Barriers to the Localness of Volunteered Geographic Information

Shilad W. Sen
Macalester College
St. Paul, MN, USA
ssen@macalester.edu

Heather Ford
Oxford University
Oxford, UK
heather.ford@oii.ox.ac.uk

David R. Musicant
Carleton College
Northfield, MN, USA
dmusican@carleton.edu

Mark Graham
Oxford University
Oxford, UK
mark.graham@oii.ox.ac.uk

Oliver S. B. Keyes
Wikimedia Foundation
San Francisco, CA, USA
okeyes@wikimedia.org

Brent Hecht
University of Minnesota
Minneapolis, MN, USA
bhecht@cs.umn.edu

ABSTRACT

Localness is an oft-cited benefit of volunteered geographic information (VGI). This study examines whether localness is a constant, universally shared benefit of VGI, or one that varies depending on the context in which it is produced. Focusing on articles about geographic entities (e.g. cities, points of interest) in 79 language editions of Wikipedia, we examine the localness of both the editors working on articles and the sources of the information they cite. We find extensive geographic inequalities in localness, with the degree of localness varying with the socioeconomic status of the local population and the health of the local media. We also point out the key role of language, showing that information in languages not native to a place tends to be produced and sourced by non-locals. We discuss the implications of this work for our understanding of the nature of VGI and highlight a generalizable technical contribution: an algorithm that determines the home country of the original publisher of online content.

INTRODUCTION

Over the past decade, *volunteered geographic information* (VGI) has transformed our relationship with the physical world. VGI datasets such as geotagged tweets, Wikipedia articles about places, and eBird observations have opened up novel avenues of research and enabled whole new classes of applications. Much of the promise of VGI has been thought to lie in its ability to engage local content producers. For instance, well-known geographer Michael Goodchild states that VGI "is putting mapping where it should be, which is the hands of local people who know an area well" [18].

The literature on VGI thus far has largely provided support for the optimism surrounding VGI and localness. In particular, research has established an association between the home locations of VGI contributors and the locations about which they create information. For instance, research has shown that Wikipedia editors tend to edit articles about places close to them [14, 16], a large proportion of Flickr photos are taken nearby photographers' homes [6], and Twitter users tend to tweet about locations near them (e.g. [4]). Many researchers see the shift in authorship from paid experts to volunteers — often described as public-participation geographic information science (PPGIS) [25] — as a way to increase local geographic information production, especially among disadvantaged populations.

This localness of VGI has implications well beyond the large group of researchers who use and study VGI. To the extent that locals see their community differently, (non-)local-ness can lead to systemic differences in how people experience and learn about places online. A search for a place name on any large search engine will prominently feature Wikipedia content, user-contributed photos, TripAdvisor reviews, and social media posts. VGI also shapes our physical experiences of places. As some have argued [11, 19], VGI does not just represent the world, but also becomes part of the world. It forms layers of code and information that augment everyday activities: it shapes where we go, what we do, and how we perceive and understand the world that we live in.

This paper is the first to study whether the localness of VGI is a *universal* property. Although research has shown that VGI has a strong overall local component at the repository level, little is known about the *geographic variation in this localness*. Is VGI about certain parts of the world more local than other parts? If so, are there patterns in where VGI is more and less local? Like much of the work on VGI localness (e.g. [14, 16]), we focus on Wikipedia, but we examine 79 different language editions of the encyclopedia. These language editions range from large encyclopedias written in global languages (e.g. English, French) to smaller encyclopedias written in regional languages from around the world (e.g. Latvian, Korean, and Georgian).

We consider localness to be a special case of what we call *geoprovenance*, or the geographic origin of information. Unlike previous research in this area, we consider

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

ACM 978-1-4503-3145-6/15/04 ...$15.00.
http://dx.doi.org/10.1145/2702123.2702170

two types of geoprovenance, each of which provides a unique lens on the localness of VGI. *Editor geoprovenance* considers the relationship between the locations of Wikipedia editors and the geographic articles they edit (e.g. "What countries do edits to articles about Iran comes from?"). *Source geoprovenance*, on the other hand, focuses on the *citations* added to Wikipedia articles. It considers the relationship between the location of cited *sources* and the geographic articles the sources describe (e.g. "What countries publish the sources cited in articles about Iran?"). Each source is produced by individuals, organisations and groups based in particular geographic regions and selected by Wikipedia editors to support informational claims in articles. Sources can therefore serve as an important signal about where (geographically) Wikipedia information comes from. To investigate source geoprovenance, we developed a novel algorithm that determines the country of the original publisher of online content with 91% accuracy.

Our analyses reveal extensive inequalities in the localness of Wikipedia VGI around the world. Through visualizations and statistical models we show that these inequalities tend to increase the local perspective in Wikipedia pages about areas with higher socioeconomic status (SES). We also find a powerful role of language in localness, a factor not been carefully considered thus far. More specifically, our findings suggest that if you read about a place in a language that is not commonly spoken in that place, you are unlikely to be reading locally-produced VGI or even VGI that references local sources.

We provide four major contributions: (1) The first global analysis of the variation in the localness of VGI. (2) We find that language can serve as an almost complete barrier to local VGI editors. (3) We find that the degree of a country's source localness primarily reflects the strength of its scholarly media networks. (4) An algorithm that identifies the home country of the original publisher of content posted on the web with 91% accuracy.

Our findings point to the existence of disconnected "VGI bubbles" that conform to language and socio-economic barriers, and suggest that VGI systems need new tools to reveal and overcome these barriers. In addition, our findings provide an empirical foundation to probe the "local = good" assumption present in the vast majority of the VGI literature. To disseminate our results and support future research, we have built a public website that visualizes the geoprovenance of information in 79 language editions of Wikipedia and released the datasets and source code for algorithms described in this paper.[1]

RELATED WORK

In addition to the work mentioned earlier that studies VGI localness at the repository level, this paper draws from research in two additional areas: research on Wikipedia citations and content biases in VGI. We address each of these in turn below.

Authors such as Luyt and Tan [21] have looked to Wikipedia's sources as a way of understanding from whose point of view and according to which principles and values it is written. They found that Wikipedia history articles are supported by few citations comprised predominantly of US government and online media news sites. Ford et al. [7] examined 500 random citations from the English Wikipedia and found that 80% of publishers of cited source material are located in a country whose primary language is English. This study extends this work beyond the English Wikipedia, considering more citations and leveraging two geoprovenance lenses.

