Digital Hegemonies: The localness of search engine results

Andrea Ballatore (Birkbeck, University of London)
Mark Graham (Oxford Internet Institute)
Shilad Sen (Macalester College)

To appear in the
Annals of the American Association of Geographers
Author copy, February 2017

Abstract

Every day, billions of Internet users rely on search engines to find information about places to make decisions about tourism, shopping, and countless other economic activities. In an opaque process, search engines assemble digital content produced in a variety of locations around the world and make it available to large cohorts of consumers. Although these representations of place are increasingly important and consequential, little is known about their characteristics and possible biases. Analysing a corpus of Google search results generated for 188 capital cities, this article investigates the geographic dimension of search results, focusing on searches such as "Lagos" and "Rome" on different localized versions of the engine. This study answers the questions: To what degree is this city-related information locally produced and diverse? Which countries are producing their own representations and which are represented by others? Through a new indicator of localness of search results, we identify the factors that contribute to shape this uneven digital geography, combining several development indicators. The development of the publishing industry and scientific production appears as a fairly strong predictor of localness of results. This empirical knowledge will support efforts to curb the digital divide, promoting a more inclusive, democratic information society.

Keywords: Internet geography; search engines; Google; localness; digital place

Introduction

"Until the lion learns how to write, every story will glorify the hunter"

- Ewe (Ghanaian) proverb

Our spatial interactions and experiences with the world are increasingly digital. The cities and towns that we live in were once constructed from bricks, concrete, glass, and steel. But now, the places we inhabit also consist of digital augmentations (Graham 2013): digital code and digital content like data from Wikipedia, photographs from Flickr, restaurant reviews from Yelp, and algorithms controlled by Google that make information visible or invisible. The digital augmentations of places, in other words, are beginning to matter as much as their material counterparts, and much research has shown how they can

have significant impacts on how we economically, socially, and politically interact with our environments. (see Graham et. al. 2015; Kitchin and Dodge 2011). Digital code and content thus does not just reflect the world, but also produces it.

This does not mean that informational augmentations of places represent anything profoundly new. Indeed, the ability to represent place has long been a domain of conflict. This geographic information (be it digital, or pre-digital) has always been produced under conditions of power (Crampton 2008), and is inherently both as a product and a producers of power relations (Harley 1989; Pickles 2004). These conditions thus tend to reinforce and legitimate the dominant and powerful (Gramsci 1971). Histories, and narratives, about the Global South, for instance, could thus be written by the coloniser rather than the colonised (Said 1978).

Because the web has long been envisaged as a participatory tool, and because its usership now approaches 3.5 billion people (almost half of the world's population), many have hoped that the construction of information geographies could become more open, participatory, and democratic. Harvard Law professor Lawrence Lessig, speaking at the World Summit on the Information Society, noted (2003): "[f]or the first time in a millennium, we have a technology to equalize the opportunity that people have to access and participate in the construction of knowledge and culture, regardless of their geographic placing." For Lessig, and other commentators such as Benkler (2007), Bruns (2008), Jenkins (2006), Tapscott and Williams (2006), and Shirky (2011) as places become more digital they could also become more participatory.

Unfortunately, many of those hopes have not been realised. As the state increasingly rolls back from the cartographic project, spatial information is increasingly controlled by for-profit companies that have entirely different motives than their public predecessors (Leszczynski 2012). The field of critical GIS, for instance, has long concerned itself with the fact that digital representations of places are rarely equitable or genuinely participatory (see Pickles 1995). Sarah Elwood (2006) has noted that little has changed for those at the bottom of the digital divide and financial and skill barriers continue to influence who gets a say in digital representations of the world and who does not (see also Craig and Elwood 1998). More recently, work has shown how user-generated digital content such as Wikipedia not just largely represents the Global North, but is also overwhelmingly produced by users in the Global North (Graham et. al. 2011, 2014, 2015, 2016).

But the key organising mechanism behind our contemporary digital architectures is not user-generated content, but rather the search engine (Hillis, et al. 2013). Most people use digital search as their gateway not just to a disembodied web, but also to their lived everyday geographies: using it to learn about a destination, shop, navigate, and perform countless other activities. This is because information search usually takes the form of a one-to-many relationship (Graham 2010). A user has a requirement for information, and the available set of published information is much larger than the user has utility for. Search engines thus not just direct a user to relevant information, but filter that information to

distinguish between more and less useful content. Although search engines only index a fraction of the dynamic and growing corpus of Web content,[1] they provide a central access point to it.

As geography becomes ever more digital, search engines thus increasingly mediate not just information, but also spatial knowledges and experiences. This paper therefore seeks to better understand the geography of information in search engines. Specifically, it focuses on Google (the world's most powerful, most dominant, information mediator) to examine the locality of content about places around the world, asking whether Google directs users to locally-produced information or non-locally produced information. It does this by focusing on search results generated in the 188 countries where Google is currently available, when searching for capital cities. In doing so, it brings novel empirical data, about one of the most practiced ways in which we access digital geographic information, to bear onto older questions of power and geography.

