The Value of Geographic Wikis

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Reid Royer Priedhorsky

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Loren G. Terveen

August 2010

Acknowledgements

First, I would like to thank my wife Erin Tatge for her consistent, unshakeable support through the many difficult years in creating Cyclopath and this thesis. Without her, this achievement would not have been possible.

Perhaps the most important thing I have learned in graduate school is the power of collaboration, and for that I thank my coauthors: Jilin Chen, Thomas Erickson, Benjamin Jordan, Shyong (Tony) K. Lam, Pamela J. Ludford, Michael Ludwig, Mikhil Masli, Katherine Panciera, Ken Reily, John Riedl, Shilad Sen, and, most importantly, my advisor, Loren Terveen, who was willing to talk to a stranger about bicycle maps and who believed in my ideas when they were vague and naïve. I have learned a tremendous amount from Loren and will always be grateful for his advice and guidance.

I am extremely thankful to the Cyclopath user community and the subjects who participated in our research. It goes without saying that this work would have been impossible without these people, and I am humbled by their numbers. Jim Nordgaard sent us patches fixing important bugs. The Cyclopath concept originated in discussions with Doug Shidell of Little Transport Press.

I also thank the Cyclopath team members, past and present: Landon Bouma, Thomas Erickson, Jordan Focht, Anthony Johnson, Michael Ludwig, Mikhil Masli, Katherine Panciera, Steve Schacht, Zach Schloss, S. Andrew Sheppard, Isaac Sparling, Loren Terveen, and Fernando Torre.

The members of GroupLens Research have provided invaluable technical and administrative help, advice, support, and feedback, particularly Angela Brandt, Rich Davies, Pam Ludford, and Katie Panciera. Similarly, the feedback of my anonymous reviewers has been of great value. I also thank my committee: Loren Terveen, John Riedl, John Carlis, and Kevin Krizek. @reidsthesis provided important comic relief.

The Minnesota Department of Transportation, the Met Council, the Land Management Information Center, and the United States Geological Survey provided geographic data for Cyclopath. Tim Starling and his colleagues at the Wikimedia Foundation made request log data available to us and provided technical support for its processing. I also thank the members of wikitech-l for their technical help.

This work was supported in part by NSF, grants IIS 03-07459, IIS 05-24851, IIS 05-34420, IIS 05-34692, IIS 08-08692, and IIS 09-64697.

Abstract

This thesis responds to the dual rising trends of geographic content and open content, where the core value of an information system is derived from the work of users. We define the essential properties of an emerging technology, the geographic wiki or *geowiki*, as well as two variations we invented: the *computational geowiki*, where user wiki input feeds an algorithm, and the *personalized geowiki*, where the system provides a personalized interpretation.

We focus on two systems to develop these ideas. First, Cyclopath, a research geowiki we founded, serves the bicycle navigation needs of cyclists. We also present analysis in the context of Wikipedia, the well-known and highly successful wiki encyclopedia, using its size and maturity to draw lessons for smaller, younger systems which are far more numerous but hope to grow.

We ask three questions with respect to this new technology. First, *can it be built?* Yes. This thesis describes the design and implementation of Cyclopath, which has grown to be a production system with thousands of users.

Second, *is it useful?* Yes. We identified a representative geographic community, bicyclists, and they both tell us that the information in the Cyclopath geowiki is useful and show us by using the system in great numbers. We also present new ways to measure value in wikis, introducing new techniques for doing so from the perspective of information consumers. In particular, user work in Cyclopath has shortened the average route by 1 km. Also, we present techniques for obtaining more contributions (familiarity matters – sometimes – and users do work beyond what they are asked to) and algorithms for increasing the value of geowiki content by personalizing it, showing that traditional rating prediction algorithms (collaborative filtering) are not effective but simple algorithms based on clustering are.

Finally, *who cares?* Many people. There are numerous communities with great interest in geographic information but limited, incomplete, or awkward access because the relevant knowledge is distributed among members of the community and otherwise unavailable. As our results demonstrate, geowikis are an effective way of gathering and disseminating geographic information, more so than previous techniques. Thus, this research has broad value.

Contents

List of Tables

List of Figures

A note on the design of this document

This thesis conforms to University of Minnesota dissertation formatting requirements. Readers preferring a book-like design (notably, single-spaced and suitable for double-sided printing) can download an alternate version from http://reidster.net/papers.html.

Chapter 1

Introduction

1.1 Wherefore geowiki?

Two major online revolutions are occurring. One is *open content*, where users produce most or all of a site's value: Wikipedia has grown to over three million articles in English alone [122], Yahoo Answers has an archive of over one billion answers [123], and YouTube users upload 24 hours of video every minute [125]. 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 taxi fare[1] to volcanoes[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. One can't hire a plumber from another country to fix some pipes – ergo, Angie's List – nor can one 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.

This thesis explores a new type of system which enhances the utility of

[1] http://worldtaximeter.com

[2] http://www.geocodezip.com/v2_volcanoBrowser.asp

[3] http://geogreeting.com

this emerging area: the geographic wiki or **geowiki**, whose properties we[4] formalized. We explore how the logical conclusion of open content – wikis, where anyone can edit anything – can be adapted to the geographic domain, and how this strange new model of mass collaboration functions within it. We also explore two extensions to this idea, which we invented: **computational geowikis** feed user geowiki contributions into an algorithm of some kind, and **personalized geowikis** add personalized interpretations of the geowiki's information content.

This work analyzes two systems in service of these goals. Cyclopath,[5] a system we created, is a web-based mapping application that serves the navigation needs of bicyclists in the Minnesota cities of Minneapolis and St. Paul, an area of roughly 8,000 square kilometers and 2.3 million people.

Cyclopath has two main roles. First, it computes bicycle-friendly routes. Second, Cyclopath is a geowiki containing information of interest to cyclists:

- The **transportation network** of streets and trails over which cyclists travel; this is represented as a graph, where street intersections are **nodes** and the segments of street and trail between them (which can have internal vertices which are not nodes) are **edges**.

- Points and regions relevant to cyclists.

- Annotations on all of the above, some shared (notes, tags, and discussions) and others private (**bikeability ratings** describing the bicycle-friendliness of the network on an edge-by-edge basis).

All of this is editable by anyone – even anonymous users – except bikeability ratings, which are private and can only be editable by their owners. No prior systems have fully implemented the complete geowiki model we propose.

[4] *A note on pronouns.* Essentially all research in this field is collaborative, and this thesis is no exception. The papers which I have adapted all have multiple authors, up to six; in each case, however, I am the lead author and led the research effort. To reflect the collaborative nature of the work reported in this thesis, I use the first person plural *we* throughout.

[5] http://cyclopath.org

We also explore Wikipedia, the famous and successful online wiki encyclopedia. This system consistently ranks among the top ten most visited sites on the web, and its information is considered to be as accurate as traditional encyclopedias [46]. While not a geowiki, Wikipedia is an example of a large and mature wiki community; thus, we use it to draw lessons relevant to the growth and limits of wiki systems, of which geowikis are an important subclass.

This thesis explores the following questions:

- *What is the rationale for the geowiki model, and how can such systems be built?* Cyclopath's success is an existence proof of the plausibility of such systems. We describe its design and implementation, and we present survey and interview data demonstrating the utility of the geowiki collaboration model. (Chapter 2.)

- *Are contributions to geographic wikis valuable?* We use content analyses and usage data to show that indeed, contributions are valuable, and the scheme has significant promise. (Chapter 3.)

- *How should the value of contributions be measured?* We introduce techniques to measure the aggregate impact of contribution and also to measure the impact of specific individual contributions, going beyond prior metrics to focus on consumer value rather than the quantity of producer actions. (Chapter 4.)

- *How can more contributions be elicited?* We report a field experiment introducing methods for eliciting geographic volunteer work and showing that geographic familiarity matters, but only for some types of work, and that users often do work beyond what they are asked to. (Chapter 5.)

- *How can geographic open content be personalized?* We present a framework for evaluating personalized routing techniques and use it to evaluate several techniques in the context of Cyclopath. (Chapter 6.)

We have created a personalized, computational geowiki called Cyclopath, producing significant design innovations and solving major implementation

challenges to do so. The results of our experiments both in Cyclopath and in Wikipedia show that this new model attracts valuable information and is a useful collaboration technique. Finally, these results are of broad interest because they affect any geographically-oriented community whose geographic information is distributed among its members, of which there are many examples. In short, our novel geowiki ideas are successful and show great additional promise. This thesis presents the results of this ongoing research theme.

1.2 Essential properties of geowikis

In this section, we define the concept of *geowiki*, along with two extensions to this concept, the *computational geowiki* and the *personalized geowiki*. First, however, we clarify exactly what a **wiki** is, an important step as the core of our argument is to take the essential properties of wikis and adapt them to the geographic domain. We believe these properties are as follows:

P1. **Open access.** Any reader, including anonymous readers if such readers have access, can be an editor, and anything can be edited (with occasional exceptions – for example, the restricted editing privileges on controversial articles in Wikipedia do make the encyclopedia a less "pure" wiki, but they don't make it not a wiki).

Importantly, this property implies that the system must be sufficiently easy to use that editing is widely accessible among readers. For example, while Wikipedia has many serious usability problems, it is good enough to have garnered millions of individual editors.

P2. **Post-review.** Changes are live immediately upon being saved, and review happens after publication, not before.

P3. **Transparent changes.** It is easy to for everyone to see how the artifact is being changed, and by whom. Specifically, a wiki needs a **recent changes list** showing at a high level the flow of changes in the artifact or parts of it, **watch lists** (the ability for a user to delegate the computer

to automatically "watch" portions of the artifact and be notified when changes occur), and **diffing** which lets reviewers see and analyze precisely what changed in a particular revision. This transparency is essential for the monitoring and review tasks which make the wiki model function.

P4. **Shared artifact.** The wiki artifact must be owned by nobody (or everybody, depending on one's viewpoint). This is partly a consequence of the first property: in order for parts of the artifact to have egalitarian open access, editing must truly be available to anyone, not just the "owners" of a particular territory.

P5. **Robust consistency.** The artifact's internal consistency (e.g., the validity of inter-article links in Wikipedia) is difficult to disrupt without explicit intent to do so; informally, small edits are not secretly large edits. For example, the disruptiveness of renaming Wikipedia articles is mitigated in several ways: the operation is only available to established users, it requires confirming a warning dialog, and users following links to the old name are automatically redirected to the new.

The existence of private data does not make a system not a wiki. For example, even a very open system like Wikipedia stores real names and addresses that are kept private. It is the presence of a substantial shared artifact that makes a system a wiki; there is no need for the shared artifact to be the only data in the system.

We extend these notions to define a **geowiki**.[6] Specifically, the *open access*, *transparent changes*, and *robust consistency* properties must be extended:

P1. **Open access.** There are three essential changes to achieve open access in a geowiki.

(a) *Web-map interface.* As does any electronic map, a geowiki has a graphical map interface with standard map navigation operations.

[6]This term is already well-established and is generally considered to mean a geographically contextualized wiki. However, we argue that, in a *true* geowiki, the geographic context itself, as well as links between geographic context and non-geographic data, enjoys wiki features – i.e., the structure of the map itself can be edited, not just items on the map.

(b) *WYSIWYG editing.* Geographic data can be edited in a more or less what-you-see-is-what-you-get style, with a reasonably complete set of editing operations. While wikitext markup works moderately well for textual data, geographic data is too complex to edit using text markup.

(c) *Full editability.* If geographic data or links involving geographic data are reasonably editable at all, they can be edited within the geowiki. For example, the names of towns and the locations and connectivity of streets must be editable, but aerial imagery need not be, because aerial images, once loaded, typically are not edited by anyone.[7]

P3. **Transparent changes.** The three transparency tools need to be made geography-aware. The *recent changes list* needs filtering capabilities appropriate for geographic concepts like "currently viewed area". *Watch lists* become **watch regions**, letting users define arbitrary geographic regions of interest and be notified of revisions which (geographically) intersect those regions. Finally, *diffing* must become graphical to let users visualize geographic changes.

P5. **Robust consistency.** Geographic data requires robust links, not those based simply on co-location. For example, text which describes a street must be linked *to the street itself* rather than to *a shape which is co-located with the street*. Co-location links are weak, breaking down both when geographic objects are moved (which they can do in a geowiki) and when multiple geographic objects are themselves co-located.

While being web-based is not an essential wiki property, this by far the most common mode of wikis, and it is critical that a wiki be a unified system which

[7]To be precise, Cyclopath is not yet a true geowiki under this definition, because green space and water bodies are not yet editable. However, this minor deviation from the precise definition has little effect from a user standpoint, and we have plans to roll these features into the *region* map feature type, which is editable.

does not require external software to do work. Thus, for web-based geowikis, all of the above requirements must be achieved within the web browser.

The **computational geowiki** further extends this notion: it is a geowiki where the shared artifact – perhaps combined with other data – feeds an algorithm.[8] In the case of Cyclopath, this algorithm is route finding: part of the shared artifact (the transportation network along with tag annotations) is combined with private data (bikeability ratings) to find bicycle routes. Other examples could include building fuel cache resupply schedules or predicting ice movement in Antarctica, automatically identifying properties of concern in a neighborhood geowiki, or predicting off-road vehicle activity with an eye towards creating a natural resource protection plan.

Finally, the **personalized geowiki** also extends the geowiki notion, this time to provide individual users with a personalized interpretation of the system's data. In Cyclopath, this takes the form of personalized route finding.

[8]Technically, search is an algorithm. But, we believe that simply searching the content of an artifact does not meet the spirit of this definition.

Chapter 2

Cyclopath: Rationale and System Description

2.1 Why study bicyclists?

The bicycling community is a good fit for the geowiki model. The information of interest to bicyclists is geographic, highly detailed, changes over time, and no comprehensive information resource previously existed. Additionally, this community has a robust tradition of sharing information. These properties led us to believe that a geowiki for bicyclists would attract an active community, and that this online community would be useful for studying a variety of topics, both those directly concerning geowikis as well as other topics.

More specifically, the bicycling community is focused around *doing*: navigating a bicycle in the physical world. This activity raises interesting challenges. First, it is inherently geographic and typically local (i.e., people mostly ride in the area where they live). Second, the *planning* task – deciding where to go, and how to get there – is hard. It is hard because cyclists must navigate a transportation network largely designed for another purpose (driving motor vehicles) and because they must do so under continually changing conditions (e.g., weather, motor traffic, construction). Third, cyclists have significant individual differences in purpose, attitude, and abilities.

The complexity of bicyclists' tasks results in complex information needs.

Cyclists have a strong tradition of sharing information, but their existing sharing practices (prior to Cyclopath) were relatively inefficient. Aside from Cyclopath, there still is no comprehensive, up-to-date information resource that helps users find routes meeting their personal preferences. This is true despite at least four major vendors offering geographic web search, automated route finding, and easy-to-use geographic web APIs.[1] Google Maps even has a cycling layer (and finds routes for bicyclists), but it simply maps the transportation network without annotations, and it is not personalized. Why hasn't the needed information resource already been built? Why add the overhead of the wiki model, with its complexity and the problems that come with it?

Importantly, the cycling community is relatively small. In the United States, cycling is two to three orders of magnitude less frequent than driving, whether measured by number of trips or number of miles traveled per person [106]. Combined with cyclists' need for detailed and continually changing information, this makes the problem too hard for hobbyists and economically unattractive for businesses and governments.

However, *wikis change the equation*. While it's still hard to gather and maintain the information, this work is distributed across many motivated users, rather than being the responsibility of the system builder.

Finally, studying a cyclist-oriented geowiki will impart general lessons. While cyclists' good fit with the geowiki model shows that at least one community exists which would benefit from this type of system, there are many others as well. Example include polar science in Antarctica, neighbors going about their daily life, and planning tasks such as monitoring off-road vehicle activity; each of these is dependent on accurate, timely geographic information which is available from distributed laypeople but not from any centralized resource. Thus, building a geowiki focused on cyclists is a good way to study geowikis and geographic open content in general. We have created such a system (Cyclopath), and in this chapter, we present survey and interview data supporting our concept of the geowiki model.

[1] Google, http://maps.google.com; Yahoo, http://maps.yahoo.com; MapQuest, http://www.mapquest.com; and Microsoft, http://www.bing.com/maps/.

We begin this chapter by surveying related work; we then present an investigation of the system requirements for a cyclist-oriented geowiki and describe how these requirements have developed into the current Cyclopath user experience. Finally, we detail the technical architecture of Cyclopath and close with a summary.

2.2 Related work

Here, we outline related work that gives general background or otherwise motivates and contextualizes our ideas, tracing the development of the geowiki notion from open content systems broadly to wikis and geowikis more narrowly. Prior work directly related to our specific studies is discussed in relevant chapters below.

2.2.1 Social and open content systems

The past few decades have seen the emergence of numerous Internet-based social media, including email lists, Usenet, MUDs and MOOs, chat, blogs, wikis, and social networking systems. Researchers have taken advantage of the opportunities created by the large user bases these media have attracted, creating a wide variety of advanced software (e.g., visualization and analysis tools [100, 107], social agents [61], and social navigation aids [111]) and conducting a broad range of empirical research (e.g., on conversational patterns [7], social interaction [23], gender [57], and group-wide patterns [112]).

Additionally, a new form of collaborative interaction has emerged. Users no longer just consume information or discuss topics; they work together in *open content systems* to produce *artifacts of lasting value* [28]. Collaborative filtering systems like MovieLens and Amazon leverage users' ratings of items (movies, books, consumer products, etc.) to enable personalized recommendations. News sites like Reddit, Digg, and Slashdot rely on user opinions to filter and order stories. Tagging systems like del.icio.us and CiteULike let users associate keywords with items, facilitating searching and exploration. Yahoo Answers and Stack Overflow let users exchange information in a structured

question-and-answer format. Media sharing sites like Flickr and YouTube let users share photos and videos with one another.

Scholarly interest in this form of interaction is intense, encompassing both studies of current sites and techniques and attempts to develop improved and novel techniques. For example, Ling et al.'s experiments based on social science theory produced design principles for online communities [73]. Lampe explored distributed moderation of comments in a discussion forum [69, 70] and Sen et al. the evolution of tag vocabularies [98]. Researchers have also studied mixed-initiative systems combining open content with automatic work, such as intelligently routing tasks to users [29] or automatically linking free-form user-contributed data to a structured database [34].

2.2.2 Wikis

Wikis take user-provided open content to its logical completion: anyone can add, edit, or delete anything. Researchers have addressed a variety of topics. For example, Kittur and Kraut explored coordination and conflict between wiki editors across thousands of wikis operated by the Wikia wiki hosting service [67], and Bruns and Humphreys explored the use of wikis for developing collaborative learning strategies [17]. Researchers have also created novel wiki software, such as Forte and Bruckman's teaching-focused wiki system [42] and Chi et al.'s wiki extensions to enhance coordination between editors [24].

The most famous and successful wiki is the Wikipedia online encyclopedia, containing over 3 million articles in English, ranking among the 10 most visited sites on the Web, and having accuracy similar to *Encyclopedia Britannica*, at least for some types of articles [46]. Wikipedia makes a great deal of its data available to anyone, including the encyclopedia's complete current historical content; this easy access to rich data has led the bulk of wiki research to focus on the encyclopedia. Researchers have studied the lifecycles of users as the progress from novices to experienced editors [18, 83], socialization into the Wikipedia community [26], the roles of elite users as gatekeepers over certain decisions [63], and the relationship of article topics to those included in traditional encyclopedias [36], among many other aspects of the encyclopedia

and its community.

2.2.3 Geography

The field of geographic information systems (GIS) is concerned with visualizing, analyzing, and manipulating geographic and spatial data [74]; traditionally, GIS work is done by experts using specialized software. The GIS community has proposed various types of collaboration [9, 78, 96].

In particular, geographic "citizen science" is becoming an established approach. For example, an annual bird census is done mostly by laypeople [14], and the National Map Corps is a volunteer program to correct and update United States Geological Survey maps [11]. Professional geographers manage these projects and vet the data submitted by citizens; often, as in the case of the National Map Corps, these professionals are a bottleneck [11].

More recently, geographers have become interested in what they call *volunteered geographic information*. Goodchild has argued for the value of average people's geoknowledge [49], and geographers held a scholarly meeting in 2007 [50] and a special issue of *GeoJournal* in 2008 [35] to set a research agenda for the area. However, geographers distrust the wiki model, perceiving tension between data quality and open content [41, 49]. On the other hand, open content is ascendant within the collaborative computing community. This is due in no small part to the great success of Wikipedia – contrasting dramatically with the failure of its predecessor, Nupedia, which had an elaborate credentials-based review process.

2.2.4 Geowikis

Many systems partially implement the geowiki model. OpenGuides[2] is a wiki travel guide, WikiMapia[3] lets users enter and edit information for places and rectangular regions, and Google Maps lets users edit the locations of searched-

[2] http://openguides.org

[3] http://wikimapia.org

for places and add new places. SeeClickFix[4] and FixMyStreet[5] let citizens collaboratively map the locations of public maintenance needs [64]. Google My Maps goes beyond this, allowing collaborative editing of geographic points, paths, and polygons, all of which can be annotated with text, images, and videos. "Digital graffiti" systems let users associate virtual information with physical geography, and this information can then be accessed *in situ* with location-aware devices [19, 37]. These systems, however, have a fundamental limitation: users cannot interact with the transportation network or anything else on the base map. Instead, they must rely on *pictures* of this base map, and any user content is a separate, weakly linked layer. The base geodata are not available for annotation or editing.

Other projects come closer to true geowikis. The most famous, Open Street Map,[6] is an ambitious wiki project building a global street map starting from scratch. However, the focus is solely on the transportation network – there are no annotations like tags, notes, ratings, or discussions. Another recent system, Google Map Maker,[7] lets users edit the transportation network as well as points of interest and monitor the edits of others, though the system is not available in North America, Europe, or many other parts of the world. While these two projects offer open WYSIWYG editing of the transportation network, key wiki monitoring features like geographically filterable recent change lists and watch regions are missing.

2.2.5 Navigation support

Efforts to enhance navigation with user contributions also exist. Some GPS devices let users enter map corrections directly into the device and subscribe to corrections made by other users [104]. Navteq, a major mapping firm, accepts user map corrections on its website, but these are checked by professionals before being offered to other users [81]. Bederson et al. have proposed a system

[4] http://seeclickfix.com

[5] http://fixmystreet.com

[6] http://openstreetmap.org

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

for augmenting automobile route-finding using data collected from and shared among users, such as a history of routes driven and notes relating personal experiences [12]. Finally, Counts and Smith have proposed a system that uses location-aware sensors to construct route information for individuals which can then be shared with others [30].

Many existing web sites aim to support cyclists' route information needs. Gmaps Pedometer,[8] Bikely,[9] and others let users manually define and share routes. Wayfaring[10] enables collaborative editing of routes and places of interest. Open Cycle Map[11] is a cyclist-focused rendering of Open Street Map data. A few offer automatic route finding, such as YTV Journey Planner for Cycling[12] and Fietsrouteplanner,[13] and others offer automatic route finding based on Open Street Map user-contributed data, such as the Cambridge Cycling Campaign's "journey planner" for cyclists.[14] Finally, Biketastic[15] is a research system which uses a mobile phone application to share information about rides, collected using both built-in sensors (pavement roughness, noise level) and user contributions (photos, videos, and text descriptions) [89].

Notably, Google recently introduced bicycle route-finding in Google Maps for many American cities [71]. This was a dramatic advance for cyclists across the country, most of whom previously had no online bicycling map or automated route-finding at all. However, the system has a number of significant limitations. Even after being available for over four months, as of July 2010, the system still makes elementary routing errors like sending bicyclists down Interstate 94 in Minneapolis. More importantly, however, Google Maps' cycling layer lacks annotations of any kind on the transportation network, the opportunity for cyclists to contribute their knowledge in a straightforward way

[8] http://gmap-pedometer.com

[9] http://www.bikely.com

[10] http://wayfaring.com

[11] http://opencyclemap.org

[12] http://kevytliikenne.ytv.fi/?lang=en

[13] http://fietsersbond.nl/fietsrouteplanner/

[14] http://www.camcycle.org.uk/map/route/

[15] http://biketastic.com

and have it show up on the map in a timely and predictable manner, and personalized routing, both in the sense of modeling individuals' preferences and letting users customize particular route requests – all features whose importance is demonstrated by our research.

While these sites recognize cyclists' need for navigational information and routes, or the potential of open content to meet these needs, or both, none have fully embraced the geowiki model in the rich and complete way that our results suggest is necessary.

2.3 System requirements

Our first task in designing the cyclist-oriented geowiki that became Cyclopath was to develop its requirements. We did this using a qualitative study of local bicyclists, conducted during the summer of 2006.

2.3.1 Method

We studied the cycling community in the Minneapolis-St. Paul metro area using surveys and interviews. We sent invitations to three local cyclists' discussion lists with about 950 total members, and we encouraged recipients to forward the invitation.

73 respondents finished the survey, and most questions had about 75 responses. 68% of survey respondents were male and 32% female; most were between 18 and 64 years of age with a roughly uniform distribution. Survey questions focused on attitudes regarding map errors and existing planning and navigation methods. All participants were over 18 and had spent at least 200 miles or 25 hours cycling within the local area during the past year.

We also used the survey to recruit for interviews. We completed 19 semi-structured interviews lasting 60 to 90 minutes each; 13 subjects were men and 6 women. Some questions elaborated on issues introduced in the survey; others explored topics more suitable to an interview setting, such as existing navigation practices and attitudes towards the wiki model. We also presented lo-fi paper mockups of three core Cyclopath features – map editing, bikeability

ratings, and watch regions. Finally, each subject sketched a map of a familiar route to provide non-verbal data on geographic thinking. Interviews were recorded by two note takers and subsequently coded.

2.3.2 Unmet needs

Using these surveys and interviews, we identified three key unmet needs. Bicyclists lacked a comprehensive and up-to-date information resource, automated route finding, and personalized bikeability ratings for roads and trails. (*Information resource* in this context means anything bicyclists use to plan routes, common examples being bicycle maps, guidebooks, and discussion forums.)

2.3.2.1 No comprehensive, up-to-date resource

To plan a route, cyclists need to know how they can travel through geographic space *now* and what they will find within the space *now*. Some of this information isn't recorded at all, and the rest is distributed across numerous electronic and non-electronic resources. For example, some bike trails and bike lanes are mapped by the state, others by municipalities, and still others by park boards.

Bicycle travel is accomplished by moving along the edges which make up the transportation network, so cyclists need to know the geometry and topology of this network. Edges clearly include dedicated bike trails and roads (though some roads are impassable to bicyclists). However, subjects suggested several additional classes of surfaces that are sometimes edges, including alleys, sidewalks, and parking lots.[16] More classes certainly exist, and it would be difficult to enumerate them even if all possible cycling routes were known. Regardless, current information resources do not include these types of edges. Furthermore, only actual edges should be included, in order to avoid overwhelming the resource and users with unimportant information. For example, most side-

[16]To be pedantic, *edge* really means "a road, trail, or other traversable facility which should be on the map", a definition subtly different from the set of such facilities that are used by cyclists. For example, the Cyclopath map includes expressways, which are not used by cyclists but which are included for orientation and context. However, this distinction is of little rhetorical or consequential import, so we ignore it.

walks should be excluded, but some, such as those that bridge highways, are key edges. Cyclists know the locations and properties of edges because they themselves travel upon them, and their memories can be enhanced with aerial photos.

The locations of *landmarks*, *resources*, and *obstacles* are also important to bicyclists. 16 out of 19 subjects mentioned landmarks. These, in addition to street names, are used for orientation and navigation, and subjects cited objects like businesses, highways, and water bodies, i.e., both point and non-point landmarks. Resources are things helpful to cyclists in some way; 11 subjects cited a wide variety of resources including restrooms, water sources, and restaurants. Obstacles cause cyclists difficulty or frustration. 12 subjects mentioned a variety of obstacles, including construction and traffic lights.

A few of these items – e.g., water bodies – were found on existing information resources, but most were not. Also, it is time-consuming to identify which landmarks, resources, and obstacles are actually important; for example, a quirk of the local cycling club's culture is that Dairy Queen restaurants are frequent landmarks. Cyclists know which of these things are important and where they are located because they themselves use or avoid them.

These observations – that existing information repositories are widely scattered and incomplete, and that cyclists themselves know the important information – motivate the distributed approach of wikis with deep editability, including the transportation network itself.

2.3.2.2 No automated route finding

Cyclists told us that they want automated route finding, i.e., "find me a good route from point A to point B": 4 of 19 subjects mentioned this desire specifically, 5 described a problem that could be solved with such a tool, and 5 expressed dissatisfaction using motor vehicle route finders for cycling (11 total). No such tool was available at the time of the experiment.[17]

Tools for automated motor vehicle route finding were successful and quite

[17]Google Maps subsequently released a bicycle route finding tool; a detailed comparison of the Google Maps and Cyclopath approaches is presented in Section 2.2.5 on page 13.

widespread, but our interviews revealed that they are unsuitable for cyclists because they do not know about all edges and they do not take into account the complexity of cyclists' routing needs. While a handful of bicycle-specific routing tools existed, most notably byCycle,[18] they suffered from the same basic problems of incomplete coverage and simplistic routes.