This research builds on work related to geographic content biases in VGI. Research has established that VGI datasets tend to have substantial biases in terms of information quantity, with some areas of the world being covered much better than others. In early research on the subject, Hecht and Gergle [15] found that most language editions of Wikipedia exhibit substantial *self-focus bias*, with more content about places where the language is spoken. Graham et al. [8] found exceptions to this general rule, with some editors creating content far away from their local areas. Neis et al. [24] made a related observation in OpenStreetMap, finding that content coverage decreased with distance from city centers in Germany. Geographic content biases have also been found in Twitter (e.g. [23, 17]), Flickr (e.g. [20, 17]), Foursquare (e.g. [17]), and geographic content on the Web more generally (e.g. [10]).

More recently, researchers have sought a more detailed understanding of the factors that are associated with increased and decreased VGI coverage. For instance, Graham et al. [9] have identified broadband Internet connectivity as "necessary, but not a sufficient" condition for high VGI coverage in Wikipedia. Li et al. [20] found that places where "well-educated people in the occupations of management, business, science, and arts" live are associated with more tweets and geotagged Flickr photos. Other researchers have identified that there is more VGI per capita in urban areas than rural areas [17], and that the accuracy of different types of VGI contributions varies with a country's socioeconomic status [12].

This paper is distinguished from work on geographic content biases in that it does not look at variation in the *quantity* of VGI content around the world, but rather is the first to look at variation in the *localness* and the *geoprovenance* more generally of that content.

EDITOR PROVENANCE DATASET

We study geoprovenance on Wikipedia through two geographic lenses: the editors who create article content and the source publishers who inform article content. This section describes the editor geography dataset.

We extracted all edits to geotagged articles[2] across all Wikipedia language editions between June 6th and

[1] `http://www.shilad.com/localness`

[2] Editors geotag an article by inserting special text specifying an article's latitude and longitude.

lang	geotagged articles	geolocated edits	URLs
EN	937,315	2,357,561	8,228,596
DE	328,625	656,206	2,757,402
FR	299,544	685,920	2,271,618
		36 more langs	76 more langs
total	4,408,100	5,355,530	31,810,797

Table 1: Basic statistics describing the editor and source datasets. Due to anonymization, the editor dataset covers fewer languages.

September 4th 2014. Our data included both anonymous and registered users, but excluded actions by bots. Editors' IP addresses were geolocated to countries using the MaxMind GeoIP database, a technique that has been shown to be 96% to 98% accurate at the country level [27]. Edits that could not be resolved to a valid nation were excluded.

Because a Wikipedia editor's IP address is considered private information, we did not access the raw geolocated edits. Instead, we analyzed an anonymized version of the dataset that aggregated the number of edits for each combination of Wikipedia language edition, article country, and editor country. For example, it indicated that in the Arabic Wikipedia, editors in South Africa contributed 20 edits to all articles about locations in Algeria. To avoid the possibility that even the aggregated dataset would not be anonymous, two anonymity filters were imposed. First, any combination of language edition and article country that did not contain at least 20 distinct articles was removed. Second, any language edition with fewer than 500 distinct articles after filtering was removed. This privacy filtering may introduce some bias in that it cuts out less active language editions and articles. Although our analyses control for language edition as a specific variable, our findings should be interpreted with this limitation in mind.

The final anonymized dataset contained 5M edits from 39 language editions (Table 1). English Wikipedia had the most geolocated edits (937K) followed by German Wikipedia (328K) and French Wikipedia (299k).

SOURCE PROVENANCE DATASET

In addition to editor location, we study geoprovenance as defined by the country of source publishers cited in Wikipedia. Since each language edition uses its own citation syntax (English has dozens of different templates), we took a language-neutral approach that extracts all URLs embedded within articles. Although this choice allows us to compare all language editions, it excludes the roughly 25% of citations without a URL[7].

Table 1 shows background statistics for the languages with the most URL citations. In total, we extracted 31,810,797 URLs from the 79 language editions of Wikipedia with geotagged articles. Because the same URL can appear in multiple articles, these 31M urls represented 9,404,072 distinct URLs. Of the 9,404,072 distinct URLs, 19.9% failed to return web content in some way (e.g. host not found, HTTP 404, etc). This paper analyzes the 7,528,431 successfully crawled URLs.

We created a novel geo-location algorithm that inferred the country of source publisher for each URL. Algorithms exist to geo-locate a specific IP address [27], and identify the geographic topics of a webpage [2]. In addition some previous algorithms used IP and country code top-level domains (CC TLDs) to identify source publisher country [22]. However, ours is the first geolocation algorithm that uses multiple signals to triangulate a web-page's source publisher, enabling it to achieve 91% precision and approach the agreement levels of human coders. The sections that follow define what we mean by "source publisher", describe the human-coded gold standard we collected, list the data sources used by the algorithm, and evaluate the algorithm's effectiveness.

Data coding procedure

To ensure that we created an accurate source provenance dataset we performed human coding to associate 198 randomly chosen URLs with their source publisher country. One member of the team first performed open coding of 90 randomly sampled citations from English Wikipedia to draft a codebook for the location of the publisher of each URL. For publications hosted on third party sites (such as the Wayback machine or a book hosted on Google Books), we looked up the location of the original publisher; for machine translated works we used the original publisher; for human-translated works, we used the translator because of their ability to shape the final source; for Google map citations, we used the copyright information provided at the resolution of the URL in order to look up copyright holder locations in addition to Google. An additional member of the team then used the codebook (which contained no specific examples) to code all 90 examples. Agreeement on these examples was 95%.[3] Both team members then coded 50 URLs chosen across all 79 language editions to ensure the coding applied robustly to non-English Wikipedias, achieving an agreement of 93%. Finally, the team indepedently coded an additional 75 non-English citations, resulting in 90 citations from English Wikipedia, and 125 citations from other languages. 17 of these URLs were no longer available (e.g. dead links) and removed from the dataset, leaving 198 URLs for evaluation purposes.

URL geolocation features

We used four signals to discern a URL's publisher.

