On the "Localness" of User-Generated Content

Brent Hecht[*] **and Darren Gergle**[*†]
Northwestern University
[*]Dept. of Electrical Engineering and Computer Science, [†] Dept. of Communication Studies
brent@u.northwestern.edu, dgergle@northwestern.edu

ABSTRACT
The "localness" of participation in repositories of user-generated content (UGC) with geospatial components has been cited as one of UGC's greatest benefits. However, the degree of localness in major UGC repositories such as Flickr and Wikipedia has never been examined. We show that over 50 percent of Flickr users contribute local information on average, and over 45 percent of Flickr photos are local to the photographer. Across four language editions of Wikipedia, however, we find that participation is less local. We introduce the spatial content production model (SCPM) as a possible factor in the localness of UGC, and discuss other theoretical and applied implications.

Author Keywords
User-generated content, volunteered geographic information, Wikipedia, Flickr, local, user behavior

ACM Classification Keywords
H.5.3. [Information Interfaces and Presentation]: Group and Organization Interfaces – collaborative computing, computer-supported cooperative work

General Terms
Human Factors

INTRODUCTION

In the bygone era of Web 1.0, a search for "Albany, California" would have returned nearly guaranteed local knowledge such as Albany's city homepage or a local newspaper. These days, however, it is likely that the information returned may be drawn from a user-generated content (UGC) repository such as Wikipedia or Flickr. UGC has become a predominant source of information about scores of geographic features (i.e., cities, towns, national parks, landmarks, etc.). In general, it has been assumed that this UGC also represents local knowledge. Geographic information expert Michael Goodchild, for instance, writes:

"…The most important value of [user-generated geographic information] may lie in what it can tell us about *local* activities… that go unnoticed by the world's media, about life at the *local* level. It is in that area that [user-generated geographic information] may offer the most interesting, lasting and compelling value" [3] (emphases added).

But who actually participates in creating these vital geographic resources? Is dominance truly local, as it is with city homepages, or are outsiders more frequently voicing their opinions? As our technologies increasingly appropriate geographic information from large-scale UGC repositories[1], the question of who is providing that information becomes fundamental.

In this paper, we investigate the assumption that participation in UGC repositories is local. We address this question across five different large-scale UGC repositories. In addition, we introduce the idea of spatial content production models (SCPMs) to describe how the particular uses and features of UGC repositories might influence the degree of "localness". This allows us to characterize, for example, the differences between the "you have to be there" model of a UGC repository like Flickr with the more easily traversable "flat Earth" model of something like Wikipedia. Finally, theoretical and applied implications are summarized, and future work is discussed.

BACKGROUND AND RELATED WORK

Geographic UGC has been shown to be incredibly useful as a resource for collaborative technologies (see Priedhorsky et al. [16] and others [11, 13, 14]). It has also been applied, for example with Flickr data, to train classifiers to georeference new photos [1, 6] and to automatically learn tagging semantics [12]. Similarly, the spatial data in Wikipedia has been leveraged to understand "self-focus" bias in UGC repositories [7] and test general theories of geographic information [8].

The in-depth study of the *nature* of geographic information in this context has been restricted to Wikipedia. Hardy [4]

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

[1] The term "volunteered geographic information" (VGI) is sometimes used to refer to geographic UGC, particularly within the discipline of geography.

shows that editors of Wikipedia follow a power law in their number of contributions of geographic UGC. Lieberman and Lin [10] demonstrate that the convex hull of edited geographic articles (specifically, the locations of the geographic entities they describe) is likely somewhat small for a large minority of registered English Wikipedia users. However, the degree of localness to the actual user is not considered.

DATA PRE-PROCESSING STAGE

In order to investigate the degree to which participation in UGC repositories is local, we draw upon data from five different UGC repositories: Flickr and four language editions of Wikipedia (English, Catalan, Norwegian, and Swedish). The following describes the processing done in order to prepare the data for analysis.

Flickr

As is evidenced by Flickr's own map interface to its photos[2], a large portion of Flickr's dataset has been geotagged by its users, either automatically through a GPS-enabled camera (such as the iPhone) or manually. We used Yahoo!'s API access to Flickr to download approximately a year's worth of geotagged photo metadata beginning in May of 2008, resulting in information about 10+ million photos.

