Designing Better Location Fields in User Profiles

Ting-Yu Wang
GroupLens Research
University of Minnesota
twang@cs.umn.edu

F. Maxwell Harper
GroupLens Research
University of Minnesota
harper@cs.umn.edu

Brent Hecht
GroupLens Research
University of Minnesota
bhecht@cs.umn.edu

ABSTRACT
Twitter, Facebook, Pinterest and many other online communities ask their users to populate a *location field* in their user profiles. The information that is entered into this field has many uses in both industry and academia, with location field data providing valuable geographic context for operators of online communities and playing key roles in numerous research projects. However, despite the importance of location field entries, we know little about how to design location fields effectively. In this paper, we report the results of the first controlled study of the design of location fields in user profiles. After presenting a survey of location field design decisions in use across many online communities, we show that certain design decisions can lead to more granular location information or a higher percentage of users that fill out the field, but that there is a trade-off between granularity and the percent of non-empty fields. We also add context to previous work that found that location fields tend to have a high rate of non-geographic information (e.g. Location: "Justin Bieber's Heart"), showing that this result may be site-specific rather than endemic to all location fields. Finally, we provide evidence that verifying users' location field entries against a database of known-valid locations can eliminate toponym (place name) ambiguity and any non-geographic location field entries while at the same time having little effect on field population rate and granularity.

Categories and Subject Descriptors
H5.m. [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Keywords
Location field, user profile, geotagging, geographic user-generated content, volunteered geographic information

1. INTRODUCTION
User profiles in online communities very often contain what is known as a *location field* [11]. This is true of many popular social media sites such as Twitter, Pinterest, Flickr, and Foursquare, but also of other types of communities like eBay and Github (see Figure 1 for examples). Moreover, the use of the location field spans Eastern and Western cultures, with popular Eastern communities like Kaixinwang, Renren, and Cyworld also incorporating these fields into their user profiles.

Researchers and operators of online communities have found location field entries to be invaluable in a number of ways. First and foremost, members of many communities are reluctant to tag individual pieces of content with their specific location; only 1.5-3.2% of tweets have geotags, for instance [19]. Location field entries can provide a rough estimate of the location of users who are not among the small group of people that frequently post or update their location. Indeed, in order to provide users of its Search API with geographic context for the 96.8-98.5% of tweets that are not geotagged, Twitter mines the location field in its users' profiles [29]. Moreover, many research projects have taken a similar approach, with the toponyms (i.e. place names) in location fields being used to assign geographic references to social media (e.g. [9,11,17]).

Location field entries can also help researchers avoid major confounds introduced by user mobility. Given that a person's geotagged social media (e.g. photos, tweets) may be widely

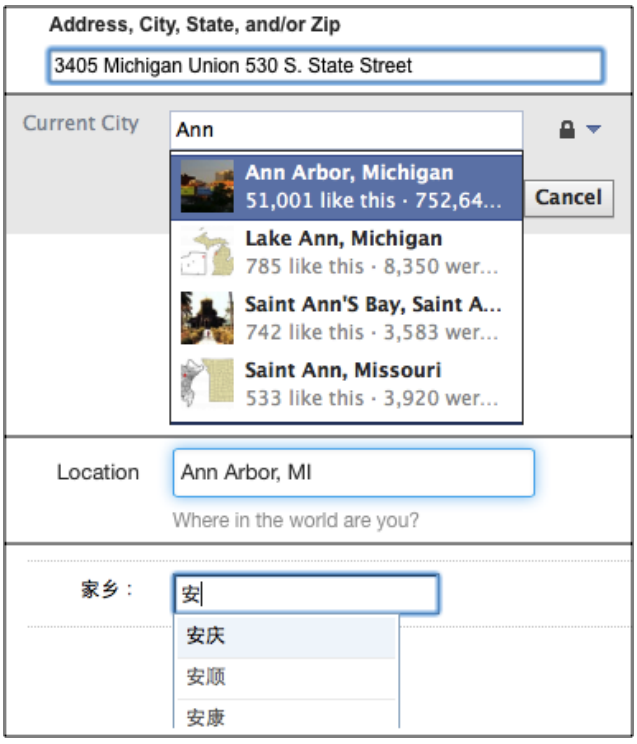

Figure 1. Examples of location fields from a number of online communities (top to bottom: Yelp, Facebook, Twitter, Kaixinwang). Note that all use different prompts, have different lengths, and take different approaches to validation. (Kaixinwang's prompt translates to "Hometown").

GROUP'14, November 9–12, 2014, Sanibel Island, Florida, USA.