Whois records: We obtained the WHOIS record that provides registration information associated with each URL in order to extract the country within them. WHOIS records are maintained by regional internet registries, and are publicly accessible. To gather these records we extracted the top-level "private" domain (TLPDs) for each URL. In total, we identified 1,015,733 distinct TLPDs and retrieved WHOIS records for all TLPDs. We constructed a hand-coded parser that identified administrative contacts in whois records using key

[3]We use simple agreement because, given the 200+ possible codings (i.e. countries), a statistic such as kappa would be extremely high and difficult to interpret.

	overall-share	webpage-language	milgov	whois-structured	whois-freetext	cc-tld	wikidata	baseline	logistic-regression
coverage	100%	96%	8%	40%	16%	49%	19%	100%	100%
accuracy	30%	61%	100%	81%	77%	98%	93%	78%	91%

Table 2: Accuracy and coverage of our web page geolocation algorithm, its component features, and a baseline algorithm.

terms (e.g. the text "admin" and "country").We refer to this data feature as **whois-structured**. When records could not be parsed using the "structured" approach, feature **whois-freetext** extracted country references from anywhere in the free-text WOHIS record.[4]

Web page languages: We detected the language of each of the 7.5M URLs using Python's langid module,[5] which detects 97 different languages with high precision. Each language was mapped to countries weighted by the number of primary and second- language speakers of the language. Our estimates for language speakers is described in the later statistical analysis of localness.

CC TLDs: For all applicable URLs, we extract the country associated with the URL's CC TLD (e.g. Russia for ".ru"). We ignore "generic" TLDs (e.g. ".com" and ".net") and CC TLDs used for their mnemonic characteristics (e.g. ".tv").[6]

Wikidata country: Wikimedia's Wikidata project is a human maintained repository of structured information about entities [29]. We identified all entities in Wikidata that contained an "official URL" and spatial coordinates (about 22K entities covering 95K URLs).

URL geolocation algorithm

We created a machine-learning classifier to infer the location of every URL resource. Country inference is a multiclass machine learning problem (one class per country). However, because training data is unavailable for many countries, we trained a country-agnostic binary classifier. The classifier was invoked for a specific url and country, and returned a single probability indicating whether a URL was published by that country. Given a URL, by repeatedly calling the classifier for all countries, we can infer a probability distribution over countries and, if desired, identify the most likely source country.

The classifier included the five previously described variables: whois-structured, whois-freetext, webpage-language, cc-tld, and wikidata. We also included two other features: overall-share, the fraction of all URLs associated with each country, and milgov, a feature that infers the United States for ".mil" and ".gov" TLDs. Each of these features is specific to a country and url. For example, webpage-language(whitehouse.gov, us) = 0.25,[7] milgov(whitehouse.gov, us) = 1, and milgov(whitehouse.org, ru) = 0. We used a logistic regression classifier to integrate individual features because of its simplicity and transparency.

[4]Names and codes from `http://download.geonames.org`

[5]`https://github.com/saffsd/langid.py`

[6]`https://support.google.com/webmasters/answer/1347922`

[7] The US represents roughly 25% of all people who have at least second-language proficiency in English.

Using the human-coded dataset of 198 URLs described above, we evaluated the accuracy of our geolocation algorithm, each individual geolocation feature, and a baseline algorithm that used the CC TLD if available and predicted US for everything else.

Table 2 shows the accuracy and coverage results. While a language was inferred for almost all webpages (96%), the accuracy of language in predicting publisher country was relatively low (61%) due to the ambiguity of mapping a language to a country (e.g. Spanish is spoken in many countries). Several features (cc-tld, milgov and wikidata) didn't apply to the majority of cases (49%, 8% and 19% coverage respectively), but had near perfect accuracy when they did. The wikidata feature proved particularly useful for locating large multi-national companies and organizations (e.g. UNESCO.org and Nestle.com) who had detailed Wikidata entries.

The machine learning classifier synthesizing these features achieved an accuracy of 91% with cross-validation, approaching levels of human agreement. The classifier also cut error rates by more than half over the baseline (91% vs 78%). More interestingly, for "difficult" urls without a clear cc tld, the classifier achieved an accuracy of 77% versus 51% for the baseline.

Given a particular URL in our human-coded dataset, we evaluated the geolocation classifier accuracy by comparing the top predicted country against the human-coded country. However, this "winner-takes-all" approach would introduce significant bias into our data. For example, for the 6% of web pages whose only applicable feature was the inferred language English, a winner-takes-all approach would always assign the US as the source country, ignoring the United Kingdom, Australia, India, etc. Instead, our data uses a proportional approach, choosing from the probability distribution estimated by the classifier (which we calibrated to be reasonably accurate). This procedure emphasizes distributional accuracy, which is critical to this work, over the accuracy of any single URL.

LOCALNESS

This section analyzes the localness of information in Wikipedia, asking *How much information about a place originates from that place, and what factors explain variations in localness?*

We analyze each country's localness at the language edition level using two lenses: editors and sources. For example, in the Arabic Wikipedia, what percentage of edits to articles about France come from editors in France? And what proportion of sources in those articles are local to France? What about in the French Wikipedia? We investigate these questions through a combination of exploratory visualizations and statistical analyses. We

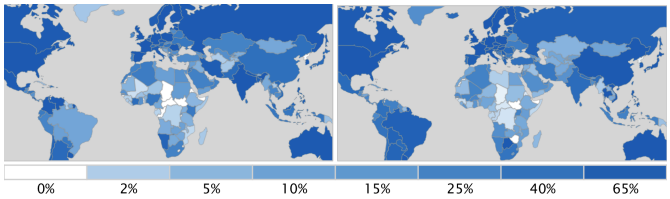

Figure 1: The overall localness of sources (left) and editors (right) across all language editions. Articles tend to have slightly more localness in the sources they cite than the editors that create them. The middle east and Africa show particularly low localness.

encourage readers to use the online visualization tool we created to explore the relationships we discuss.

Exploratory analysis of localness

Figure 1 shows each country's localness, aggregated across all language editions.[8] In this map we see our first evidence of inequalities in the localness of sources and editors. The countries with the highest percentage of local sources tend to be countries with large economies and "nation-states" responsible for a large share of worldwide speakers for their primary language: Sweden (93.0%), the U.S. (90.6%), The Czech Republic (87.9%) and France (86.3%).The countries with the highest percentage of local edits overlap substantially, but the list seems to favor countries that represent the majority of speakers of their primary language: Netherlands (89.2%), Germany (86.8%), and Latvia (86.1%).