However, for the purposes of the studies described below, we also needed data about the location of the Flickr users who took these photos. We again accessed the Flickr API to download photographer information using the photographer ID tags included in each of the 10 million photos' metadata. We were particularly interested in the photographer's self-specified location, an optional field in Flickr user profiles. While a small percentage of users did provide this information, it was text-based and often quite colloquial in nature (i.e., "Grand Rapids, U S & A", "Minneapolis-St. Paul, Twin Cities"). This created a problem, as there is no formal gazetteer, to our knowledge, that is capable of handling this type of vernacular spatial data.

Fortunately, Wikipedia has a rich set of this data in the form of Wikipedia redirects, which effectively form a massive mapping table designed to "redirect" users who search for, say, "San Fran" or "San Francisco, USA", to the "San Francisco" article. As such, we were able to leverage these redirects and the process described in the following subsection to identify the latitude and longitude location of a large number of Flickr users. To supplement this process, we also performed a Wikipedia-only Yahoo! Search API query on each colloquial location, and if the first result was identified as a geotagged Wikipedia article, we applied the geotag to the user's location. In the end, we were able to successfully geocode 14,295 photographers who took 185,871 geotagged photos.

[2] http://www.flickr.com/map/

Wikipedia

Our data gathering and preparation approach borrows from existing work, particularly [4] and [7]. We started by using WikAPIdia[3], the hyperlingual Wikipedia API described in [7], which allowed us access to basic Wikipedia data and metadata in a concept-aligned fashion. To attach explicit spatial information to the multilingual data set, we utilized the massive number of geotags provided by Wikipedians themselves, compiled in [9] and also used by Hardy [4].

The above methodologies mine the spatial footprints of Wikipedia articles. What about the locations of contributors? Wikipedia contributors can be broadly split into two classes, anonymous and registered users. While we can mine the IP address of anonymous contributors and use these in IP geolocation, it is extremely difficult or even impossible to discover the position of large numbers of registered Wikipedia users. As such, we omit them from our studies, admittedly a drawback given that they produce a large portion of the content that is read by Wikipedia consumers. Anonymous users are responsible for about 26 percent of content read by visitors to the English Wikipedia [15]. However, analyzing the patterns of this much more spatial data-rich subset of users has merit especially in the context of the "localness" question.

The Problem of Scale

Spatial UGC all too frequently suffers from the "Geoweb Scale Problem" [8]. Wikipedia and Flickr are no exception. In this context, the Geoweb Scale Problem occurs when spatial data schemas only support point-based spatial data representations. For example, the entire state of Alaska is represented in Wikipedia as a single point. Many first-order administrative districts and countries have been reduced to zero dimensions in the same way. We avoid much of this problem by using simple name matching to mostly remove first-order administrative districts and countries from our dataset (e.g., the Wikipedia article "United States" and Flickr users who specified their home location as "England, United Kingdom"), but second-order districts (e,g., counties) and even cities with large areal extent (e.g., "Houston, Texas") still cause difficulties at scales around that of the "radius" of a typical city or county (~50km in the U.S., less elsewhere).

Compounding the issue when IP addresses are considered is the accuracy of IP geolocation. The IP geolocation software used in our study[4] performs at 68 – 79 percent accuracy within 25 miles in the predominant home countries of the languages in our study (and at nearly 100% accuracy on a country scale). Despite our filtering efforts, the end result of the GSP and IP issues is to essentially make random distance estimates at very local scales. To reflect this, we report all of our data accordingly.

[3] http://collabolab.northwestern.edu/wikapidia/

[4] http://www.maxmind.com/app/geolitecity

STUDY OF CONTRIBUTOR BEHAVIOR

For our five repositories, we calculated for each contributor the *mean contribution distance* (MCD). A contributor's MCD is defined as:

$$MCD = \sum_{i=1}^{n} \frac{d(C,c_i)}{n}$$

where C is the specified location of the contributor, and the location of each of C's n contributions is denoted by c_i. This metric has a large benefit over that used in [10] in that each location is effectively weighted by the number of times a contribution is made, an important fact considering Lieberman and Lin's discovery that many Wikipedia users edit their "pet geopages" very frequently. Our distance function d is that of the great circle distance[5].