Search engines, mediation, and power

Although they portray themselves as neutral aggregators of information, search engines have created an informational infrastructure with precise characteristics, logics, and biases (Ballatore, 2015). Scanning, interpreting, and organizing large volumes of online information to be served to users, global mediators such as Google and Bing play a crucial role in determining which websites, news, blogs, videos, and photographs become visible to whom. Although it is impossible to estimate how much of existing digital content is actually captured by search engines, they provide a highly visible access point to the "surface Web" crawled by their bots. This mediation between content producers and consumers remains opaque, unfolding behind closed doors with far-reaching consequences. On this point, Grimmelmann (2010) states that "search engines are the new mass media ... capable of shaping public discourse itself" (p. 436).

While a variety of web search services and technologies exist, the search market is controlled by a handful of large actors. A comScore report[2] shows that majority of desktop searches in the US in 2015 were executed on Google (64%), Bing (20%), and Yahoo! (13%), capturing a staggering 97% of the market. The dominant position of Google is even more pronounced in the UK, where it attracts 88% of searches, and in other European countries where it commands similar market shares.[3] Other search engines firmly dominate large markets, notably Baidu in China, Yandex in Russia, and Naver in South Korea, but they are confined to their home country. By contrast, Google still has the lion's share of the global search market, attracting a large majority of searches in most countries, and collecting 54% of the global search advertising market in 188 countries. This monopoly has not gone unnoticed and has attracted the attention of European antitrust agencies and that of a variety of critics, worried about such a concentration of information, power, and capital (e.g., Vaidhyanathan 2011, Grimmelmann 2013).

[1] https://www.nasa.gov/jpl/deep-web-search-may-help-scientists
[2] http://www.comscore.com/Insights/Market-Rankings/comScore-Releases-January-2015-US-Desktop-Search-Engine-Rankings
[3] http://theeword.co.uk/info/search_engine_market

Social scientists and humanists have analyzed search engines from cultural, cognitive, and political viewpoints, pointing out how these tools exert a powerful influence on society (Wouters & Gerbec, 2003; Spink and Zimmer, 2008; Halavais, 2009; Mager, 2012; Brossard & Scheufele 2013; Hillis, et al. 2013; Graham, et al., 2014; König and Rasch, 2014). The importance of these new mediators is observable in the industry of search engine optimization (SEO), which reflects how these tools are now at the core of the media landscape, representing a considerable portion of global advertising markets. The main activity of SEO consists of shaping and adapting web content to make it more visible on specific engines, reverse-engineering their algorithms. In this sense, content producers are influenced by the mediation of search engines as much as consumers; and biases in results have real effects (Vaughan & Thelwall, 2004).

Political analyses of search engines focus on the forms of power they exert. Notably, Epstein & Robertson (2015) identified what they term search engine manipulation effect (SEME) as the influence that biased search results can have on political choices. In different contexts, engines might help groups spread counter-narratives and fringe ideologies (Ballatore 2015) or, by contrast, further entrench dominant positions (Introna & Nissenbaum 2000).

The personalization of results is another researched aspect of search engines. Based on sophisticated and rich personal profiles, Google Search produces different content for different users, increasing and decreasing the visibility of links in order to deliver more relevant results. This process, according to Pariser (2011), might result in a "filter bubble," in which users are systematically exposed only to content that matches their political and cultural inclinations: in 2011, different users searching for "Tahrir Square" were shown either news reports about the revolution that started there, or websites of travel agencies that did not engage with the political context. The Google search personalization was strongly criticized for its lack of transparency, and the company recently introduced an option to disable it. Although precise quantification is difficult, a study suggests that on average about 12% of results differ because of personalization, and this applies only to users logged into their Google account (Hannak et al., 2013).

Research on search engine effects has raised valid concerns, but most studies have overlooked the spatiality of the information retrieval process. In this sense, digital representations of physical places play an increasingly important societal and cultural role (Graham et al., 2014; Ballatore, 2014). No prior research has addressed the representation of places in search engine results, analysing where the web content assembled by the algorithms is generated from.

The geography of Google search results

Google systematically collects massive amounts of data worldwide, and serves it to billions of users in 188 countries. The characteristics and effects of Google's presences and absences can be investigated from approaches developed in the sub-field of Internet geography: by, for example, studying the spatial

layout of its large network infrastructure, and the patterns of usage of its services across the globe (Dodge & Zook, 2009). This study focuses, in particular, on Google Search, the company's search engine and most-used service, analyzing the geography of its content.