Africa and the Middle East stand out as having particularly low localness. For example, many central African countries have less than 0.2% localness in both sources and editors (Chad, South Sudan, Central African Republic, the DRC). Figure 1 points to SES characteristics affecting VGI localness. Many countries in sub-Saharan Africa have per-capita GDPs below $1000 U.S. per year, among the lowest in the world. We return to these ideas in the statistical analysis that follows.

The apparent role of language in these results suggests that we should look more closely at individual language editions of Wikipedia. Figure 2 does so by comparing the localness of the Middle-East and North African (MENA) countries along two axes. The left column shows results for the Arabic Wikipedia while the right column represents the French Wikipedia. The top row shows editor localness, while the bottom row conveys source localness. For example, Algeria's dark color in the top-left map indicates that in the Arabic Wikipedia, a large percentage of edits to articles about Algeria come from editors in Algeria (75%). Comparing the result to the other three maps, we can see that Algeria's editor localness in the Arabic Wikipedia is substantially higher (75%) than its editor localness in the French Wikipedia (39%,

[8]All map visualizations use an identical color scheme. Thematic cartography teaches that, while there are some accepted best practices, choosing an appropriate color scheme for a given map can involve both art and science [28]. We attempted to communicate the percentage associated with each country by using cubed values mapped to a gradient from white to blue. Values above 0.65 receive maximum saturation. This transformation balances between highlighting large and small effects.

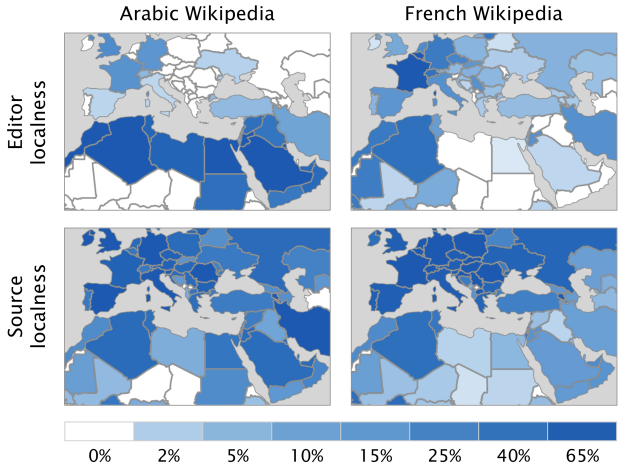

Figure 2: Across all articles about a country, the localness of editors (top) and sources (bottom) for the Arabic Wikipedia (left) and the French Wikipedia (right). For example, Algeria's dark color in the top-left shows that in Arabic Wikipedia, a large percentage of edits to its articles come from editors in Algeria (75%).

upper right) and also higher than its source localness in both language editions (44% and 40% for the Arabic and French Wikipedias in the bottom two maps).

The comparison between the French and Arabic Wikipedias reveals how Wikipedia is embedded into a broader system of knowledge. The language that an editor is writing in — and the place that that language originates from — matters immensely to what information can be produced on Wikipedia. If there are few Arabic speakers in Greece, few people in Greece will create content about the country in the Arabic Wikipedia (e.g. none in our dataset). We can see the Arabic Wikipedia's *language barrier* preventing local edits through much of Europe and central Africa. Similarly, local edits in the French Wikipedia are constrained to Europe and former French colonies such as Algeria and Morocco.

While language also shapes source localness, the effects are substantially dampened. Editors proficient in one language seem to be able to draw on a broader array of digital and digitized knowledge in other languages. For example, editors of French Wikipedia can still draw on local sources when authoring articles on Saudi Arabia (12.3%), even though less than 1% of editors are local to Saudi Arabia. This effect persists broadly across our data. Across all 847 project-specific country localness values that have at least 100 edits and sources, 82% have higher source than editor localness.

Overall, we see support for both socio-economic status (SES) and socio-linguistic effects on localness. SES appears to play the primary role in overall localness (Figure 1). However, when considering individual language editions, language also plays an important role, and appears to serve as a near-total barrier to local editors.

Statistical analysis of localness

The visualizations above suggest that language and SES factors both play roles in the localness of content. We performed a statistical analysis to systematically iden-

Localness of sources			
	coeff	odds-ratio	z-value
intercept	-0.13	0.87	-0.6
log(journals)	0.34	1.41	23.4 ***
log(population)	-0.24	0.78	-13.5 ***
lang-share	1.24	3.45	3.6 ***

Localness of edits:			
	coeff	odds-ratio	z-value
intercept	-4.10	0.01	-6.6 ***
is-native	1.76	5.83	7.8 ***
log(internet)	0.12	1.12	2.9 **
lang-share	3.04	20.80	4.3 ***

Table 3: Models of source and edit localness for articles in a particular country. Socio-economic variables play a stronger role in sources, followed by language. The reverse is true for editor localness. This paper uses * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$

tify country characteristics that explain differences in localness. We included a variety of variables that may correlate with a country's ability to support people editing Wikipedia (population, GDP, broadband penetration [1]), and produce sources that can be cited in Wikipedia (the number of journal articles[9] and the number of newspapers [1] published by a country). Across all explanatory variables, we used raw counts over per-capita numbers because this better reflects the total quantity of volunteer work and citeable information available in a country. Our models also include log-transformed variables for attributes with long-tailed distributions (e.g. population, GDP, num-journals). This (and all future) analysis uses stepwise forward variable selection to isolate the most explanatory model variables.

Based on the language patterns we saw above, the statistical models also included two language variables. First, `is-native` is a boolean indicator that is 1 if a country is a native speaker of the Wikipedia language edition (e.g. Saudi Arabia for the Arabic Wikipedia). Second, `lang-share` represents a country's share of the global population of speakers for the Wikipedia language. For example Saudi Arabia contains 8% of the world's Arabic speaking population, so `lang-share(ar-wiki, Saudi Arabia) = 0.08`. We estimated both variables using data from the Geonames Gazetteer.[10]

Our analysis uses logistic regression to model the fraction of sources or edits that are local. Every data point in the regression is the localness of edits for a particular country, within a particular language edition of Wikipedia. For example, our dataset includes a single record for all 1882 edits to articles about Algeria in the Arabic Wikipedia. The dependent variable in this case would be 0.75 (the localness of edits is 75%). This record's value for GDP would be Algeria's GDP, the `is-native` indicator would be 1, and `lang-share` would be 0.11 (roughly 11% of Arabic speakers live in Algeria).