ACM 978-1-4503-3043-5/14/11…$15.00
http://dx.doi.org/10.1145/2660398.2660424

dispersed due to vacations, business trips, and other forms of travel, the information in the location field is frequently used as a simple heuristic for determining the location(s) in which a social media user can be considered a "local" [10]. For instance, this approach has been utilized for studying the demographic makeup of social media communities (e.g. [12,18]), understanding geographic patterns in social networks (e.g. [14,15,22]), and inferring the home locations of the online community members from the content (e.g. tweets) they produce (e.g. [20,23]).

However, data from location fields also has a number of important disadvantages. Research has shown that the toponyms in location fields tend to be of relatively course geographic *granularity*, with most field entries being city names [11,20]. In addition, far from all users fill out their location field (i.e. there is a low field *population rate*) and many users input non-geographic information like "Justin Bieber's Heart" and "preferably anywhere but here" [11] (i.e. low *geographicness*).

The goal of this paper is to inform the design of location fields that can minimize these disadvantages, thereby providing more and higher quality location information to online community operators and researchers alike. Despite the importance of the information entered into location fields, no work has examined the relationship between location field design and the quality of the information entered into them. Indeed, as we will show, there is extensive variation across online communities in location field design, with sites using different prompts, having different verification strategies, and using fields of highly varied length, among other differences (Figure 1).

Below, we report the results of a series of controlled experiments targeted at identifying the most effective approaches to location field design. Our objective was to understand the relationship between the design of location fields and the three major limitations of location field data that have been identified in the literature:

1) *Population Rate*, or the percent of users who fill out the location field in their user profile.
2) *Granularity*, or the geographic scale of the location information that is entered (e.g. city-level, address, country-level).
3) *Geographicness*, or the percent of location field entries that contain valid geographic information rather than non-geographic entries like 'Justin Bieber's Heart' and 'preferably anywhere but here'.

Through these experiments, we are able to establish that simple changes in location field design can increase the amount or the granularity of location field entries, but that online community operators must generally negotiate a trade-off between the two. In addition, we show that concerns about geographicness in location fields may only be valid in certain online communities rather than being endemic to location fields as a whole. However, we also report findings that can inform the design of location fields in communities where geographicness is indeed a problem. Namely, validating users' entries against a database of legal places had no effect on population rate or granularity, but completely removes concerns about geographicness. Validation has the added benefit of a-priori disambiguation of place names, reducing the need for and error introduced by geocoding.

Below, we begin with a discussion of related work. In the subsequent section, we discuss the results of a survey of location field design decisions made in 18 different online communities. Next, we introduce our overall experimental approach and walk the reader through our three controlled studies on location field design. Finally, we conclude with a discussion of design implications and future work.

2. RELATED WORK

The work most related to our research can largely be grouped into two areas: (1) studies of location disclosure behavior and (2) research that utilizes location field data.

The rapid increase in the popularity of location-aware technologies like smartphones has led to a strong interest in location disclosure behavior in the literature. For instance, Consolvo et al. [6] and Wiese et al. [30] used questionnaires to understand why and with whom people share their locations. Tsai et al. [28] identified that feedback can improve user comfort levels with location sharing. Other researchers have studied the effect of incentives (e.g. [27]) and place naming strategies (e.g. [16]) on location sharing behavior.

No existing work has investigated the effect of location field design on location disclosure in location fields. In addition to the applied utility of a study directly targeted at location fields as outlined above, our work also sheds light on location sharing behavior in the context of a categorically different type of location information: low temporal resolution location information. Several schema of location information in online communities exist, and all of them distinguish between *high temporal resolution* information and *low temporal resolution* information. High temporal resolution information (e.g. Hecht and Gergle's [10] "Contribution" location type and Schultz et al.'s [24] "Tweet Location" type) is the focus of existing location disclosure work, which tends to look at phenomena like Foursquare check-ins and the sharing of real-time locations. Location fields, on the other hand, are infrequently updated and contain low temporal resolution location information (e.g. Hecht and Gergle's "Contributor" location type and Schultz et al.'s "User's Residence" location type).

Location field data is used in a wide variety of research projects from a number of different disciplines. In addition to the work noted above on geographic social networks, demographic analysis, and location inference, other researchers have used location field data to, for instance, study online activism surrounding major political events (e.g. [9,17]), examine the prevalence of local perspectives in user-generated content (e.g. [10]), and monitor public health [3,4,7]. Further, the inference attack problem – in which a user's location is predicted from her social media – has attracted considerable interest beyond the papers cited above, with location field data frequently serving as ground truth (e.g. [2,5,11,13,21]). Much of this location field-based inference work involves using location field entries to ground *geographic topic models*, which have a number of other uses such as modeling linguistic variation across space [8]. Finally, it is important to note that *any* study (or application) that uses the Twitter Search API implicitly uses location field information.