Automated bicycle route finding can use the same basic approach as motor vehicle routing: compute a minimum-weight path through the transportation network. However, while motor vehicle routing uses simple factors like distance and travel time to calculate edge weights, effective bicycle weights are based on many additional factors. Subjects cited factors both objective and subjective, including the locations of hills, presence and quality of pavement, motor vehicle traffic levels, motorist attitude, and numerous others.

2.3.2.3 No personalized bikeability

Furthermore, bicyclists' ratings of and preferences for any given edge are a matter of personal taste: people do not agree on which quality factors should be considered and what their relative importance should be. This led 8 of 19 subjects to question the utility of existing generic bikeability ratings (in the cases where they are available), expressing either a general concern that their own notion of what made for a good edge might differ or else that they had actually encountered ratings they disagreed with. Additionally, existing resources offered bikeability ratings for only a subset of edges. What cyclists really wanted was a way to obtain personalized ratings for *all* edges.

2.3.3 Cyclists' culture of sharing

An open content system is of no value if users willing to share their knowledge are not available. Our surveys and interviews highlighted cyclists' existing culture of sharing and their openness to technology-assisted sharing.

Subjects expressed a strong propensity to share what they knew with other cyclists. 83% of survey respondents reported asking other cyclists for route

[18] http://bycycle.org

Willing to spend	**If others saw corrections ... in six months**	**immediately**
1 minute or more	67%	96%
5 minutes or more	44%	73%
10 minutes or more	21%	43%

Table 2.1: Willingness to correct map errors.

planning help. They were also willing to spend substantial effort correcting map errors, especially if their work was available to others immediately; see Table 2.1. Furthermore, some cyclists already spend considerable effort on helping one another. In 2006, the local recreational cycling club's 100 volunteer ride leaders led over 1,400 rides [5]. The duties of a ride leader are to obtain or create a route, scout the route regularly, maintain and distribute maps and turn lists, and lead rides along the route – many hours of effort on behalf of other cyclists.

In our interviews, we asked subjects if they would share routes with the general public; 17 of 19 said yes, and 14 mentioned textual information they would share. When asked if they would rate the bikeability (i.e., bicycle-friendliness) of roads and trails to help other cyclists, 13 subjects said yes. When asked if they believed that they *knew how to* correct map errors they had encountered themselves, 14 said yes, and when asked why they would fix map errors, 7 gave helping others as a reason.

We also observed cyclists using existing collaborative technology, however cumbersome, to share information. For example, the following routing help request appeared on a local cycling web forum:

> How do I get to Khan's in Roseville, from the St. Paul campus [of the University of Minnesota], without being killed in traffic?
>
> Khan's Mongolian Barbeque
> 2720 Snelling Ave N
> Roseville, MN 55113

This request generated 8 responses, including (a) a detailed Gmaps Pedome-

ter[19] route posted within about 7 minutes, (b) another Gmaps Pedometer route recommended for use after dark, (c) an endorsement of the second route, and (d) a warning that a specific section contained many potholes. Another thread began with a warning that a particular bridge was closed, and a third, titled "streets to avoid", had 24 posts.

These results reflect the existing tradition of information sharing within the bicycling community. Cyclopath leverages this tradition and adds tools addressing cyclists' unmet needs, allowing for even more effective sharing.

2.4 The Cyclopath user experience

Our fundamental challenge was to produce a system that rivaled Google Maps in quality of user experience and performance while also supporting the two key roles of Cyclopath: finding bicycle-friendly routes and offering a fully-featured geowiki information resource for cyclists. Importantly, Cyclopath derives much of its value from the work of its user community (a finding explored in detail throughout this thesis). We call this **geographic volunteer work**, to emphasize the active role of end users and in contrast with the geographic community's term *volunteered geographic information* [49].

Figure 2.1 on the facing page is a screenshot of the Cyclopath user interface. Key interface elements are tagged, and we reference these elements throughout this section.

The right-hand side of the interface is taken up by the map. Basic display and navigation work in the now-standard "slippy map" style, similar to Google Maps and its peers. Geographic features (i.e., the core map feature types of edge, point, and region, as well as green space and water) are distinguished visually using color, and users navigate by dragging the map, using the pan/zoom controls (B), or searching for specific locations (C). A hideable map key (H) reminds users how colors and other markings correspond to data, and users can choose one of eight aerial photo underlays instead of the flat

[19]Gmaps Pedometer (http://gmap-pedometer.com) is a website which lets anyone create and then share an arbitrary polyline shape overlay on a Google map.

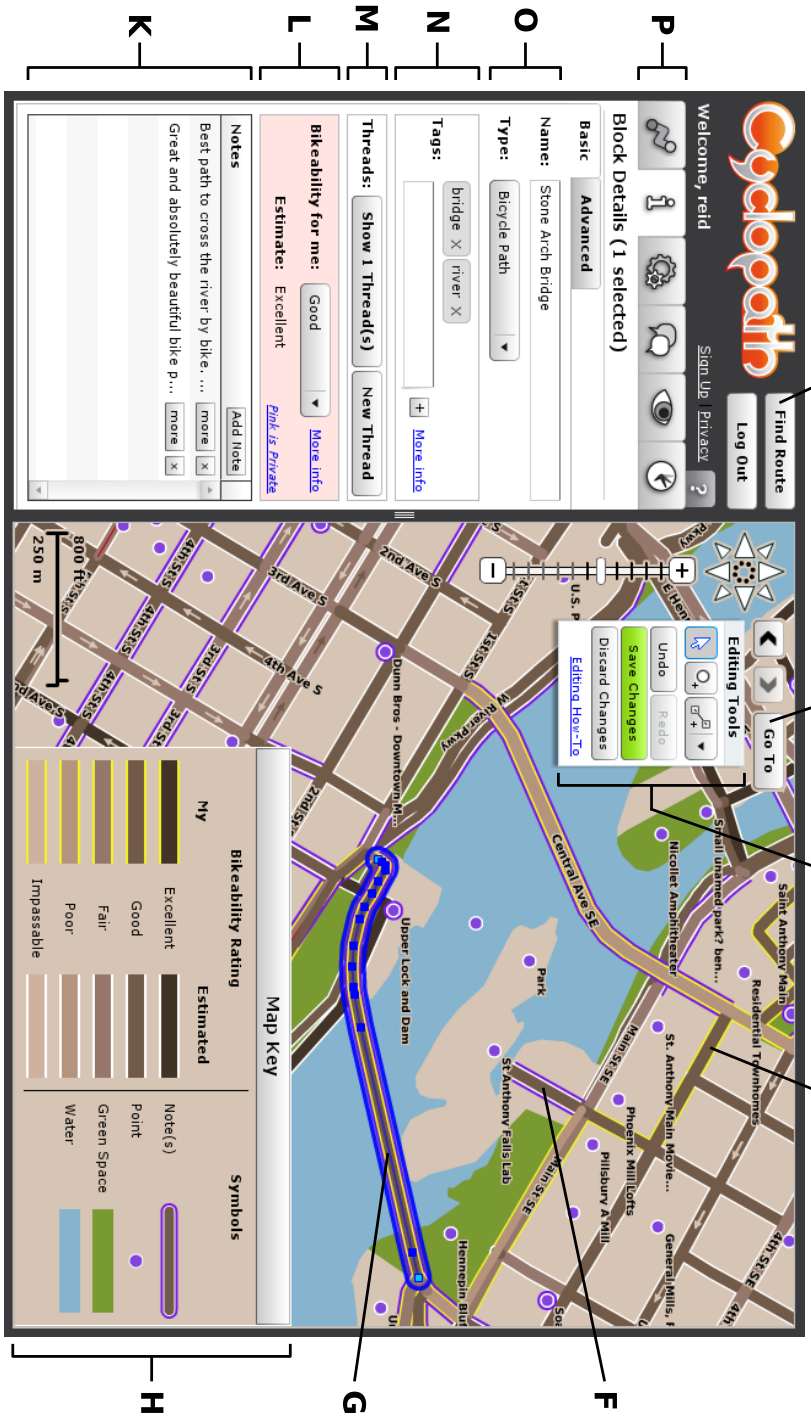

Figure 2.1: The Cyclopath user interface. Tagged elements of interest are discussed in the text.

background.

The left-hand side is a tabbed interface used for information display. From left to right, the six tabs (P) are:

1. A list of found routes (the system can display multiple routes simultaneously).

2. Information about the selected map object(s), in this case an edge (G), which the interface refers to as a *block*. The edge's name and type (O) as well as the four types of annotation (K-N), discussed in more detail below, are shown. The advanced tab (not shown) contains additional attributes for the edge.

3. Preferences such as display options and filters to display only certain points and regions.

4. A discussion forum.

5. *Watch regions*; users can define regions of interest and receive notice when changes occur within those regions.

6. A filterable recent changes list.

The remainder of this section discusses the Cyclopath user experience in more detail. We first discuss route finding, as it is the most common use of the system, and then move on to deeper interactions with Cyclopath's information content: the four types of map annotations supported by Cyclopath, editing the map, and monitoring the editing of others.

Finding routes. To find a route, users click the Find Route button (A), producing the dialog box in Figure 2.2 on the facing page, where they enter starting and ending locations (street addresses, intersections, points of interest, or regions) and input route preferences. The slider controls how much bikeability affects edge weights, and users can also set edges with particular tags to have a bikeability bonus or penalty or be avoided altogether. A simpler route finding interface is also available both on the Cyclopath home page and

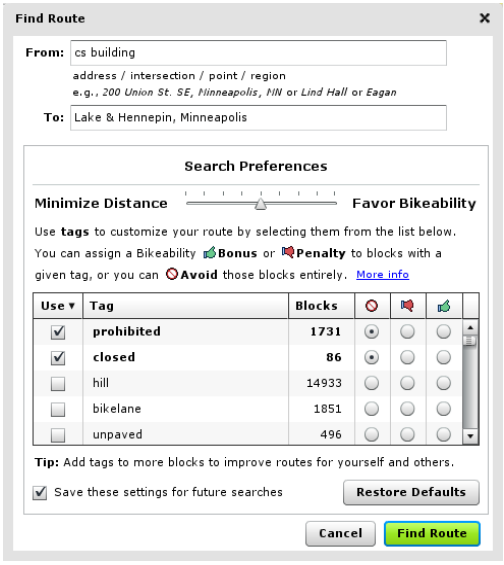

Figure 2.2: Cyclopath's route finding dialog.

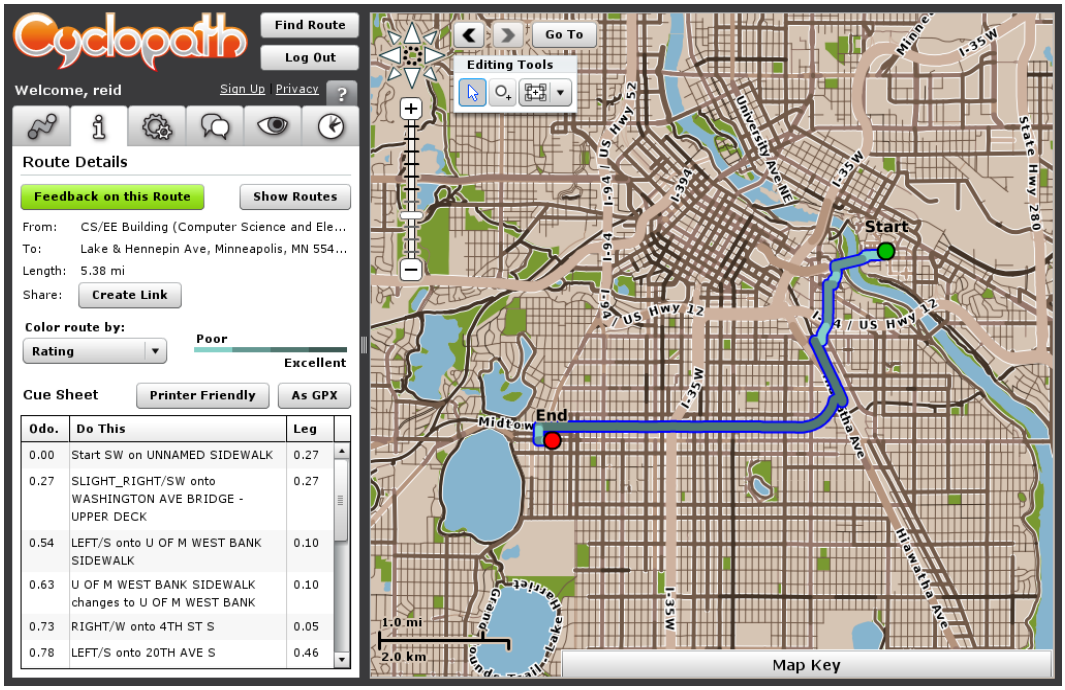

Figure 2.3: The route found using the settings in Figure 2.2. On the right is the route on the map, while on the left is the map as a list of turn-by-turn directions (a *cue sheet*) in addition to controls for changing the route's visualization, giving feedback, printing out the cue sheet, and other tasks.

in widgets that third-party website owners can embed on their own web pages. The result (a route) is shown in Figure 2.3 on the previous page.

Annotations. In addition to the geographic objects themselves (edges, points, and regions), Cyclopath supports annotations on the objects. These annotations are attached to map features using ID-based links, not co-location; this makes the links robust in the face of changing or overlapping map objects. The four types of annotations are:

- *Notes* (K) – free-form text of arbitrary length. They are shared, meaning that anyone can add, edit, or remove them, and they have a many-to-many relationship with edges – i.e., one note can be applied to one or many edges and vice versa. Currently, notes have a one-to-one relationship with points and regions; each point or region can have exactly zero or one notes.[20]

- *Tags* (N) – short text snippets. These are also shared, and they have a many-to-many relationship with edges, points, and regions (or some mix of the three).

- *Discussion threads* (M) – conversations. They are visible to all, but messages are "owned" by whoever posted them and cannot be edited. Individual messages have a many-to-many relationship with edges, points, and regions, or a mix, and messages can also have no links at all. A thread's set of links is the intersection of the links in all the messages and can thus also be empty.

- *Bikeability ratings* (L) – ratings of the bicycle-friendliness of edges, on a 5-level scale from "Impassable" to "Excellent". These are private to individual users (and thus only available when logged in). A user can apply at most one rating to an edge, and a rating applies to only one edge;

[20]The astute reader will notice that the relationships between annotations and map objects have different form between the different types of annotations or objects. The annotation features were added at different times, and these inconsistencies are by oversight rather than by design. We are working to correct them.

for example, rating the four-block stretch of Hennepin Ave. between 20th and 24th St. involves four ratings. We provide advanced selection tools to speed the process of rating many edges at once [76].

These annotations are reflected in the cartography of the map. Edges and points with notes, tags, or discussions are drawn with a purple border (e.g., F), while edges that have been rated by the logged-in user are drawn with a yellow border (E). Edges are also drawn according to their rating: from dark brown for Excellent to light brown for Impassable. If the logged-in user has not rated an edge, the rating is estimated using the techniques described in Section 2.5.3 on page 31.

Editing the map. As a geowiki, Cyclopath provides in-browser editing of points, regions, and (notably) the edges and nodes of the transportation network. The *Editing Tools* palette (D) contains tools to add new map features and manipulate their geometry. The interface works similarly to standard drawing programs and is fully WYSIWYG; Figure 2.4 on the next page illustrates the process of adding a new edge. Editing of attributes and annotations is done using the fields in the item details tab (P). Changes are managed in the client and are not live until they are sent to the server in a batch by clicking the *Save Changes* button (or *Post* for discussion messages, not shown); this batch of saved changes forms a **revision** of the map.

Monitoring changes. Cyclopath provides three key features useful for monitoring the edits of others.

- The **recent changes list** enumerates changes to the map. Each time a user saves a revision, his or her username (or IP address if anonymous), the date and time, an optional comment, and a polygon summarizing the geographic extent of the revision are added to this list. Other users can then filter this list of revisions based on geography, username, and other factors to see an overview of the map's editing activity.

- Users can also define **watch regions**, which adapt the wiki watch list to

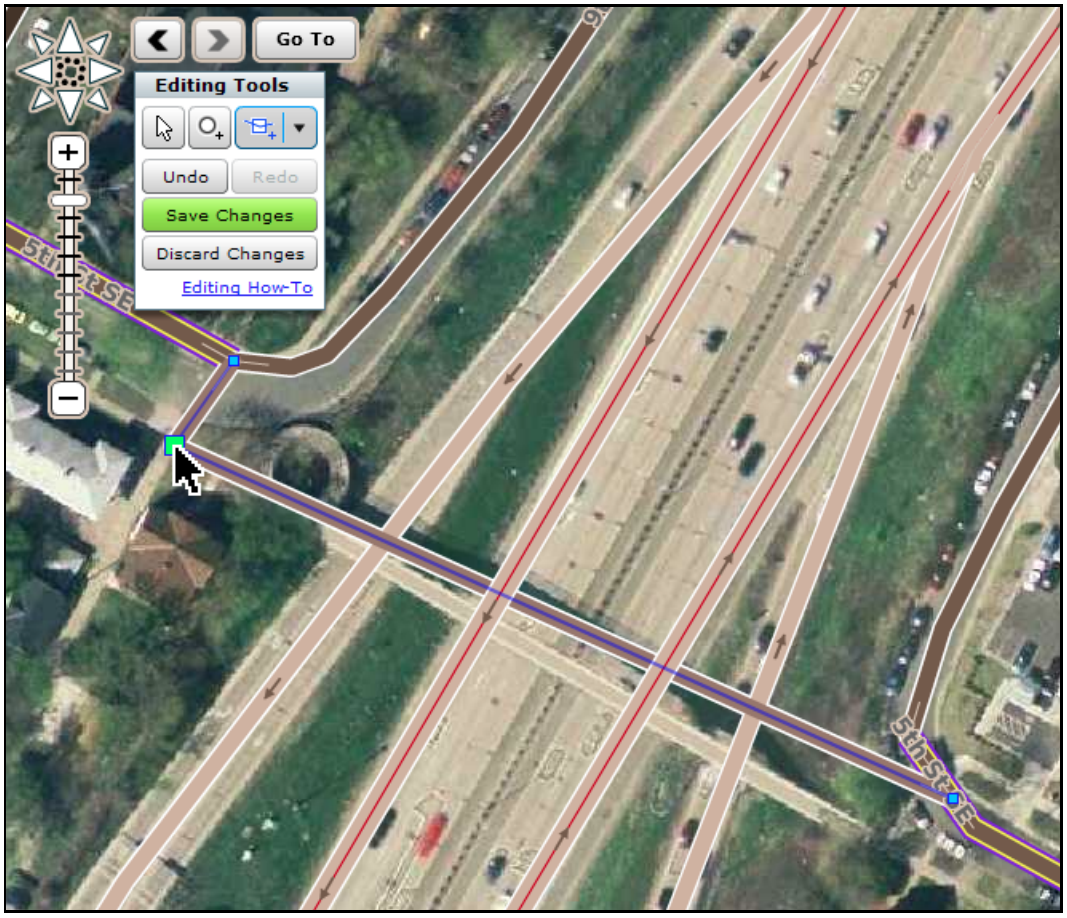

Figure 2.4: Editing the geometry of a pedestrian bridge over a highway, using aerial photos for assistance. This edge has three vertices; the two endpoints (in light blue) and the one currently being dragged by the mouse (green). The light-colored lines extending from each endpoint indicate topological connections (i.e., graph nodes).

the geographic domain. Rather than listing objects to watch, each user can create one or more polygonal geographic regions of interest and then be notified when objects within those regions are modified; these watch regions are private and not visible to other users. Users can also choose to watch existing public regions (e.g., the Ventura Village neighborhood).

- **Geographic diffing**, shown in Figure 2.5 on the facing page, lets users visualize how revisions affect the map.

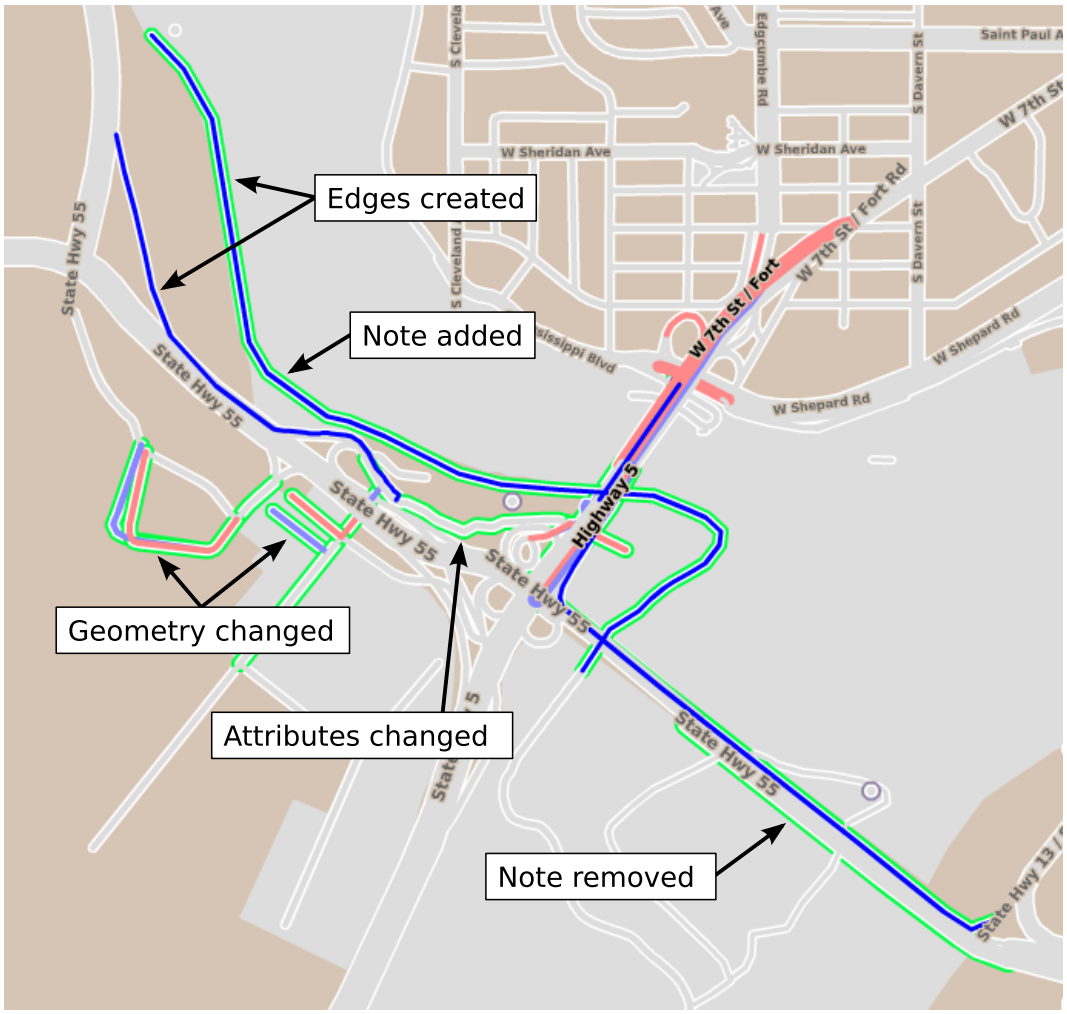

Figure 2.5: Cyclopath's geographic diffing visualization. New edges are shown in dark blue, changed geometry is indicated with light red (old geometry) or light blue (new), and changes to non-geometric attributes and annotations yield a green outline. In the live system, users can flip between this complete view, just the old geometry, and just the new geometry, and they can also click on objects with non-geometric changes to see a summary of those differences. This revision added 7 edges and changed 19 (7 geometrically, 6 non-geometrically, and in 6 both), added 5 notes to 5 blocks, and removed 4 notes from 3 blocks.

2.5 System architecture

The fundamental implementation challenge in realizing our design for Cyclopath was to create a web-based geowiki system with a good user interface and acceptable performance. This section details these challenges and how we solved them.

2.5.1 Architecture overview

Cyclopath's architecture is illustrated in Figure 2.6 on the facing page. At a high level, Cyclopath consists of a web-based mapping interface which fetches and saves map data over the Internet using HTTP. We implemented this web client in Adobe Flex (using the ActionScript and MXML programming languages); it runs under the widely deployed Flash Player virtual machine. We evaluated other technologies, notably AJAX, SVG, and Java, but found them too slow or not widely supported; also, the user experience quality was significantly better with Flex.

There are three data paths used by this web client ("flashclient"). First, vector map data (e.g., the location and attributes of edges) and miscellaneous data (e.g., user preferences) are exchanged with server software ("pyserver") written by us in Python, a common technology for Internet servers, which is embedded in the Apache web server using the mod_python glue toolkit. This data exchange uses a lightweight custom XML serialization protocol on top of HTTP. The pyserver fetches and stores most of this data directly in our PostGIS[21] database using SQL. The exception is requests for routes – these are passed on using a raw Internet socket to a third program we wrote ("routed"), which is a stand-alone daemon that keeps the complete current transportation network in memory. routed obtains this network from the PostGIS database using its own independent SQL connection.

Second, rasterized map tiles are obtained using Tiled Web Map Service (Tiled WMS) [82], a standard HTTP-based protocol for transferring such map

[21] http://postgis.refractions.net

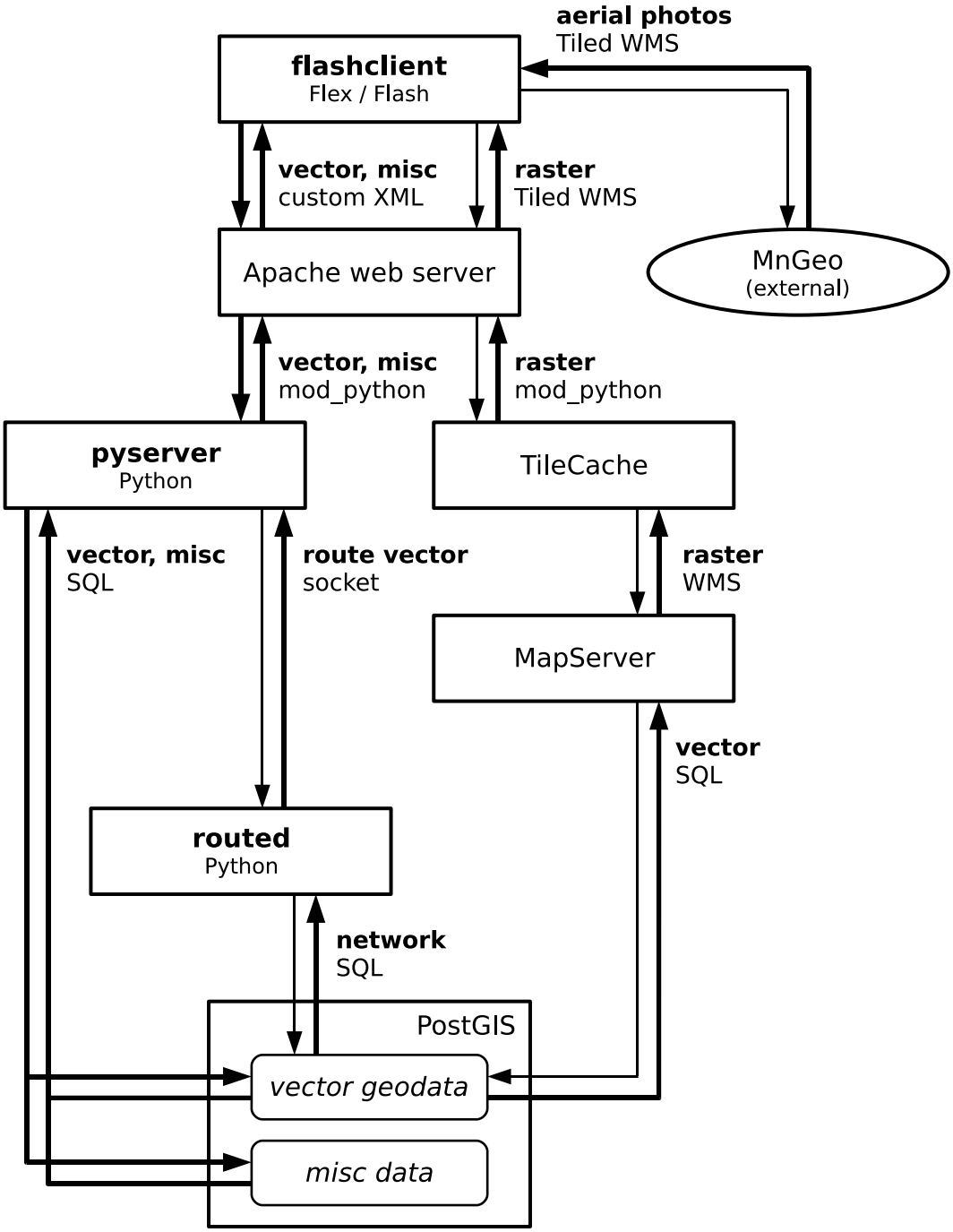

Figure 2.6: Architecture of the Cyclopath system. Heavy arrows indicate data transfer, while light arrows indicate requests only. Bolded box labels indicate software we created.

images; these requests are routed to TileCache,[22] a Tiled WMS cache server, and (for cache misses) on to MapServer,[23] which renders tiles based on vector map data obtained from the database using a third SQL connection.