Table 3 shows the results of the logistic regression model. Both three-variable equations explain significant portions of the overall deviance in localness (27% for source localness, 47% for editor localness). Although both models select similar variables, the importance they place on each of them differs. The statistical model of source localness selects the number of journal articles as the most important factor. Interestingly, though `log(journals)` and `log(population)` are strongly correlated ($\rho = 0.72$), the "non-population part" of the number of journal articles signal seems to matter. A country's share of the Wikipedia project's language also plays an important role, with all three variables being strongly significant ($p < 0.001$). The effect sizes of these variables are substantial. For example, for every doubling in the number of journals, the odds of a country's sources being cited increase by 41%.

The statistical model for editor localness reverses the importance of the two types of variables. A country whose population natively speaks a language corresponding to a Wikipedia language edition (e.g. Saudi Arabia for the Arabic Wikipedia) has 483% higher odds of a particular edit being local. Interestingly, the most predictive SES variable is broadband penetration, which captures an editor's ability to access the Wikipedia site itself. For every doubling in the number of broadband users for a country, its odds of having a local editor increase by 12%.

To summarize, we identify substantial barriers to localness in VGI that explain differences of several orders of magnitude. Source localness is shaped most strongly by the quantity of scholarly publications that a country produces, with language barriers playing a smaller role. However, these language barriers are a huge factor in editor localness with SES characteristics (in particular Internet access) a secondary effect.

Overall, our findings reflect previous findings that SES correlates with the quantity of VGI about places. However, we do find two key differences. First, we illustrate language's key role as a barrier to editors. Although this finding is intuitive, the scale of the effects we see are dramatic, explaining differences in editor localness of several orders of magnitude. Second, localness about a place and the quantity of information about a place need not be correlated. For example, the U.S. is one of the most heavily represented countries in Arabic Wikipedia [8] despite having less than 1% editor localness.

[9] `http://www.scimagojr.com/countryrank.php`

[10] This work focuses on second-language proficiency language speakers who could plausibly make a substantive contributions to a language edition. For example, although only 0.02% of India's population are native English speakers, the over 200 million Indians with second-language proficiency in English represent the fourth most prolific group of editors to English Wikipedia (behind the U.S, U.K., and Canada). Because no dataset of worldwide second-language proficiency exists, we estimated it using the Geonames dataset (`http://www.geonames.org/`), which orders languages by number of speakers. We created an asymptotic function that estimated second-language speakers by combining a country's Geonames' language rank with the country's population. Though this seemed to perform well in practice, more accurate estimates of world-wide second language proficiency may improve the accuracy of our models, and represent a valuable area of future work.

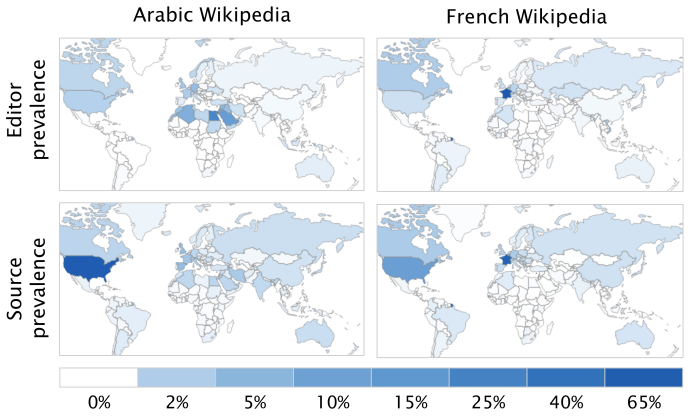

Figure 3: The overall percentage of sources and editors that come from each country. The images visualize differences along two dimensions: editors (top) vs sources (bottom) and Arabic Wikipedia (left) vs French Wikipedia (right). For example, the left images show that while editors from MENA countries constitute the majority of editors to Arabic Wikipedia (top), U.S. sources are most heavily cited (bottom).

Our results indicate that when we experience the world through VGI — either online or in augmented environments – our experience is shaped by many of the historical barriers and inequalities that existed before VGI.

GEOPROVENANCE

Localness tells us how much content is produced and sourced from a particular country, but it does not tell us where non-local content originates. We next move beyond localness to geoprovenance more generally. Specifically, we ask: *What is the worldwide network of sources and editors for articles about a particular country? What factors explain differences between countries and language editions?*

Our analysis of geoprovenance is divided into two parts. First, we study the overall prevalence of source and editor countries within each language edition (e.g. What proportion of edits in the Arabic Wikipedia come from people in Tunisia?). In doing so, we isolate language edition-wide effects due to language, population, demographics, etc. Next, controlling for these top-level language edition effects, we drill down to focus on the results for articles about a specific country, and the worldwide distribution of editors and sources for that country. (e.g. For articles about Iraq in the Arabic Wikipedia, what is the worldwide distribution of sources and editors?)

Top-level effects between language edition and country

We begin by studying the effect of language edition on the overall prevalence of sources and editors from each country (e.g. What proportion of all edits in the Arabic Wikipedia come from people in Tunisia?). Figure 3 shows four example distributions corresponding to the same two axes in the previous section. As an example, the left images show that while editors from MENA countries constitute the majority of editors across all articles in the Arabic Wikipedia (top), U.S. sources are most heavily cited (bottom).

Several clear patterns emerge that resonate with earlier findings about localness. The top row shows that editors in a language edition come almost exclusively from countries that speak the language. Editors of the Arabic Wikipedia come primarily from Egypt (23%), Saudi Arabia (13%), Jordan (8%), Algeria (7%) and Kuwait (5%). Editors of the French Wikipedia come from French speaking countries: France (88%), Belgium (3%), Vietnam (2%), Canada (2%), and Switzerland (1%). Editors from the U.S., for example, who account for 20% of all edits in our dataset, account for less than 1% of edits to both the Arabic and French Wikipedias.

While language appears to serve as an almost total barrier to editor contributions, socio-economic status (SES) factors appear to explain substantial variation within that barrier. For example, the Arabic Wikipedia dramatically under-represents native speaking low-GDP countries such as Sudan and Yemen, which constitute 10% and 7% of worldwide Arabic speakers but only 1% of Arabic Wikipedia edits. The same pattern appears in the French Wikipedia for the Democratic Republic of Congo. The DRC, where 40% of the population speaks French, account for 13% of French speakers worldwide but represent essentially no edits to the French language edition. We do not see the reverse trend occurring. Relatively wealthy, native-speaking countries (e.g. Egypt and Saudi Arabia in the Arabic Wikipedia and Belgium in the French Wikipedia) do not significantly exceed their "fair share" of edits.

As with localness, language does not constitute an absolute barrier for source use. Instead, traditionally dominant places in the geography of media persist, with the U.S. accounting for 61% of Arabic Wikipedia sources, followed by Spain (4%), and the U.K. (3%). While French Wikipedia is dominated by countries with large French speaking populations such as France (54% of sources), Belgium (7%), and Canada (2%), the U.S. also plays a notable role, accounting for 11% of all sources.

The role of the U.S. as the dominant publisher of Arabic Wikipedia sources seems surprising at first glance. This finding reflects the large share of total Arabic Wikipedia citations that occur in articles about the United States (39%) compared to Arabic countries like Egypt (2.5%) and Saudi Arabia (1.6%). This finding also resonates with Graham and Hogan [8], who find that a significant number of editors from MENA countries choose to write about places in North America rather than the MENA region. However, we will see in the next section that this finding also reflects the U.S.'s role as the primary publisher of sources about many Arabic-speaking countries — even in the Arabic Wikipedia.

To unpack these affects, we statistically analyzed the prevalence of source and editor countries for each Wikipedia project (Table 4). The forward stepwise variable selection procedure identified `lang-share` and `log(journals)` as the most explanatory pair of variables for both source publisher country and editor coun-

Source publisher country prevalence:

	coeff	odds-ratio	z-value
intercept	-1.79e+1	6.08e-7	-14.1 ***
log(journals)	1.08e+0	2.19e+0	10.6 ***
lang-share	4.27e+0	5.61e+1	10.0 ***

Editor country prevalence:

	coeff	odds-ratio	z-value
intercept	-10.50	2.75e-5	-5.3 ***
lang-share	7.46	1.75e+3	10.4 ***
log(journals)	0.45	1.57e+0	3.1 **

Table 4: Model of top-level effects between a language-edition and source or editor country. As with localness, our model for sources identifies SES attributes as primary factors and language as a secondary (but important) effect. The reverse is true for editors.

Country of Article

Country of Publisher	French Wikipedia					Arabic Wikipedia				
	Overall	Egypt	Iran	Iraq	Tunisia	Overall	Egypt	Iran	Iraq	Tunisia
Algeria	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%
Egypt	0%	4%	0%	0%	0%	1%	34%	0%	1%	1%
Iran	0%	0%	10%	0%	0%	2%	0%	73%	1%	0%
Iraq	0%	0%	0%	1%	0%	0%	0%	0%	9%	0%
Saudi Arabia	0%	0%	0%	0%	0%	1%	2%	0%	1%	1%
Tunisia	0%	0%	0%	0%	33%	0%	0%	0%	0%	40%
France	55%	40%	19%	37%	45%	3%	2%	1%	2%	17%
Belgium	7%	1%	1%	1%	1%	0%	0%	0%	0%	1%
Canada	2%	1%	3%	2%	2%	1%	2%	2%	2%	1%
UK	2%	5%	3%	7%	1%	4%	5%	2%	10%	4%
US	12%	37%	46%	37%	9%	62%	37%	13%	44%	16%
Other	22%	12%	18%	16%	8%	24%	17%	9%	29%	20%

;

Country of Article

Country of Editor	French Wikipedia					Arabic Wikipedia				
	Overall	Egypt	Iran	Iraq	Tunisia	Overall	Egypt	Iran	Iraq	Tunisia
Algeria	0%	0%	0%	0%	0%	7%	0%	0%	1%	1%
Egypt	0%	0%	0%	0%	0%	24%	65%	36%	4%	8%
Iran	0%	0%	17%	0%	0%	0%	0%	9%	1%	0%
Iraq	0%	0%	0%	0%	0%	4%	3%	24%	35%	1%
Saudi Arabia	0%	0%	0%	0%	0%	13%	5%	8%	7%	3%
Tunisia	0%	0%	0%	0%	12%	1%	1%	1%	1%	51%
France	88%	96%	70%	88%	56%	2%	0%	1%	1%	18%
Belgium	3%	2%	3%	1%	1%	0%	0%	0%	0%	0%
Canada	2%	0%	1%	3%	0%	1%	2%	1%	2%	2%
UK	0%	1%	1%	1%	0%	3%	3%	4%	13%	0%
US	0%	0%	1%	0%	0%	2%	1%	1%	4%	2%
Other	7%	1%	8%	6%	31%	44%	21%	14%	31%	14%

Figure 4: Distribution of source country (left) and editor country (right) for four countries (Egypt, Iran, Iraq, and Tunisia) for both French and Arabic Wikipedias. The table includes 12 key source publisher countries. "Overall" lists the percentage of all sources in the language edition that are published in that country (i.e. the top-level effect modeled at the beginning of this section).

try prevalence. This simple model explains the majority of total deviance in the data (61% and 80% for source and editor country respectively).

Once again, SES indicators play the primary role in the location of source publishers, while language plays the critical role in the location of editors. However, the effects appear more pronounced. For example, for every 10% increase in a country's share of the worldwide speakers of the Wikipedia edition's language, its odds of producing a source increase by 461% and the odds of a particular editor coming from it are 175 times higher.

The model's selection of journal articles as the key predictor of source publisher suggests that the volume of content produced by traditional media industries plays a critical role in a country's prominence as a publisher of sources in Wikipedia. Since journal article numbers correlate with many other SES indicators we probed this finding by replacing `log(journals)`, with other variables (e.g. GDP, Internet, etc.). When we did so, we saw residual deviation increase by 12% to 33%, suggesting that a country's scholarly media environment does indeed capture its likelihood of being a Wikipedia source publisher better than conventional SES indicators.

To summarize, we find that a country's prevalence among both sources and editors at the Wikipedia language edition level can be explained by the size of the country's media network and the fraction they represent of the world's fluent speakers of that language.

Focusing on articles in particular countries

We now shift our focus from broad project-level effects to articles about a specific country, and the worldwide network of sources and editors those articles draw upon.

To provide some intuition for this effect, Figure 4 shows the worldwide distribution of editors and sources for all geotagged articles in four specific MENA countries (Egypt, Iraq, Iran, and Tunisia) in both the Arabic and French Wikipedias. The table columns ("Overall", "Egypt", ...) indicate article country. The rows ("Algeria", "Egypt", ... , "U.S.", "Other") are associated with publisher countries, and the numbers in the cells indicate the percentage of citations in French Wikipedia associated with that country. The "Overall" column shows each country's share of all citations in French or Arabic Wikipedia (analyzed in the previous section).

The left table shows that language plays a large role in the geoprovenance of sources. Sources in France constitute a large share of citations for three of the four countries in French Wikipedia, with 40%, 37%, and 45% of citations for Egypt, Iraq, and Tunisia respectively. Iran serves as a noticeable outlier, having only 19% French citations. In the Arabic Wikipedia, where native speakers are not concentrated in any single country, we see a sharp rise in local sources across all MENA countries (34% to 73%) with Iraq serving as an outlier with only 9% localness. Iraq's low source localness may reflect its lack of web infrastructure; the CIA Factbook ranks it at #218 in number of Internet hosts, with 26 hosts.

Interestingly, although the U.S. does not publish many sources overall in French Wikipedia (12%), it constitutes a large share of citations for articles about many countries in the Middle East. We see similar effects for U.S. sources in Arabic Wikipedia. These findings provide insight into the U.S.'s prevalence in the Arabic Wikipedia broadly, extending results that we saw earlier (Table 3). Across language editions, U.S. sources constitute a large share of citations in articles about MENA countries.

The right table shows similar patterns for editor geoprovenance, with exaggerated language effects. France represents 70% or more of edits to every country except for Tunisia, while the Arabic Wikipedia is dominated by editors from MENA countries. In contrast to our results

Source geoprovenance:

	coeff	odds-ratio	z-value	
intercept	-4.16	0.01	-155	***
top-level(wplang, B)	6.07	432.90	91	***
is-same(A,B)	3.42	30.84	95	***
log(migrants A->B)	0.09	1.10	29	***

Editor geoprovenance:

	coeff	odds-ratio	z-value	
intercept	-4.18	0.01	-44.0	***
top-level(wplang, B)	6.33	563.57	53.0	***
is-same(A,B)	2.50	12.12	14.3	***
log(migrants A->B)	0.03	1.03	2.4	*
lang-overlap	-0.58	0.56	-2.3	*

Table 5: Given an article about country A, factors influencing the likelihood of a source or editor coming from a country B. top-level(wplang, B) is the overall prevalence of sources / editors from country B in the language edition (analyzed in the previous section).

for publishers, editors from the U.S. do not play a significant role in either language edition, representing 0% and 2% of editors. Analyzing the outliers in Figure 4 also provides insight into factors that affect geo-provenance. Iran exhibits notable source localness in French Wikipedia (10%), and exceptionally strong source localness in Arabic Wikipedia (73%) despite not being a native language country for either encyclopedia. These effects may be attributed to Iran's educational system, which ranks #7 in total books produced in the world, and a literacy rate of 97% among young adults. However, it does not appear to be solely Iranian editors who are contributing these local sources; Iranians only have an editor localness of 9% in Arabic Wikipedia. Tunisia, whose inhabitants speak both French and Arabic shows high localness in both language editions and source and editor lenses.

Table 5 shows the results of a statistical analysis that seeks to identify underlying factors that explain the worldwide geoprovenance network. More specifically given an article about country A, the model identifies factors influencing the likelihood of a source or editor coming from a country B. Because we seek to study the specific relationship between two countries, we control for the effects of source and editor localness at the Wikipedia language edition and article country level by including a dummy variable for it: `top-level(wplang,B)`. We also include `is-same(A,B)`, a dummy variable for localness that is 1 when A = B.

The statistical models for both publisher and editor identify a strong effect of localness. Migration also seems to be associated with source publishers. For every doubling in the number of people who emigrate from country A to country B, the odds of the country B being a publisher for articles about country B increase by 10%. This can statistically explain relationships such as Tunisia and French Wikipedia in Table 4. However, determining the underlying cause requires more research. Is it migrants who are actually responsible for these edits, or is this a signal of cultural overlap along many dimensions (spatial, religious, lingual, ethnic) that leads to increased knowledge and interest in another country?

To summarize, high-level effects of language edition on geoprovenance show that large differences exist in where different language editions get their content from. Our findings at the project-level mirror our findings for localness. Countries that fluently write in the language of the Wikipedia project contribute more sources and dramatically more edits. SES factors (in particular, the number of journal articles) contribute to both lenses of geoprovenance, but serve as the driving factor for sources. When honing in on the pairwise relationship between an article country and source or editor country, we find language overlap, localness (a boolean indicator), and migration to be indicators of the likelihood of one country to produce sources and editors for another country.

CONCLUSION AND DISCUSSION

This paper presents the first large-scale study of the forces shaping the localness of VGI. We find inequalities in the localness of content of several orders of magnitude, and statistically identify socio-economic and linguistic explanatory factors. The countries most-often cited in VGI are those that are strong publishers of traditional scholarship. For VGI editors, language serves as an almost total barrier, but within that barrier, SES factors (such as Internet penetration) explain large differences in localness. More broadly, our findings suggest that if a place faces serious SES obstacles, or you read about a place in a language that is not spoken in that place, you are unlikely to be reading locally-produced VGI or even VGI that references local sources.

Our findings have important implications for VGI. First, the primary role that SES factors play in localness suggests that VGI may not be the equalizing panacea that it has been portrayed as being. Instead, VGI appears to follow many of the same structural barriers to equality as traditional expert-curated systems, although research must determine whether the strength of these barriers in VGI is different. Barriers to localness may form the contours of "VGI bubbles" (similar to what Eli Pariser called "filter bubbles" [26]). However, our findings indicate these VGI bubbles exhibit unusual topologies. For example, U.S. source publishers seem to straddle nearly all language editions and countries. Future research should more deeply chart the shape of these barriers.