All of the above work (including studies and applications that use the Twitter Search API) could benefit from more and higher quality location field entries, the end goal of this paper. For instance, with more granular entries, those who study the geographic properties of online social networks would be able to understand these properties at a more local scale. The same can be said for research that takes a geographic approach to understanding the demographics of online community members

and uses tweets to monitor public health. Along the same lines, the Twitter Search API could provide more precise geographic context for more tweets if more Twitter users populated their location fields with more granular toponyms. In addition, increasing the geographicness of location field entries could open up new applications for location field data. Researchers have eschewed the use of location field data out of concern for geographicness in a number of areas (e.g. emergency management [25]).

3. LOCATION FIELD DESIGN SPACE

Our first step in understanding the effect of location field design decisions on the field's population rate and the granularity and geographicness of its entries was to survey the design space of location fields in a variety of online communities. Examining 18 communities that are popular in Eastern and Western cultures, we identified five key location field design dimensions:

- *Prompt:* the text that appears to the top or the left of the location field.
- *Length:* the length of the field in number of characters, as measured by the number of "0" characters one can enter before one of the "0" characters is not fully visible.
- *Verification*: whether or not entries are validated against a dataset of known-valid locations (e.g. using a gazetteer).
- *Visibility*: whether the information placed in the location field is public, private, or whether users have control over the extent to which the information is shared with others.
- *Number of fields*: whether the user profile contained a single or multiple location fields. For instance, Twitter uses a single field ("Location") while Kaixinwang uses two ("Current City" and "Hometown"[1]).

Table 1 describes the design choices made by each of the 18 surveyed communities (on their non-mobile websites) along each of these dimensions as of Fall 2013. The table reveals that there is a great deal of diversity in all five dimensions. For instance, Yelp prompts its field with "Address, City, State, and/or Zip", while Twitter uses "Location" (see Figure 1). Location field lengths also range widely, for example, with Twitter adopting a 29-character field and Pinterest using a 49-character field. A similar lack of consensus can be seen with regard to whether or not location fields are verified against known place names, how widely location field information is shared within an online community, and the number of location fields in a user profile.

Using the results of our survey of the location field design space, we developed three experiments to identify the design choices that result in (1) the highest location field population rates, (2) the most granular location information, and (3) the highest degree of geographicness. In these experiments, which are described immediately below, we evaluated the effect of a range of design decisions along all of the key design dimensions outlined above with the exception of the number of fields. Most of the online communities whose location field data has been used in the literature utilize single fields (e.g. Twitter, Foursquare) and we anticipate that our conclusions about single fields will also apply to individual fields on multiple-field profiles.

4. EXPERIMENTS

All experiments were performed in MovieLens[2], a movie-focused online community that has over 100,000 users and, as of Fall

Community	Entry Prompt	Field Length	Verification	Visibility
Pinterest	Location	49	No	Public
Twitter	Location	29	No	Public
Yelp	(1) Address, City, State, and/or Zip (2) My Hometown	(1) 49, (2) 45	(1) Yes, (2) No	(all) Public
Facebook	(1) Current City, (2) Hometown	(1) 27, (2) 27	(1) Yes, (2) Yes	(all) User-Controlled
Foursquare	Location	23	No	Public
LinkedIn	Postal Code	39	Yes	Public
Flickr	(1) Your Hometown, (2) City you live now, (3) Country, (4) 3 letter Airport Code	(1) 26, (2) 26, (3) 26, (4)7	(1) No, (2) No, (3) No, (4) No	(1) Public, (2) (3) User-Controlled, (4) Private
G+	Place lived	57	No	User-Controlled
Bitbucket	Location	49	No	User-Controlled
MeetUp	(1) ZIP, (2) Hometown	(1) 17, (2) 20	(1) Yes, (2) No	(all) Public
Lang-8	Location	Drop-down	Yes	User-Controlled
Ebay	(1) Address, (2) City, (3) Postal Code	(1) 40, (2) 40, (3) 20	(1) No, (2) Yes, (3) Yes	(all) Private
Renren	(1) Location, (2) Hometown	(1) Drop-down, (2) Drop-down	(1) Yes, (2) Yes	(all) User-Controlled
Cyworld	Current City	31	No	User-Controlled
Weibo	Location	Drop-down	Yes	Public
Ameba	(1) Hometown, (2) Haunt, (3) Region In Which You Live	(1) 39 and Drop-down* (2) 39, (3) Drop-down	(1) Hybrid*, (2) No, (3) Yes	(all) User-Controlled
Skype	(1) City, (2) State/Province	(1) 31, (2) 31	(1) No, (2) No	(all) Public
Kaixinwang	(1) Current Location, (2) Hometown	(1) 20, (2) 20	(1) No, (2) No	(all) Public

Table 1. The variation in location field design decisions across 18 online communities. In communities with multiple fields, each field is numbered. Lang-8, Renren, and Weibo use drop-down menus that allow users to select cities in China (drop-down menus implicitly use verification by our definition). *The Ameba "Hometown" field is a hybrid field, with a 39-character free text field and a prefecture(state)-level drop-down menu.