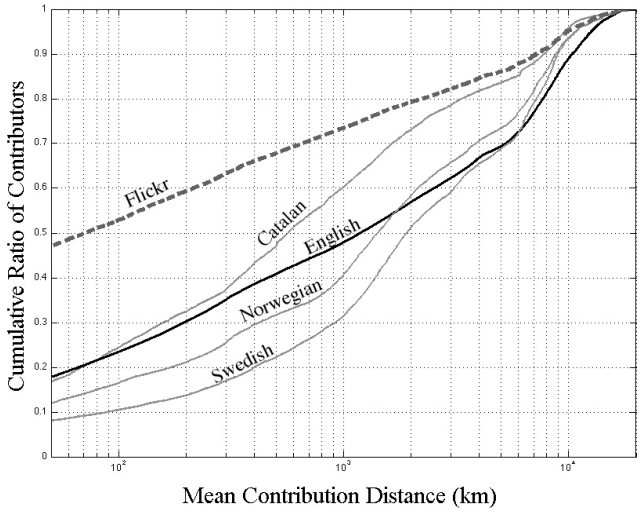

Figure 1. The empirical cumulative distribution of MCDs for each dataset examined, or *cdf*(MCD). Note that the x-axis is on a log scale.

Figure 1 shows the empirical cumulative distribution function of contributors' MCDs. While ~53 percent of Flickr users contribute, on average, content that is 100km or less from their specified home location, this number drops quite a bit for Wikipedia users. The equivalent number for the English Wikipedia, for example, is ~23 percent, although this number is subject to errors in IP geolocation.

Why does this difference between Wikipedia and Flickr exist? We hypothesize that the answer to this question lies in the spatial content production models of each repository. In Wikipedia, "the encyclopedia anyone can edit", contributors simply must posses the desire to add/edit/delete content. In the spatial domain, this means that there exists "total time-space compression" [5], and as such, we can categorize Wikipedia's SCPM as a "flat Earth" model. In other words, it is just as easy for someone in Albany, CA to edit the "Albany, California" page as it is for that person to edit the "Chicken, Alaska" page. This fact is reflected in the much smaller percentage of contributors who edit locally on *average*.

Flickr's "you have to be there" SCPM, on the other hand, more or less requires that contributors have visited the location about which they are contributing. This creates a MCD pattern that begins to resemble offline spatial behavior models, and therefore creates a repository in which local participation is much greater.

While Wikipedia has less local participation compared to Flickr, it is important to note that distance still matters a great deal on Wikipedia's "flat Earth". The data in Figure 1 elaborates on [10], which found that the convex hull of edited spatial articles tends to be somewhat small for a large minority of English users.

STUDY OF THE LOCALNESS OF EACH REPOSITORY

We now turn our attention to the "localness" of participation across entire repositories, rather than individual contributors' behaviors. In other words, we indirectly incorporate the power law found by Hardy (and confirmed to apply to Flickr) into our analyses. Figure 2 is similar to Figure 1 except that instead of showing the distribution of contributors' MCDs, it shows the distribution of distances from all contributions to their respective contributors.

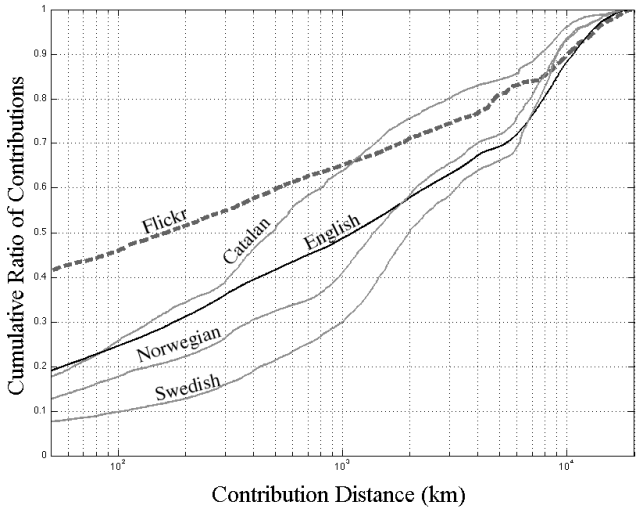

Figure 2. Empirical cdfs of contributor-to-contribution distances for all contributions. In other words, *cdf*(*d*(*C*,*c*)) for all (*C*,*c*) pairs in each repository.

Figure 2 demonstrates that the spatial contribution behavior is relatively independent of number of contributions. If it were not, we would see a significant difference between Figures 1 and 2. For instance, ~47 percent of Flickr photos are taken within 100km from their photographer (Figure 2), while we saw that ~53 percent of photographers take photos within 100km on average (Figure 1). There is one small exception: the Catalan Wikipedia line crosses the Flickr line

[5] We use a spherical Earth assumption to speed the calculation of great circle distance. For the purposes of this paper, the errors introduced by doing so are minimal.

at ~1000km. Analyzing this phenomenon in detail is a subject for future work.