To use an example, when searching for information about the city of Ankara on Google by typing the text string "Ankara" into a search box, a user obtains a set of links that includes Wikipedia articles, government websites, tourist guides, and news stories (Figure 1). This Google page represents a highly visible entry point to obtain information about the city's geography, economy, politics, history, and culture. Despite recent efforts by Google in promoting transparency,[4] it is extremely hard to know with any precision how and why these Web resources are selected over competing ones, as hundreds of signals about the relevance of resources are combined into ranking presented to users. Indeed, it is worth mentioning that Google Search actually uses the geography of Web content in calculating its relevance. This is visible in the option in Google's interface that allows user to filter results by country, when using a localized version of the product. For example, when searching for "Ankara" on the UK version (google.co.uk), the system allows users to choose between results from "any country" and results exclusively generated from the UK.

Default options, however, exert a powerful influence when users face complex choices, and most users do not alter the default settings of software tools, including search engines (Nielsen, 2005). This friction against changing pre-set options is referred to as the "default effect" by psychologists, and has been identified in a wide variety of contexts (Dinner et al., 2011). Default search results are therefore what the vast majority of users searching for "Ankara" will see. For this reason, it is important to pay particular attention to the default results returned by Google Search, even though users have a small degree of control over them. The effect of personalization should thus not deter research into Google results, as only about 12% of links differ from the non-personalized, default results (Hannak et al. 2013).

The central question of this study is: *Where is the web content that is returned by Google Search produced?* In our case of Ankara, some URLs point to local content produced and hosted in Turkey, while others refer to content from the United States and other countries. In other words, the search results have a certain degree of localness that can be quantified and can reveal crucial facets of Google's information geographies. For this empirical investigation, we consider the representation of the world's capital cities. The general workflow of the study is outlined in Figure 2. The remainder of this section illustrates the study's methodology, data collection, and analysis.

Methodology

To investigate the localness of Google search results in a systematic way, we collected search results for capital cities, adopting the methodology outlined by Ballatore (2015). This approach consists of extracting search results at different times and at different geo-locations, reducing the effects of personalization and spatio-temporal biases in the data. The repetition and randomization of the queries

[4] The Google Transparency Report is available at https://www.google.com/transparencyreport (last updated in June 2014).

produce results that are more stable and reproducible than individual observations at a specific location and time. To closely simulate a user's experience on Google, we extracted results from Google's HTML search results pages (i.e. the page that a user sees), approaching the typical usage of the service, rather than from the Google Custom Search API, which is known for returning different results.[5]

Google Search accepts many parameters that allow a user to increase the relevance of results. This study focuses on the effects of three parameters related to the geography of search: the country-specific localization setting for the search engine (e.g., *google.com* or *google.it*), the desired language of results (e.g., English or Italian), and the capital city query text ("Rome" or "Roma"). Other parameters were left to their default options. For each country, we generated a set of searches for the capital of the country. We also generated one US-centric Google query for the US version of Google (*google.com*) with the capital's English name, abbreviated henceforth as "US Google".

Second, we generated queries for the capital city in each language officially supported by Google in each country, which we refer to as "local Google" queries. For example, because Arabic and French are official languages in Morocco, we included two local Google queries with results in Arabic and in French respectively. The local Google data only includes languages that are supported in the target country, and not any other languages. Table 1 shows a sample of these Google queries, with one query on the US Google and two on local versions of Google. One query ("Washington, DC", in English) is present in both the US and local datasets.

The set of search queries used in this study is shaped by variations in different countries' access and representation within Google and the web more broadly. While some countries have unlimited access to the search engine (e.g., United States) others face total censorship if proxies are not used (e.g., Iran). Google provides localized websites for 188 countries (e.g., *Google.it* in Italy or *Google.co.ke* in Kenya), supporting 103 languages, all of which are included in the study. The list of countries and languages was extracted directly from the Google website on February 2015, while the capital city names in different languages were obtained from the multilingual gazetteer GeoNames. At the time of the data collection, Google was not available in China, Iran, North Korea, and Cuba, and therefore these countries were excluded from the study.

Data collection and validation

At the core, Google Search is an online service that, given a set of input parameters, returns search engine result pages (SERP). The parameters are passed to the service through a Uniform Resource Locator (URL), visible in the user's browser (an example of this would be https://www.google.com/search?q=sri+lanka). To collect results systematically, it is therefore essential to generate relevant and stable URLs. "Relevant" refers to results that are about the intended topic (e.g. about the French capital Paris and not about the American celebrity Paris Hilton), and "stable" refers to the degree to which the results are the same regardless of the search location. To identify such URLs, we

[5] https://support.google.com/customsearch/answer/70392?hl=en

ran a series of trials, in which we manually executed a sample of queries from different geo-locations (UK, US, France, and Italy) and compared the results. The URL that showed the highest stability and relevance has four URL query parameters: "q", which specifies the textual query (e.g. *q=Ankara*), "hl", which specifies the query's results language (*hl=en* for English), "gl", which specifies the provenance of results (*gl=us* for US results), and "oe" which specifies the text encoding (*oe=utf8*).[6] The text encoding was set to "utf8" for all queries in order to make the results consistent and easily machine-readable, while the other parameters were changed for each query.