2013, received about 20 registrations per day. When a user signs up for MovieLens, they are invited to input information for their online profile. We manipulated the design of the location field in this process (Figure 3). All experiments were performed between November 2013 and February 2014 and a total of 1,673 users took part.

In our first experiment, we examined the effect of prompt and length on the quality metrics outlined above: population rate, granularity, and geographicness. Next, we looked at the effect of verification against a database of known-valid locations. Finally, we conducted a third experiment that examined the role of visibility on the quality metrics.

When analyzing the location field entries recorded during our experiment for granularity and geographicness, we followed the approach of Hecht et al. [11] in which two coders independently assessed granularity and geographicness. Each entry was assigned granularity codes from the following set: {address, neighborhood, city, intrastate region, state, interstate region, country}. To afford ordinal analysis, the granularity codes were ranked from 0 (address) to 6 (country). Geographicness was treated as a binary. Interrater agreement was above 97% in all cases for both granularity and geographicness. Conflicts were resolved through negotiation between the two coders.

4.1 Experiment 1: Length and Prompt

We analyzed the effect of location field length and prompt on the quantity and quality of location field entries using a between-subjects 3x3 experiment. The three levels of the length factor were set to 30 characters, 50 characters, and 70 characters so as to explore the short, medium, and long areas of the field length spectrum. The levels of the prompt factor were "Location" (e.g. Twitter, Pinterest, Foursquare), "Current City" (e.g. Facebook, Cyworld), and "Address, City, State, and/or Zip" (Yelp). These levels were selected to (1) cover the most common prompts (i.e. the first two) and (2) to span the spectrum of requested granularity. New users of our experiment online community were randomly assigned to one of the nine conditions.

The online movie community received 663 new registrations during the one-month period Experiment 1 was active[3]. Looking at our data a high-level, we saw that overall, 44.0% of users entered a value into the location field. 67.1% of geographic entries were at the city-level. Country was the next most common granularity (16.3%), followed by state (6.0%). 4.6% of users who entered a valid geographic location entered an address.

While research has shown that many users input non-geographic information into the Twitter location field [11,26], we observed a much smaller percentage of non-geographic entries. A total of *four* entries (0.7%) were non-geographic (e.g. "Sears", "here") as opposed to the 16% reported by Hecht et al. [11]. As we will discuss below, this result was not limited to Experiment 1; we saw very little in the way of non-geographic location field entries from the over 1,600 users in our entire study. Because there were so few non-geographic location field entries, we did not consider geographicness in our subsequent analyses.

To understand the effect of field length and field prompt on population rate and granularity, we performed logistic regressions with length and prompt as independent variables. A nominal

[3] The location field entries from six users had to be omitted from granularity and geographicness analysis due to data corruption likely related to character encoding issues.

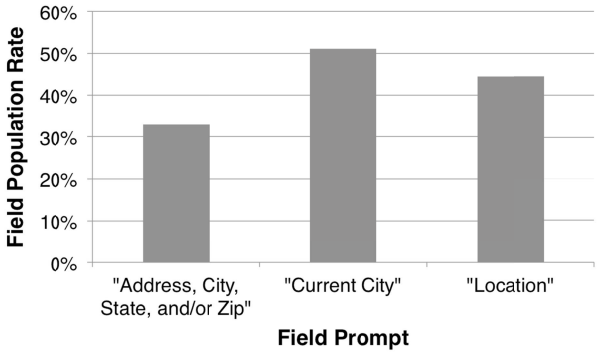

Figure 2. The population rates for each prompt considered. The Facebook-style "Current City" had a rate approximately 1.5 times higher than that for the Yelp-style "Address, City, State, and/or Zip".

logistic regression with field population (has entry / does not have entry) as the dependent variable indicates that prompt has a significant effect on population rate ($\chi^2(2,N=663) = 17.97$, $p < 0.01$). No significant effect could be detected for field length ($\chi^2(2,N=663) = 0.30$, $p = 0.86$) or for a prompt by length interaction ($\chi^2(4,N=663) = 2.43$, $p = 0.66$).

Looking at the relationship between prompt and population rate more closely (Figure 2), a clear trend emerges: while the "Current City" and "Location" prompts have population rates of around 50%, the population rate for "Address, City, State, and/or Zip" – the prompt that requests the most granular information – is only 32.8%.

With regard to granularity, an ordinal logistic regression indicates that prompt has a significant effect on ordinal granularity ($\chi^2(2,N=283) = 73.68$, $p < 0.001$). The regression also indicates a marginal effect for both length ($\chi^2(2,N=283) = 4.80$, $p = 0.09$) and a prompt by length interaction ($\chi^2(4,N=283) = 8.83$, $p = 0.07$).

It is not unexpected that prompts that request different levels of granularity get location field entries of different granularities. For instance, *all* of the address-scale entries in this experiment came from the "Address, City, State, and/or Zip" prompt, making up 19.4% of entries for this prompt. Similarly, 95.1% of "Current City" entries were at the city-scale. "Location" received a much more balanced distribution.

While the prompt main effect may not be surprising, it is revealing of an important larger implication. Namely, while designers of online communities can significantly increase location field granularity using the Yelp-style "Address, City, State, and/or Zip" prompt over the Twitter-style "Location" and Facebook-style "Current City" prompts, our results related to input rate indicate that by requesting an address, input rates will drop. As such, designers of online communities must negotiate a trade-off between granularity and input rate and choose a location field approach that maximizes the outcome that is most important to them.

Returning to the results of the granularity regression, the marginal main effect for length can likely be explained by the drop-off of country-scale entries as soon as length gets longer than 30 characters. 20.8% of locations entered into the 30-character field were country-scale while 13.2% and 14.3% were country-scale for the 50- and 70-character fields, respectively. We also saw a steady increase in addresses as the field length got longer, although the numbers are sufficiently small to prevent us from drawing major

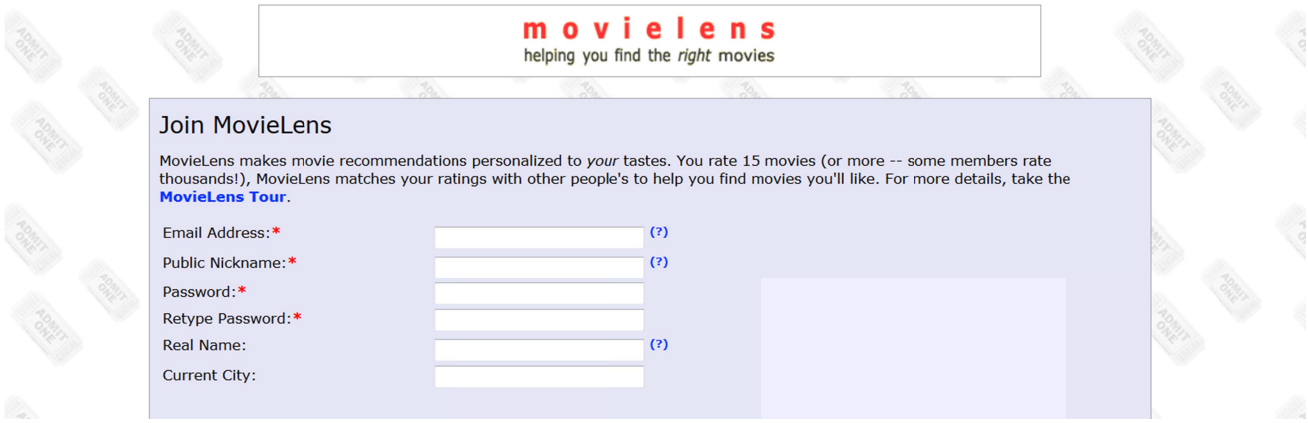

Figure 3. An example of the MovieLens registration page under the 30-character, "Current City" condition in Experiment 1. (Note: By the time of publication, MovieLens will have undergone a major redesign).

conclusions. Both of these findings can likely be explained by there simply being enough room to write a full address in a 70-character field, while a country will nearly always fit in a 30-character field. The marginal interaction effect is likely due to the fact that the country drop-off occurs almost entirely in the "Location" prompt and the address increase occurs entirely in the "Address, City, State, and/or Zip" prompt.

4.2 Verification

In our second experiment, we looked at the effect of verification on the quantity and quality of location field entries. Verification, which is employed by Facebook, Yelp and others, involves checking location field entries against a database of locations (e.g. cities, addresses, etc.) known to be valid. This process is typically executed using a drop-down auto-complete functionality, with users not being allowed to save entries that do not match a valid location.

The necessary result of verification is that 100% of location field entries will be of a geographic nature ("Justin Bieber's Heart" is not likely included in any database of valid locations). However, the effect of verification on population rate and granularity is not clear. For instance, users may shy away from entering information if they cannot enter a colloquial name for a location (e.g. "Singa" for Singapore) or if they are forced to fully disambiguate their entries (e.g. writing "Springfield, IL" versus just "Springfield"). Similarly, they may change the information they enter due to verification, for instance writing "Illinois" instead of "Springfield, IL".

To test the effect of verification on location field population rate and granularity, we conducted a 2x3 between-subjects experiment similar to Experiment 1. The verification factor had two levels: verification and no verification. Our verification implementation was modeled closely on Facebook's but we extended the database of known valid locations to include location types other than cities (e.g. addresses, states, countries) using Google's Geocoding API[4]. We also considered the prompt factor (and the same three levels) in this experiment due to its strong effects in the previous experiment. Length was fixed at 50. Our verification experiment ran for one month, during which time 819 users registered for the online movie community.

[4] https://developers.google.com/maps/documentation/geocoding/

Examining the effect of verification on location field population rate, we performed a nominal logistic regression with field population (has entry / does not have entry) as the dependent variable and both field prompt and verification as independent variables. No significant effect could be detected for either verification ($\chi^2(1,N=819) = 0.31$, $p = 0.58$) or a verification by prompt interaction ($\chi^2(2,N=819) = 2.95$, $p = 0.23$). (A significant main effect for prompt was found again, providing additional support for our conclusions from the first experiment, $\chi^2(2,N=819) = 9.76$, $p < 0.01$.) At a descriptive level, we found that the population rate was 58.6% for the no verification condition and about 1.4% lower (57.2%) for the verification condition.

We identified a similar result with regard to the relationship between verification and granularity. An ordinal logistic regression with ordinal granularity as the dependent variable and the same two independent variables revealed no significant effect for verification ($\chi^2(1,N=329) = 1.70$, $p = 0.19$) or for a verification by prompt interaction ($\chi^2(2,N=329) = 0.71$, $p = 0.70$). For instance, considering the "Location" prompt, in the verification condition, 44.8% of users provided city-level entries and 48.2% of users provided country-level entries. The equivalent numbers for the no verification condition were 48.2% and 46.6%, respectively. We did see a moderate (non-significant) effect for verification in one prompt: in "current city" fields, 72.9% of entries were city-level or more local in the verification condition, while 90.1% were in the no verification condition.

These findings have potentially important implications for the design of location fields. Most obviously, they indicate that operators of online communities may be able to achieve 100% geographicness without drastically affecting the location field population rate or the granularity of location field entries, although research on a larger online community is needed to increase confidence in this conclusion.

If indeed verification has little effect on population rate and granularity, the benefits would extend beyond geographicness. Namely, verification eliminates the issue of *toponym ambiguity* (i.e. place name ambiguity), a well-known challenge in the natural language processing and geographic information retrieval communities. A critical preprocessing step in the use of location field entries by both researchers and practitioners involves the use of a *geocoder*, which converts place names into machine-readable geospatial representations (e.g. latitude and longitude coordinates). One of the fundamental challenges in the development of geocoders is handling toponym ambiguity. For instance, if a

member of an online community enters "London" into their location field (as many of our users did), the geocoder must figure out whether they are referring to the London in England, the London in Canada (with over 300K people), or one of the many other places named "London" around the world. The problem gets even worse with place names like "Danville", which is an entry by a user in this experiment. There are over a dozen cities named Danville in the United States, and none of them are an obvious first choice for a geocoding system.

Verification allows the system designer to require users to disambiguate their location field entries manually. A user who types in "Danville" is forced to choose among a list of possible senses of the toponym in, for instance, a drop-down menu (employed in both Facebook and our implementation of verification). In other words, verification reduces to zero the error introduced by geocoders due to toponym ambiguity. As such, our results suggest that, like non-geographic entries, toponym ambiguity in the processing of location field entries may be able to be entirely eliminated without drastic effects on population rates or granularity.

4.3 Visibility

Our final experiment examined the effect of the visibility of location field information on population rate and granularity. In this experiment, we tested whether users would change their location field entries if they were told these entries would be shared. This experiment had three visibility conditions: *publicly visible*, *not visible*, and *no information*. In the *publicly visible* condition, text appeared below the location field box that informed the user that the information in the field would be publicly accessible on their user profile. In the *not visible* condition, the text instead informed the user explicitly that the location information would remain private. Finally, in the *no information* condition, users were not told anything related to the visibility of the location information. The experiment ran for one week and 191 users participated. In order to capture any interaction effects with location field prompt, we also varied the prompt using the same three levels as before.

Just as was the case with verification, visibility played little role in population rate. A nominal logistic regression predicting whether or not a user entered information into the field (field population) revealed no significant main effect for visibility ($\chi^2(2,N=191) = 0.78$, $p = 0.68$) or for a visibility by prompt interaction ($\chi^2(4,N=191) = 4.04$, $p = 0.40$).

We did, however, see marginally significant results when examining the relationship between visibility and granularity. An ordinal logistic regression indicated that visibility ($\chi^2(2,N=86) = 4.53$, $p = 0.10$) and a visibility by prompt interaction ($\chi^2(4,N=86) = 9.11$, $p = 0.06$) have a marginally significant effect on the granularity of location field entries[5].

Examining the main effect more closely, we saw that the *publicly visible condition* had more low-granularity entries. For instance, while 54% percent of entries in that condition were less granular than the less city-level (e.g. country-level), the equivalent numbers for the *not visible* and *no information* conditions were 27% and 38% respectively. This result echoes what has been seen with high temporal resolution location sharing, where Lin et al. [16] found that participants will share less granular locations with people with whom they have fewer connections, e.g. strangers.

[5] We again saw a significant main effect for prompt ($\chi^2(2,N=191) = 15.45$, $p < 0.001$).

5. DISCUSSION AND FUTURE WORK

As discussed above, entries in Twitter's location field have proven essential to numerous studies and systems (including Twitter's Search API). As such, it is a useful thought experiment to reflect on the effects of Twitter adopting the design implications of each of our experiments. Our results suggest that if Twitter were to change its location field prompt ("Location") to one that requests more granular information like Yelp's ("Address, City, State, and/or Zip"), the large group of researchers and practitioners who use Twitter's location field data either directly or implicitly through the Twitter Search API would have access to more granular information to incorporate into their studies and systems. On the other hand, our results also suggest that this would reduce the percent of users who fill out the location field.

Our results also suggest that the incorporation of verification into Twitter's location field (like Facebook has done) would not have an enormous cost in terms of granularity and field population rate, but would eliminate non-geographic location field entries (16% of entries on Twitter [11]) and all issues with toponym ambiguity. This would result in a large increase in the accuracy of geocoders when they are applied to Twitter location field entries. Since the application of a geocoder is a nearly universal step in the pre-processing of these entries, verification would result in significant improvements to the many research projects and technologies that rely on Twitter location field data.

This paper takes a traditional (non-critical) geographic information perspective on location field design. That is, it is concerned with increasing the quantity and quality of location field entries so that they may be more useful for a wide variety of studies and systems. However, designers of certain online communities may want to consider factors other than quantity and quality of geographic information. For instance, some designers may not want to disallow users from entering non-geographic information like "Justin Bieber's Heart" into their location fields, for instance to allow for greater self-expression in user profiles. Examining users' motivations for entering non-geographic information and developing approaches to support this behavior while reducing the large problems related to non-geographic information in location fields [11] (e.g. geocoders' tendency to return real latitude and longitude coordinates for non-geographic entries) is an important direction of future research.

As our work is the first investigation of location field design, there are several additional important directions of future work. Existing location disclosure research on high temporal resolution location information has found that people's location sharing preferences vary depending on the group of people with whom their location is shared (e.g. [1,16]). It would be useful to see whether the same occurs with low temporal resolution information, and if so, whether the behaviors are different than those that have been observed with high temporal resolution information. The online movie community that supported this study does not have social network features, but our study could be easily repeated and extended to look at difference audiences in online communities such as Facebook (e.g. share with "Public", "Your Friends) and Google Plus (e.g. share with certain circles versus others).

Finally, another important area of future work relates to multiple location fields. Many sites include multiple location fields in their user profiles, and this study did not examine the interaction between these fields. Does having more than one field affect

population rates? Granularity? Geographicness? In addition, some of these multiple location field communities often request information that may be outside of the current temporal context (e.g. Flickr and Facebook's "Hometown" field). Examining the effect of "currentness" would shed additional light on location field design.

7. CONCLUSION

In this paper, we demonstrated that the design of a location field in a user profile has an effect on the field's population rate and the granularity of its entries, which are critical to many systems and studies. In particular, through a series of controlled experiments, we demonstrated that the choice of location field prompt can result in higher granularity or higher field population rates, but that there is a trade-off between the two. We also saw evidence that designers of online communities can include verification in location fields without having a large negative effect on population rate or granularity. This suggests that toponym ambiguity and non-geographic entries can be eliminated without huge costs. Finally, as opposed to what has been found on Twitter, we identified only a few location field entries that were non-geographic in nature, suggesting that the geographicness issue found by Hecht et al. [11] is online community-specific rather than endemic to location fields in general.

8. ACKNOWLEDGEMENTS

The authors would like to thank our colleagues in GroupLens Research, and particularly Loren Terveen, Joe Konstan, and the MovieLens development team. This work was supported in part by NSF IIS-0808692, a 3M Non-Tenured Faculty Award (NTFA), and a Yahoo! ACE Award.

9. REFERENCES

1. Benisch, M., Kelley, P.G., Sadeh, N., and Cranor, L.F. Capturing Location-privacy Preferences: Quantifying Accuracy and User-burden Tradeoffs. *Personal Ubiquitous Comput. 15*, 7 (2011), 679–694.
2. Bergsma, S., Dredze, M., Durme, B.V., Wilson, T., and Yarowsky, D. Broadly improving user classification via communication-based name and location clustering on twitter. *NAACL-HLT '13*, (2013).
3. Broniatowski, D.A., Paul, M.J., and Dredze, M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLoS ONE 8*, 12 (2013), e83672.
4. Burton, S.H., Tanner, K.W., Giraud-Carrier, C.G., West, J.H., and Barnes, M.D. "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information. *Journal of Medical Internet Research 14*, 6 (2012), e156.
5. Cheng, Z., Caverlee, J., and Lee, K. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *CIKM '10: 19th ACM International Conference on Information and Knowledge Management*, (2010).
6. Consolvo, S., Smith, I.E., Matthews, T., LaMarca, A., Tabert, J., and Powledge, P. Location Disclosure to Social Relations: Why, When, & What People Want to Share. *CHI '05*, (2005), 81–90.
7. Dredze, M., Paul, M.J., Bergsma, S., and Tran, H. Carmen: A Twitter Geolocation System with Applications to Public Health. *AAAI-13 Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, (2013).
8. Eisenstein, J., O'Connor, B., Smith, N.A., and Xing, Eric P. A Latent Variable Model for Geographic Lexical Variation. *EMNLP '10: 2010 Conference on Empirical Methods in Natural Language Processing*, (2010), 1277–1287.
9. Gaffney, D. #iranElection: quantifying online activism. (2010).
10. Hecht, B. and Gergle, D. On The "Localness" of User-Generated Content. *CSCW '10: 2010 ACM Conference on Computer Supported Cooperative Work*, (2010), 229–232.
11. Hecht, B., Hong, L., Suh, B., and Chi, E.H. Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles. *CHI '11: 29th ACM Conference on Human Factors in Computing Systems*, (2011), 237–246.
12. Java, A., Song, X., Finin, T., and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, ACM (2007), 56–65.
13. Kinsella, S., Murdock, V., and O'Hare, N. "I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, ACM (2011), 61–68.
14. Kulshrestha, J., Kooti, F., Nikravesh, A., and Gummadi, K.P. Geographic Dissection of the Twitter Network. *ICWSM '12: Sixth International AAAI Conference on Weblogs and Social Media*, (2012).
15. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. Geographic routing in social networks. *Proceedings of the National Academy of Sciences 102*, 33 (2005), 11623–11628.
16. Lin, J., Xiang, G., Hong, J.I., and Sadeh, N. Modeling People's Place Naming Preferences in Location Sharing. *UbiComp '10*, (2010).
17. Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., and Boyd, D. The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions. *International Journal of Communication 5*, 0 (2011), 31.
18. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J.N. Understanding the Demographics of Twitter Users. *ICWSM '11: 5th International AAAI Conference on Weblogs and Social Media*, (2011), 554–557.
19. Morstatter, F., Pfeffer, J., Liu, H., and Carley, K.M. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *ICWSM '13: Seventh International AAAI Conference on Weblogs and Social Media*, (2013).
20. Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., and Almeida, V. We Know Where You Live: Privacy Characterization of Foursquare Behavior. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM (2012), 898–905.
21. Popescu, A. and Grefenstette, G. Mining User Home Location and Gender from Flickr Tags. *ICSWM '10: 4th International AAAI Conference on Weblogs and Social Media*, (2010).
22. Quercia, D., Capra, L., and Crowcroft, J. The Social World of Twitter: Topics, Geography, and Emotions. *ICWSM '12: Sixth International AAAI Conference on Weblogs and Social Media*, (2012).

23. Rout, D., Bontcheva, K., Preotiuc-Pietro, D., and Cohn, T. Where's@ wally?: a classification approach to geolocating users based on their social ties. *HT '13*, (2013).

24. Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., and Mühlhäuser, M. A Multi-Indicator Approach for Geolocalization of Tweets. *ICWSM '13: Seventh International AAAI Conference on Weblogs and Social Media*, (2013).

25. Starbird, K., Muzny, G., and Palen, L. Learning from the Crowd: Collaborative Filtering Techniques for Identifying On-the-Ground Twitters during Mass Disruptions. *ISCRAM '12*, (2012).

26. Takhteyev, Y., Gruzd, A., and Wellman, B. Geography of Twitter networks. *Social Networks 34*, 1 (2012), 73–81.

27. Tang, K.P., Lin, J., Hong, J.I., Siewiorek, D.P., and Sadeh, N. Rethinking Location Sharing: Exploring the Implications of Social-driven vs. Purpose-driven Location Sharing. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ACM (2010), 85–94.

28. Tsai, J.Y., Kelley, P., Drielsma, P., Cranor, L.F., Hong, J., and Sadeh, N. Who's Viewed You?: The Impact of Feedback in a Mobile Location-sharing Application. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2009), 2003–2012.

29. Twitter Inc. Twitter Streaming API. *Twitter Developers*, 2013. https://dev.twitter.com/docs/api/1.1/get/search/tweets.

30. Wiese, J., Kelley, P.G., Cranor, L.F., Dabbish, L., Hong, J.I., and Zimmerman, J. Are You Close with Me? Are You Nearby?: Investigating Social Groups, Closeness, and Willingness to Share. *Proceedings of the 13th International Conference on Ubiquitous Computing*, ACM (2011), 197–206.