CONCLUSION

The results shown above, combined with those from related work, have many applied and theoretical implications. If Goodchild is correct in his statement that the main benefit of volunteered geographic information is local knowledge, designers of geographic UGC communities will want to learn from the differences between Wikipedia and Flickr. One suggestion would be to adopt SCPMs that "decompress time-space" in content production, as is naturally done in the process of taking photographs. For instance, consider geospatially-oriented Wiki UGC communities such as OpenStreetMap, [13] and [16]. If these applications wish to ensure more local knowledge and avoid the Wikipedia phenomenon, they could require that users upload information from GPS units rather than allowing them to encode their knowledge using a web interface.

Additionally, despite Goodchild, many papers on Flickr in the strict data mining space view Flickr as a tourist photo database. Our results (particularly Figure 2) suggest that the "tourist assumption" about Flickr is false: ~47 percent of Flickr photos are within 100km of the photographer's home location.

An important theoretical direction that must be investigated involves the importance of UGC repositories as sources of *place* information [2]. The degree to which these repositories are defined by locals versus outsiders is an important question in this respect. While we have answered this question in the Flickr context, the dynamics of Wikipedia participation make this more difficult. A deeper inspection of the *content* on Wikipedia pages is warranted. We were only able to measure *participation* (the relationship between the two has been a subject of much research, e.g. [15]). Of course, another major question in this area necessitates looking more deeply at contributors rather than simply classifying them as local or non-local (e.g., socioeconomic status).

Before concluding, it is important to note the limitations of this work. We only use a very small set of Flickr data for which we were able to definitively locate both the photo and the photographer. We are also subject to the very valid criticisms of IP geolocation, particularly with regard to poor accuracy. We have accounted for these issues where possible and discussed the implications where it is not. Future work will attempt to confirm these findings on a larger scale and with greater precision.

ACKNOWLEDGMENTS

We would like to thank Lauren Scissors, Alan Clark, Nada Petrović, and Brian Keegan. This work was supported in part by a Royal E. Cabell fellowship, National Science Foundation grant #0705901, and the Robert and Kaye Hiatt fund.

REFERENCES

[1] Crandall, D., Backstrom, L., Huttenlocher, D. and Kleinberg, J. (2009). Mapping the World's Photos. *WWW '09*.

[2] Dourish, P. (2006). Re-space-ing place: "place" and "space" ten years on. *CSCW '06*, 299-308.

[3] Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. GeoJournal, 69 (4). 211-221.

[4] Hardy, D., (2008). Discovering behavioral patterns in collective authorship of place-based information. *Internet Research 9.0*.

[5] Harvey, D. (1989) The *Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Wiley-Blackwell.

[6] Hays, J. and Efros, A.A. (2008). IM2GPS: estimating geographic information from a single image. *CVPR '08*.

[7] Hecht, B. and Gergle, D. (2009). Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. *Communities and Technologies 2009*. 11-21.

[8] Hecht, B. and Moxley, E. (2009). Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. *COSIT '09*. 88-105.

[9] Kühn, S. WikiProjekt Georeferenzierung, 2009.

[10] Lieberman, M.D. and Lin, J. (2009). You are where you edit: Locating Wikipedia users through edit histories. *ICWSM '09*. 106-113.

[11] Ludford, P.J., Priedhorsky, R., Reily, K. and Terveen, L.G. (2007). Capturing, Sharing, and Using Local Place Information. *CHI '07*, 1235-1244.

[12] Moxley, E., Kleban, J., Xu, J. and Manjunath, B.S. (2009). Not All Tags are Created Equal: Learning Flickr Tag Semantics for Global Annotation. *ICME '09*.

[13] Mummidi, L.N. and Krumm, J. (2008). Discovering points of interest from users' map annotations. *GeoJournal*, 72 (3-4). 215-227.

[14] Naaman, M., Nair, R. and Kaplun, V. (2008). Photos on the Go: A Mobile Application Case Study. *CHI '08*. 1739-1748.

[15] Priedhorsky, R., Chen, J., Lam, S.T., Panciera, K., Terveen, L.G. and Riedl, J. (2007). Creating, Destroying, and Restoring Value in Wikipedia. *GROUP 2007*. 259-268.

[16] Priedhorsky, R. and Terveen, L.G. (2008). The computational geowiki: what, why, and how. *CSCW '08*. 267-276.