The URL query parameter *gl*, which emphasizes results from a particular country (e.g., *gl=us* for US results), plays a particularly important role in this study. Without this parameter, Google prioritizes results from the current geo-location of the user. For example, searching for English results about "Rome" from a machine located in France returns mostly results from France, regardless of the other parameters. Setting the *gl* parameter to "US" enables the observation of the typical results that a North-American user sees when searching for "Rome" in the US, accessing more stable, representative results as opposed to transient, personalized results (Ballatore, 2015). The same results could be generated from machines physically located in the target countries, obtaining extremely similar results. To obtain a more spatially diffuse sample, the URLs were accessed through the Tor network, obtaining a different IP address for each URL.

As a result of this process, 188 URLs were generated for US Google (one for each country), and 357 for Local Google for each country in the languages supported by Google (1.9 languages per country on average), for a total 545 URLs. As the first page of results attracts more than 91% of clicks,[7] for each query we collected the results on the first page, typically varying between 8 and 12 URLs. To reduce the temporal bias of the results, each query was executed four times over three months, obtaining four separate snapshots of the 545 queries.[8] In total, 33,736 result URLs were collected. Over time, as observed in Ballatore (2015), the composition of the results varies up to 20%, particularly with respect to news stories. By aggregating the snapshots over time, the more stable results are reinforced, as they tend to occur in all of the four snapshots, while more transient results occur in fewer snapshots.

After the collection phase, the resulting dataset was analyzed for quality control. To reduce noise in the data, the results of the 545 queries were inspected to ensure that they captured the object of the study, i.e., the typical results of Google queries in different countries. This inspection showed that, out of 545 queries, 24 had invalid results, mainly because of errors in the multilingual data from GeoNames.[9] As a result of this validation, these 24 cases (4.4%) were removed from the dataset, resulting in 32,327 search results. The validated dataset can therefore be considered reliable for this study, and is summarized in

[6] Example of the US Google query for Zimbabwe:
https://www.**google.com**/search?q=**Harare**&hl=**en**&gl=**us**&oe=utf8
[7] https://chitika.com/google-positioning-value
[8] The data collection was executed on 2015-03-24, 2015-04-15, 2015-04-22, and 2015-05-09.
[9] We considered results to be invalid when the the query was formulated inconsistently with the experimental design, for example if the query language was Italian but the string was "Rome" instead of "Roma".

Table 2. The linguistic composition of the searches on local versions of Google broadly reflects the size of linguistic communities online, with the notable exception of China that was not included in the study: English (27.4%), French (9.8%), Spanish (6.5%), Arabic (5.1%), Russian (3.5%), Portuguese (2.3%), German (2.2%), while all other languages combined amount to 44.4%.

The dataset also provides insight into the websites that dominate the representation of places in Google. Table 3 shows the top 15 websites that obtain the highest visibility in both the US Google and the local versions of Google, showing a combination of crowdsourced reference websites (Wikipedia, Wikitravel, Wikivoyage), social media platforms (Facebook), newspapers (New York Times and The Guardian), and travel agencies (Booking.com). Wikipedia has the strongest presence, occupying about 10% of the URLs. The tourist websites Wikitravel and TripAdvisor rank second and third highest in the US, but only 6th and 8th in the rest of the world. Overall, the top 15 websites provide between 35% (US Google) and 22% (local Google) of the content, with a tail of less prominent websites. Figure 3 summarizes the global distribution of URLs at the country level, highlighting the strong influence of the US and Western Europe as web content producers.

Geo-location of web pages

Our analysis requires us to accurately identify the geographic origin of each web page returned as a google search result. We operationalize this by using Sen et al.'s (2015) notion of the *geoprovenance* of a URL, or the country primarily responsible for publishing the information on a particular web page. Although defining geoprovenance of some pages can be difficult or ambiguous, Sen et al. found that human coders agreed 93% of the time on the geoprovenance of a web page, indicating that the specific definition used in this paper provides a reliable measure of the geographic origin of information.

To scale the identification of URL geoprovenance to all 32,327 URls in our dataset, we used Sen et al.'s geoprovenance inference algorithm, which accurately predicts the country that published a specific URL, adopting its reference implementation.[10] In summary, the algorithm relies on five signals predictive of the geoprovenance of a particular URL:

1. The administrative contact's mailing address from a whois[11] database search for the URL.
2. A search for any known organizational headquarters locations associated with the domain using the Wikidata database (for instance, the IBM website is linked to the company's headquarters in Armonk, New York).[12]
3. Identification of suffixes associated with specific countries in a URL's top-level-domain such (e.g. ".uk" for "bbc.co.uk").
4. Geolocation of the country physically housing the server hosting facility for a particular URL, obtained by geo-coding a URL's IP address.

[10] https://github.com/shilad/geo-provenance/tree/master/py
[11] Whois is a protocol which allows access to a store of the addresses of people and firms that register every domain name (https://whois.icann.org).
[12] https://www.wikidata.org

5. The language of the content of a specific URL is used to identify candidate countries that have proficiency in the language.

When combined using a machine learning algorithm, these five different signals achieve 91% accuracy, approaching human levels of agreement. In addition to predicting the country that published a URL, the geoprovenance inference algorithm assigns a *confidence indicator* to each URL prediction ranging from 0.0 (no confidence) to 1.0 (full confidence). The confidence indicator is calibrated against Sen et al.'s human-created ground truth dataset to estimate the probability a predicted publisher country is correct. Thus, we use this confidence indicator to validate the quality of the inferred publisher countries in our dataset.

Across our dataset, the average confidence for a set of search results ranges from 0.42 to 0.95 with a median 0.88, indicating high confidence in publisher country predictions for the vast majority of cases, with few outliers with low values, such as Nigeria and Mali. The map in Figure 4 shows the confidence in the classification spatially, as the mean probability per country of a correct classification. We consider these values appropriate for this study, with a probability $p > 0.8$ for most major countries.

Localness of Google search results

This analysis quantifies the localness of search results for each country. It considers results for both the US version of Google and all local versions. To study this geography of content, we define a localness indicator *L* as the ratio between local results and the total number of results. Hence, *L* ranges from 0 (all search results are non-local) to 1 (all search results are local). More formally we define localness *L* of a country *c* and URLs *U* as the ratio of URLs originated from country *c* U_c and total URLs:

$$L(c, U) = \frac{|Uc|}{|U|}$$

L is a simple ratio that assumes equal weight of the first page results. This simplified assumption makes *L* easy calculate and apply across datasets. While more complex, weighted indicators could indeed be closer to the actual prominence of links, they would also be less interpretable. As our analysis seeks to understand variation at the country level, the unit of analysis is a single country, and we do not place more importance on a country with a large population than one with a small population. We did take several pre-processing steps to make the dataset less biased. While we weight countries equally, there are a large number of very small countries, such as micronations in the Pacific, with fewer than 1 million inhabitants. To prevent these countries from influencing our analysis (micro-nations tend to have a great deal of incomplete data), we excluded them, leaving 144 countries and 99 languages in the analysis.

We also took several steps to clean the search results data. An inspection of the geo-locations of the URLs revealed that all URLs of images were pointing to Google cache services, making them unreliable. For this reason, all URLs to images were removed from the computation of the localness indicator (12,056). Similar caution is needed when considering the effect of Wikipedia URLs in the results. Wikipedia is a large international crowdsourcing project with a complex content geography that cannot be reduced to its main host country, the United States (Sen et al. 2015). However, as one Wikipedia page

occurs in every single result set, its effect on the localness indicator can be safely ignored. Hence, Wikipedia URLs were not removed from the dataset.

We computed the indicator *L* for the 144 countries. As each case includes a Google query at four different times, the localness for a country is expressed as *mean L* of the four temporal snapshots and the corresponding standard deviation (SD). As shown in Figure 5, the value of *L* varies widely across the dataset. The figure compares the distribution of *L* in Google US (median = 0.22) with local Google results (median = 0.37). The gap between the distributions reflects the expected fact that results for US Google in English are less local than the localized versions of Google in different languages. The distribution of local Google is particularly wide, ranging from 0 to 1 without large gaps, indicating that cases exist for every level of localness. As shown in Table 4, Google US data is less local, more skewed, and has four outliers, corresponding with high-income, English-speaking countries (US, Canada, UK, and Australia). Localized versions of Google are more local, less skewed, and without noticeable outliers. Table 5 shows the countries grouped in 5 categories of localness, ranging from very low to very high, showing how cases are spread uniformly across the spectrum.

The global variation in localness *L* can be observed through a regional lens. Figure 6 shows the distribution of *L* for local Google grouped by the seven World Bank regions. To make the countries comparable, countries with multiple languages are aggregated into one point, considering the average *L* across languages. While North America (median $L = 0.9$) and South Asia (median $L = 0.36$) are tightly clustered, all the other regions show considerable variation, having both countries with low localness ($L < 0.3$) and countries with high localness ($L > 0.6$). Europe & Central Asia has the second highest localness after North America (median $L = 0.68$), followed by East Asia & Pacific and Latin America & Caribbean (median $L \sim 0.45$). Substantially lower *L* are observable for Sub-Saharan Africa (median $L \sim 0.27$) and Middle East & North Africa (median $L = 0.24$). Figure 7 and 8 show the same distributions spatially, at the country level. These maps are complemented by Figure 4, which represents the uncertainty in the classification, ranging from 0 (no certainty in the countries of origin of the URLs) and 1 (total certainty), based on the probability of a correct classification.

To corroborate this indicator, we also measured the diversity of results in terms of countries of origin. While some countries receive results from a few dominant countries, others receive results from a broader range of sources. To quantify diversity at the country level, we draw an ecological analogy comparing species diversity in an ecosystem with the diversity of countries of origin in the URLs. Hence, we compute a widely-used indicator of diversity based on Shannon entropy (Levine & HilleRisLambers 2009) on the URLs, obtaining an index ranging between 0 (low diversity) and 2 (high diversity). Figure 9 shows the variation of this entropy-based diversity globally. It is possible to notice that, while African and South and Southeast Asian countries tend to obtain low levels of local content, their results also tend to be produced from a wider range of places than those of countries in the Global North.

We also observed the country that generates the highest number of URLs in a result set for a country. The map in Figure 10 illustrates the ability of Google to capture local content in North America, most of South America, Europe & Central Asia. By contrast, most of Africa, Middle East, and South-East Asia are

dominated by URLs from the United States and France. Given the dominant position of the US in this arena, Figure 11 represents the proportion of URLs from the US in local versions of Google, showing the country's global but deeply uneven reach in this geography of content.

Explaining localness at country level

The localness of Google search results varies substantially around the world. While this variability is largely expected, we seek to understand the factors that influence it. In this section, we build an explanatory model of localness at the country level that includes a variety of socio-economic indicators. Henceforth, the mean localness *L* is the dependent variable, while all the other variables are considered to be explanatory. For explanatory variables, we consider a wide range of variables related to the robustness of digital infrastructure (e.g. population with internet access), education levels, and other socio-economic indicators published by the World Bank[13] that may relate to the ability for individuals within a certain country to produce searchable content. We used the World Bank datasets from 2011, the most recent complete dataset. In addition to socio-economic indicators, we also include indicators related to scholarly publication from the Spanish research group SciMago[14] that past research has shown to be strongly correlated with the geographic provenance of information on the web (e.g. Sen et al. 2015). The bibliometric data collected by SciMago draws upon over 21,000 journals in the Scopus database,[15] and captures the impact of scientific publications around the world.

Table 6 shows a complete list of variables in our analysis and the correlation coefficients between localness *L* and the explanatory variables, both using Pearson's *r* and Spearman's *rho*. The SciMago indicators on the publishing industry show substantially higher correlations with the localness indicator than the World Bank indicators. In particular, the *h-index* of a country, calculated as the maximum *h* such that *h* articles within the country have been cited at least *h* times, exhibits strong correlations (> 0.55 for both coefficients). Although there are differences between US and local Google Pearson's *r*, the differences appear greatly when using Spearman's *rho*. This suggests that there are non-linear relationships in the US data that are mostly linear in the local Google data. This may reflect the large inequalities that result in long tailed or skewed distributions across a variety of socio-economic indicators (Ostry et al, 2014). For example, the top 10 countries by GDP account for 66% of the world's GDP, and statistical analyses that do not normalize GDP data may be overly influenced by countries with the largest GDPs (e.g. the U.S. and China).

Localness regressions

Next, we develop separate regression models that explains localness of results from the US and local Google datasets. As many columns have missing variables, we filter out the 35 observations missing 10 or more explanatory variables, leaving 171 complete observations for US Google and 340 for local

[13] http://data.worldbank.org
[14] http://www.scimagojr.com
[15] https://www.scopus.com

Google. We use random forests to impute values for any remaining missing values (Stekhoven, 2012), with the imputed values achieving a relatively low normalized root mean squared error (NRMSE) of 0.18 (NRMSE values range from 0.0 to 1.0). Because the dependent variable L represents a proportion, and not an absolute value, we use a general linear model with a logistic link function (Long, 1997).

To begin, we study to what extent US Google localness can be explained by the World Bank socio-economic indicators and SciMago publishing metrics. A forward variable selection process that starts with no explanatory variables and iteratively adds the "best" unused variable identifies a 10-variable model that explains 58% of the overall deviation in *L*. This implies that most of the variation in search localness does in fact follow topologies that reflect measurable characteristics. While the full 10-variable model explains a substantial amount of the variation in *L*, the collinearities highlighted in Table 5 make the model difficult to interpret. To provide a more interpretable model M1, we identified four variables using forward selection that explain 53% of the total variance: h-index, region (a categorical variable with seven World Bank regions), the percentage of internet users, and a boolean factor indicating whether or not English is considered a "local" language in the country by Google:

M1: $L \sim$ h-index + region + Internet users + English is a local language

The terms in this model can be interpreted as follows. The h-index term indicates a country's level of activity and impact in scholarly publishing. Internet users reflects the level of digital infrastructure in a country that supports both the creation and consumption of web resources. For example, Italy and Japan exhibit similar h-indices, neither has English as a local language, but Japan has more Internet users (79% vs 55%) and also higher US localness (16% vs 12%). The coefficients of this model, along with the results of an ANOVA test on the regression results, including degrees of freedom and deviance, are shown in Table 7 and 8. The residuals of the regression are shown spatially in Figure 12.

Unsurprisingly, countries whose official languages include English exhibit higher US localness. For example, both Norway and Ireland are in the same region (Europe), exhibit high levels of Internet users (93% and 75%, respectively), and have relatively high h-indexes (439 and 364). Despite Norway's modest advantage in both Internet users and h-index, Ireland's use of English as an official language correlates with its much higher levels of localness (49% vs 29%). Finally, as shown in Figure 6, region-level localness exhibits large variations reflecting a many possible region-level linguistic, economic, and other cultural patterns.

We repeated the same procedure for local query results. A seven-variable model explains 53% of the deviation, while a four variable model explains just under half (49%). Forward selection identifies the same four variables as most explanatory (see Tables 9 and 10). Figure 13 shows the residuals spatially.

Discussion

The general similarities between the US and local Google datasets may reflect the variety in "local" languages supported by Google for each country, and Google's frequent support for English for a country

even though it is not an official language. For example, for Switzerland Google supports five languages: German, English, French, Italian and Romansh (and all five appear in our Google local dataset). However, English is not one of the official languages of Switzerland. Similarly, 72 records appear in the local Google dataset for countries where Google supports English but it is not an official language of the country. Thus, the overlap between the two datasets may partly explain the similarity in the regression models above.

While our analysis selected the same four explanatory variables as most important for the local and US analyses, a more carefully analysis finds that the models differ significantly. To determine this, we compared two nested models on all 511 records in the combined US and local dataset. The *reduced model* included all ten variables. The *full model* included the ten variables, along with interactions between those variables and a boolean variable indicating whether the record was either a US search result or local search result. An ANOVA comparing the models found that the full model explained significantly more variance (68% as opposed to 54%), and the models differed significantly ($p < 0.0001$). This indicates that the role of the explanatory variables *does* differ between the US and local datasets.

A closer inspection of the nested model found two key differences between local and US interactions. First, in the full model, the *L* values for North America were significantly higher relative to other regions in the US Google results compared to the local Google results. This reflects the United States' close cultural and geographic relationship to Canada (the only North American country in the US Google results). Since the local results were viewed through a less US-centric lens, regions coefficients were similar across regions. Second, the h-index exhibited a coefficient twice as large for the local search results. We verified this finding in a minimal model with an interaction between the h index and the boolean US versus local factor. This result indicates that while a country's scholarly publishing network is a critical feature for predicting *L*, the importance of a country's scholarly network is dampened when viewed through a US search lens.

To summarize, we find that geographic region, strength of scholarly publishing, internet infrastructure, and language barriers all play important roles in shaping the geography of Google search results. These four variables explain roughly half of the variation in *L*, leaving much room to evaluate the role of other country-specific attributes. We note that these explanatory variables do exhibit significant collinearities making it difficult to neatly unpack the roles of individual factors. We also find significant differences between the explanatory models for the US and local results; countries *not* in North America exhibit much higher *L* values in local results, and a country's scholarly network is more predictive of *L* in local results.

Conclusions

This investigation of the geography of Google search results shows that wealthy and well connected countries tend to have much more locally-produced content that is visible about them than poor and poorly connected countries. Even cities located in countries with huge populations such as Lagos show a

tendency towards having relatively little local content about them in Google Search results. This means that a user in the US or Germany searching for cities is far more likely to be given access to locally-produced content than a Tanzanian or Cambodian.

In our empirical study, the results of only eight countries in Africa (and four low-income countries, Tajikistan, Madagascar, Burkina Faso, and Tanzania) have a majority of content that is locally produced. This gives rise to a form of digital hegemony, whereby producers in a few countries get to define what is read by others. The US in particular is a dominant content producing force, even when excluding Wikipedia which is a highly visible US-based but globally assembled resource (Figures 10 and 11). In the results for 61 countries, the US supplies over half of the first page content on Google. This means that not only are American internet users surrounded by an extremely locally-produced internet, but American-produced content is highly visible in much of the rest of the world. However this does not necessarily mean that the US is an informational hegemon everywhere in the world. France has a somewhat smaller sphere of influence, mainly limited to countries in Africa, whereas Russia produces a visible effect only on results about Kyrgyzstan (see Figure 10).

It is important to note that, as our dataset focussed only on capital cities, caution should be taken when extending the results to higher spatial granularities. Our localness indicator does not take into account the actual interaction of users with search results, and the variety of devices and media across which individuals currently access search engines. Despite the precautions that we took to access representative samples of search results, some noise is still present and some results might show high volatility. Much more empirical work is needed to study finer patterns within countries, and to build more accurate models to investigate the consumption of geographic information on search engines in different geographic locales.

More broadly, the point remains that most countries in the Global South continue to be defined by a diverse range of sources originating from a diverse range of places. The issue here is not that Internet users are exposed to a diverse range of sources from a diverse range of places -- indeed, as Pariser (2011) notes, there are significant concerns for people and media-ecosystems that lack access to such diversity. The issue is rather that that diversity itself has a particular bias and those sources tends to be almost entirely from the Global North, and very few of the sources come from anywhere in the Global South. For instance, while the search results for Google's Ghanaian page for its capital "Accra" include pages from six countries, five of them are firmly located in the Global North.[16] When looking at countries in the Global North, the results for Denmark's capital are similarly diverse, with five out of six source countries also being located in the Global North.[17] By contrast, a country like the United States suffers from the inverse problem: having almost no exposure to geographic representations made by non-locals.

[16] The geographic composition of the 183 URLs returned for the capital of Ghana (Accra) is as follows: US: 67, Ghana: 43, Netherlands: 29, Switzerland: 20, United Kingdom: 20, Sweden: 4.
[17] This is the geographic composition of the 77 URLs returned for the capital of Denmark (Copenhagen): Denmark: 32, Faroe Islands: 22, US: 15, Namibia: 4, Poland: 3, Norway: 1.

The key question then is why. What explains this informational hegemony, or the dominance of the Global North in producing digital representations about not just themselves, but also about much of the Global South? Interestingly, our explanatory models indicate that network connectivity and economic development in a country are not enough to make that content about that place more local in Google search results. The presence of a strong publishing industry, using SciMago publication data as proxy, is the strongest predictor of the production of visible online content. The importance of the h-index in the model also shows that the impact of scientific publications is a better predictor of localness than the mere number of publications. Thus, we suggest that socio-economic systems that produce high-quality research also tend to produce highly visible online content. There are no countries in the Global South that score well on such metrics, and there are consequently no countries in the Global South that play a major role in constructing contemporary Internet geographies.

Having moved a first step in this direction, more quantitative and qualitative research is needed to better understand why exactly scientific knowledge production explains so much of the variance in Google's local digital representations. More relational variables and different spatial granularities will have to be considered. But, until then, we hope that the finding that wealth or network connectivity alone are not sufficient factors to be worth demonstrating, especially for internet activists who hope to bring about more genuinely participatory and representative digital environments. This point increasingly matters because places are ever-more defined by their digital presences, and the ways that places are represented digitally increasingly shapes how people understand and reproduce those very places (Graham et. al. 2015). Google plays an enormous role in constructing these digital representations of places. Because of their dominant role in mediating a majority of the world's internet use and the fact that few people ever explore beyond a first page of search results, they essentially determine which digital augmentations of place are made visible or invisible, with tangible effects in the physical world.

This paper demonstrated that Google search results are actively reproducing new forms of informational hegemony around the globe. A few countries in the Global North play an inordinately large role in defining the digital augmentations of the Global South. Google's methods for ranking and representing are notoriously opaque (Vaidhyanathan 2011; Graham et. al. 2014), but we do know that two key factors come into play. First, much of the reason of the lack of local voice in the South is likely simply because the production of Internet content happens at a much lower rate as compared to the North (Graham et. al. 2015). Second, since the company's creation in 1998, Google's algorithms have tended to favour highly central web content: pages linked to by a lot of other pages are prioritised, and those largely ignored are demoted in the rankings. This creates a worrying situation whereby it becomes difficult for those on the information peripheries to break out of their digital marginality.

Building on earlier research looking at the geographies of information, this paper has analysed not just where digital content comes from, but how it is ranked in the world's most powerful digital mediator. Much more will need to be done to understand not just the ways in which people are afforded voice about their own communities and countries, but also the myriad factors that serve to amplify or constrain it. Until then, we hope that other research can use this paper as a beginning to ask not just

why some parts of the world are denied locally-produced representations, but how we might bring about more representative and participatory digital augmentations of place.

Acknowledgements

The idea for this article emerged at the Specialist Meeting on Spatial Search at University of California, Santa Barbara. The authors thank the Center for Spatial Studies for organizing and funding the event.

Funding

We received funding from the European Research Council under the European Union's Seventh Framework Programme for Research and Technological Development (FP/2007–2013) / ERC Grant Agreement n. 335716. This work is also funded by National Science Foundation grant #1527173 and a grant for computational resources from Amazon.com, Inc.

References

Ballatore, A. (2015). Google chemtrails: A methodology to analyze topic representation in search engine results. First Monday, 20(7), 1–20. doi:http://dx.doi.org/10.5210/fm.v20i7

Ballatore, A. (2014). The myth of the Digital Earth between fragmentation and wholeness. Wi: Journal of Mobile Media, 8(2). Retrieved from http://wi.mobilities.ca/myth-of-the-digital-earth

Ballatore, A., Hegarty, M., Kuhn, W., & Parsons, E. (2015). Spatial Search, Final Report. Santa Barbara, CA. Retrieved from https://escholarship.org/uc/item/33t8h2nw

Benkler, