes requested by Cyclopath users from August 2008 through April 2009 and analyzed them as follows. For a given route request (start/end pair), we compared the quality of routes obtained at four different *analysis instants*.

1. *Before user input.* The first instant was at system initialization, after the Cyclopath base maps were loaded but before any user work.

2. *Before experiment.* The second instant was immediately before the experiment began. By this point, users had entered 36,541 ratings and made 32,894 block additions, edits, and deletions.

3. *After experiment.* The third instant was immediately after the experiment ended. During the experiment, users entered 4,700 ratings and made 3,974 block changes.

4. *Now.* The final instant was on May 1, 2009, after an additional 2,144 ratings and 7,020 block changes.

For each analysis instant, we restored the transportation network to the state it was in at that instant. We then issued each of the 800 route requests and recorded the length of the route obtained. To the extent that routes became shorter over the course of the four analysis instants, we could conclude that user input benefitted route finding.[9]

[9] This account hides two details. First, when requesting a route, users can specify that distance, bikeability, or some mix be optimized. Our analysis issued two requests for each route at each instant, one with the default setting – balanced distance/bikeability priority – and one that prioritized distance only. Second, route cost can be measured by total bikeability rather than length; we did this, too. However, in all four cases (two route-request settings, two improvement measures), the results are nearly identical; therefore, for

Time	Distance (kilometers)						X
	mean	min	Q1	Q2	Q3	max	
init	14.8	0.3	6.5	11.7	19.6	74.0	32
before	13.8	0.3	6.5	11.5	18.6	63.5	10
after	13.8	0.3	6.5	11.4	18.6	63.5	8
now	13.8	0.3	6.5	11.4	18.5	63.8	10
i − n	1.00	-3.9	-0.01	0.03	0.39	32.4	
i − b	0.97	-3.9	-0.01	0.02	0.36	32.4	
b − a	0.01	-1.9	0	0	0	1.0	
a − n	0.02	-0.8	0	0	0	3.6	

Figure 13. Summary statistics of sample routes at four key analysis instants and improvements due to the editing in each interval (positive means improvement). We also report the number of route requests for which no route could be obtained (*X*).

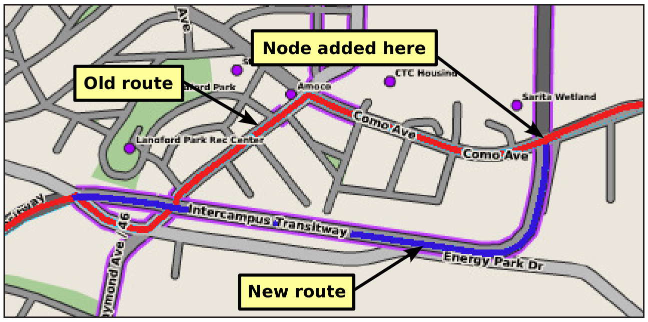

Figure 14. Excerpt of a route improved from 15.6 km to 15.0 km due to work during the experiment (a node added to the network at the indicated location).

Effect on routes

Figure 13 summarizes the effect of user input on routes. The central finding is that user input improved the quality of routes obtained (t-test without log transformation, p=0.03, df=1517). Before any user input, the average length of a route was 14.81 km; after the last of our analysis instants, the average length was 13.82 km: thus, the typical route became one kilometer shorter!

There are several additional interesting observations. First, user work had more effect prior to the experiment, and not just because more work was done. While this period had on the order of 10 times as many edits as the during- and after-experiment periods together, it was closer to 100 times as impactful. We speculate that this is because that edits that happened early in the history of the system captured “low hanging fruit”. For example, one route’s length decreased from 54 to 22 kilometers! This was because the start and destination were on opposite sides of a river, but our initial dataset was missing four bicycle-accessible bridges over a 15 km stretch of the river. Users added these bridges and connected them to the rest of the transporation network.

Second, user input had little effect on many routes, while greatly improving some and worsening (fewer) others. Figure 13 illustrates this relationship. For the first quartile of routes (the 200 routes for which there was least improvement), there was a reduction of about 9 meters, while for the third quartile of routes, there was a reduction of 387 meters. Figure 14 shows a specific example of a route improved by 600 meters due to user work; on the other hand, an erroneous one-way setting added to a bridge over a highway added 1.9 km to a route that depended on the link.

Third, our figures underestimate the benefits of user input. 38 (of the 800) route requests could not be satisfied at one or more of the four analysis instants; that is, the destination was not reachable from the start through the transportation network. Since the length at one or more instants was effectively infinite, we excluded these routes from our analysis. Of the 32 route requests that could not be satisfied prior to user input, 28 could be satisfied at the last analysis instant. On the other hand, the reverse happened for 6 routes. In other words, users “fixed” 28 routes – 3.5% of the sample – but “broke” 6 others – 0.75% of the sample.

CONCLUSION: IMPLICATIONS AND SUMMARY

Our results suggest several interesting implications and issues for the design of techniques to elicit user work.

Intelligent edit monitoring. A key issue in open content systems is directing user attention where it is most helpful. We explored one aspect of this, eliciting work. However, there is another aspect which has received more attention (at least in Wikipedia): monitoring for erroneous or malicious edits. Watch lists and the recent changes feed are powerful tools for this task, but there is room for improvement: most edits are good and don’t need scrutiny, and some of the few edits that do need scrutiny don’t receive it.

Our results suggest a heuristic for identifying edits in Cyclopath that need user attention: those that have non-minimal impact on routes (either positive or negative). The system could flag these edits for extra scrutiny, and existing monitoring mechanisms could be extended accordingly. For example, users could subscribe to “recent *significant* changes”.

To generalize, this heuristic works because users edits in Cyclopath *influence the results of a computation*. There is no direct analogue of this in Wikipedia, because people, not algorithms, are the consumers of Wikipedia articles. While algorithms that measure edit properties such as the number of characters or what proportion of an article changed are useful, these are only rough proxies for the impact of an edit; they don’t distinguish edits that change the meaning of an article from mere “wordsmithing”. Thus, this heuristic is most directly applied to other systems where user edits are input to a computation [21].

Matching elicitation to the type of work elicited. We tried two techniques to elicit user work: visual prompts and familiarity. Familiarity had a powerful effect on ratings entered, but none on nodes repaired. On the other hand, visual prompts affected both work types, but the effect was more significant for node repairs. We conjecture that visual prompts helped reduce the inequality of the distribution of geographic editing. We think the difference in effects results from how specific properties of the two work types aligned

simplicity of presentation, we report only the length change metric for routes obtained using the distance-only setting.

with the two techniques. Rating bikeability requires personal experience and a specific memory of that experience. Thus, familiarity is key, and visually highlighting a block that a cyclist is not familiar with does no good. On the other hand, visual prompting completely transformed the node repair task. Without visual prompts, identifying a missed node is a difficult perceptual recognition task; visual prompts make it easier. Aerial photos help users decide whether a node exists – no personal experience required! In other words, visual prompts reduce a user's *personal cost* of doing work [15].

Utility of different work types in an open content system. There was some diversity among the leading editors for different types of GVW in Cyclopath. For example, about 25 individuals were among the top 10 editors for at least one of the four types of work we considered. Having different types of user work has several benefits. First, each work type may require different skills, thus appealing to different types of people. In Wikipedia, for example, spell-checking, adding content to articles, and ensuring adherence to policies like *Neutral Point of View* require different kinds of knowledge. Second, different types of work vary in accessibility to new users. In Cyclopath, we speculate that it will be easy and engaging for new users to add points representing key places in their neighborhoods and notes to explain dangerous, scenic, or otherwise noteworthy road or trail segments. Third, each work type can give rise to a different social comparison (e.g., one leaderboard for ratings and another for node repairs), multiplying the effect of this powerful motivator.

Summary. We implemented two techniques to elicit and focus work in the Cyclopath bicycle bicycle route-finding system. The techniques succeeded, resulting in large increases in work done. Additional analysis revealed factors that contributed to the success of each of the techniques. We also quantified the distribution of work in Cyclopath and the effect our techniques had on it. Finally, we showed that user work led to demonstrable benefits: user input has decreased the length of the average Cyclopath route by 1 km.

ACKNOWLEDGEMENTS

We thank the members of GroupLens Research, the Cyclopath team, the Cyclopath user community and subjects who volunteered, and Tom Erickson. This work was supported in part by the National Science Foundation, grants IIS 05-34692 and IIS 08-08692.

REFERENCES

1. Adamic, L. A. and Huberman, B. A. Zipf's law and the Internet. *Glottometrics*, *3* (2002), 143–150.
2. Atkinson, A. B. On the measurement of inequality. *Journal of Economic Theory*, *2*, 3 (1970), 244–263.
3. Bryant, S. L. et al. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proc. GROUP*. 2005.
4. Butler, B. S. Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures. *Information Systems Research*, *12*, 4 (2001), 346–362.
5. Cosley, D. et al. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proc. CHI*. 2006.
6. Cosley, D. et al. SuggestBot: Using intelligent task routing to help people find work in wikipedia. In *Proc. IUI*. 2007.
7. Drenner, S. et al. Crafting the initial user experience to achieve community goals. In *Proc. RecSys*. 2008.
8. Festinger, L. A Theory of Social Comparison Processes. *Human Relations*, *7*, 2 (1954), 117–140.
9. Frey, B. and Meier, S. Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *Amer. Econ. Review*, *94*, 5 (2004), 1717–1722.
10. Golder, S. A. and Huberman, B. A. The structure of collaborative tagging systems. *CoRR*, *abs/cs/0508082*.
11. Goodchild, M. F. Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*.
12. Harper, F. et al. Social comparisons to motivate contributions to an online community. In *Persuasive Technology*. 2007, 148–159.
13. Hill, W. et al. Recommending and evaluating choices in a virtual community of use. In *Proc. CHI*. 1995.
14. Hoffmann, R. et al. Amplifying community content creation with mixed initiative information extraction. In *Proc. CHI*. 2009.
15. Karau, S. and Williams, K. Social loafing: A meta-analytic review and theoretical integration. *Personality and Social Psychology*, *65*, 4 (1993), 681–706.
16. Kittur, A. and Kraut, R. E. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *Proc. CSCW*. 2008.
17. Kittur, A. et al. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. alt.CHI*. 2007.
18. Ling, K. et al. Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, *10*, 4 (2005), 00.
19. Locke, E. A. and Latham, G. P. Building a practically useful theory of goal setting and task motivation. a 35-year odyssey. *Amer. Psych.*, *57*, 9 (2002), 705–717.
20. Ortega, F. et al. On the inequality of contributions to Wikipedia. In *Proc. HICSS*. 2008.
21. Priedhorsky, R. and Terveen, L. The computational geowiki: What, why, and how. In *Proc. CSCW*. 2008.
22. Priedhorsky, R. et al. How a personalized geowiki can help bicyclists share information more effectively. In *Proc. WikiSym*. 2007.
23. Sutton, L. Using Usenet: Gender, power, and silence in electronic discourse. In *Proc. Berkeley Linguistics Society*. 1994.
24. Whittaker, S. et al. The dynamics of mass interaction. In *Proc. CSCW*. 1998.