One way to reduce these barriers is by supporting, identifying and encouraging contributions that cross linguistic and cultural lines. For example, the Wikidata project aims to make factual units of information available in all language editions. Hale's work on cross-lingual blog linking [13] suggests that multimedia can also help in this respect, and the Wikimedia Commons project provides a good technical platform for this strategy. VGI barriers could also be reduced by making readers and editors aware of them. For example, a visualization such as Omnipedia [3] could alert readers to differing perspectives in other language representations of a particular artifact.

Several limitations of this study suggest areas for future work. While we analyze a single snapshot of geoprovenance, future research should explore the longitu-

dinal dynamics of this it. In particularly, the effects of translation on localness deserves further study, given English Wikipedia's dominant role as a source of translated Wikipedia articles [30]. We also consider content and people from the same country as "local." We believe this is reasonable given the scale of our analysis and the accuracy of source country prediction, but we would like to study variations in localness at other scales. Following the vast majority of the VGI literature, our work also equates "local" with "good". However, it would not be good if 100% of the sources in, for example, North Korea were local. There is much to be gained from an understanding of the benefits of diversity in creating strong knowledge platforms and conversations. As Elwood stated, "legitimizing VGI based on its locality or the nearness of observer to observed further elevates already-powerful cultural conventions... as valid ways to perceive what is true about the world [5]." Future research could propose frameworks that help us reason about optimal levels of locality.

Finally, we believe the URL geoprovenance algorithm introduced in this paper represents an important new tool for researchers. In addition to supporting analyses of the geographic diversity of viewpoints, it could be used to identify differences in how media from different countries discuss particular issues or it could track geographic flows in online media networks. To support research on improved geoprovenance algorithms, we are releasing our gold standard dataset and a reference implementation of our algorithm.[11]

ACKNOWLEDGEMENTS

We would like to thank Dario Taraborelli and the Wikimedia Foundation for helping to craft the anonymized editor geoprovenance dataset. We would also like to thank Matthew Zook for his advice about geocoding publishers. This research was generously supported by the National Science Foundation (IIS-0964697, IIS-1421655, and IIS-0808692), the John Fell Fund (Oxford Univ.), and an Amazon Web Services Research Grant.

REFERENCES

1. *World Development Indicators 2012*. World Bank, 2012.
2. Amitay, E., Har'El, N., Sivan, R., and Soffer, A. Web-a-where: geotagging web content. In *SIGIR* (2004).
3. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. Omnipedia: bridging the wikipedia language gap. In *CHI* (2012).
4. Cheng, Z., Caverlee, J., and Lee, K. You are where you tweet: A content-based approach to geo-locating Twitter users. In *CIKM* (2010).
5. Elwood, S., Goodchild, M. F., and Sui, D. Z. Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the AAG 102*, 3 (2012), 571–590.
6. Fischer, E. Locals and tourists (Flickr), June 2010.
7. Ford, H., Sen, S., Musicant, D. R., and Miller, N. Getting to the source: where does Wikipedia get its information from? In *WikiSym* (2013).
8. Graham, M., and Hogan, B. Uneven openness: Barriers to mena representation on wikipedia. *SSRN* (2014).
9. Graham, M., Hogan, B., Straumann, R. K., and Medhat, A. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the AAG* (2014), 1–19.
10. Graham, M., and Zook, M. Augmented realities and uneven geographies: exploring the geolinguistic contours of the web. *Env. and Planning A 45*, 1 (2013), 77 – 99.
11. Graham, M., Zook, M., and Boulton, A. Augmented reality in urban places: contested content and the duplicity of code. *Trans. Inst. of British Geog. 38*, 3 (2013), 464–479.
12. Haklay, M. How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design 37*, 4 (2010), 682 – 703.
13. Hale, S. A. Net increase? Cross-lingual linking in the blogosphere. *JCMC 17*, 2 (2012), 135–151.
14. Hardy, D., Frew, J., and Goodchild, M. F. Volunteered geographic information production as a spatial process. *IJGIS 26*, 7 (2012), 1191–1212.
15. Hecht, B., and Gergle, D. Measuring self-focus bias in community-maintained knowledge repositories. In *C&T* (2009), 11–19.
16. Hecht, B., and Gergle, D. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *CHI* (2010), 291–300.
17. Hecht, B., and Stephens, M. A tale of cities: Urban biases in volunteered geographic information. In *ICWSM* (2014).
18. Helft, M. Online maps: Everyman offers new directions. *The New York Times* (Nov. 2009).
19. Kitchin, R., and Dodge, M. *Code/space: Software and everyday life*. MIT Press, 2011.
20. Li, L., Goodchild, M. F., and Xu, B. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *CAGIS 40*, 2 (2013), 61–77.
21. Luyt, B., and Tan, D. Improving Wikipedia's credibility: References and citations in a sample of history articles. *JASIST 61*, 4 (2010), 715–722.
22. McCurley, K. S. Geospatial mapping and navigation of the web. In *WWW* (2001).
23. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. Understanding the demographics of twitter users. In *ICWSM* (2011).
24. Neis, P., Zielstra, D., and Zipf, A. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet 4*, 1 (2011), 1–21.
25. Obermeyer, N. J. The evolution of public participation GIS. *CAGIS 25*, 2 (1998), 65–66.
26. Pariser, E. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.
27. Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. IP geolocation databases: Unreliable? *ACM SIGCOMM Comp. Comm. Review 41*, 2 (2011), 53–56.
28. Slocum, T. A., McMaster, R. B., Kessler, F. C., and Howard, H. H. *Thematic Cartography and Geovisualization*. Prentice Hall, N.J., USA, 2009.
29. Vrandečić, D. Wikidata: A new platform for collaborative data collection. WWW '12 Comp. (2012).
30. Warncke-Wang, M., Uduwage, A., Dong, Z., and Riedl, J. In search of the ur-wikipedia: universality, similarity, and translation in the wikipedia inter-language link network. In *WikiSym* (2012).

[11] `http://www.shilad.com/localness`