Finally, aerial photos are also obtained using Tiled WMS, but from a third party: the Minnesota Geospatial Information Office[24] provides this data as a public web service.

The Cyclopath software is made up of approximately 60,000 lines of code, roughly two-thirds in the client and one-third in the server.

2.5.2 Performance of the interactive map

Achieving good performance for a geowiki is difficult, more so than for Google Maps and other systems that do not present the edges of the transportation network as interactive objects. Not only must a web browser manage hundreds or thousands of interactive, clickable objects, geodata must be delivered from the server to the client in a form manipulable by users and software. In other words, it seems impossible to serve pre-rendered image tiles.

However, we make a critical observation: in practice, geographic data objects cannot be manipulated by users if they are too small on-screen. In other words, if a map is zoomed out, users can't visually distinguish or accurately point to specific geographic features. We exploit this observation by implementing a two-part scheme for serving geodata. When zoomed out beyond a map scale of about 1:24,000 (a region roughly 4 km square in a 1024×768-pixel window), we serve pre-rendered tiles; when the user zooms in, reducing the number of geo-objects in the the visible portion of the map to a tractable level, we switch to serving individual objects containing vector geometry and full attributes.

The performance savings gained by this scheme are significant. For example, at the most-zoomed-in level where we serve raster tiles, a typical view of an urban area might contain 3-4,000 geographic objects comprising about

[22] http://tilecache.org

[23] http://mapserver.org

[24] http://www.mngeo.state.mn.us/chouse/wms/wms_image_server_description.html

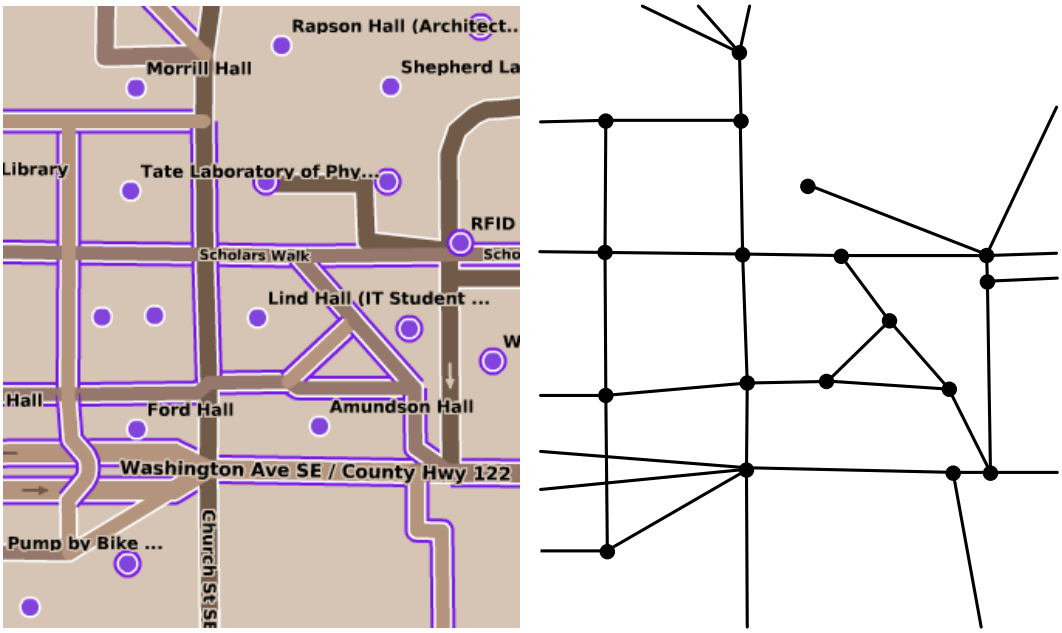

Figure 2.7: A portion of the Cyclopath web map covering the University of Minnesota (left) and its representation as a graph (right).

250 kB of either compressed vectors or raster tiles. However, the interface would be unacceptably slow if it had to render this many discrete objects, and when zooming out n levels, space consumption increases by $O(n^2)$ in vector mode but is roughly constant in raster mode.

Additionally, we have implemented another optimization. When the map is panned or zoomed, data already within the application which remains on-screen is retained; only vectors or tiles which are in the newly-exposed region but not the existing view are fetched from the server (this could result in an L-shaped or donut-shaped request for data). Data which has been moved off-screen is discarded. This technique ensures that only data which is actually needed is fetched, and data which is no longer needed is discarded.

2.5.3 Finding routes

As noted above, the the streets and trails travelled by cyclists form Cyclopath's transportation network. Figure 2.7 shows a small portion of this network as it is displayed in the system and the underlying graph representation.

Cyclopath finds routes using A* search [55] to compute the minimum-cost path through the transportation network. Edge weights are controlled using the route finding dialog shown in Figure 2.2 on page 23; two elements are of interest for this technical discussion. First, the slider from *Minimize Distance* to *Favor Bikeability* controls a constant k that determines how much edge weight depends on bikeability, from not at all ($k = 1$) to fairly heavily ($k = 0$). Second, if a particular tag's checkbox is checked, then the bikeability score b of each edge having that tag is increased or decreased, producing its modified bikeability b': if any tags with the *Avoid* radio button are checked, then $b' = 0$; otherwise, tags with *Bonus* increase the bikeability and *Penalty* decrease it. Specifically, $b' = \frac{b+5n+0.5m}{1+n+m}$, where n is the number of *Bonus* tags on the edge and m is the number of *Penalty* tags. Thus, while the range of b is 0 (Impassable) to 4 (Excellent), the range of b' is 0 to 5.

For an edge of length d meters, its weight $w = kd + (1 - k)3d(1 - b'/5)$. The constant 3 was chosen so that if the slider is in the middle (the default position), traveling 4 km on edges with maximum bikeability is the same cost as traveling 1 km on edges with zero bikeability. Additionally, edges with bikeability less than 0.5 are excluded from consideration, and turns are given costs as well: a right turn costs 40 meters, a left 100 meters, and going straight through an intersection 20 meters.

For each edge, bikeability is obtained as follows, using the first technique that is able to provide a bikeability rating:

1. The user's actual rating for that edge.

2. The mean of other users' ratings on that edge, if at least one other user has rated it (algorithm *Edge-Global-Mean.1*, explained in Section 6.5.2 on page 109).

3. An objective rating algorithm taking into account properties of the edge (algorithm *Objective-CBF* in Section 6.5.1 on page 108).

4. A simpler objective algorithm, which is guaranteed to succeed (algorithm *Objective-Simple* in Section 6.5.1 on page 108).

id	version	valid revision id starting	valid revision id before	name	speed limit
995236	1	4803	4884	Maine St	25
995236	2	4884	4906	Main St	25
995236	3	4906	∞	Main St	35

Table 2.2: Example version history of a Cyclopath edge. Edges have several additional data fields, including geometry and topology, but they are omitted for clarity.

These personalized edge weights, the transportation network, and the start and end points are then fed into the A* graph search algorithm, which computes a minimum-cost route which is then returned to the user.

2.5.4 Versioning

Versioning of objects in Cyclopath works as follows. This scheme is somewhat similar to how other wikis work, but there is a single global sequence of revisions rather than one for each article. This is important in a geowiki, where inter-object relationships (e.g., which edges connect to which other edges) are as important as the objects themselves.

Each time an object is changed, its complete data is copied, the copy updated to reflect changes, and both new and old versions marked with metadata identifying when each version is current. In Cyclopath, a *revision* occurs each time a user clicks the *Save Changes* button; all outstanding changes in the client – which typically affect more than one map object – are saved atomically. The revision's metadata (user, timestamp, change comment, etc.) are saved, and a new *version* of each affected object is created and the metadata of now-old versions updated.

Table 2.2 shows an example edge with three versions. The edge was created in revision 4803 as Maine St with a speed limit of 25 miles per hour. (This revision could have also added, changed, or deleted other objects not shown here.) In revision 4884, the edge's name was updated to Main St, and in revision 4906, the speed limit was changed from 25 to 35. This is the current version of the edge. Thus, from when version 4803 was saved until the instant

id	version	valid revision id starting	valid revision id before	name	speed limit
995236	1	4803	4884	Maine St	25
995236	2	4884	4906	Main St	25
995236	3	4906	5220	Main St	35
995236	4	5220	∞	Maine St	25

Table 2.3: Information stored for the example edge in Table 2.2 after revision 4884 was reverted by revision 5220.

before 4884 was saved, version 1 of this edge was current, and before version 4803 was saved, the edge did not exist. The current version is version 3, since no new version has been created to supersede the one saved in revision 4906.

To generalize: to fetch the map as it existed at time t, the following steps are followed. First, look up the revision ID i with the most recent timestamp not greater than t. Then, fetch map object versions having a valid-starting revision ID less than or equal to i and a valid-before revision ID greater than i. For any given map object, this will yield at most one version (or zero versions if the object did not exist at that time). For the current map, this can be simplified to fetching object versions with valid-before revision ID equal to infinity.

Cyclopath also supports **reverts**, letting users reverse the effect of undesirable edits; a key property of reverts is that a revert of a revert is a no-op. To revert a revision, Cyclopath finds the version of each object added, changed, or deleted in that revision, copies the immediately prior versions, and makes those current. (As a special case, if an object was created by the reverted revision, a new, deleted version is created.) Table 2.3 is an example of this process. Revision 4884 has been reverted by revision 5220, so a new version 4 of the edge – which is identical to version 1 – was created and the metadata for version 3 updated.

2.5.5 Seed data

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, and a list of approximately 100 bicycle shops obtained from the Twin Cities Bicycle Club – but even these unified data had 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.

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 real-world observations – i.e., that the best data available to us have significant flaws – support our conclusions from Section 2.3.2.1 above: that comprehensive cyclist-relevant data is difficult to obtain except from cyclists. This further motivates the deep editing implemented in Cyclopath, and we elaborate in later chapters on why this is critical to the quality of routes that Cyclopath can generate.

2.6 Summary

We identified a representative community – bicyclists – who are a good match for the geowiki model and are enthusiastic about its possibilities (and who have subsequently embraced Cyclopath as it has become an established, production system); many other communities of users are similarly concerned with timely, accurate geographic data which is not available from any centralized resource. Cyclopath is itself a significant contribution: other geographic open content systems, even major efforts like Open Street Map, have to our knowledge not yet implemented a full-featured geowiki as we describe and as we have designed Cyclopath, including full editing and monitoring features. The architecture we have outlined is general, and could be useful in other geographic domains as well, as might the software itself (and thus Cyclopath is open source under the Apache software license[25]); this conjecture is supported by our survey and interview data in a representative community.

Work on Cyclopath continues. We are generalizing the code in order to make it easier to adapt to non-bicycling domains, and we have a number of new features planned, such as a new, unified search interface that lets users search more generally and visualize on the map where search results are located, and saving and editing routes, so users can create routes manually (not just with the route finder) and build up a shared library of useful routes.

We next turn to research focusing on the use of geowikis: the value of contributions, obtaining more contributions, and personalizing the resource.

[25] http://cyclopath.org/wiki/Tech:Open_Source

Chapter 3

The Plausibility of Valuable Contributions

3.1 Introduction

Open content systems involve much overhead. Critical non-productive tasks include creating, debugging, and updating software which enables volunteer work by users of varying levels of expertise, monitoring user work for malice or error and repairing any damage, and the significant task of building and maintaining a community with critical mass. Other, non-open-content methods of data collection and management do not have this overhead. In other words, the wiki model has significant costs, so it is important to evaluate whether its benefits outweigh those costs.

Accordingly, we propose the following framework for determining whether an open content system is useful. *An open content system is useful when users (a) have and share information which is (b) useful and (c) not available otherwise.* In this chapter, we present evidence that these conditions are met for geowikis, using our representative community, bicyclists. Specifically, this chapter explores two research questions:

- *What do people contribute to geowikis?* Our content analysis of Cyclopath contributions reveals that they contribute information both di-

rectly relevant to the core task (bicycling) as well as information which is irrelevant to the task but is of interest to the cycling community.

- *Are these contributions valuable and unique?* Cyclists' responses to our task-based interviews, as well as their usage of and comments on Cyclopath as a production system, show that geowiki contributions are valuable and contain information available only from cyclists.

Additionally, we present a comprehensive enumeration of the different types of *damaging* work, in the context of Wikipedia, which is more rigorous than previous categorizations.

In this chapter, we explore the question of whether contributions are valuable, using self-reports, usage data, and content analyses, first in Cyclopath and then in Wikipedia. These types of evidence suggest that there *is* value in geowiki contributions but cannot quantify it; thus, in the next chapter, we present deeper analyses which do quantify this geowiki value.

3.2 Related work

The value of various types of user contributions is well established. For example, Cedergren explored the dynamics which lead to value in open content news systems [21], and Agichtein et al. produced a system which automatically identified high-value content in Yahoo Answers [3]. Hill et al. showed that user movie ratings in a collaborative filtering recommender system outperformed recommendations from professional movie critics [58]. The value of computer-based volunteer work has been quantified financially; for example, Black Duck Software estimated that $65 billion could be saved annually in the private sector by eliminating software development redundant with open source efforts [15]. Finally, theoretical work by Cosley et al. suggests that users are as effective as experts in reviewing other users' work [27], and that the wiki model and traditional review-before-publication result in the same quality, but the wiki model achieves it faster [28].

The issue of vandalism and damaging work has also attracted attention.

For example, Slashdot uses a socially-based moderation system that is effective in making it easy to ignore uninteresting or offensive comments [70]. This has been most visible in the context of Wikipedia, where there have been a number of high-profile cases of vandalism. For example, Adam Curry allegedly altered pages about the history of podcasting in order to promote his role and diminish that of others [118], Jeffrey Seigenthaler's article stated falsely for months that he was involved in the John F. Kennedy assassination [119], and the comedian Stephen Colbert has even conducted humorous tutorials on how to vandalize Wikipedia [121]. Researchers have also explored the issue; for example, Viégas et al. created an anecdotal categorization of types of vandalism [108], and Geiger and Ribes traced the development of anti-vandalism practices and tools [45].

We go beyond prior work to explore the plausibility of valuable contributions in a new type of open content system, the geowiki, and we present a more rigorous categorization of damage types in Wikipedia than was available previously.

3.3 Content of Cyclopath contributions

In this section, we present evidence that cyclists find Cyclopath valuable. We do this by establishing that all three parts of the utility framework introduced in Section 3.1 are met, using two methods. First, we analyze contributions made in a lab setting in response to representative tasks to establish that cyclists have knowledge, that this knowledge is useful to other cyclists, and that this knowledge is unavailable except from cyclists. Second, these results have been validated by subsequent usage of the live Cyclopath system, and we present usage statistics detailing this support.

3.3.1 Method

We conducted a task-based lab study. Cyclists came to our offices and used an early prototype of Cyclopath to do four tasks, including to enter knowledge; these data became part of the initial content of the live system.

We recruited active cyclists using online methods, posting invitations on several mailing lists and forums frequented by cyclists as well as Craigslist. This skewed our subject pool towards cyclists who are comfortable on computers, which seems reasonable because we are evaluating computer-based support tools. Subjects were 18 or older and had spent at least 3 hours or 25 miles bicycling in the Uptown neighborhood of Minneapolis, Minnesota during the year preceding the study. They were compensated with a gift certificate to a local bike shop.

After conducting 6 pilot interviews, we ran the experiment with 29 subjects.[1] 17 of the subjects were men and 12 women. 17 reported riding "nearly every day", and 19 rode at least 50 miles per week.

The basic form of an interview was a series of four tasks, each followed by a mini semi-structured interview. The four tasks were:

T1. **Route finding.** The subject used our system to compute a familiar route of their choosing. *We asked:* whether the subject liked or disliked the route, how they currently learned new routes, and how useful these learning methods were (on a scale of 1 to 5).

T2. **Entering places.** The subject entered at least 4 new points (places) that he or she wanted to share with other cyclists. *We asked:* why the new places were chosen, whether it was difficult to think of places to enter, how subjects currently found information about cycling-relevant places, and how well these information-finding methods worked.

T3. **Editing comments.** The subject entered or modified at least 4 comments about places or edges. *We asked* questions analogous to Task 2. Additionally, when any of the last 17 subjects read an existing comment, we asked them if it was useful.

[1] We conducted a 30th interview, but it failed. Specifically, communication between subject and experimenter was highly ineffective, and only a small portion of the interview plan was completed. We believe that this interview contained little or no useful data and was unrepresentative, so we exclude it.

T4. **Rating bikeability.** The subject rated the bikeability of 12 or more edges on a 5-star scale. *We asked* questions analogous to Task 2.

Points and comments edited by one subject were visible to subsequent subjects (and all users of Cyclopath, once the system went live), while routes and ratings were not. Keeping routes and ratings private is consistent with standard practice, as this information is considered personal. We concluded interviews by asking questions about privacy, the usefulness of several different information resources, and satisfaction with current information-gathering methods.

3.3.2 Results

Subjects had no difficulty entering the information they were asked to, often entered additional information, and reported that it was easy to think of this information. Furthermore, our content analysis reveals a rich diversity of entered information – "subcultural community" resources, personal experiences and advice, and detailed cycling-relevant information. Finally, this information is clearly useful to cyclists, and it is difficult to obtain except from cyclists.

Summary statistics. *Places:* Subjects entered a total of 129 new places. 28 of 29 subjects entered at least the 4 requested places, and nine entered at least 5. *Comments:* Subjects edited a total of 224 comments (71 on places, 153 on edges), with 32 edited by at least two subjects (19 on places, 13 on edges). All subjects edited at least the 5 requested comments, and seven edited 12 or more. *Ratings:* Subjects entered 828 bikeability ratings for edges. 26 subjects entered at least the 13 requested ratings, nine entered at least 23, and the top two entered 88 and 121 ratings.

Knowledge self-reports. A majority of subjects reported that it was easy to think of comments (25 of 29) and ratings (17); 10 subjects reported ease in thinking of new places. However, majorities in all three tasks thought that it would be easy if they were not under the pressure of an interview (16, 26, and

Category	**#**	**Examples**
Lifestyle/community	*53*	
Community resource	22	post office ; Como Zoo ; YWCA – uptown
Retail (non-food)	17	Arise Bookstore ; Target ; Chicago Lake Liquor
Park	14	Hidden Beach ; Como Zoo ; Community Garden
Arts	14	Orchestra Hall ; Soap Factory [an art gallery]
Food/drink	*47*	
Restaurant	23	Black Forest restaurant ; Longfellow Grille
Coffee shop	17	Espresso Royale
Groceries	8	Kowalski's ; Byerly's ; Penzie's Spices
Water fountain	2	water fountain
Cycling-specific	*30*	
Landmark (explicit)	13	Obnoxious Billboard ; Old Grain Belt Brewery
Road/trail	8	... Bridge under greenway to 37th Ave.
Cycling-specific (other)	7	franklin ave hill ; difficult intersection to ride
Meeting spot (explicit)	3	Critical Mass gathering area ; bike trick hang out
Other	*6*	

Table 3.1: Categorization of the 129 places entered by subjects.

19 subjects on places, comments, and ratings, respectively) and that they had more information they could share (24, 22, and 20).

3.3.2.1 Cyclists have knowledge

The core of our argument that cyclists have knowledge is a content analysis of the places (Task 2), comments on places (Task 3), and comments on edges (Task 3) entered by subjects. We coded these three groups of data separately, considering both the data themselves and relevant interview notes. Two coders independently defined categories for each group, then met to produce consensus categories. The same two coders then independently applied these categories, resolving disagreements by discussion. In general, categories were non-exclusive – i.e., one item can be in multiple categories at each level – except for the *Other* categories. Therefore, membership counts cannot be summed.

Places. Table 3.1 on the facing page summarizes our content analysis of places. We found three major categories. *Lifestyle/community* places relate to everyday urban life, e.g. post offices, zoos, bookstores, parks, and art galleries. *Food/drink* places are where these two items can be obtained. *Cycling-specific* places are directly related to the practice of bicycling, including landmarks, shortcuts, meeting places, and big hills. A surprising observation is that few places are cycling-specific: this category accounts for only 23% of the total. We first discuss these places, then the other two categories.

The largest sub-category of cycling-specific places is *landmarks*. In principle, nearly any place could serve as a landmark; however, we categorized a place as such only if a subject described it that way explicitly (e.g., to "key off of") or if this use was clear from their interview comments. *Meeting spots* are analogous. *Road/trail* information would ideally have been attached to a edge; that is, if subjects had understood the system fully, these would have been edge comments instead of place comments. Some of these cycling-specific places strikingly illustrate our hypothesis that the knowledge cyclists have is difficult to obtain; e.g., that cyclists gather at a particular place to do bike tricks, or that a particular lighthouse is a critical landmark on certain routes. Cycling-specific places can play an important role in route finding; e.g., relevant ones – perhaps chosen in a personalized way – could be added to a route description to aid orientation and make it easier to identify turns.

Turning our attention to non-cycling-specific places (76% of the total), we assume that subjects followed task instructions, believing that fellow cyclists would find these places interesting even though they weren't about cycling per se. Why would they think this? We conjecture that cyclists know their peers well enough to know what else – beyond cycling – they tend to like. The places they entered form a cultural snapshot of the local bicycling community; i.e., they mark information of interest *to cyclists*, but not necessarily of interest *while cycling*. This result, consistent with [90], suggests that information resources for cyclists, and perhaps for many other groups, should support off-topic conversation as a useful means to build community, rather than suppressing it as undesirable noise.

Category	**#**	**Examples**
Objective place information	*60*	
Description	57	collectice [sic] bookstore with activist literature
Events	6	Starting point for Messenger Challenge ...
Subjective place information	*47*	
Review	44	Wonderful Asian food ; Best priced bike repairs
Advice	6	Lock your bike here ...
Personal narrative	3	Every time I ride this street, I see ...
Cycling-specific	*23*	
Bike parking	9	Plenty of bike parking in front.
Bikeability notes	8	useful exit/entrance off/onto the Greenway
What cyclists do at place	7	Lock your bike here ... ; Starting point for ...
Other	*3*	

Table 3.2: Categorization of the 71 place comments edited by subjects.

Place comments. The tendency to enter non-cycling-specific information continued for place comments, although not as strongly; Table 3.2 summarizes these comments. The most popular categories were *objective place information* – essentially factual descriptions of places – and *subjective place information*, typically comprising brief free-form reviews. 92% of place comments fell into one of these two categories, while 32% contained cycling-specific information. (Note that there was considerably more overlap in categories for comments than for places.)

However, here the critical distinction is between objective and subjective information. This shows that geowikis must record *opinion* as well as fact. Some wikis already provide for non-factual discussion; for example, each Wikipedia article comes with an associated "talk" page where users can discuss the article's content. However, these talk pages are for discussion of what the facts are and whether Wikipedia conventions are being followed – *not* the expression of subjective opinions about the subject matter. Our data suggest that objective fact and subjective opinion both deserve a first-class role. Therefore, we believe an organization similar to that of product reviews on e-commerce sites is more appropriate: both fact (features, price, etc.) and user opinion are

Category	#	**Example(s)**
Description	*108*	
Lane/facility type	61	pretty wide ; Bike lane on right side
Description (other)	40	Several road crossings with 4 way stop signs
Surface quality	22	smooth pavement ; a lot of potholes
Snow removal	10	Plow conditions are not that great here
Current conditions	6	not currently plowed well (12/10/07)
Worries/annoyances	*100*	
Motor traffic (quantity)	48	heavy traffic ; Quiet
Worries/annoyances (other)	39	many stop signs ; sketchy at night (crime)
Motor traffic (behavior)	20	can be dangerous because of turning cars
Hazards	19	possible broken glass hazards
Construction	9	Construction is ongoing
Subjective information	*41*	
Advice on route choice	35	great place to enter the greenway ...
Personal narrative	5	Just recently discovered
Scenery	3	... woods and hills that surround the trail
Other	*6*	

Table 3.3: Categorization of the 153 edge comments edited by subjects.

present, and clearly distinguished. This result also shows the desirability of letting users rate places.

Edge Comments. Comments on edges are summarized in Table 3.3. The major categories are *descriptions*, including basic properties of a edge like its width and surface type; *worries/annoyances*, e.g. quantity and quality of motor traffic or construction; and *subjective information*. In contrast to places and place comments, virtually all edge comments relate directly to cycling. We conjecture that this is because when subjects focused on where they actually cycled – the paths and roadways – they naturally thought of information useful *while cycling*. Edge comments are useful in both evaluating and following routes; e.g., a comment that a edge has many potholes could lead a cyclist to choose a different route or to ride that route with better preparedness.

Intuitively, the type of information revealed by our content analysis seems

useful: cyclists "obviously" want to know the places they can go, what those places are like, and what to expect on their way there. We next present our data, which support this intuition.

3.3.2.2 This knowledge is useful

Subjects found information from other cyclists useful, both when asked about specific entered information and when asked about the general utility of such information.

Other cyclists' comments were useful. When a subject read a comment edited by another subject, we asked if the comment was useful (for the first 6 comments read per subject). 50 of 64 comments read (70%) were judged useful.

Other cyclists were considered useful. We asked subjects about the utility of various sources of cycling information. When queried after each task about current information sources, subjects frequently mentioned other cyclists, giving an average utility rating of between 3.5 and 3.9 out of 5, depending on the task. Also, in the final set of questions, subjects rated the utility of other cyclists' bikeability opinions as a mean of 4.1. This is slightly higher than (though not statistically different from) their rating of "objective bikeability factors" at 4.0. We suspect that these figures actually understate the utility of other cyclists' opinions; two thirds of subjects said that individual cyclists differ in their bikeability assessments. Therefore, we conjecture that a system which computes *personalized* routes using only the ratings of like-minded cyclists will lead to other cyclists' opinions having a higher perceived value – one of the motivations for the personalized routing work discussed in Chapter 6.

Social connections may mean better knowledge access. Many types of information flow through social networks; our results hint that this is true for cycling-related information as well. The concluding survey asked subjects to respond to the following two statements on a 5-point Likert scale, from

Source	**Task 1**		**Task 2**		**Task 3**		**Task 4**	
Word of mouth	18	3.5	23	3.8	23	3.9	19	3.8
Trial & error	24	4.2	8	4.3	18	4.3	14	4.6
Internet forums	7	3.9	6	3.4	14	3.3	11	3.2
Paper maps	15	3.8	5	4.0	2	2.5	15	3.3
Online maps	17	3.6	3	4.3				
Internet search			11	4.2	4	3.5		
Specific websites			6	3.7	4	2.3		
Newspapers			4	2.9	2	3.8		
Phone book			6	2.3				
Advertisements			3	2.3	1	2.0		
Other	5	2.8	4	3.1	3	3.0	3	2.7

Table 3.4: Information sources currently used by subjects for finding routes (Task 1), places (Task 2), properties of places and edges (Task 3), and bikeability (Task 4). We show the number of subjects who mentioned each source and the mean usefulness for each source on a 5-point scale.

"strongly disagree" to "strongly agree": (1) *It is hard to find other cyclists who have the specific information I need* and (2) *I am satisfied with my existing methods of gathering information for planning rides and selecting routes.* The correlation between responses to these two items was -0.40. In other words, having weak links with other cyclists may make it harder to find cycling knowledge. This argues for a geowiki, which collects and distributes knowledge and supports social ties.

3.3.2.3 This knowledge is available only from cyclists

After each task, we asked subjects how they currently obtained each type of information they entered into the system and how useful their methods were. Table 3.4 details the responses. We highlight two key findings:

- "Trial and error" was most useful for every task (4.2 to 4.6 out of 5), and was mentioned second most often. In other words, subjects told us the best way to get useful knowledge was to go out find it themselves. While effective, this is time-consuming and particularly unhelpful when

answers are required immediately.

- "Word of mouth" was mentioned most often and was rated as quite useful (3.5 to 3.9). That is, cyclists try to benefit from each other's experience whenever they can. While not as useful as personal experience, this method is less time-consuming and has a wider reach. Also, as we suggest, a system that matches cyclists with similar opinions is likely to improve the utility of word of mouth.

The results of this section strongly support the utility of a geowiki for cyclists. Cyclists have much knowledge to enter, they judge the knowledge contributed by others to be useful, and this knowledge is not readily available elsewhere.[2]

3.3.3 Real-world usage

Usage data from Cyclopath further support the idea that cyclists have knowledge and it is useful. Table 3.5 on the facing page presents summary statistics of Cyclopath contributions as of July 2010. These provide evidence that the geowiki model is useful for cyclists in two ways: the contribution statistics (production) show that cyclists have and are willing to share information, and the usage statistics (consumption) show that cyclists find this information useful.

Additionally, the results of a survey conducted by Katherine Panciera in March 2010 provide further evidence that cyclists find Cyclopath useful. Of the 303 people who responded to the question *How do you feel you have benefitted from using and/or contributing to Cyclopath?*, 89% responded indicating that they had benefitted, making comments such as "I can make better choices about my routes", "I feel I have reliable local info", and "it has helped me find bike routes that I wouldn't have otherwise found".

[2]This experiment took place before Google released its bicycle routing mode for its Maps product in March of 2010. We believe that had the system been available at the time of the experiment, it would have mentioned frequently, though its utility ratings are harder to predict. Google Bike vs. Cyclopath is discussed in more detail in Section 2.2.5 on page 13.

Activity	Count
places added	2,466
notes added	2,144
note applications	6,626
tags added	265
tag applications	2,091
revisions	12,136
bikeability ratings	70,133
registered users	2,105
logged-in users per day	*150*
total routes computed	64,527
routes computed per day	*175*

Table 3.5: User contributions and usage activity in Cyclopath. Data for tags and tag applications excludes the five tags we automatically applied, and per-day data (italic) is for peak season and is approximate.

3.3.4 Discussion

In this section, we established that geographic wikis are a plausibly useful information resource for a representative community, bicyclists. First, cyclists have information and are willing to share it; both in a lab study and using Cyclopath system, cyclists contributed much information. Second, cyclists find the information shared by other cyclists useful. In a lab study, they told us this explicitly, and in the live system, this is evident in the system's robust usage and positive survey comments with respect to benefit. In particular, we found that two types of information often considered not valuable are in fact useful: off-topic conversation enhances the community surrounding a resource, and should thus be supported rather than shunned as noise, and opinions are useful, so wikis and other open content systems should allow users to contribute opinions as well as facts. Finally, this useful information does not appear to be available except from cyclists: subjects named no other information resources which were both useful and widely used. This point was reinforced when we actually built Cyclopath and encountered incompleteness and error in the best available seed data, a process detailed in Section 2.5.5 on page 35.

3.4 Anti-value: Types of wiki damage

The previous section explored the properties of "good" work in a wiki: contributions which were consistent with the system's goals. On the other hand, due to their highly open access policies, wikis are also vulnerable to "bad" work which is contrary to those goals. This section explores the properties of such damaging activity. We ask: *What types of damage occur, and how often?*

To understand the impact of different types of damage, it is meaningful to define different types of damage from the reader's perspective and provide estimates of how often each type of damage occurs. This section departs from Cyclopath, exploring damage in the context of Wikipedia. Cyclopath is not yet large enough to have a meaningful problem with damaging work (in fact, we have observed it on only a handful of occasions, and we have never observed apparently malicious activity of any kind); on the other hand, Wikipedia is extremely well-known and well-trafficked, thus attracting significant damaging activity. It therefore provides a "worst-case" scenario useful in considering the development of smaller open content systems

While most discussion of these issues focuses on *vandalism*, we use the more general concept of *damage*. This is for two reasons: classifying damaging work as vandalism requires discerning malicious intent, which can be prohibitively difficult in an Internet-based system, and the effect on users is the same: damage is damage, regardless of whether it was created maliciously or innocently.

This section is a data analysis study. We present a categorization of damage types, empirically classify damage incidents into those categories, and explore the implications of damage in each category.

3.4.1 Method

Based on observations made while judging Wikipedia revisions as damaged or not damaged (results explored in detail below in Section 4.5 on page 73), we developed a set of features exhibited by Wikipedia damage, aiming for comprehensiveness. These features, with comparisons to the anecdotal categories

of Viégas et al. [108], are as follows:

- **Misinformation.** Information which is false, such as changed dates, inappropriate insertion of "not", or stating incorrectly that a public figure is dead. (No analogue in Viégas.)
- **Mass delete.** Removal of all or nearly all of an article's content. (Same as Viégas.)
- **Partial delete.** Removal of some of an article's content, from a few sentences to many paragraphs. (No analogue in Viégas.)
- **Offensive.** Text offensive to many users, such as obscenities, hate speech, attacks on public figures, etc. This is a broad category, ranging e.g. from simple "you suck" to unexpected pornography. (Includes Viégas' *offensive copy.*)
- **Spam.** Advertisements or non-useful links. (No analogue in Viégas.)
- **Nonsense.** Text that is meaningless to the reader, for example "Kilroy was here", characters that do not form words, obviously unrelated text, and technical markup leaking into formatted pages. (Includes Viégas' *phony copy.*)
- **Other**: Damage not covered by the other six types.

Viégas' *phony redirection* is not included above because we observed only one instance (and it was better described as Offensive), and we believe that *idiosyncratic copy* ("text that is clearly one-sided, not of general interest, or inflammatory") better describes disputed content, not damage.

As does Cyclopath, Wikipedia supports the notion of *revert*: when an article's text is returned to an earlier state, the effect of the intervening revisions is removed, and those revisions are said to be reverted. We took a random sample of 493 sequences of reverted revisions; of those, 308 were judged as damaged in the analysis for Section 4.5, and these 308 revision sequences form the basis of this section's analysis. We analyzed revision sequences rather than

Feature	%	Agreement			Reliability	
		3v0	**2v1**	**1v2**	**PF**	**Ja**
Nonsense	53	108	56	70	0.66	0.46
Offensive	28	57	30	29	0.66	0.49
Misinformation	20	28	34	64	0.45	0.22
Partial Delete	14	35	7	20	0.83	0.56
Spam	9	25	3	6	0.89	0.74
Mass Delete	9	23	5	3	0.82	0.74
Other	5	1	15	21	0.06	0.27

Table 3.6: Distribution of damage features. *%* is the percentage of damage sequences where the feature applies (determined by majority vote), while the *Agreement* columns list how many times all (*3v0*), two of the three (*2v1*), and only one of the judges (*1v2*) believed the feature applied. (Percentages do not sum to 100 because features are not mutually exclusive.) *PF* (*proportion full*) gives the proportion assigned unanimously (i.e. $PF = 3v0/(3v0 + 2v1)$), while *Ja* gives the Jacquard statistic: the number of times all judges assigned the feature divided by the number of times any assigned the feature, i.e. $Ja = 3v0/(3v0 + 2v1 + 1v2)$.

revisions because we observed that revision sequences that were later reverted as damage generally formed a coherent single incident.

After calibration on a different sample of damaged edit sequences, three judges independently classified the sequences, applying as many of the damage features as were appropriate. We used a "majority vote" procedure, i.e., our analysis applies a feature to a revision sequence if at least two of the three judges believed it applicable.

3.4.2 Results

Table 3.6 summarizes our results. It is not surprising that agreement was highest for Spam, Mass Delete, and Partial Delete, since these features do not require much judgement. On the other hand, what's offensive or nonsense is somewhat subjective, and misinformation can be subtle. Finally, the low number of damage sequences labeled Other indicate that our categories are relatively comprehensive.

3.4.3 Discussion

From the perspective of an open content system, all damage is serious because it affects the credibility of its content. However, there are specific factors that we can use to assess more precisely the implications of our results. First, how *common* is a given type of damage? If a particular type is infrequent, we need not worry as much. Second, what is the potential *impact* on readers? If there is little harm, we need not worry as much even if occurrence is frequent. Finally, how easy is it to *detect* automatically? Even if damage is not automatically repaired, automatic notification of human editors can speed repair.

With this in mind, Mass Delete and Nonsense are low-impact types of damage. The former is relatively uncommon and trivial to detect automatically. The latter, while common, damages only presentation, not content, except in cases where its sheer bulk overwhelms content. For example, one incident consisted of the insertion of thousands of repetitions of a string of Korean characters into the article "Japan". (Interestingly, the characters formed hate speech, but we classified the incident as Nonsense because few readers of the English Wikipedia understand Korean.) Spam and Partial Delete are somewhat higher impact, because they are tricky to detect automatically (useful edits introduce links and remove text all the time); also, Spam wastes readers' time and Partial Delete may cause the omission of important information.

Offensive damage is troublesome because it is common (28% of incidents) and potentially highly impactful – offensive content damages the reputation of a system and drives away readers. Automatic detection of offensive damage is plausible in some cases (e.g., detecting obscenities) but harder in the general case due to the complexity of offensive speech and the difficulty of analyzing images automatically.

Misinformation may be the most pernicious form of damage. It is both common (20% of incidents) and difficult to detect. Automatic detection is essentially impossible because it requires understanding the content of the page, and people who visit a page are typically there to learn about its topic, not because they understand it well. An intriguing and subtle example is that of the "Uchi-soto" article, which discusses a specific facet of Japanese

language and social custom. A (presumably well-meaning) editor changed the translation of the word *uchi* from *inside* to *house* – both are correct, but *inside* is the one appropriate for this article. This error could only be detected by a reader with sophisticated knowledge of Japanese.

Finally, computational geowikis like Cyclopath raise an interesting possibility for automatic detection of damage: because user work feeds an algorithm, the effects of particular revisions on the results of the algorithm can be measured, and revisions with notably large effect flagged (whether that effect is positive or negative, as "too good to be true" may also indicate damage). This notion of value measurement is explored in detail in Section 4.3 below.

3.5 Summary

In this chapter, we show that the geowiki forms a powerful knowledge sharing tool in a representative domain, bicycling. Using our utility framework, we demonstrated that cyclists have and share information (they shared much knowledge both in the laboratory in the live system) that is useful (they tell us that it is useful, and they come frequently to Cyclopath to get it) and not available otherwise (cyclists could tell us of no other appropriate information resources). These results show that geowikis are more effective than previous information gathering and exchange tools, and they support the general promise of geowiki technology in the growing domain of geographic content.

Additionally, we presented a comprehensive enumeration of damaging work types in wikis, of both high and low concern, validated using quantitative coding of Wikipedia damage incidents. These results are again of general utility, as open content systems vulnerable to such abuse are multiplying rapidly.

In the following chapter, we continue this exploration to quantify the value of wiki and geowiki contributions and the impact of damaging edits.

Chapter 4

Measuring the Value of Contributions

4.1 Introduction

The previous chapter established that geowiki contributions are valuable; users say that contributions they see are valuable, and they come actively to Cyclopath to consume them. However, we can and should go beyond that, and *measure* the value of contributions. This lets us both quantify the impact of work and see whose work is the most impactful.

We measure value from the perspective of information consumers. Why? It is certainly easier to count the actions of producers, as they leave a much richer trail of activity than consumers. But in the real world, there are no points for effort: the value of an information resource is the value of its information to the people who use it – consumers. Even though it is more difficult, it is essential to measure value from a consumer perspective. This chapter introduces several ways to bridge producer actions (which are easy to count) to consumer value (which is what we really want). In particular, this work is the first to measure value from the consumer perspective in a wiki context, and we introduce a particularly powerful way to measure the value of wiki work: feed it into an algorithm, and measure the effect on that algorithm's results.

In other words, this chapter is about metrics, which we propose in several

contexts. In systems like Cyclopath, where user input feeds an algorithm, we propose that impact on algorithm output is a key metric, and we present the details of such a metric in Cyclopath, showing that user work has reduced the length of computed routes by an average of 1 kilometer. In other systems, there is no such algorithm; in these cases, we propose a metric based on readership, achieving the same end goal – the effect of user contributions on consumers of the resource. We discovered that measured by consumer value, contributions in Wikipedia are even more unequal than is usually encountered in social systems: just 0.1% of users contribute 44% of the value. Finally, because wikis are vulnerable to malicious or erroneous contributions, we propose a metric to estimate the impact of damaging activity, focusing again on the effects on consumers; using this metric, we estimate that 0.7% of views in Wikipedia are of articles with some kind of damage.

We begin by exploring related work, and then we present three experiments, exploring value in Cyclopath, who creates the value in Wikipedia, and the impact of damage in Wikipedia. We close with a summary.

4.2 Related work

The questions we address – how to measure value, and who contributes it – have been widely researched in different social systems. For example, Sen et al. created techniques for identifying high-quality tags based on interaction patterns and explicit ratings [99], Harper et al. showed that an open access policy where anyone can contribute answers leads to higher-quality answers in Q&A sites [54], and Hawn has shown that social media such as blogs and social networks can improve health care [56].

One core result is the highly unequal nature of participation in these communities. When users' participation levels are visualized, the graph nearly always looks exhibits a *long tail* or looks like a "hockey stick". More rigorously, it often follows a power law [1]. Such relationships arise in many different kinds of online communities, including Usenet discussions [112], tagging [47], and blog links [79].

Researchers have also addressed the question of measuring value in the context of Wikipedia; for example, Adler and Alfaro developed a scheme to compute reputation of authors based on whether their edits persisted or were removed and used this metric to estimate the credibility of each section of an article [2], and Stvilia et al. proposed several metrics to compute articles' quality based on article properties such as frequency of reverts and number of links [102]. Viégas et al. addressed the persistence of vandalism, measuring that certain types of vandalism were repaired in an average time of 2.8 minutes [108].

The issue of which editors contribute Wikipedia's content has been a matter of some dispute. Jimmy Wales, one of the founders of Wikipedia, has stated that "2% of the users do 75% of the work" [110], while Swartz has argued that the work is more distributed [103]. Voss provided data on this question by counting the number of edits; unsurprisingly, the data showed a power law distribution [109].

Kittur et al. [66] analyzed the amount of content contributed by different classes of editors, finding that elite users (10,000 or more edits) accounted for over 50% of edits in 2002 but only 20% by mid-2006, due to increasing participation by users with less than 100 edits. Furthermore, Kittur and his colleagues explored whether elite and low-edit groups accounted for different amounts of content. By measuring the total number of words added and deleted in edits, they found that elite editors accounted for a higher proportion of content changes (around 30%) and were more likely than low-edit users to add (rather than delete) words.

Our work forms a significant advance in that our metrics measure the value of wiki content from the perspective of content *consumers*, either by measuring how much it affects core algorithms (route-finding in Cyclopath) or by estimating how much the content is viewed (words in Wikipedia). There is no real value in content that is never consumed, even if there is a lot of it; conversely, content that is consumed frequently or with large effect has high value, regardless of how much of it there is. Thus, our metrics match the notion of the value of content better than previous metrics.

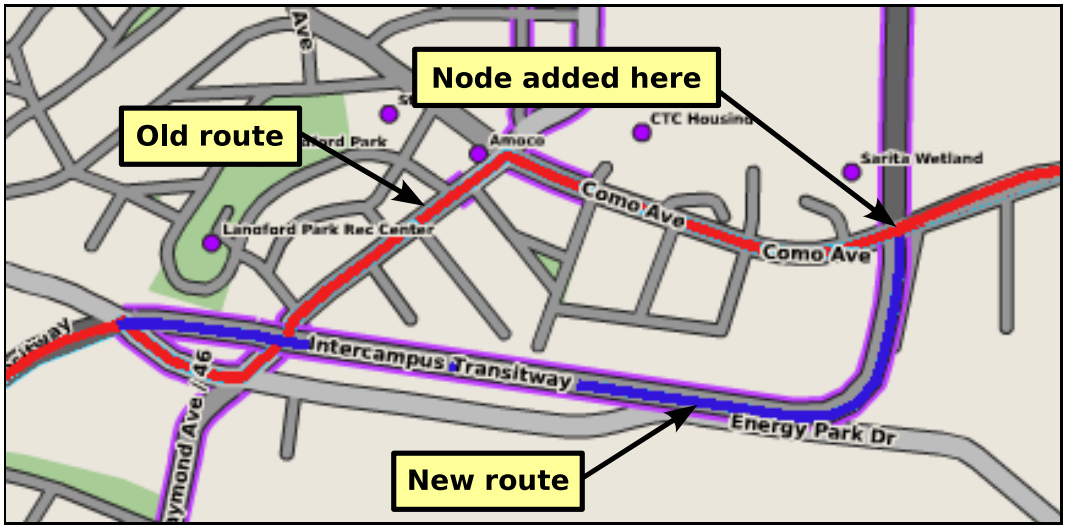

Figure 4.1: Excerpt of a route improved from 15.6 km to 15.0 km due to user work; a node was added to the network at the indicated location.

4.3 Measuring impact of production

Cyclopath's primary service is generating bicycle routes. Contributions made by users of the geowiki affect the routes that are generated; i.e., user contributions influence the result of a computation – routes generated for other users. For example, Figure 4.1 shows a route shortened by 600 meters due to user work; on the other hand, a bridge over a highway erroneously marked for one-way travel added 1.9 km to a route that depended on the link. Occasionally, these effects are quite dramatic. For example, one route's length decreased from 54 to 22 kilometers during Cyclopath's first year of existence! 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 40 km stretch of that river. Users added these bridges and connected them to the rest of the transportation network.

This link between user work and computational results leads to a quantitative technique for measuring that user work: we can compute routes, add user work, and then recompute routes, measuring how they were changed. This is the experiment discussed in this section. It turns out that user work in Cyclopath has shortened the average route by 1 km.

4.3.1 Method

We randomly selected 800 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 computed using the map's state at two different instants in Cyclopath history: (a) system initialization in May 2008, after the Cyclopath base maps were loaded but before any user work, and (b) after one year's time and significant user work – 43,385 bikeability ratings and 43,888 edge additions, edits, and deletions.

For both, 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.[1]

4.3.2 Results

Table 4.1 on the next page summarizes the effect of user input on routes. The central finding is that user input improved the quality of routes – i.e., they were shortened (t-test, 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!

User input had little effect on many routes, while greatly improving some and worsening (fewer) others. The first quartile (i.e., the 200th-least-improved

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

Time	**Distance (kilometers)**						**∅**
	mean	**min**	**Q1**	**Q2**	**Q3**	**max**	
initially	14.8	0.3	6.5	11.7	19.6	74.0	32
after user work	13.8	0.3	6.5	11.4	18.5	63.8	10
improvement	1.00	−3.9	−0.01	0.03	0.39	32.4	22

Table 4.1: Summary statistics of sample routes at the two analysis instants with improvements due to the editing in the interval (positive means improvement). We also report the number of route requests for which no route could be obtained (∅). Note that statistics for the *improvement* line refer to the improvements themselves, rather than the difference between the figures for the two analysis instants. For example, while the shortest route remained 0.3 km at both, the minimum improvement was -3.9 km – i.e., one route was made 3.9 km longer by user editing.

route) was a reduction of about 9 meters, while the third quartile improvement was a reduction of 387 meters.

Also, our figures underestimate the benefits of user input. 38 of the 800 route requests could not be satisfied at one or the other of the two analysis instants, i.e., the destination was not reachable from the start via the transportation network. Because the length at one or more instants was effectively infinite, we excluded these routes from our analysis. Of the 32 route requests unsatisfiable prior to user input, 28 could be satisfied afterwards. 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).

4.3.3 Discussion

This section reports two significant findings. First, if user work in an open content system feeds a computation, that computation can be leveraged to create a powerful impact metric. Second, we found that geographic volunteer work can have significant benefits. In our case, the computation is route finding, and we measured a strong effect on its result: user work improved the average route by 1 km, or about 7%.

These two observations lead to a potential useful application: *intelligent edit monitoring*. An aspect of open content systems that has received much

attention (both scholarly and otherwise) is 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 need no scrutiny, and some of the few edits which do need scrutiny do not 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 useful in any open content system where user edits are input to a computation.

4.4 Who creates the value?

In addition to measuring the impact of user contributions, it is useful to measure the value of each user's work and see which users are the most impactful. This section, like Section 3.4 above, explores this question in the context of Wikipedia, because Wikipedia is a mature system with a well-developed community, and our results in this context can provide guidance for users of smaller systems as they grow. Thus, the question addressed by this section is: *Who contributes Wikipedia's value?* Is it the handful of people who edit thousands of times, or is it the thousands of people who edit a handful of times?

Like Wikipedia, many open content systems do not feed user input into an algorithm which can be leveraged to build a value metric, but it is still important to estimate value from a consumer's perspective rather than simply counting producer actions. This section builds on prior work by developing

a new approach to estimating the value of Wikipedia, based on how many people view a particular change to an article (and thus are affected by it). Specifically, this work was the first to estimate the *value* of edits in terms of how many user views they receive.

Estimating this is hard, so we first detail how it is done. We then present our analysis method, followed by our results and their implications.

4.4.1 Estimating article views

We define a few critical terms as follows. The *act* of making and saving changes to an article is an **edit**, and the history of an article forms a sequence of content states called *revisions* – i.e., edits are transitions between revisions. Further, there is a special kind of edit, called a *revert*: reverting an article means restoring its content to some previous revision, removing the effects of intervening edits.

4.4.1.1 Why measure views?

We measure the value of contributions or the impact of damage in terms of number of **views**. Importantly, this information is required for every point in time during the period of Wikipedia's history under study. It's not enough to know how many views an article receives "now". Instead, we need to know how many views it received (for example) between 9:17 and 9:52 on July 8, 2005 – perhaps because that is the period from one revision to the next, and we need to give "credit" for the changes made in that revision.

There is a reason prior analyses haven't used view data: it was not available, because the relevant logs did not exist. However, we have access to several datasets that let us estimate view data.

A word of caution. We assume that one serving of an article by a Wikipedia server is a reasonable proxy for one view of that article by a user. While a human may request an article but not read all of it, or web caching schemes may cause one Wikipedia serving of an article to correspond to many user views of that article, we believe that these factors do not materially affect the

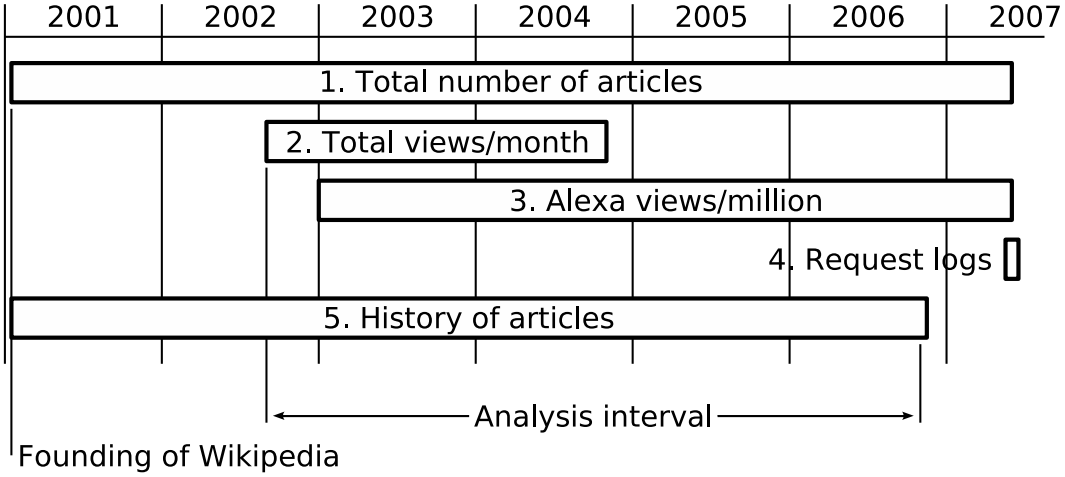

Figure 4.2: Wikipedia datasets used in this thesis. We analyzed the period September 1, 2002 to October 31, 2006.

accuracy of our view estimates, in particular since we are most interested in comparisons between articles.

4.4.1.2 Datasets

We use five datasets in this work, illustrated in Figure 4.2:

1. *Total number of articles* in Wikipedia over time [120] is provided by Wikipedia itself.

2. *Total views per month* is also provided by Wikipedia [115], but only for the period August 2002 to October 2004.

3. *Alexa views per million* over time is compiled by Wikipedia [117] from data recorded by Alexa based on use of its Alexa Toolbar browser plugin, a "search and navigation companion" [4]. These data measure, of each million page requests made by users with the plugin installed, how many requests were for pages under wikipedia.org, including all languages. Users self-select to install the plugin, but Alexa data are among the best available estimates of Web page views.

4. *Request logs* from Wikipedia server activity. The Wikimedia Foundation granted us access to a sampled log of all requests served by its web servers and internal caching system. This includes all Wikipedia languages plus several other wiki projects; our current analyses make use only of the English language Wikipedia data. The logs contain the timestamp and URL of every 10th HTTP request.[2]

   Our analyses consider log data between April 12 and May 11, 2007. During an average day in this time period, Wikimedia served 100 million English Wikipedia article requests (and a total of 1.7 billion HTTP requests). Even these 10% sampled logs are huge, comprising 10-15 GB of data per day.[3]

5. *History of articles.* Wikipedia provides a historical archive of its content, i.e. the text of all articles with complete edit history. We analyzed the 1.2 TB archive containing changes through the end of October 2006.

A tempting proxy for article views is article edits. However, we found essentially no correlation between views and edits in the request logs. Therefore, we must turn to a more elaborate estimation procedure.

4.4.1.3 Computing article views

We need to compute the number of article views during a specific interval – how many times was article X viewed during the interval t_1 to t_2? We do this by computing the article's *view rate* – i.e., how many views per day did article X have at time t – from which it is straightforward to compute views. Specifically, we compute:

$$r(X,t) = r(X,\text{now}) \times \frac{R(t)}{R(\text{now})} \times \frac{Z(\text{now})}{Z(t)}$$

[2] 1.5% of the log data were lost between Wikimedia and us due to dropped packets and interruptions of our collector process. Our estimates compensate for this small loss.

[3] This log collection process continued to the present and is ongoing. As of July 2010, we receive roughly 40 GB of log data from Wikimedia daily and have accumulated 38 TB in total. We believe that this complete picture of Wikimedia-operated wiki usage is a research resource unavailable anywhere else.

where $\boldsymbol{r(X,t)}$ is the view rate of article X at time t, $\boldsymbol{R(t)}$ is the view rate of Wikipedia as a whole at time t (i.e., $R(t) = \sum_a r(a,t)$ for each article a), and $\boldsymbol{Z(t)}$ is the number of articles in Wikipedia at time t.

Intuitively, to compute the historical view rate of an article, we take its current view rate, *shrink* it to compensate for shrinkage of Wikipedia's view rate as a whole, and then *expand* it to compensate for the smaller number of articles. This computation is based on the assumption that the view rates of articles, relative to each other, are fairly stable.

The values of the five terms in the above formula are themselves computed as follows. $\boldsymbol{r(X,\text{now})}$ and $\boldsymbol{R(\text{now})}$ are average values from the request logs (data set 4). "Now" is the center of the request log data period, i.e. April 24, 2007, while $\boldsymbol{Z(t)}$ and $\boldsymbol{Z(\text{now})}$ are interpolated from the total number of articles (data set 1).

$\boldsymbol{R(t)}$ is the most complex to compute. If t falls within the period covered by the total views-per-month data from Wikipedia (data set 2), it is interpolated from those data. If not, we interpolate it from *scaled* Alexa data (data set 3). Intuitively, what we want is a scaling of the Alexa data so that it matches well both the old views-per-month data and the new $R(\text{now})$ computed from request logs. This is done by taking the linear regressions of the log of the Alexa data and the log of the views-per-month data during the period of overlap, then scaling up the Alexa data until the linear regressions intersect at the center of this period. This also results in a close match between the scaled Alexa data and $R(\text{now})$: 97 and 104 million views per day, respectively. This scaled Alexa data is what we use to interpolate $R(t)$.[4]

If an article had aliases at time t, $r(a,t)$ is computed for each alias a as well, and the final view rate is the sum of the view rates of the article and each of its aliases. Historical view rates for articles or aliases which no longer exist will be zero, but we believe this does not materially affect our results because few articles and aliases are deleted, and those that are are obscure.

For example, suppose that we wish to calculate the number of views of

[4]The time period between Wikipedia's founding and the beginning of the views-per-month data is excluded from analysis in the present work.

article Y, which has no aliases, from June 9 to June 14, 2004. Suppose that $r(Y, \text{now})$ is 1,000 views per day, and recall that $R(\text{now})$ is 104 million views per day. We first compute the views per day of article Y on June 12, the center of the period.

1. $Z(\text{now}) = 1{,}752{,}524$, interpolated from the values for April 16 (1,740,243) and May 1, 2007 (1,763,270).

2. $Z(\langle \text{June 12, 2004} \rangle) = 284{,}240$, interpolated from the values for May 13 (264,854) and July 10 (302,333).

3. $R(\langle \text{June 12, 2004} \rangle) = 5{,}019{,}355$, interpolated from the Wikipedia views-per-month data of 2,400,000 views per day on May 15 and 5,300,000 on June 15.

Applying the formula, $r(Y, \langle \text{August 12, 2004} \rangle) = 298$. Because the period is six days long, we estimate the number of views as six times this value, or 1,785 views. The calculation would proceed similarly for the period November 9 to November 14, 2004, except $R(\langle \text{November 12, 2004} \rangle)$ would be interpolated from scaled Alexa data, because the last Wikipedia views-per-month datum is for October 15, 2004.[5]

4.4.2 Method

4.4.2.1 Persistent word views

As a proxy for the encyclopedic value contributed by an edit, we use the **persistent word view** (PWV) – the number of times any given word introduced by an edit is viewed. PWV builds on the notion of an article view: each time an article is viewed, each of its words is also viewed. When a word written by editor X is viewed, he or she is credited with one PWV.

Two key insights drive this metric. First, authors who write content that is read often are empirically providing value to the community. Second, if a

[5]These computations are actually done in floating-point seconds, but we simplify the presentation here for clarity.

#	Editor	Article text
1	Carol	alpha bravo charlie delta
2	Denise	alpha alpha bravo delta charlie
3	Bob	alpha bravo charlie echo delta
4	Bob	alpha bravo echo foxtrot delta
5	Alice	alpha delta echo foxtrot

Figure 4.3: Example revision history.

contribution is viewed many times without being changed or deleted, it is likely to be a valuable. Of course, this metric is not perfect: the concept of value is dependent on the needs and mental state of the reader. One might imagine a single fact, expressed in only a few words, that provides enormous value to the one reader who really needs that fact. In a large pseudonymous reading community like Wikipedia, capturing a notion of value that depends on the specific information needs of the readers is outside the scope of this work.

For example, see Figure 4.3. Assuming that each page is viewed 100 times after each edit, Carol has accrued at least 1,200 PWVs: 400 from *bravo* (because she wrote it and it was present for 4 edits), 300 from *charlie*, and 500 from *delta* (even though it was moved several times).

The case of *alpha* is problematic because it is ambiguous. When Bob deleted an *alpha*, whose did he delete: Carol's or Denise's? Words carry no identifier, so it is impossible to tell for certain without understanding the text of the two edits. In these cases, we choose randomly. Carol could have 1,300 or 1,700 PWVs depending on whether or not her *alpha* was chosen. Amortized over the trillions of PWVs analyzed, these random choices have little effect on our results.

4.4.2.2 Calculating PWVs

PWVs are calculated per-article, and the final score for each editor is the sum of his or her scores over all articles. The "owner" of each PWV is determined by comparing the text of subsequent article edits, data contained in the history of articles (data set 5 above); specifically, we:

1. Remove punctuation (except hyphens) and wiki markup from the texts of the old and new edits.
2. Eliminate letter case.
3. Remove stopwords, because very common words carry little information. We use the same word list used by Wikipedia until it began using the Lucene full-text search engine [113].
4. Sort each list of words, because we analyze the appearance and disappearance of words, not their movement within an article.
5. Compare the new and old word sequences to determine which words have been added and which deleted.
6. The editor who made the new edit begins accruing PWV credit for added words, and editor(s) who wrote the deleted words stop accruing PWV credit for them.

Our software does not track persistent words if text is cut and pasted from one article to another. If an editor moves a block of text from one article to another, PWVs after the move will be credited to the moving editor, not to the original editors. This problem is challenging, because edits are per-article, making it difficult to detect where the text moved to, or even if it moved to only one place.

An editor can work either anonymously, causing edits to be associated with the IP address of his or her computer, or while logged in to a pseudonymous user account, causing edits to be associated with that pseudonym. We exclude anonymous editors from some analyses, because IPs are not stable: multiple edits by the same human might be recorded under different IPs, and multiple humans can share an IP.

4.4.2.3 Dealing with reverts

Editors who revert do not earn PWV credit for the words that they restore, because they are not adding value, only restoring it; rather, the editors whose words they restore regain credit for those words.

Reverts take two forms: *identity revert*, where the post-revert revision is identical to a previous revision, and *effective revert*, where the effects of prior edits are removed (perhaps only partially), but the new text is not identical to any prior revision. Identity reverts are unusually common, because Wikipedia includes a special mechanism through which any editor can easily revert a page to a previous edit, and because the official Wikipedia guide to resolving vandalism recommends using this mechanism. Kittur et al. [65] report that of identity reverts and effective reverts which could be identified by examining edit comments, 94% are identity reverts. There are probably other effective reverts, because some editors do not clearly label their edits, but detecting these is challenging because it requires understanding the *intent* of the editor. In this chapter, we consider only identity reverts.

4.4.3 Results

We analyzed 4.2 million editors and 58 million edits. The total number of persistent word views was 34 trillion, or, excluding anonymous editors, 25 trillion. 300 billion PWVs were due to edits before the start of our analysis and were excluded. 330 billion PWVs were due to bots – autonomous or semi-autonomous programs that edit Wikipedia.[6][7]

Figure 4.4 on the next page shows the relative PWV contributions of editors divided by edit count decile. From January 2003 to February 2004, the 10% of editors with the most edits contributed about 91% of the PWVs. Then, until February 2006, Wikipedia slowly became more egalitarian, but around February 2006, the top 10% re-stabilized at about 86% of PWVs. Growth of PWV share increases super-exponentially by edit count rank; in other words,

[6]We identified bots by taking the union of (a) editors with usernames ending in "bot", followed by an optional digit, that had at least 100 edits and (b) users listed on Wikipedia's list of approved bots [116].

[7]Some damage-reverting bots had a bug causing a few reverts to become non-identity reverts. Because our software could not detect these reverts, it treated the situations as removal of all text and replacement with entirely new text. Effectively, these bots "stole" PWVs from their rightful owners. Our measurements show that these bugs resulted in only about 0.5% of PWVs being stolen.

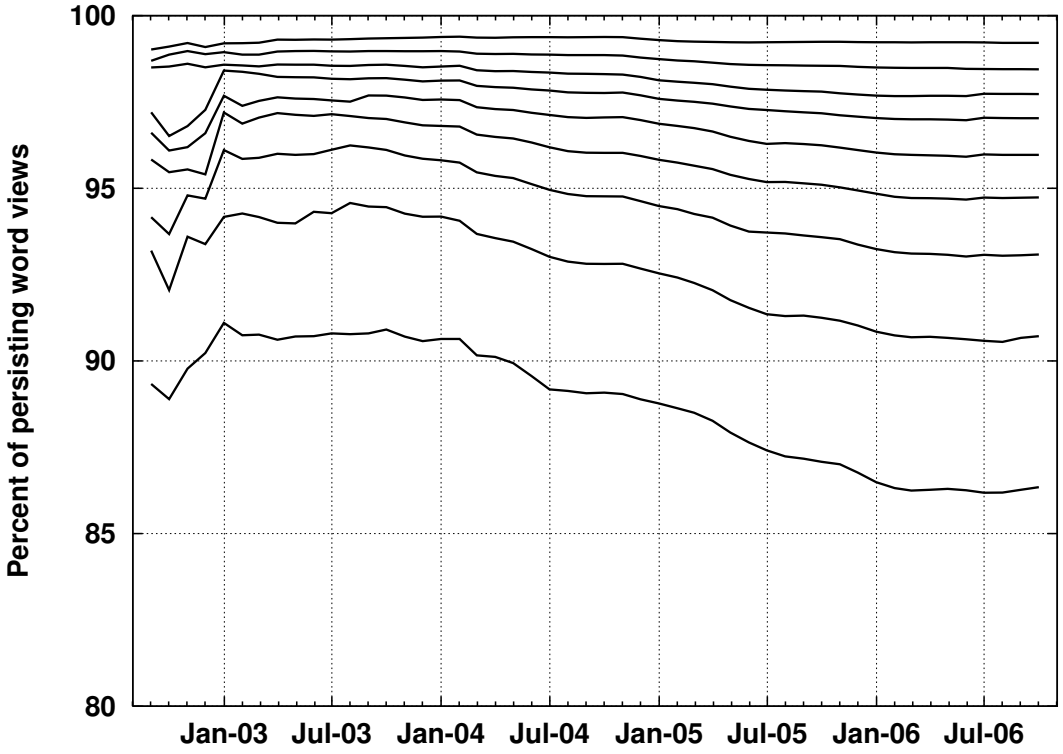

Figure 4.4: PWV contributions of editors. Percentage of PWVs according to the decile of the editor who contributed them.

elite editors (those who edit the most times) account for *even more value* than they would given a power-law relationship.

Figure 4.5 on the facing page zooms in; editors with the top 0.1% of edits (about 4,200 users) have contributed over 40% of Wikipedia's value. Collectively, the ten editors with the most PWVs contributed 2.6% of all the PWVs.

4.4.4 Discussion

Editors who edit many times dominate what people see when they visit Wikipedia. The top 10% of editors by number of edits contributed 86% of the PWVs, and top 0.1% contributed 44% – nearly half! The domination of these very top contributors is increasing over time.

Of the top 10 contributors of PWVs, nine had made well over 10,000 edits.

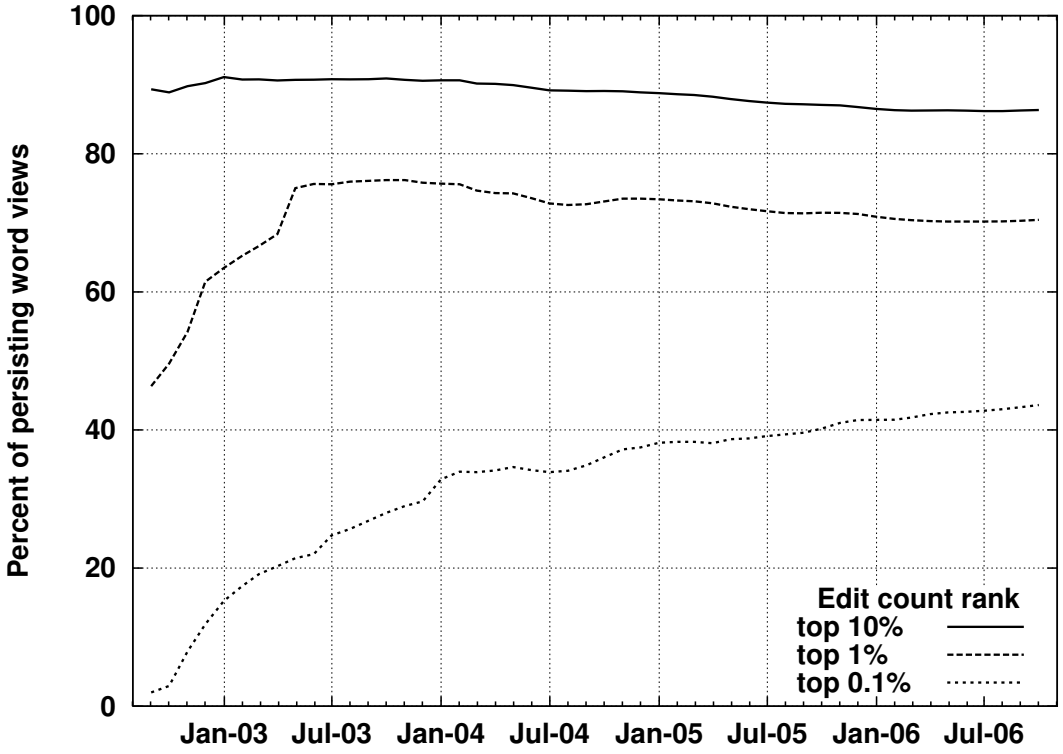

Figure 4.5: PWV contributions of elite editors.

However, only three of these users were also in the top 50 ranked by number of edits. The number one PWV contributor, Maveric149, contributed 0.5% of all PWVs, having edited 41,000 times on 18,000 articles. Among the top PWV contributors, WhisperToMe (#8) is highest ranked by number of edits: he is #13 on that list, having edited 74,000 times on 27,000 articles.

Exploring the list of top editors by edit count, we notice something interesting: the list is filled with bots. They occupy the top 4 slots, 9 of the top 10, and at least 20 of the top 50. One the other hand, the list of top editors by PWV is filled with humans: only 2 bots appear in the top 50, and none in the top 10. Thus, while bots edit frequently and have an important role (for example, as explored in the next section), humans dominate when measured by value added.

This work has set the scientific study of Wikipedia – and, by extension, study of other open content communities – on a firmer basis than previously

available. Most fundamentally, we offer a better way to measure the phenomena people care about. Others have used author-based measures, counting edits to approximate the value of contributions. We use reader-based measures, approximating value by estimating the number of times contributions were viewed.

These view-based metrics let us both sharpen previous results and go beyond them. Others have shown that 1% of Wikipedia editors contributed about half of edits [25]. We show that *1/10th of 1% of editors contributed nearly half of the value*, measured by words read.

What are the implications of our results? First, because a very small proportion of Wikipedia editors account for most of its value, it is important to keep them happy, for example by ensuring that they gain appropriate visibility and status. However, turnover is inevitable in any community. Wikipedia and other open content systems should also develop policies, tools, and user interfaces to bring in newcomers, teach them community norms, and help them become effective editors.

A similar analysis in Cyclopath would be straightforward. One possibility is an analysis using some metric analogous to the persistent word view. There are multiple options for the granularity of consumption actions, corresponding to *view* in our Wikipedia analysis. Cyclopath records the viewport whenever a user pans or zooms, so we know fairly specifically where a user is looking, and we could apply heuristics giving more "points" to objects closer to the center of the screen, yielding a more precise notion of view than is possible in Wikipedia. Also, clicks on objects are a signal of further interest, and the Cyclopath software could be instrumented to track mouse hovers or other similar signals. Another notion of consumption interest for edges is inclusion of routes – *one view* could become *included in one route*.

Similarly, there are multiple options for the granularity of what is contributed, corresponding to *word* in our Wikipedia analysis. In Wikipedia, articles can naturally be broken down into a bag of words. In Cyclopath, is it not so clear. One option is map objects: each time a map object is changed, credit for "views" of that object go to whoever changed it last. Other granularity

possibilities include the attribute, or, for text fields, the word.

Another possibility is to rely wholly on the value metric introduced in Section 4.3: whenever a revision is saved, recompute (perhaps a sample of) all saved routes. The impact of a revision is the total number of meters or bikeability points saved. These two analyses are complementary, since there is significant value in Cyclopath that does not directly affect routing (for example, a note warning of potholes).

4.5 Anti-value: Impact of damage

Wikis afford destructive work as well as constructive work. This is a common and very important criticism, but to what degree is this a real problem that affects consumers of the resource? To answer this question, it is necessary to go beyond simply counting damage incidents. We ask: *What is the impact of damage such as nonsensical, offensive, or false content?* How quickly is it repaired, and how much of it persists long enough to confuse, offend, or mislead consumers? Again, we ask this question in the context of Wikipedia, as it is a system large and mature enough to attract significant destructive activity, unlike Cyclopath.

This analysis builds on the analysis presented in Section 4.4 above, which reasoned what the impact of particular types of damage might be. Here, we *compute* a quantitative estimate of the impact of damage in Wikipedia, though we do not break this impact down along the specific classes enumerated above.

4.5.1 Method

Like the previous section, this section analyzes activity in Wikipedia after the fact. We use human judges to calibrate a heuristic which classifies revisions as damaged or not damaged, and combine that with our view estimator to estimate how many views articles received while damaged.

4.5.1.1 Damaged article views

The relationship between the number of words changed, added, or removed in a damaging revision and the impact of that revision is weak. In some cases, changing only a few words can result in a worse-than-useless article: for example, adding or removing the word "not" or changing a key date can mislead a reader dangerously. On the other hand, voluminous damage such as repeated words can have only minor impact. Therefore, we classify a revision simply as "damaged" or "not damaged" rather than counting the number of damaged words. The key metric in this analysis is the **damaged article view** (DAV), which measures the number of times an article was viewed while damaged (estimating views using the techniques introduced in Section 4.4.1 on page 62 above).

4.5.1.2 Calculating DAVs

We use the opinions of the Wikipedia community in deciding which revisions are in a damaged state. Note that we focus on revisions that are damaged, *not* the edits which repair the damage, because these are the revisions which lead to damaged article views. Revisions that are subsequently reverted (using an identity revert) are considered as candidates for the "damaged" label. To distinguish damage repair from disagreement or other non-damaging behavior, we look for edit comments on the reverts that suggest intent to repair damage.

We assume that for the purpose of damage analysis, most incidents of damage are repaired by identity reverts. This assumption is motivated by two factors. First instructions on Wikipedia itself have, since the beginning of our analysis period, recommended the use of identity reverts for damage repair [114]. Second, repairing damage using the wiki interface to make an identity revert is easier than manually editing away the damage.

It is important to note that our method is not foolproof, as editors sometimes make mistakes or place overheated rhetoric into the comments of reverts, labeling each other vandals when a neutral reader would consider the situation simply a content dispute. We also cannot discover damage which was not yet repaired by the end of the article history. Nonetheless, as our results below

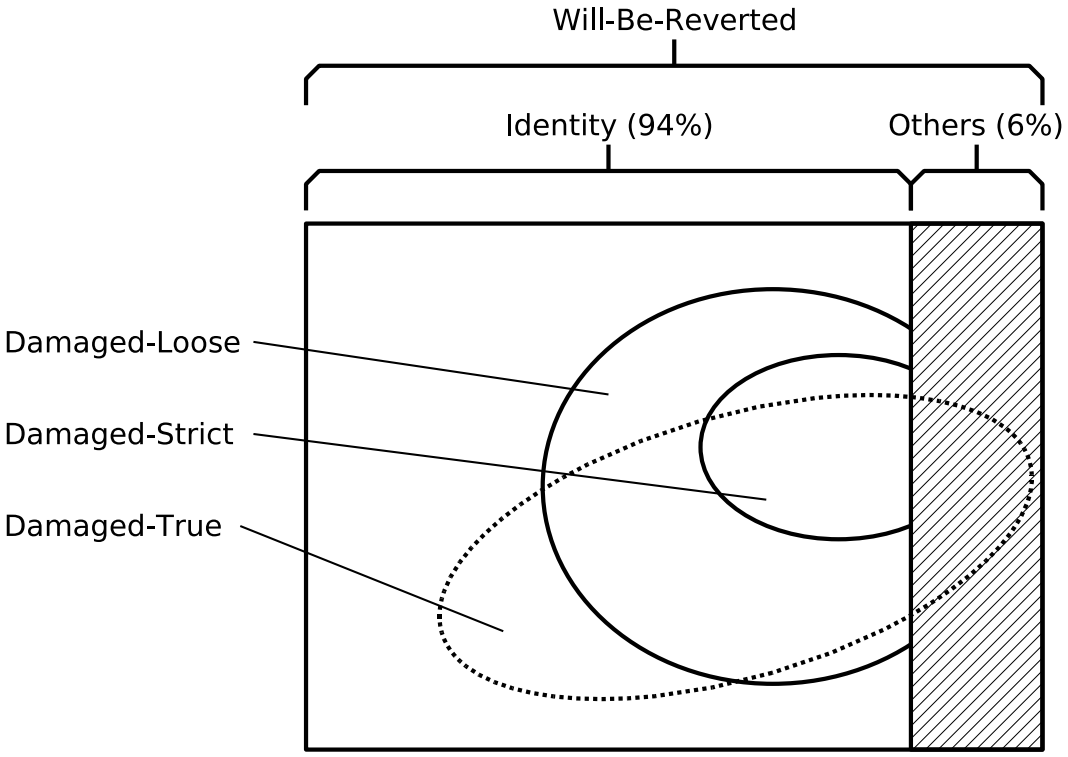

Figure 4.6: Classes of revisions. Damaged-Truth is the conjectured true set of damaged revisions, while Damaged-Loose and Damaged-Strict are increasingly restrictive subsets of Will-Be-Reverted designed to approximate Damaged-True.

show, our method is essentially sound.

At any given instant, a revision is in zero or more of the following states. State membership is determined by how future editors react to the revision, so it can only be determined in retrospect. The edit comments and timestamps necessary for these computations are available in the history of articles (data set 5, noted in Section 4.4.1.2 on page 63). Figure 4.6 shows the relationships of the various states informally.

- *Will-Be-Reverted (WBR)*: A revision which will reverted by a future edit. Several revisions in a row might be reverted by the same revert; we refer to such a group of revisions as a *WBR sequence*. We detect reverts by comparing the MD5 checksums of the texts of each revision.

- *Damaged-Loose (D-Loose)*: A WBR revision where the future revert's

#	Time	MD5	Editor	Edit comment
1	9:19	`24bd`	Carol	new article
2	10:04	`6f59`	Denise	clarify
3	10:19	`2370`	Bob	arrrrr!!!
4	10:37	`02ac`	Bob	shiver me timbers!!!
5	10:56	`6f59`	Alice	revert vandalism

Figure 4.7: Example revision history.

edit comment suggests either (a) explicit intent to repair vandalism *or* (b) use of revert-helper tools or autonomous anti-vandalism bot activity. Specifically, the revert's edit comment matches a regular expression for criterion (a) or another for (b).

- *Damaged-Strict (D-Strict)*: A D-Loose revision that matches criterion (a). This more-selective state is intended to trade some of the recall of D-Loose for greater precision.

- *Damaged-True (D-True)*: A revision that is damaged. The damage may have appeared in this revision, or it may persist from a prior one.

For example, consider the revision history in Figure 4.7. Revision 5 reverts back to revision 2 – we know this because the MD5 checksums are the same – discarding the effects of revisions 3 and 4. Therefore, revisions 3 and 4 are in state WBR; additionally, because Revision 5's edit comment matches the criteria for D-Strict, these two revisions are also in D-Loose and D-Strict. Finally, if each revision were viewed 10 times, there would be 20 DAVs generated by this sequence.

Both D-Loose and D-Strict limit the distance between the first damaged revision and the repairing revert to 15 revisions. This is to avoid false positives due to a form of damage where someone reverts an article to a long-obsolete revision and then marks this (damaging) revert as vandalism repair. We assume that essentially all damage is repaired within 15 revisions.

The purpose of states D-Loose and D-Strict is to be proxies for the difficult-to-determine state D-True. To evaluate the two metrics for this purpose, three

human judges independently classified 676 WBR revisions, in 493 WBR sequences selected randomly from all WBR sequences.[8] Classification included a best effort to figure out what was going on, which often included a minute or two of research to verify information or clarify unfamiliar topics, words, or links. The edit comment of the final revert was hidden in order to avoid biasing the judges. Revisions were classified into the following three classes:

- **Vandalized-Human** (V-Human): WBR revisions that introduce or persist clearly deliberate damage. We attempted to follow the Wikipedia community definition of vandalism, which emphasizes *intent*.
- **Damaged-Human** (D-Human): WBR revisions which introduce or persist damage (a superset of V-Human).
- **Other**: All other WBR revisions. Frequent examples were content disputes or editors changing their minds and reverting their own work.

Determining whether a revision was V-Human or just D-Human is difficult because it requires assessing the intent of the editor. Indeed, despite written guidelines and calibration by judging together a smaller independent set of WBR revisions, there were considerable differences between the judges. Due to this lack of convergence, and because from the reader's perspective, the intent behind damage is irrelevant, we consider further only D-Human.

We use these judgements to evaluate the effectiveness of the classes D-Loose and D-Strict as proxies for D-True. Of the 676 revisions judged, all three judges agreed on 437 (60%), while the class of the remaining 239 (35%) was determined by 2-1 majority. We assumed that revisions judged D-Human by a majority of judges, and no others, were in class D-True.

By this measure, 403 revisions (60%) were in D-True. The automatic D-Strict classifier had a precision of 0.80 but a recall of only 0.17, i.e., within the judged revisions, 80% of D-Strict revisions were in D-True, but only 17% of D-True revisions were in D-Strict; it is therefore a poor proxy for D-True.

[8] This is the same sample used in Section 3.4.

On the other hand, the precision and recall of D-Loose were 0.77 and 0.62 respectively. Clearly, D-Loose suffers from both false negatives and false positives. The former arise when editors revert damage but do not label their actions clearly, while the latter can be seen in content disputes, as described previously. While imperfect, D-Loose is a reasonable proxy for D-True. Thus, the remainder of this section will consider D-Loose only.

4.5.2 Results

We found 2,100,828 damage incidents (i.e., D-Loose sequences). 1,294 overlapped the end of our study period, so there were 2,099,534 damage-repair reverts. No incidents overlapped the beginning of the study period. These incidents comprised 2,955,698 damaged revisions, i.e., an average sequence comprised 1.4 damaged revisions before repair. The study period contained 57,601,644 revisions overall, so about 5% of revisions were damaged.

During the study period, we estimate that Wikipedia had 51 billion total views. Of these, 188 million were damaged – 139 million by anonymous users – meaning that the overall probability of a typical view encountering damage was 0.0037. In October 2006, the last month we analyzed, it was 0.0067. Figure 4.8 on the facing page illustrates the growth of this probability over time. In particular, the data through June 2006 fits the exponential curve $y = e^{0.70x-8.2} - 0.0003$.

Figure 4.9 on page 80 illustrates the rapidity of damage repair. 42% of damage incidents are repaired essentially immediately (i.e., within one estimated view). This result is roughly consistent with the work of Viégas et al. [108], which showed that the median persistence of certain types of damage was 2.8 minutes. However, 11% of incidents persist beyond 100 views, 0.75% – 15,756 incidents – beyond 1000 views, and 0.06% – 1,260 incidents – beyond 10,000 views. There were 9 outliers beyond 100,000 views and 2 beyond 500,000; of these, 8 were false positives (the other was the "Message" incident discussed below). The persistence of individual incidents of damage has been relatively stable since 2004, so the increasing probability of damaged views indicates a higher rate of damage.

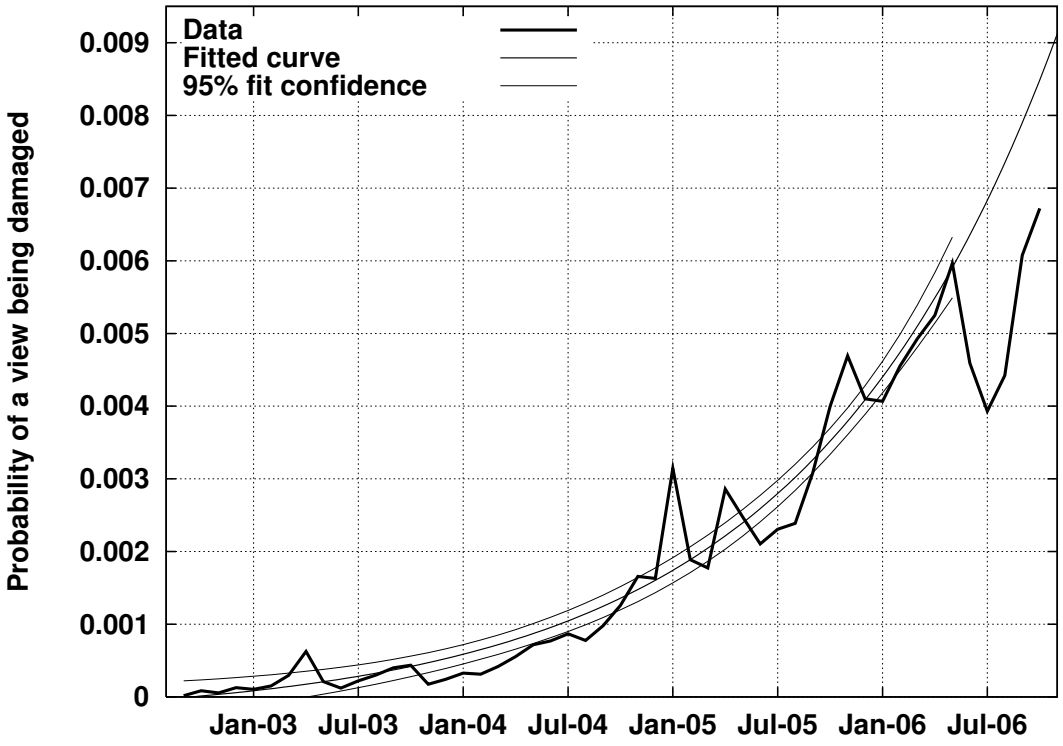

Figure 4.8: Probability of a typical view returning a damaged article. Both our data and a fitted exponential curve (with boundary lines representing the 95% confidence interval) are shown. We fitted through June 2006, when widespread use of autonomous vandalism-repair bots began.

A possible cause for this increase is that users may be writing edit comments differently, increasing the number of edit comments that fit the D-Loose pattern. To test this hypothesis, we judged the precision and recall of D-Loose for a sample of 100 WBR sequences (containing 115 revisions) from 2003 and earlier, using three judges and majority opinion as above. We found that the recall of D-Loose over this sample was 0.49, compared to 0.64 for a sample of 100 sequences (134 revisions) from 2006. Thus, commenting behavior has changed, and this change explains about one-third of the increase in probability of a damaged view. Damage was only 14 times more impactful in 2006 than 2003, not 18 times.

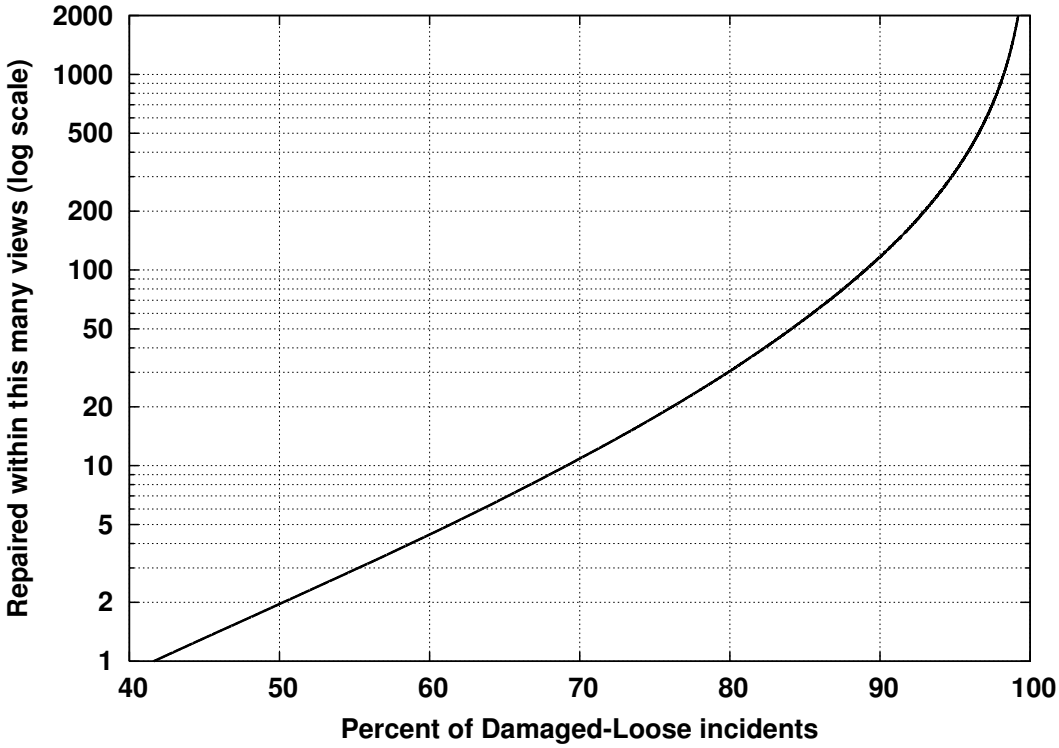

Figure 4.9: Rapidity of damage repair. 42% of damage incidents are repaired within one estimated view, meaning that they have essentially no impact.

4.5.3 Discussion

While the overall impact of damage in Wikipedia was low during the study period, it was rising. The appearance of vandalism-repair bots in early 2006 seems to have halted the exponential growth (note the dramatic drop in view probability after June 2006), but this analysis does not show what the lasting impact has been or will be.

Many of the editors who performed the reverts we analyzed appear to have treated vandalism and other damage the same way. For instance, it is common to find edit comments asserting vandalism repair on reverts of revisions which are apparently users practicing editing on a regular article, something which is not welcome but is (by policy) explicitly not vandalism. This makes sense, as from the reader's perspective, damage is damage regardless of intent.

The most viewed instance of damage was deletion of the entire text of the

article "Message", which lasted for 35 hours from January 30 to February 1, 2006 and was viewed 120,000 times. The popularity of this article is partly explained by the fact that it was (and remains today) the top Google search result for "message". It may seem odd that such a high traffic article was not fixed more quickly. However, it was not edited very frequently, only about once every 19 days. (This reminds us that there is no correlation between view rate and edit rate.) Put another way, the tens of thousands of people who viewed the article while it was damaged simply may not have included any Wikipedia editors. Further, maybe this type of damage does not invite a fix as much as others, such as obscenities.

One example of a high-traffic article which suffered greatly from damage was "Wiki" in October 2006. Of the 3.1 million estimated views during that month, 330,000 were damaged, over 10%. This page was bombarded with damaging edits having no apparent pattern, most of which were repaired rapidly within minutes or hours but which had dramatic impact due to their sheer numbers. Another interesting example is "Daniel Baldwin," an article about an American actor, damaged by a single incident which lasted for three months. In October 2005, someone vandalized the article by deleting the introduction and replacing it with with text asserting that Baldwin was a college student and frequent liar. The next day, someone removed the bogus text but failed to restore the introduction; this situation persisted through five more revisions until someone finally made a complete repair in February 2006.

We have two policy suggestions for combating damage, both based on distributing repair work to humans. The first is to ensure that revisions are reviewed by n humans within a few seconds of being saved. The second is to ensure that each article is on at least n users' watch lists. Assuming an edit rate of 280,000 edits per day (the average rate we observed in our log analysis), and assuming it takes 30 seconds to determine if an average revision is damaged, schemes like these would require about $n \times 28{,}000$ reviewers averaging five minutes of daily work.

Developments subsequent to this experiment have largely followed these suggestions, though with a slightly different angle. Geiger and Ribes described

the anti-vandalism landscape as of 2009 [45]; the bulk of this activity takes the form of humans using specialized anti-vandalism software which makes it easy to quickly bring up article diffs, determine if the revisions constitute damage, and revert the offending revisions if so. The software uses annotations within Wikipedia (on user talk pages) to coordinate activity between users doing anti-vandalism work and make it easy to identify repeat offenders (whose edits are then assumed to be suspect).

While Cyclopath is not yet mature and visible enough to have attracted significant destructive activity, we expect that as geowikis mature, similar problems will emerge. We believe the solution is the same: strong transparency and tools which make reversing undesirable edits quick and easy. Our existing watch region, recent changes, and geodiffing tools were designed with this in mind, and they could be further enhanced. As noted in Section 4.3.3 above, the system could make heuristic guesses based on routing impact to flag revisions which deserve particular scrutiny, and simple actions like including a diff or geodiff in watch region notification emails, with one-click links to endorse or reject the change, could also help.

4.6 Summary

For any information resource, it is essential to measure impact on consumers, not just the effort of producers. This is not easy, and we propose three techniques for doing so. If user content feeds an algorithm, measure the effect of user work on that algorithm; in Cyclopath, we showed that user work has improved computed routes by shortening them an average of 1 km. If not, use viewing activity as a proxy for value. This is particularly difficult in Wikipedia, which is a large system of broad interest. We introduce techniques for doing so, showing that 0.1% of Wikipedia contributors create 44% of the value, i.e., just 4,200 editors produce nearly half the encyclopedia's value.

Finally, open content systems are vulnerable to anti-value: damaging work which is contrary to the goals of the system. We propose measuring the impact of damage by counting the number of times a damaged resource is viewed,

estimating that 0.7% of article views in Wikipedia are of articles damaged in some way.

These results are of general interest, as they are useful for any systems that accept user contributions and seeks value by presenting them to other users. Only by correctly identifying the value of contributions can more valuable contributions be elicited.

Our efforts have not completed this research direction. For example, simply requesting an article is a coarse proxy for interest and value: the reader could realize that an article is unhelpful only after seeing it or be interested only in a small part of a large article. Other factors like viewing time or click paths when browsing within a resource could be informative, and more granular measures of viewing, such as recording which part of a resource is in view [32], can sometimes be introduced. These notions should be validated with user-based experiments, which may be able to also produce correction factors to improve these value proxies.

Chapter 5

Eliciting More Contributions

5.1 Introduction

Above, we established that user contributions are valuable in a geographic open content system; thus, it benefits the community to obtain more of this *geographic volunteer work* (GVW). It is also beneficial to focus this work where it is most needed. Eliciting participation in open content systems is an ongoing challenge; many online communities fail [20], and even those that succeed need to focus user work. For example, of the many techniques that Wikipedia uses to encourage participation, 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.

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., at the time of the experiment reported in this chapter, 423 logged-in users had saved a least one revision, but only 7 had 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.

This leads to our research question: *What techniques lead to increased GVW contributions?* We carried out a field experiment in Cyclopath, finding 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, but only for certain types of work; and (c) users do significant "extra" work beyond what is highlighted.

These findings are of interest to operators of open content systems who wish to elicit work in their system. Identifying which types of contributions require familiarity and which do not is useful, as it affects who should be asked to do a particular task. The possibility of "extra" work means that it may be plausible to ask users to do work other than what is actually needed (perhaps because it is more initially appealing), knowing that they will also do the needed work.

In this chapter, we first present related work, and follow that with a description of the design of our field experiment, our results, and a summary.

5.2 Related work

Social science offers insights into the problem of eliciting user work. Karau and Williams [62] 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 [40, 43] and setting specific and challenging goals [75].

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. describe several experiments that evaluated techniques built on factors like uniqueness and goal-setting [73], and Harper et al. evaluated techniques based on social comparisons [53].

Most relevant to our work, Cosley et al. have developed several automated techniques for eliciting work. In one, the researchers wanted users of the MovieLens movie recommendation site to edit information about movies [28]. In the other, they wanted Wikipedia editors to edit articles [29]. 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*).

5.3 Method

Types of work elicited. The Cyclopath database contains thousands of errors. Specifically, automated analysis of one class of errors revealed 7,000 *missing X nodes* – places where two edges cross one another geometrically but no network node exists – and 6,300 *missing T nodes* – places where a dead-end edge came within 20 meters of intersecting another edge. While these potential nodes can be identified automatically, human judgment is required to determine 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.

We also asked users to enter ratings. The Cyclopath ratings database is very sparse, with about 43,000 ratings for 150,000 edges at the time of the experiment. More ratings would improve the accuracy of the route finding engine's edge evaluations, thus enabling it to compute better routes.

Hypotheses. We hypothesized that the following three factors would lead to increased geographic volunteer work.

H1. **Familiarity.** *Users will do more GVW in areas they are familiar with.* In

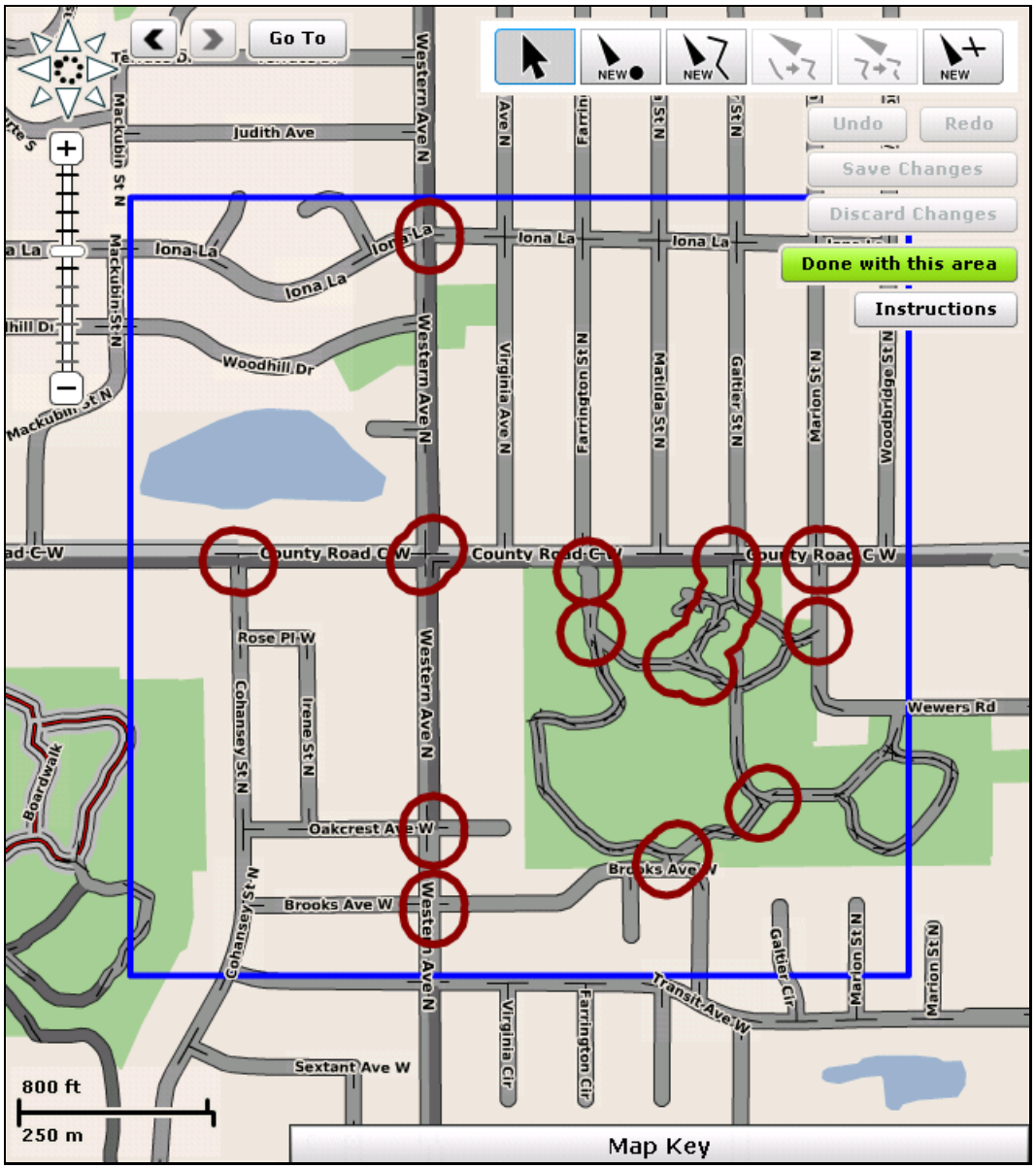

Figure 5.1: What a subject might see in the Visual Prompts / Node Repair condition. The blue square is the trial viewport – where the subject has been asked to do node repair work – and the maroon circles highlight potentially missed nodes within the viewport. In the No Visual Prompts condition, these visual highlights would be absent, but the blue square would still be present. In the Ratings condition, edges 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.

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 edges in an area is a fairly underspecified instruction. Thus, we used visual highlights to focus user attention on specific objects that may need work: edges that need ratings are colored blue, and potentially missing nodes are indicated by maroon circles (see Figure 5.1 on the previous page).

H3. **Work Type-Familiarity.** *The familiarity effect of H1 will be stronger when users are asked to rate edges than when they are asked to repair missed nodes.* We thought that the familiarity effect would be stronger for ratings because rating a edge 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).

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 edges or 75% of the edges 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 5.1 on page 87 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 edges 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 5.2 on the facing page shows the work done in one selected 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.

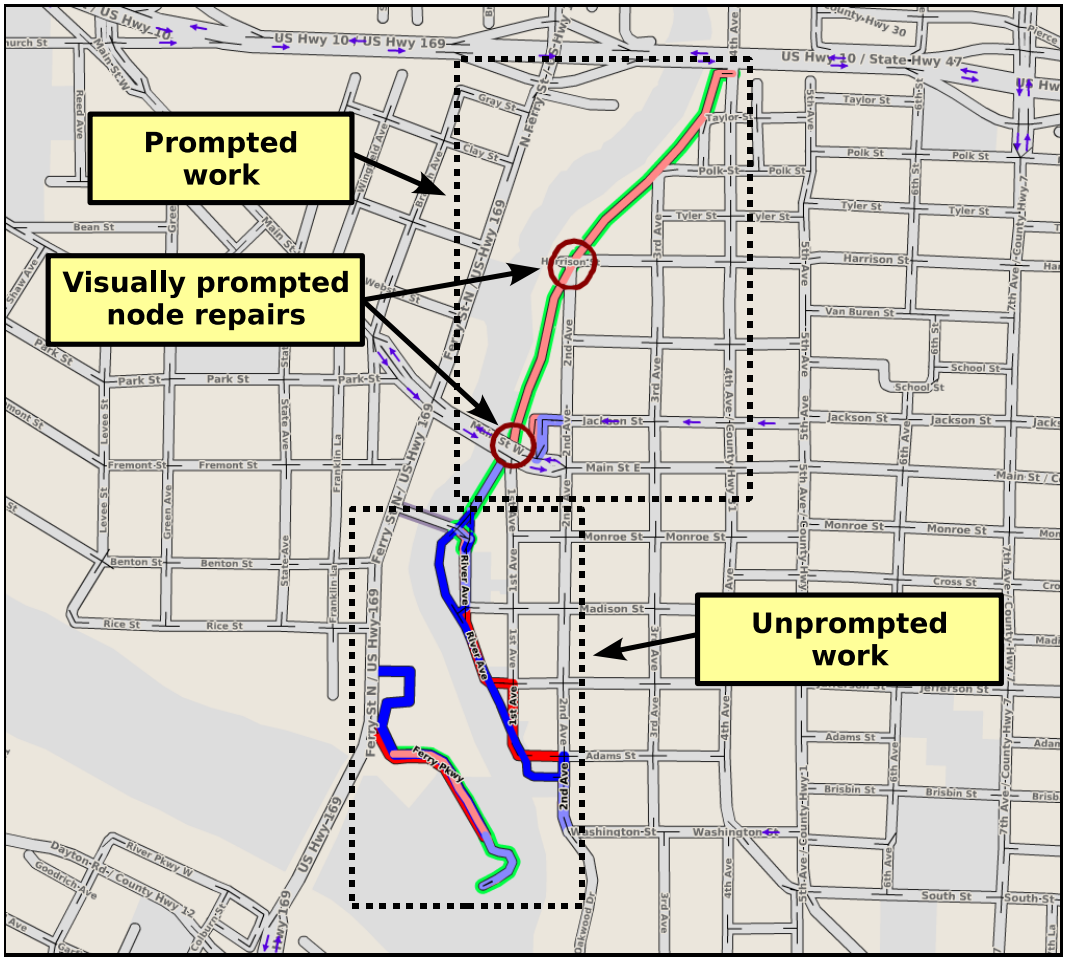

Figure 5.2: Work done in a selected trial. Red and blue show the changed edges. In this case, the subject has made edge edits both related and unrelated to the prompted node repairs.

5.4 Results

Our research question is 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, and H3: Work Type-Familiarity. The experimental trial is the basic time interval for counting units of work. It also is useful to aggregate work, notably to the level of all trial 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

Measure	Condition		*P*
	familiar	random	
Total edges	112.0	53.0	$***$
Rateable edges	102.7	52.8	$***$
Node repairs	2.89	2.32	$**$

Table 5.1: Available work in an average trial viewport in the Familiar and Random 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.)

in that condition.

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

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

Additionally, we counted *edge 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 edge 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 edge edits was more tractable and still gave us a reasonable metric for quantifying work.

Finally, it is useful to distinguish edge edits made 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 edges passing within 80 meters of a prompted node repair to be prompted edge edits. Figure 5.2 on the previous page shows work done in a selected trial, highlighting prompted and unprompted edge edits.

When analyzing our data, we found an unanticipated correlation between

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

Table 5.2: Total work completed during experiment. For each type of work, we compare the Familiar (n=57 subjects completing at least one Familiar trial) and Random (n=50) conditions (df=67, 101, 101, and 70, respectively).

Familiarity and work availability: as Table 5.1 on the facing page shows, the Familiar viewports we generated had more available work than Random viewports. This is because Cyclopath 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 edges rated) and counts normalized by the amount of available work (e.g., proportion of available edges 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. Table 5.2 and Table 5.3 on the next page 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 edge edits was virtually identical in the two conditions. And while many more notes were edited in the

Measure	Condition	mean	median	Q3	max
Ratings	random	0.55	0	0	45
	familiar	7.19	0	5	119
Node repairs	random	0.62	0	0	8
	familiar	0.58	0	1	7
Edge edits	random	3.90	0	0	186
	familiar	3.57	0	3	69
Note edits	random	0.02	0	0	3
	familiar	0.13	0	0	25

Table 5.3: Work per trial, comparing Familiar (n=372 trials) and Random (n=355).

Familiar conditions, the total number were small (48 vs. 7), and the difference was only marginally significant.

We speculate that very different effect of familiarity for the different work types is explained by the original rationale for H3: rating a edge 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 appropriate. In contrast, in previous work that found familiarity to be the basis of effective work elicitation techniques [28, 29], 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, Hoffmann et al. had subjects verify that certain information was present in a Wikipedia article [59]. 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 5.3 on the facing page shows the total amount of work of each type done in the two conditions, and Table 5.4 on page 96 provides more detailed comparisons of work done in

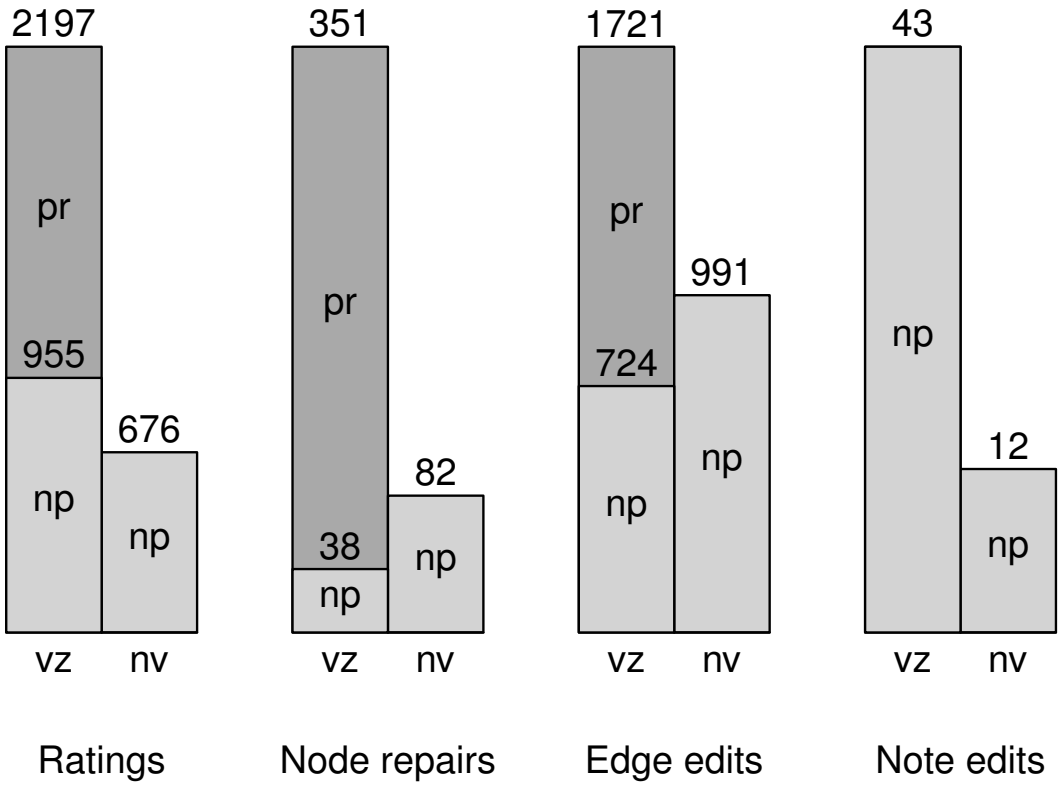

Figure 5.3: Total work completed in experimental trials. This figure shows what was done in the Visual Prompts (*vz*) and No Visual Prompts (*nv*) conditions, divided into prompted (*pr*) and not prompted (*np*). For ratings and node repairs, prompted work included visually highlighted edges rated and node repairs resolved; for edge edits, prompted work includes all edge 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.

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.

Specifically, the advantage of Visual Prompts was largely not because subjects did more work *per trial* – they did about the same amount of work per trial in the Visual Prompts and No Visual prompts condition (Table 5.5) –

Measure	Condition	Total	Per subject mean	Per subject Q3	Per subject max	*P*
Ratings	no visual prompts	676	21.1	21.8	224	∘
	visual prompts	2197	68.7	96	537	
	unprompted	955	38.8	46.8	416	
	prompted	1242	29.8	50.8	121	
Node repairs	no visual prompts	82	2.6	1.5	40	*
	visual prompts	351	11.0	7.3	108	
	unprompted	38	1.2	0.3	11	
	prompted	313	12.0	12.3	102	
Edge edits	no visual prompts	991	31.0	10.8	499	
	visual prompts	1721	53.8	33.8	420	
	unprompted	724	22.6	15.3	226	
	prompted	997	31.2	25.5	272	
Note edits	no visual prompts	12	0.4	0	7	
	visual prompts	43	1.3	1	26	

Table 5.4: Total work completed during experiment. For each type of work, we compare Visual Prompts (n=32 subjects) to No Visual Prompts (n=32 different subjects) (df=44, 48, 55, and 50, respectively); further, we show work done in Visual Prompts that was actually prompted and not prompted. *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*) subjects. *P* gives the P-value code comparing Visual Prompts and No Visual Prompts.

but because they requested over three times more trials (Table 5.6). In other words, the *visual focus* vs. *visual clutter* distinction was not particularly supported; instead, we speculate that the visual highlights simply made the task more fun.

A second surprise was that in addition to subjects doing much work they were prompted to do, they also did much *unprompted* work. Figure 5.3 shows that subjects in the Visual Prompts condition entered more ratings of unprompted edges than No Visual Prompts subjects did in total (though this result is not statistically significant).

Measure	Condition		n	mean	Q3	max
	work type	**visual prompts**				
Ratings	ratings	no visual prompts	80	7.03	5	105
	ratings	visual prompts	224	6.00	3	119
Node repairs	node repairs	no visual prompts	81	0.93	1	7
	node repairs	visual prompts	268	1.30	2	8
Edge edits	node repairs	no visual prompts	81	10.90	11	186
	node repairs	visual prompts	268	5.90	7.3	80
Note edits	*both*	no visual prompts	161	0.075	0	3
	both	visual prompts	566	0.076	0	25

Table 5.5: Work units completed per trial in Visual Prompts and No Visual Prompts conditions, when asked to do Ratings or Node Repairs. 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.

Condition	Total	Per subject			P
		mean	Q3	max	
no visual prompts	161	5.03	5.3	25	*
visual prompts	566	17.69	18	140	

Table 5.6: Number of trials requested by subjects (df=49).

5.5 Summary

We tried two techniques to elicit and focus geographic volunteer work: familiarity with the geographic region where work was requested and visual prompts highlighting needed work. 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 with the two techniques. Rating bikeability requires personal experience and a specific memory of that experience. Thus, familiarity is key, and visually highlighting a edge 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 [62].

In short, we produced two surprising findings of general applicability beyond our techniques for obtaining more geographic volunteer work: that familiarity has no effect for certain types of work (i.e., users can be asked to do work anywhere that it is needed), and that users do a significant amount of work beyond what they are asked to do (i.e., it is not always necessary to ask users to do the specific type of work that is needed, as they may do it anyway). These are important lessons for any system that solicits user contributions.

Chapter 6

Personalizing Open Content

6.1 Introduction

Personalized interpretation of data sets is common; **recommender systems** are widely used for finding *items* like books and movies. Companies like Amazon and Netflix depend on recommender systems to achieve their business goals, and the field has created mature research platforms like MovieLens. Other applications include finding news stories [31] and online friends [22].

Bicyclists also want personalized interpretation of their information resources. Specifically, they want a system which can compute personalized routes; for example, people might prefer a slightly longer route (in time or distance) if it is one that is scenic. They told us this in our foundational surveys and interviews when designing Cyclopath (see Section 2.3.2.3 on page 18), and previous work has also shown the value of route finding which takes into account subjective preferences [12, 60, 80]. In this chapter, we explore ways to produce better personalized routes.

Route requests consist of a starting location (address, point of interest, etc.) and ending location; the system then computes a minimum-cost path. Our goal is: rather than returning the path which has minimum cost according to some objective metric like distance or time, return the path which is *most subjectively desirable* – i.e., most preferred by the requesting user.

However, recommender system ideas have not yet been applied to the do-

main of route finding, which is interestingly different for three reasons: the structure of the rating datasets are different (items dramatically outnumber users rather than the other way around, and ratings are much sparser), items have structured relationships and the output is a lengthy sequence of items (a route), and 100% coverage is essential (no point pairs can be unrouteable due to lack of rating data).

This is the gap which this chapter seeks to close. We introduce a three-stage framework for evaluating recommender systems in the context of route finding and evaluate 11 algorithms using the first two stages of our framework.

We first survey related work and present evidence suggesting the need for personalized routing. We then describe our evaluation framework, the results of the algorithms we tested under that framework, and close with a summary.

6.2 Related work

6.2.1 Recommender systems

Established collaborative filtering approaches include user-based [91], item-based [94], and content-based [8] algorithms, with later refinements such as dimensionality reduction using singular value decomposition [44] and ensemble recommenders which combine the output of multiple algorithms [13].

More recently, researchers have begun to explore geographically aware recommender systems. Some systems recommend items near a user's current or future location, such as restaurants [84, 124] or tourist destinations [6, 92]. Others assume that a user's location contains information about taste preferences. For example, Li et al. found that similarity of two users' location histories predicts the strength of the social tie between them [72]. Zheng et al.'s GeoLife2 social networking system does both, drawing on location to infer user similarity and by recommending social activities between potential friends who are near each other [126]. Our work resembles these types of systems in that it predicts user satisfaction for edges in the transportation network, but we go beyond prior work to explore how these individual predictions affect their assembly into a complete route.

Finally, researchers have also studied recommending collections of items. Ziegler et al. showed that user satisfaction with a set of recommended items was enhanced when diversity of items within the set was considered, rather than simply minimizing the total prediction error [127]. Felfernig et al. explored recommending product configurations, where different components required or conflicted with other components [39]. Our work builds upon these ideas by addressing "collections" of large size (the sequence of edges which forms a route) and the new types of structure relevant to route-finding applications.

6.2.2 Route finding

Modern route-finding systems are supported by a large scholarly literature. For example, geographers have explored the effects of transportation on human beings: the meaning of mobility and distance, economic and political influences, and where people choose to live and travel within their communities [93]. Computer scientists have studied the technical aspects of route finding, exploring issues like algorithm performance and scalability [33], uncertain costs [68], and multi-modal routing (e.g., trips by both car and bus) [16].

Researchers have begun to explore *personalized* routing. Bederson et al. motivate personalization by pointing out the need for "subjective human experience" in route-finding, proposing a system for recording such experiences [12]. Personalized algorithms that accommodate these differences in routing preferences have been shown to outperform traditional routing algorithms. The OurWay system created by Holone et al. asked nine users in wheelchairs to rate route segments to order to receive personalized route recommendations, and these outperformed traditional algorithms [60]. McGinty and Smith used case-based reasoning to find a route by drawing on data from other users with similar route preferences [80].

Our work improves upon these personalized routing ideas in two ways: we draw upon a much larger dataset from a mature production route-finding system, and we apply established algorithms from the recommender systems community to this problem.

6.2.3 Route finding for cyclists

Numerous websites exist addressing various aspects of bicycle route finding, most notably Google Maps, which provides map-based routing in many American cities. Section 2.2.5 on page 13 discusses these systems in detail.

A variety of metrics for computing the bikeability of roads exist, but they are generally impractical because they require an unrealistically large number of attributes for each edge; for example, Sorton's bicycle stress level requires 6 to 14 [101], and an analysis at Macalester College identified 16 [77]. In Cyclopath, we currently use a metric developed by the Chicagoland Bicycle Federation [10], which requires 4 attributes; of these, we have moderately complete data for only 3 (speed limit, daily traffic volume, and shoulder width) and guess the last one (lane width).

6.3 Routing and Recommending

This section discusses how routing in Cyclopath works from the perspective recommender systems. Section 2.5.3 on page 31 above details the technical aspects of route finding in Cyclopath; specifically, recall that for analysis, we translate Cyclopath's word ratings (a five-level scale from "Impassable" through "Excellent") to a numeric scale from 0 to 4 "stars"

6.3.1 Evidence for personalization

In addition to the qualitative results mentioned above, cyclists' ratings give quantitative evidence that they need personalized route finding. As of April 2010, Cyclopath has 66,474 ratings by 464 users, with a mean rating of 2.78 and standard deviation of 1.13; in MovieLens, these numbers are 3.51 and 1.06, respectively – Cyclopath ratings have equal or higher variance than MovieLens. Also, of the 101,130 instances when two users rated the same edge: they gave the same rating 53.6% of the time, disagreed by one star 34.2% of the time, and disagreed by two or more stars 12.2% of the time. In other words: nearly half the time, users disagree on the appropriate rating for an edge, and about

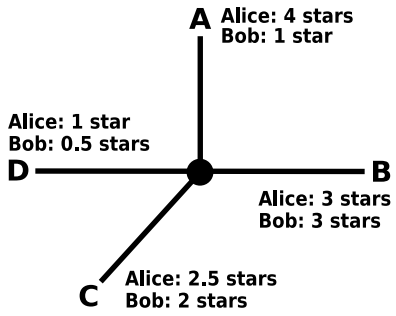

Figure 6.1: A hypothetical node in the transportation network. Bikeability ratings by two users are shown.

one eighth of the time, this disagreement is at least two stars (the difference between "fair" and "excellent" or "poor" and "good").

Another way to quantify disagreement is: to what degree would different users' ratings actually lead to different decisions by the graph search algorithm? We can explore this by examining the preference order of edges rather than actual rating values; the intuition is that even if two users give different ratings, but graph search would make the same choices regardless, the two users' ratings are effectively the same.

Figure 6.1 shows a hypothetical graph node with four adjacent edges, A through D, and bikeability ratings from two users, Alice and Bob. The predicted preference order of the four edges is ABCD according to Alice but BCAD according to Bob. For example, Edge B has the same rating (3 stars) from both users, but only for Alice is it the most preferred edge at this node.

This is relevant because graph search algorithms such as A* wander through a graph, choosing at each node the best edge to wander next. For example, if a graph search algorithm arrived at this node via Edge C, it would next choose Edge A if using Alice's bikeability ratings, but Edge B using Bob's ratings (assuming that other considerations such as edge length are the same).

The results of this analysis (performed by Ryan Kerwin) are as follows. Given adjacent edges A and B and users U and W who have each rated both edges, in 87.3% of the 81,903 observations, U and W agree: either all four ratings were the same (84.7% of observations) or both users prefer A to B (2.6%). On the other hand, 12.7% of the time, they disagree; either U prefers

A to B and W rates them equally (12.0%) or U prefers A to B but W prefers B to A (0.7%). At first glance, this result is cause for skepticism – users give a different preference order only one eighth of the time. However, the average computed route contains 76 edges, the graph search to produce this average route evaluates thousands of edge pairs, and these results imply that each set of 23 edge pairs is 95% likely to contain a disagreement ($0.873^{23} = 0.044$). Thus, even this fairly low level of disagreement will result in computations that encounter many points of disagreement, where the computed paths might diverge, whether in major or minor ways or not at all.

These three analyses suggest that route personalization is valuable. Accordingly, we now turn to efforts to improve it.

6.3.2 Challenges for recommendation

As noted earlier, there are three core challenges in applying recommender systems technology to route finding in systems like Cyclopath, which we have overcome. First, universal coverage is required. In contrast to systems that recommend products (e.g., 17% of movies in MovieLens are unrecommendable), graph search algorithms need edge weights for every edge in the graph in order to produce a path for arbitrary starting and ending points.

To be pedantic, this is not strictly true. It isn't necessary to predict all edge weights; instead, it is necessary to predict edge weights for all the "important" edges – those which have a realistic chance of being considered for routes (not just included in the result). However, 100% coverage is achievable with good performance (see below), and this is a simple way to meet this stricter criterion. Otherwise, one must produce and justify a quantitative definition of "important", a difficult task.

Second, the structure of the user/item rating matrix is less favorable. As noted in Table 6.1 on the facing page, when considering all edges (not just those that are rated), as is needed to address the first challenge, the Cyclopath dataset is 12-44 times more lopsided than active MovieLens or the Netflix Prize dataset, and in the opposite direction; for example, MovieLens has 7.6 users per item, but Cyclopath has 331 items per user. Furthermore, the Cyclopath

Dataset		Users	Items	Aspect	Ratings	Sparsity
MovieLens		102,757	13,608	7.6 u/i	17,488,416	0.013
Netflix Prize		480,189	17,770	27 u/i	100,480,507	0.012
Cyclopath	rated edges	464	34,817	75 i/u	66,474	0.0041
	all edges		153,841	331 i/u		0.00093

Table 6.1: Descriptive statistics for MovieLens, Netflix Prize [51], and Cyclopath ratings datasets. We include only Cyclopath users who have made at least one rating, in order to permit comparison with other systems; for the same reason, we present the number of rated edges as well as the total number of edges in Cyclopath. The latter is of greater interest in our analysis due to the 100% coverage requirement.

dataset is 14 times sparser.

Finally, both items and the output of the system have significant constraints. Items (edges) have highly structured relationships defined by the topology of the transportation network, and rather than outputting one or a few individual items, recommenders for route finding must produce a structured sequence of items – a route.

6.4 Evaluation framework

Recommender algorithms predict ratings for individual edges, which are then used by the graph search algorithm to produce the complete route. Our framework analyzes performance at three stages – individual edges, node-level graph search decisions, and the final route – testing algorithms at each stage to see if differences remain meaningful.

6.4.1 Stage 1: Predicting edge ratings

The goal of this stage is to assess the accuracy of edge rating predictions when compared against actual user ratings; it corresponds to the "predict" task in traditional recommender systems [95]. Here, items are the edges in the transportation network: given an edge and a user, predict the user's rating of that edge. The procedure is straightforward: (a) define and implement different prediction algorithms, (b) evaluate them using standard k-fold cross-

validation, and (c) report coverage and error for each algorithm, using standard metrics like mean absolute error (MAE) and root mean squared error (RMSE).[1]

6.4.2 Stage 2: Comparing node-level decisions

This level considers whether edge weights created by different prediction algorithms would lead to different decisions by a graph search algorithm. Similarly to the discussion of user ratings in Section 6.3.1 above, the intuition is that even if Algorithm A has better accuracy than Algorithm B, but graph search would make the same decisions regardless, there is no effective difference between the two. It corresponds to the "recommend" task in traditional recommender systems [95], where a system tries to choose the best available item (e.g., "find me the movie I would most like"); the key difference is that in recommending routes, this decision is made thousands of times per high-level operation (find a route), rather than a few times (recommend a few movies).

For this analysis, we introduce a metric called *best edge agreement* (BEA): given two algorithms and a set of nodes, the BEA is the percentage of nodes where the two algorithms select (from the edges adjacent to the node) the same edge as best. This metric is hard to operationalize, however, because it requires both algorithms to predict *all* edges adjacent to a particular node in order to assess their BEA at that node; as we show below, many algorithms have poor coverage, so this requirement is often not met. Thus, we take a pairwise approach, introducing a variant metric, *pairwise best edge agreement* (PBEA): given two algorithms and a set of pairs of edges such that both edges (a) are adjacent to the same node and (b) have predictions from both algorithms, the PBEA is the percentage of pairs where both algorithms select the same edge as best.

[1] Given a sequence of predictions $p_1..p_n$ and corresponding true values $t_1..t_n$, these two metrics are defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|p_i - t_i| \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - t_i)^2}$$

PBEA can be used to estimate BEA. For example, suppose that Algorithm X's predictions for four edges adjacent to a node give preference order ABCD; in particular, Edge A has a higher predicted rating than B, C, and D. The probability that Algorithm Y chooses the same best edge (A) is the probability that its own rating prediction for Edge A is also higher than for the other three edges, i.e., Y must agree with X on the predicted preference order of three edge pairs: AB, AC, and AD. Assuming that edge predictions are independent,[2] the probability of the algorithms agreeing three times is the probability that they agree once raised to the third power: we estimate BEA on this four-edge node as PBEA cubed. Cyclopath nodes have on average 3.2 edges (excluding dead-ends, which don't affect graph search decisions), so overall we estimate BEA $\approx$ PBEA$^{2.2}$.

6.4.3 Stage 3: Comparing paths

This stage evaluates how much *paths* produced with different prediction algorithms differ. The motivation is similar to Stage 2: even if different algorithms produce different low-level graph search decisions, if the resulting paths are the same, the algorithms are effectively equivalent. Differences in paths are evaluated in two ways: (a) objectively using some metric such as counting the number of shared edges and (b) subjectively with a user study, e.g., by showing users the alternative routes and asking if they are *meaningfully* different. For example, two parallel roads in a residential neighborhood consist of objectively different edges but may be effectively equivalent to a human cyclist.

This is the first stage where 100% coverage is needed, because it is the first where rating predictor and graph search algorithms are combined. As our results below show, generally, an ensemble predictor is needed, because there is a tradeoff between accuracy and coverage.

We next turn to our application of this framework in Cyclopath. We pro-

[2]This assumption is probably untrue – i.e., ratings of adjacent edges are probably correlated – so it's likely that this technique underestimates BEA. However, if we assume that the ratings of adjacent edges are positively (and not negatively) correlated, this underestimation is bounded, as the true BEA must then lie between PBEA and the estimated BEA.

vide an evaluation through Stage 2, because our focus is on comparing algorithms with each other – Stage 3 is beyond the scope of our current inquiry, and we look forward to pursuing it in the future.

6.5 Algorithms tested

We tried 11 algorithms, or 23 including variations, which we divide into four classes: objective algorithms which do not consider user ratings, algorithms which average user ratings in simple ways, collaborative filtering algorithms, and algorithms which build clusters of edges.

6.5.1 "Objective" algorithms

These two algorithms, which make use of only objective edge attributes rather than ratings, represent the general prediction approach favored by the transportation community. They are unaffected by ratings propagation.

We first consider *Objective-CBF*, a bikeability metric developed by the Chicagoland Bicycle Federation which considers speed limit, daily traffic volume, lane width, and shoulder width [10]. While simpler than other metrics, it is the only one we could find whose input attributes had sufficient coverage in Cyclopath, which is based on the best available data (and data availability for our area is typical for American cities). Our implementation has minor modifications to make it usable in Cyclopath (motivated by data availability and my expertise as a bicyclist), specifically: (a) expressways are always rated 0 stars (Impassable); (b) bike lane width is included in shoulder width; (c) lanes are assumed to be 3.6 meters wide, which is the state standard, for the 99.7% of edges where we have no value for this attribute; (d) rating adjustments for shoulder width and bike lane were altered – shoulder width greater than 1.2 meters earned a 1.5-star bonus (not 2 as in original CBF), no additional bonus was given for very wide shoulders, and the presence of a bike lane earned a 0.5-star bonus (none in original CBF); and (e) a 1-star penalty for unpavedness was added.

Bike paths and sidewalks are not addressed by this metric, so we estimate ratings for these based on the edge's length: bike paths less than 100 meters, 2.5 stars; over 100 meters, 3.5 stars; over 500 meters, 4 stars; and sidewalks 1 star less. This is again based on my intuition as a cyclist and feedback from users. Finally, we clamp the output to the range 0-4.

We created *Objective-Simple* to achieve universal coverage. This ad-hoc metric predicts roads based on their type: expressways 0 stars; other highways 1 star; major road, unknown, and other, 2 stars; and all other types 3 stars. We then apply several corrections as in *Objective-CBF*: adjust for shoulder width, bike lanes, and unpavedness and clamp to 0-4 stars.

6.5.2 Simple averaging algorithms

We next turn to algorithms that average ratings in a simple ways. First, we tested two very simple baseline methods. *Global-Mean* predicts the mean of all ratings in the system, while *User-Mean* predicts the mean rating of the user for whom the prediction is being made.

We also tried a family of algorithms which predict the mean of other users' ratings on a particular edge; for example, if an edge is rated 2 stars by User A and 3 by User B, the prediction for other users would be 2.5 stars. We varied the threshold of ratings required for prediction, only considering a prediction valid if there are more than n ratings on an edge. We tried $n = 1, 2, 3, 5, 10$, and we refer to these algorithms as *Edge-Global-Mean.n*.

6.5.3 Collaborative filtering

Collaborative filtering (CF) is one of the most powerful recommendation techniques in other contexts [13]. We expected CF to give accurate ratings but have problems with coverage, given our extremely sparse dataset. We implemented collaborative filtering according to [97] and [95], with several variations: user-based vs. item-based, different distance metrics (Pearson correlation, Euclidean distance, cosine similarity), and adjusting or not adjusting ratings by the user average. We used a similarity threshold of just one co-rating in an

effort for more coverage (except for the Pearson distance metric, which is not defined for one co-rating).

6.5.4 Cluster-based personalized algorithms

Finally, we tried content-based methods; the intuition is to group edges into clusters based on attributes and then build personalized user profiles using average ratings within each cluster.

One simple way to cluster is to use the 9 edge types (major road, bike path, etc.). *Type-Global-Mean* predicts the mean of all users' ratings on edges of the same type, while *Type-User-Mean* predicts the mean of the requesting user's ratings within edges of the same type.

We also tried automatic clustering, using Weka [52] to cluster edges with the DBSCAN algorithm [38], with $\epsilon = 0.9$ and minimum cluster size 100, on all available attributes (most of which have extensive null values): type, one-way status, speed limit, lane width, shoulder width, number of lanes, daily motor vehicle traffic, presence of bike lane, unpavedness, presence of "hill" tag, and presence of "prohibited" tag. This yielded 27 clusters ranging in size from 97 to 79,982 edges; 996 edges were not in any cluster. Algorithm *Cluster-Global-Mean* predicts the mean of all users' ratings on edges in the same cluster, while *Cluster-User-Mean*, which predicts the mean of the requesting user's ratings within a cluster.

6.6 Stage 1: Predicting ratings

We begin by analyzing the coverage and accuracy of 11 rating prediction algorithms, to find which algorithms are promising in the context of route finding.

6.6.1 Method

We implemented the algorithms in Python and evaluated their accuracy with 10-fold cross-validation, using Cyclopath ratings as of April 2010. This dataset

contains 66,474 ratings by 464 users on 34,817 of 153,841 edges.[3]

We report accuracy using MAE and RMSE. Coverage is computed statistically: we took a random sample of 5,000 user,edge pairs and tried to predict each, counting how many times we succeeded. We use the full dataset for this computation, i.e., without cross-validation, which is not needed because we are not using prediction values.

We report coverage for users who have made at least one rating. Cyclopath also computes routes for users who have made no ratings, but we report as we do for two reasons. First, it allows for comparison with other recommender systems. Second, in general, it is impossible to compute coverage including users who have rated no edges, because this set contains an unknown and unbounded number of anonymous users. However, some algorithms can make predictions for a user without any ratings from that user: in these cases, coverage for both classes of users is equal. We point out such algorithms in our discussion below. In particular, when we claim an algorithm has "universal" coverage, that means it can predict any user,edge pair even if that user has rated no edges.

We were concerned about the sparsity of our dataset (recall that only 0.093% of user,edge pairs actually had a rating, 14 times sparser than MovieLens or Netflix). We tried to address this by inferring some ratings. Our approach is similar to Good et al., who inferred rating profiles based on movie attributes such as genre and keywords [48]. The intuition is that unrated edges "nearby" rated edges are likely to have similar ratings. Figure 6.2 on the next page gives an example; two ratings are shown, one on Scholars Walk and one on Church St. We can reasonably infer that ratings on topologically adjacent edges of the same street have the same rating, as do edges adjacent to those, and so on. The example shows two iterations of this propagation – we have inferred 7 ratings, increasing the number of rated edges from 2 to 9.

We did this for 10 iterations, creating 62,237 inferred ratings for a total of

[3]We removed 1,400 ratings on expressways made before the system was complete, when rating was the only way to prevent certain expressways from being offered in routes. We reason that working around this missing feature was unrepresentative of typical rating behavior.

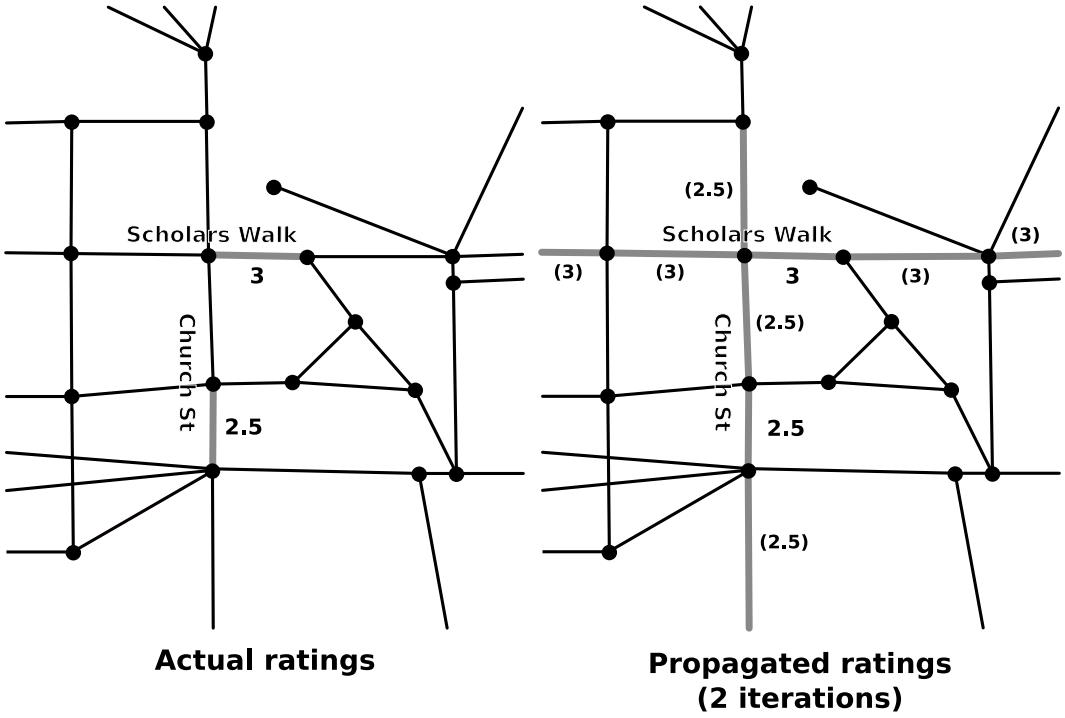

Figure 6.2: Ratings propagation. This illustration shows our (failed) technique to address the sparsity of ratings.

128,711; at least 49,004 edges then had at least one rating, up from 34,817. This reduced sparsity from 0.00093 to 0.0018, still 7 times more sparse than MovieLens or Netflix and much less improvement than we expected. We speculate that this is for two reasons. First, some ratings were near the "end" of a street, so it was not possible to propagate for the full 10 iterations. Second, rated edges are not uniformly distributed. For example, if a user has rated 20 edges, they are frequently in an adjacent sequence already (e.g., Hennepin Ave. between 20th and 40th St.), so propagation only expands the first and last ratings in the sequence – yielding closer to 20 inferred ratings than 400. Unsurprisingly in light of this, the technique turned out to be not very helpful: coverage was improved only modestly, and accuracy was damaged. Details are given below as algorithm results are summarized.

Algorithm	Without propagation MAE	RMSE	Cov.	With propagation MAE	RMSE	Cov.
Objective algorithms						
Objective-Simple	0.766	1.033	1.000 ⋆		no effect	
Objective-CBF	0.706	1.029	0.883 ⋆		no effect	
Simple averaging algorithms						
Global-Mean	0.888	1.126	1.000 ⋆	0.910	1.127	1.000 ⋆
User-Mean	0.804	1.018	0.996	0.823	1.026	0.997
Edge-Global-Mean.1	0.683	0.972	0.206 ⋆	0.688	0.962	0.293 ⋆
Edge-Global-Mean.2	0.630	0.867	0.076 ⋆	0.650	0.879	0.143 ⋆
Edge-Global-Mean.3	0.594	0.821	0.040 ⋆	0.626	0.843	0.086 ⋆
Edge-Global-Mean.5	0.556	0.777	0.018 ⋆	0.606	0.820	0.035 ⋆
Edge-Global-Mean.10	0.443	0.653	0.003 ⋆	0.566	0.776	0.011 ⋆
Collaborative filtering						
CF-User.pearson	0.617	0.897	0.030	0.637	0.907	0.058
CF-User.euclid,unadj	0.554	0.847	0.074	0.494	0.765	0.101
CF-User.euclid,adj	0.626	0.918	0.074	0.629	0.919	0.106
CF-User.cosine,unadj	0.674	0.958	0.077	0.677	0.946	0.128
CF-User.cosine,adj	0.610	0.887	0.050	0.623	0.889	0.086
CF-Item.pearson	0.415	0.642	0.007	0.468	0.670	0.021
CF-Item.euclid,unadj	0.577	0.792	0.079	0.616	0.810	0.132
CF-Item.euclid,adj	0.577	0.792	0.079	0.616	0.810	0.132
CF-Item.cosine,unadj	0.632	0.855	0.069	0.674	0.880	0.116
CF-Item.cosine,adj	0.565	0.784	0.060	0.591	0.793	0.109
Cluster-based personalized algorithms						
Type-Global-Mean	0.683	0.902	1.000 ⋆	0.704	0.908	1.000 ⋆
Type-User-Mean	0.507	0.726	0.632	0.520	0.736	0.637
Cluster-Global-Mean	0.644	0.864	0.995 ⋆	0.656	0.868	0.994 ⋆
Cluster-User-Mean	0.440	0.656	0.516	0.455	0.666	0.562

Table 6.2: Results of predicting edge ratings. Coverage includes only users who have made at least one rating. Algorithms which can make predictions for a user without any ratings from that user are marked with ⋆. For these algorithms, coverage for the (unbounded) set of no-rating users is equal to what is listed; for the others, coverage is undefined for no-rating users.

6.6.2 Results

Table 6.2 on the previous page lists our full results, divided into sections according to algorithm class. Below, we discuss key results and interpretations for each algorithm.

Objective algorithms. *Objective-CBF* has coverage of 88% and MAE of 0.706, while *Objective-Simple* has universal coverage (its motivation) and MAE 0.766. While both of these algorithms have excellent coverage, their accuracy is uncompetitive compared with other algorithms.

Simple averaging algorithms. *Global-Mean* has universal coverage and MAE of 0.888, while *User-Mean* has 99.6% coverage (the gap from universal arising from a few users having rated only edges which were later deleted from the system) and MAE of 0.804.

Accuracy of *Edge-Global-Mean.n* is surprisingly good; even copying a single other user's rating (*Edge-Global-Mean.1*) is slightly more accurate (MAE of 0.683) than a moderately sophisticated objective algorithm (*Objective-CBF* with MAE 0.706). *Edge-Global-Mean.2* and higher are competitive with the more complex algorithms below, with MAE from 0.630 to 0.443. However, coverage is poor. *Edge-Global-Mean.1* can be stretched to 29% using ratings propagation with essentially no penalty in accuracy (MAE declines by 0.005), but other algorithms are superior in both coverage and accuracy, and a higher threshold quickly leads to very poor coverage (*Edge-Global-Mean.2* has just 14% coverage even with ratings propagation). Furthermore, *Edge-Global-Mean.1* is vulnerable to trivial attacks – for most of the edges in the system, one user can destroy the ability of the system to predict ratings for that edge by entering bogus ratings (which then become the predicted ratings for everyone).

Collaborative filtering. Our hypotheses regarding the CF algorithms we tried were confirmed, more dramatically than we expected – these algorithms are essentially useless. One variant (*CF-Item.pearson* without ratings propa-

gation) is the most accurate of anything we tried (MAE of 0.415), but coverage is extremely poor, just 0.7%. Other variants have better, though still poor, coverage (at most 13%), but other algorithms have much better coverage and better accuracy.

Cluster-based personalized algorithms. *Type-Global-Mean* achieves universal coverage and an MAE of 0.683, while *Type-User-Mean*'s coverage is 63% and MAE 0.507. *Cluster-Global-Mean*, has 99.5% coverage and MAE of 0.644, while *Cluster-User-Mean*, 52% coverage and MAE 0.440.

These algorithms perform very well. The *-User-Mean* algorithms have excellent accuracy and good coverage, while the *-Global-Mean* algorithms have good accuracy and universal coverage. Additionally, they have several further advantages. They are amenable to an easy cold start process – simply rate 1 edge in each cluster – and users might not even need to rate any actual items if the *Type-* algorithms are used (e.g., "what rating do you typically give to bike paths?"). They are difficult to attack because users' ratings do not influence other users in *-User-* and have weak influence in *-Global-*. Finally, the models are very simple and can be easily communicated to lightweight client software like web applications; in Cyclopath, this would be useful as ratings could be predicted even for newly created edges that have not yet been saved, without a round-trip to the server.

Summary. Figure 6.3 on the next page summarizes the performance of algorithms we tried. Both simple averaging algorithms and collaborative filtering were failures, due to poor coverage and/or inferior accuracy. We do note that Cyclopath's existing ensemble predictor – try *Edge-Global-Mean.1*, *Objective-CBF*, and *Objective-Simple* in turn, falling through until a prediction is obtained – is in the correct order. We identify five plausible algorithms which have sufficient coverage and accuracy to be of use: *Objective-CBF* and all four clustering algorithms. However, as we note above, accuracy differences at Stage 1 might not be meaningful at Stage 2. We next explore this question.

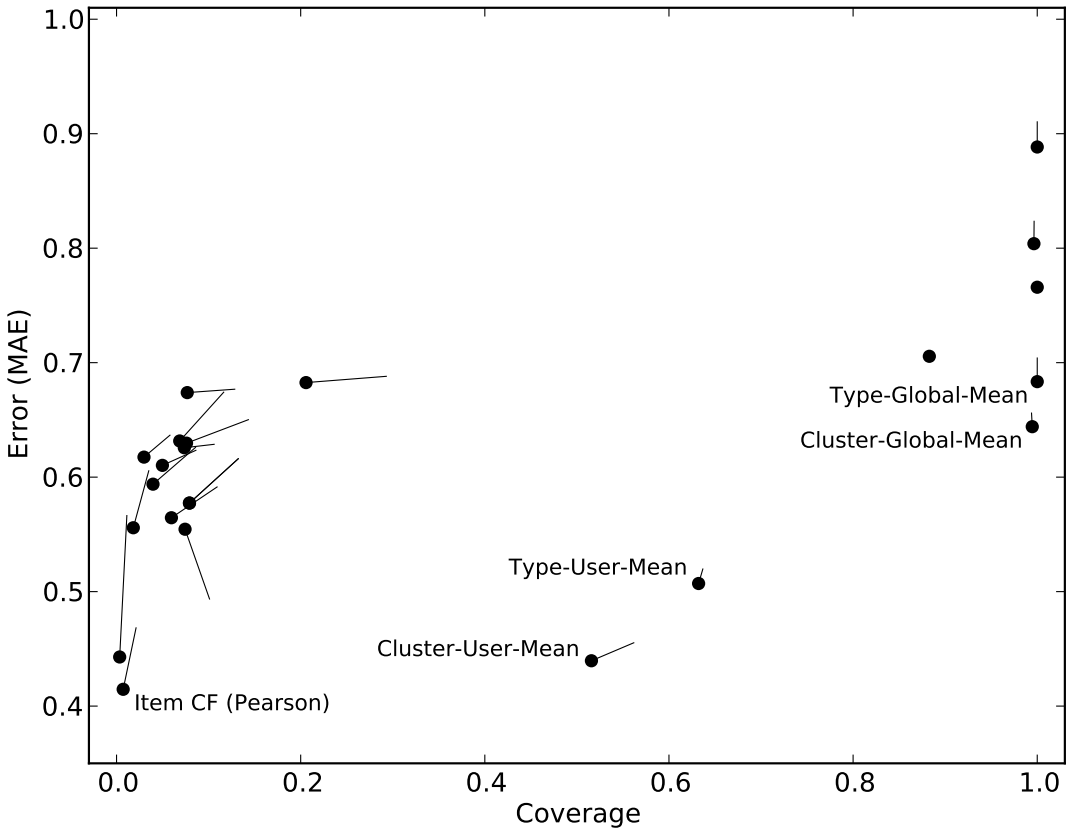

Figure 6.3: Scatter plot of prediction algorithm coverage vs. error. Algorithms lower and more to the right are more desirable. Dots are results without ratings propagation, while the flags show the effect of propagation. Key algorithms are labeled.

6.7 Stage 2: Comparing node-level decisions

Prediction accuracy is irrelevant if different algorithms do not lead to different graph search decisions. This section describes our evaluation of the agreement of the five plausible algorithms in this context, using the metrics and ideas explained in Section 6.4.2 on page 106.

Stage 1 above produced as set of user,edge,value predictions for each fold in the 10-fold cross-validation. We implemented a Python program which, in each fold and for each two algorithms, counted the number of pairs of adjacent edges where the two algorithms gave the same predicted rating ordering and the number where the orderings were different. This yielded the pairwise best edge agreement (PBEA) for the two algorithms.

	Objective-CBF	*Type-Global-Mean*	*Cluster-Global-Mean*	*Type-User-Mean*	*Cluster-User-Mean*
Objective-CBF		0.74	0.71	0.73	0.70
Type-Global-Mean	0.74		0.85	0.96	0.83
Cluster-Global-Mean	0.71	0.85		0.82	0.92
Type-User-Mean	0.73	0.96	0.82		0.85
Cluster-User-Mean	0.70	0.83	0.92	0.85	

Table 6.3: Pairwise edge order agreement (PBEA) between "plausible" rating prediction algorithms.

Our results are summarized in Table 6.3. *Objective-CBF* and the cluster-based algorithms have PBEA at most 74% (and estimated BEA of 52%, suggesting that graph search algorithms would make the same low-level decisions a least 52% of the time), a large difference. Similarly, the two *Type-* algorithms vs. *Cluster-* have PBEA up to 85% (BEA 70%), a smaller but we believe meaningful difference. Interestingly, however, the agreement between the *-User-* and *-Global-* algorithms is rather high. *Type-Global-Mean* and *Cluster-Global-Mean* have PBEA 96% (BEA 91%), while *Type-User-Mean* and *Cluster-User-Mean* have slightly lower PBEA of 92% (BEA 85%). These results are similar to the PBEA-like computation in Section 6.3.1, where users' pairwise agreement was 87%. Given that graph search algorithms will often consider thousands of nodes during a route request, we expect many points of disagreement to be encountered, even between algorithms with quite high BEA, leading to potentially different routes.

Future research should explore whether these differences result in different routes (Stage 3 of our framework). Another consideration, however, is that even if *-User-* results in no quantitative benefit over *-Global-*, perhaps the former should be implemented anyway, because personalized routing may be useful in attracting users.

6.8 Summary

These results represent a step forward in two distinct ways. First, we have shown the value of personalization in a representative open content community and demonstrated specific techniques for creating a personalized interpretation of highly structured user-contributed content. Second, our framework and algorithm evaluations are a significant advance for recommender systems research. Specifically, we identified simple edge rating prediction algorithms which work excellently (the four cluster-based personalized algorithms introduced in Section 6.5.4) and showed that the collaborative filtering algorithms we tried, which work very well in other contexts, are useless in this one. In practice, route recommender systems should use an ensemble predictor [13] in order to leverage more accurate algorithms but still achieve the necessary 100% coverage (in our case, combining the four cluster-based algorithms).

There is much future work. Most obviously, our algorithms should be analyzed at Stage 3 to determine if the Stage 2 differences we identified translate to (meaningfully) different routes, and to find a mapping from PBEA to path differences (as minimizing the effort of Stage 3 would be useful). Other machine learning techniques should be tried, in particular additional clustering algorithms beyond DBSCAN. Also, rating predictions based on sensing could be quite fruitful; for example, bicyclists could carry accelerometers to measure bumpiness, power sensors, physiological sensors to measure stress, etc.

Furthermore, route finding presents a number of additional complications. First, the cost of a path may differ from the sum of its edge weights. We have noted above that turns have costs as well, and that if any edge on a path has a rating below some threshold, the entire path may be unacceptable even if its total cost is the best. Further extensions may be appropriate; for example, each additional unit of poorly-rated edge distance may have exponential additional cost. This notion is similar to the finding of Ziegler et al. that the desirability of a set of books differed from the average desirability of the individual books in the set [127].

Second, users have asked us several times for loop routes; e.g., "I am at Point A and I want a one-hour loop that I would enjoy". This is not a tra-

ditional graph search, and what does one minimize? Finally, we might not want to compute a minimum-cost path, but perhaps instead the one with the smallest distance on poorly-rated edges, the highest poorest rating, or some other metric.

In general, we believe that there are significant nonlinearities present which we have not accounted for. Future research should explore this issue, and new graph search algorithms may be required.

Chapter 7

Implications

In judging any new technology, it is necessary to ask three critical questions: Can it be built? Is it useful? And who cares? Accordingly, we do this for the geowiki system we created and the ideas supporting it.

Can it be built? Yes, geowikis and their extensions, computational geowikis and personalized geowikis, can be built. Cyclopath, a system we designed, created, and maintain, is an existence proof of this. It is a successful geowiki, computational geowiki, and personalized geowiki, a system which has served the real-world information needs of thousands of bicyclists over the past two years. We describe in this thesis the design, implementation, and user experience of Cyclopath and the significant challenges and innovations involved in this process. Further, Wikipedia, a much larger and more mature system, is evidence that even very large wikis can be successfully constructed.

Is it useful? Yes, geowikis are useful. We identified a representative geographic community – bicyclists – and our analysis of this community shows that geowikis are powerful geographic knowledge sharing tools which are more effective than previous techniques. Cyclists told us in surveys and interviews that Cyclopath is useful (both when described as a concept and with respect to the live system), and they have also demonstrated this perception of utility by their robust usage of Cyclopath.

Furthermore, we have quantified this utility, introducing three new ways of measuring value in wikis, which go beyond prior methods by focusing on the

value of information to those who consume it rather than counting producer actions. In systems like Cyclopath where user input feeds an algorithm, we measure the value of an edit by measuring its impact on the output of the algorithm; in particular, user work has shortened the average route computed using Cyclopath by 1 kilometer, or about 7%. Other systems, such as Wikipedia, do not have such an algorithm. In these cases, we measure value by using view rate as a proxy, using this metric to estimate who contributes Wikipedia's value: 0.1% of the editors contribute 44% of its value, nearly half. Similarly, we measure the impact of damage by estimating views; 0.7% of Wikipedia article views can be expected to be of articles in a damaged state.

Additionally, we present strategies for increasing the utility of geowikis. First, we demonstrate techniques for eliciting more user contributions and focusing this work where it is needed. We show that not only does visually highlighting specific work opportunities increase the amount of work done (not because it eases the task but because it seems to make it more compelling), users also do much additional work beyond what they are prompted to do. Also, we show that familiarity increases the quantity of work elicited, but only for some types of work; in others (such as our node repair task), it does not matter at all, i.e., anyone who is available is equally likely to do the task.

Second, we increase the value of geowiki information by personalizing it. Both subjective interview and survey responses from cyclists and analysis of their expressed ratings show that a personalized interpretation of Cyclopath's wiki information (specifically, personalized routes) is important. We created a framework for evaluation personalized route finders in a geographic context and evaluated several algorithms using the framework, showing that classic rating prediction algorithms successful in other contexts (collaborative filtering) were of no use, while simple clustering-based algorithms were accurate and effective.

Who cares? Geography matters to many people. There are numerous communities whose core activities depend on geographic information. We have previously noted Antarctic polar science, neighbors' daily life, and natural resource managers; additional possibilities are disaster response, doctors tracking epidemics, and many others. The common thread is that geographic

information is critically important, but access to this information is limited, incomplete, or awkward because it is distributed among the community members themselves and otherwise unavailable. But geowikis can help. Our results show that they are an effective way of gathering and disseminating such distributed geographic information; thus, our results are broadly applicable.

We identify some general design implications. First, algorithms matter beyond the value of their output, as the impact of user work on algorithm output can be used to measure value. Thus, it may be useful for system designers to introduce algorithms which consume user work even if the output is not itself useful. Second, top users contribute far beyond their numbers, so they need to be kept happy. However, turnover is inevitable, so it is also important to keep new users flowing in and help them learn to become productive (and perhaps top users). This avoids system collapse when key users leave. Third, our results have implications for eliciting more work. When soliciting work, it isn't always necessary to ask users to do the work actually needed; sometimes, one can ask for other (perhaps more appealing) work, and the needed work will be done anyway. And some types of work require familiarity, while others don't; knowing which is which can help route work requests to appropriate users. Finally, the rate of damaging work increases exponentially in a popular system, but this growth can perhaps be stemmed using strategies like automatic damage-repair software, and heuristics like using high impact on computational geowiki algorithm output can flag interesting (and potentially troublesome) edits.

Future work is, as always, quite extensive. Obviously, Cyclopath's coverage should be expanded; we are beginning to do this, having recently (July 2010) created a new instance of the system serving the Denver metro area. Also, as geowikis are a new collaboration technology, it is useful to identify whether prior results hold in this new context. For example, the Cyclopath team is already exploring user lifecycles [83] and tagging behavior [105] in order to compare them to other contexts. It is also useful to test geowikis in additional domains. A particular domain of interest is the neighbor-centric geowiki, which seeks simply to provide a geographically contextualized online space for neigh-

bors. This may not require editability of the transportation network, which would yield much simpler software (with correspondingly greater reach).

An intriguing extension of the computational geowiki model (i.e., driving system-defined algorithms with user-contributed data) is to drive user-contributed algorithms with user-contributed data; for example, users could contribute a new route finding algorithm. This certainly fits the spirit of the wiki model and appears useful. However, it introduces significant challenges. In general, algorithms must run on the server to avoid downloading the entire database, but they must do so in a sandbox which insulates user-defined algorithms from the rest of the system, as arbitrary software can be arbitrarily broken or malicious. Private data must be kept private, and performance of the main system should not suffer noticeably.

These research ideas, as well as non-research geowiki implementations created simply because the technology is useful, show great promise. In short, the geographic wiki model introduced in this thesis forms a significant step forward, offering meaningfully better techniques for collecting, maintaining, and sharing geographic data.

Appendix A

References

This thesis includes much or most of [85, 86, 87, 88]. Also, Chapter 6 is a working version of a paper in preparation with Shilad Sen and Loren Terveen.

[1] Lada A. Adamic and Bernardo A. Huberman. Zipf's Law and the Internet. *Glottometrics*, 3:143–150, 2002. http://www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf.

[2] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. WWW*, pages 261–270, 2007.

[3] Eugene Agichtein et al. Finding high-quality content in social media. In *Proc. Web Search and Web Data Mining*, pages 183–194, 2008.

[4] Alexa Internet, Inc. Quick tour, 2007. http://www.alexa.com/site/help/quicktour.

[5] Katie Angle. Annual recognition dinner 2007. *Twin Cities Bicycling Club News*, 7, 2007. http://www.pdewebworks.com/tcbc/news/.

[6] Liliana Ardissono et al. Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence*, 17(8):687–714, 2003.

[7] Jaime Arguello et al. Talk to me: Foundations of successful individual-group interactions in online communities. In *Proc. CHI*, pages 959–968, 2006.

[8] Marko Balbanovic and Yoav Shoham. Combining content-based and collaborative recommendation. *CACM*, 40:66–72, 1997.

[9] Shivanand Balram and Suzana Dragićević, editors. *Collaborative Geographic Information Systems*. Idea Group Publishing, 2006.

[10] Ed Barsotti and Gin Kilgore. The road network is the bicycle network: Bicycle suitability measures for roadways and sidepaths. In *Proc. Transport Chicago*, 2001. http://bikelib.org/roads/roadnet.htm.

[11] Morgan J. Bearden. The National Map Corps, 2007. http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Bearden_paper.pdf.

[12] Benjamin B. Bederson et al. Enhancing in-car navigation systems with personal experience. In *Proc. Transportation Research Board*, pages 1–11, 2008.

[13] Robert M. Bell and Yehuda Koren. Lessons from the Netflix Prize challenge. *SIGKDD Explorations*, 9(2):75–79, 2007.

[14] Bird Studies Canada. Christmas bird count, 2008. http://www.bsc-eoc.org/national/cbcmain.html.

[15] Black Duck Software, Inc. Estimating the development cost of open source software, Apr 2009. http://www.blackducksoftware.com/development-cost-of-open-source.

[16] Joel Booth et al. A data model for trip planning in multimodal transportation systems. In *Proc. Extending Database Technology*, pages 994–1005, 2009.

[17] Axel Bruns and Sal Humphreys. Wikis in teaching and assessment. In *Proc. WikiSym*, pages 25–32, 2005. http://doi.acm.org/10.1145/1104973.1104976.

[18] Susan L. Bryant et al. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proc. GROUP*, pages 1–10, 2005.

[19] Jenna Burrell and Geri K. Gay. E-graffiti: Evaluating real-world use of a context-aware system. *Interacting with Computers*, 14(4):301–312, 2002.

[20] Brian S. Butler. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information Systems Research*, 12(4):346–362, 2001. http://isr.journal.informs.org/cgi/content/abstract/12/4/346.

[21] Magnus Cedergren. Open content and value creation. *First Monday*, 8(8), 2003. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1071/991.

[22] Jilin Chen et al. Make new friends, but keep the old: Recommending people on social networking sites. In *Proc. CHI*, pages 201–210, 2009.

[23] Lynn Cherny. *Conversation and Community: Chat in a Virtual World*. CSLI Publications, 1999. http://portal.acm.org/citation.cfm?id=520127.

[24] Changyan Chi et al. Dandelion: Supporting coordinated, collaborative authoring in wikis. In *Proc. CHI*, pages 1199–1202, 2010.

[25] Ed Chi. Long tail of user participation in Wikipedia, 2007. http://asc-parc.blogspot.com/2007/05/.

[26] Boreum Choi et al. Socialization tactics in Wikipedia and their effects. In *Proc. CSCW*, pages 107–116, 2010. http://doi.acm.org/10.1145/1718918.1718940.

[27] Dan Cosley et al. How oversight improves member-maintained communities. In *Proc. CHI*, pages 11–20, 2005.

[28] Dan Cosley et al. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proc. CHI*, pages 1037–1046, 2006.

[29] Dan Cosley et al. SuggestBot: Using intelligent task routing to help people find work in Wikipedia. In *Proc. IUI*, pages 32–41, 2007.

[30] Scott Counts and Marc Smith. Where were we: Communities for Sharing Space-Time Trails. In *Proc. ACMGIS*, 2007.

[31] Abhinandan S. Das et al. Google News personalization: Scalable online collaborative filtering. In *Proc. WWW*, pages 271–280, 2007.

[32] Uri Dekel. A framework for studying the use of wikis in knowledge work using client-side access data. In *Proc. WikiSym*, pages 25–30, 2007. http://doi.acm.org/10.1145/1296951.1296954.

[33] Daniel Delling et al. Engineering route planning algorithms. In *Algorithmics of Large and Complex Networks*, pages 117–139. Springer-Verlag, 2009.

[34] Sara Drenner et al. Insert movie reference here: A system to bridge conversation and item-oriented web sites. In *Proc. CHI*, pages 951–954, 2006. http://doi.acm.org/10.1145/1124772.1124914.

[35] Sarah Elwood. Volunteered geographic information: Key questions, concepts and methods to guide emerging research and practice. *GeoJournal*, 72(3-4):133–135, 2008. Special issue.

[36] William Emigh and Susan C. Herring. Collaborative authoring on the Web: A genre analysis of online encyclopedias. In *Proc. HICSS*, 2005.

[37] Frederik Espinoza et al. GeoNotes: Social and navigational aspects of location-based information systems. In *Proc. Ubicomp*, pages 2–17, 2001.

[38] Martin Ester et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. KDD*, pages 226–231, 1996.

[39] Alexander Felfernig et al. Personalized user interfaces for product configuration. In *Proc. IUI*, pages 317–320, 2010.

[40] Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954.

[41] Andrew J. Flanagan and Miriam J. Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4):137–148, 2008. http://www.springerlink.com/content/t77154837870p37t/?p=d89da9dfbf6d4a5ba8152a825d92996f&pi=1.

[42] Andrea Forte and Amy Bruckman. From Wikipedia to the classroom: Exploring online publishing and learning. In *Proc. Learning Sciences*, pages 182–188, 2006. http://portal.acm.org/citation.cfm?id=1150061.

[43] Bruno Frey and Stephan Meier. Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review*, 94(5):1717–1722, 2004. http://dx.doi.org/10.1257/0002828043052187.

[44] Simon Funk. Netflix update: Try this at home, 2006. http://sifter.org/simon/journal/20061211.html.

[45] R. Stuart Geiger and David Ribes. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proc. CSCW*, pages 117–126, 2010. http://doi.acm.org/10.1145/1718918.1718941.

[46] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

[47] Scott A. Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *CoRR*, abs/cs/0508082, 2005. http://arxiv.org/abs/cs/0508082.

[48] Nathan Good et al. Combining collaborative filtering with personal agents for better recommendations. In *Proc. AAAI*, 1999.

[49] Michael F. Goodchild. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69, 2007.

[50] Michael F. Goodchild and Rajan Gupta. Workshop on volunteered geographic information, 2007. http://www.ncgia.ucsb.edu/projects/vgi/.

[51] Ilya Grigorik. Dissecting the Netflix dataset, Oct 2006. http://www.igvita.com/2006/10/29/dissecting-the-netflix-dataset/.

[52] Mark Hall et al. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[53] F. Maxwell Harper et al. Social comparisons to motivate contributions in an online community. In *Proc. Persuasive Technology*, pages 148–159, 2007. http://dx.doi.org/10.1007/978-3-540-77006-0_20.

[54] F. Maxwell Harper et al. Predictors of answer quality in online Q&A sites. In *Proc. CHI*, pages 865–864, 2008. http://doi.acm.org/10.1145/1357054.1357191.

[55] Peter E. Hart et al. A formal basis for the heuristic determination of minimum cost paths. *TOSSC*, 4(2):100–107, 1968.

[56] Carleen Hawn. Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Affairs*, 28(2):361–368, 2009. http://www.ncbi.nlm.nih.gov/pubmed/19275991.

[57] Susan C. Herring. Gender and democracy in computer-mediated communication. In *Computerization and Controversy: Value Conflicts and Social Choices*, pages 476–489. Academic Press, 2 edition, 1996.

[58] Will Hill et al. Recommending and evaluating choices in a virtual community of use. In *Proc. CHI*, pages 194–201, 1995. http://doi.acm.org/10.1145/223904.223929.

[59] Raphael Hoffmann et al. Amplifying community content creation with mixed initiative information extraction. In *Proc. CHI*, pages 1849–1858, 2009. http://doi.acm.org/10.1145/1518701.1518986.

[60] Harald Holone et al. Aspects of personal navigation with collaborative user feedback. In *Proc. NordiCHI*, pages 182–191, 2008.

[61] Charles Lee Isbell, Jr. et al. Cobot in LambdaMOO: A social statistics agent. In *Proc. AAAI*, pages 36–41, 2000.

[62] Steven Karau and Kipling Williams. Social loafing: A meta-analytic review and theoretical integration. *Personality and Social Psychology*, 65(4):681–706, 1993. http://www.sciencedirect.com/science/article/B6X01-46T9BSD-2S/2/1ee754dc608f2c941f46a88d8621fd1a.

[63] Brian Keegan and Darren Gergle. Egalitarians at the gate: One-sided gatekeeping practices in social media. In *Proc. CSCW*, pages 131–134, 2010. http://doi.acm.org/10.1145/1718918.1718943.

[64] Stephen F. King and Paul Brown. Fix my street or else: Using the Internet to voice local public service concerns. In *Proc. Theory and Practice of Electronic Governance*, pages 72–80, 2007. http://doi.acm.org/10.1145/1328057.1328076.

[65] Aniket Kittur. He says, she says: Conflict and coordination in Wikipedia. In *Proc. CHI*, 2007.

[66] Aniket Kittur et al. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. alt.CHI*, 2007. http://www-users.cs.umn.edu/~echi/papers/2007-CHI/2007-05-altCHI-Power-Wikipedia.pdf.

[67] Aniket Kittur and Robert E. Kraut. Beyond Wikipedia: Coordination and conflict in online production groups. In *Proc. CSCW*, pages 215–224, 2010.

[68] William H.K. Lam and G. Xu. A traffic flow simulator for network reliability assessment. *Advanced Transportation*, 33(2):159–182, 1999.

[69] Cliff Lampe et al. Follow the reader: Filtering comments on Slashdot. In *Proc. CHI*, pages 1253–1262, 2007. http://doi.acm.org/10.1145/1240624.1240815.

[70] Cliff Lampe and Paul Resnick. Slash (dot) and burn: Distributed moderation in a large online conversation space. In *Proc. CHI*, pages 543–550, 2004. http://doi.acm.org/10.1145/985692.985761.

[71] John Leen. It's time to bike, 2010. http://google-latlong.blogspot.com/2010/03/its-time-to-bike.html.

[72] Quannan Li et al. Mining user similarity based on location history. In *Proc. ACMGIS*, pages 1–10, 2008.

[73] Kimberly S. Ling et al. Using social psychology to motivate contributions to online communities. *Computer-Mediated Communications*, 10(4), 2005. http://dx.doi.org/10.1111/j.1083-6101.2005.tb00273.x.

[74] Chor Pang Lo and Albert K. W. Yeung. *Concepts and Techniques of Geographic Information Systems*. Prentice Hall, 2nd edition, 2006.

[75] Edwin A. Locke and Gary P. Latham. Building a practically useful theory of goal setting and motivation: A 35-year odyssey. *American Psychologist*, 57(9):705–717, 2002. http://view.ncbi.nlm.nih.gov/pubmed/12237980.

[76] Michael Ludwig et al. Path selection: A novel interaction technique for mapping applications. In *Proc. CHI*, 2009.

[77] Macalester College. Collecting bikeways data in Minnesota, 2006. http://www.macalester.edu/geography/civic/BikewaysProject_Geog364_Spring2006.pdf.

[78] Alan M. MacEachren. Cartography and GIS: Facilitating collaboration. *Progress in Human Geography*, 24(3):445–456, 2000.

[79] Kevin Marks. Power laws and blogs, 2003. http://homepage.mac.com/kevinmarks/powerlaws.html.

[80] Lorraine McGinty and Barry Smyth. Shared experiences in personalized route planning. In *Proc. FLAIRS*, pages 111–115, 2002.

[81] NAVTEQ, Inc. NAVTEQ map reporter, 2007. http://mapreporter.navteq.com/.

[82] OSGeo. WMS tiling client recommendation, Jul 2010. http://wiki.osgeo.org/index.php?title=WMS_Tiling_Client_Recommendation&oldid=48480.

[83] Katherine Panciera et al. Lurking? Cyclopaths? A quantitative lifecycle analysis of user behavior in a geowiki. In *Proc. CHI*, pages 1917–1926, 2010. http://portal.acm.org/citation.cfm?id=1753326.1753615.

[84] Moon-Hee Park et al. Location-based recommendation system using Bayesian user's preference model in mobile devices. In *Proc. Ubiquitous Intelligence and Computing*, pages 1130–1139, 2007.

[85] Reid Priedhorsky et al. Creating, destroying and restoring value in Wikipedia. In *Proc. GROUP*, pages 259–268, 2007.

[86] Reid Priedhorsky et al. How a personalized geowiki can help bicyclists share information more effectively. In *Proc. WikiSym*, pages 93–98, 2007.

[87] Reid Priedhorsky et al. Eliciting and focusing geographic volunteer work. In *Proc. CSCW*, 2010.

[88] Reid Priedhorsky and Loren Terveen. The computational geowiki: What, why, and how. In *Proc. CSCW*, 2008.

[89] Sasank Reddy et al. Biketastic: Sensing and mapping for better biking. In *Proc. CHI*, pages 1817–1820, 2010.

[90] Yuqing Ren et al. Applying common identity and bond theory to design of online communities. *Organization Studies*, 28(3):377–408, 2007.

[91] Paul Resnick et al. GroupLens: An open architecture for collaborative filtering of netnews. In *Proc. CSCW*, 1994.

[92] Francesco Ricci. Travel recommender systems. *IEEE Intelligent Systems*, 17(6):55–57, 2002.

[93] Jean-Paul Rodrigue et al. *The Geography of Transport Systems*. Routledge, 2006.

[94] Badrul Sarwar. Item-based collaborative filtering reccomendation algorithms. In *Proc. WWW*, pages 285–295, 2001.

[95] J. Ben Schafer et al. Collaborative filtering recommender systems. In *The Adaptive Web*, chapter 9, pages 291–324. Springer-Verlag, 2007.

[96] Wendy A. Schafer et al. Designing the next generation of distributed, geocollaborative tools. *Cartography and Geographic Information Science*, 52(2):81–100, 2005.

[97] Toby Segaran. *Programming Collective Intelligence*. O'Reilly, 2007.

[98] Shilad Sen et al. tagging, communities, vocabulary, evolution. In *Proc. CSCW*, pages 181–190, 2006. http://doi.acm.org/10.1145/1180875.1180904.

[99] Shilad Sen et al. The quest for quality tags. In *Proc. GROUP*, pages 361–370, 2007. http://doi.acm.org/10.1145/1316624.1316678.

[100] Marc A. Smith and Andrew T. Fiore. Visualization components for persistent conversations. In *Proc. CHI*, pages 136–143, 2001.

[101] Alex Sorton and Thomas Walsh. Urban and suburban bicycle compatibility street evaluation using bicycle stress level. In *Proc. Transportation Research Board*, 1994.

[102] Besiki Stvilia et al. Assessing information quality of a community-based encyclopedia. In *Proc. Information Quality*, pages 442–454, 2005. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.6243&rep=rep1&type=pdf.

[103] Aaron Swartz. Who writes Wikipedia?, Sep 2006. http://www.aaronsw.com/weblog/whowriteswikipedia.

[104] TomTom. Get to know TomTom MapShare, 2007. http://www.clubtomtom.com/general/get-to-know-tomtom-mapshare%E2%84%A2/.

[105] Fernando Torre et al. bumpy, caution with merging: An exploration of tagging in a geowiki. In *Proc. GROUP*, 2010.

[106] U.S. Department of Transportation. National household travel survey, 2001. http://nhts.ornl.gov/2001/.

[107] Fernanda B. Viégas and Judith S. Donath. Chat circles. In *Proc. CHI*, pages 9–16, 1999.

[108] Fernanda B. Viégas et al. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. CHI*, 2004.

[109] Jakob Voss. Measuring Wikipedia. In *Proc. Scientometrics and Infometrics*, 2005. http://hapticity.net/pdf/nime2006_180-works_cited/MeasuringWikipedia2005.pdf.

[110] Jimmy Wales. Jimmy Wales talks Wikipedia, Dec 2005. http://writingshow.com/?pageid=91.

[111] Alan Wexelblat and Pattie Maes. Footprints: History-rich tools for information foraging. In *Proc. CHI*, pages 270–277, 1999.

[112] Steve Whittaker et al. The dynamics of mass interaction. In *Proc. CSCW*, pages 257–264, 1998.

[113] Wikimedia Foundation. Stop word list, 2006. http://meta.wikimedia.org/w/index.php?title=Stopwordlist&oldid=313397.

[114] Wikipedia. Vandalism in progress, 2002. http://en.wikipedia.org/w/index.php?title=Wikipedia:Archive/Wikipedia:Vandalisminprogress/History&oldid=188844.

[115] Wikipedia. Page requests per day, 2004. http://stats.wikimedia.org/EN/TablesUsagePageRequest.htm.

[116] Wikipedia. Bots/status, 2006. http://en.wikipedia.org/w/index.php?title=Wikipedia:Bots/Status&oldid=133390929.

[117] Wikipedia. Awareness statistics, 2007. http://en.wikipedia.org/w/index.php?title=Wikipedia:Awarenessstatistics&oldid=129505430.

[118] Wikipedia. Podcast, 2007. http://en.wikipedia.org/w/index.php?title=Podcast&oldid=133330735.

[119] Wikipedia. Siegenthaler controversy, 2007. http://en.wikipedia.org/w/index.php?title=Seigenthaler_controversy&oldid=132296396.

[120] Wikipedia. Size of wikipedia, 2007. http://en.wikipedia.org/w/index.php?title=Wikipedia:SizeofWikipedia&oldid=127300877.

[121] Wikipedia. Wikipedia in culture, 2007. http://en.wikipedia.org/w/index.php?title=Wikipedia_in_culture&oldid=133473824.

[122] Wikipedia. Wikipedia, Jul 2010. http://www.wikipedia.org/.

[123] Yahoo! One billion answers served!, May 2010. http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/.

[124] Fan Yang and Zhi-Mei Wang. A mobile location-based information recommendation system based on GPS and WEB2.0 services. *WSEAS Transactions on Computers*, 8(4):725–734, 2009.

[125] YouTube. Oops pow surprise, 2010. http://youtube-global.blogspot.com/2010/03/oops-pow-surprise24-hours-of-video-all.html.

[126] Yu Zheng et al. GeoLife 2.0: A location-based social networking service. In *Proc. Mobile Data Management*, pages 357–358, 2009. Demo.

[127] Cai-Nicolas Ziegler et al. Improving recommendation lists through topic diversification. In *Proc. WWW*, pages 22–32, 2005.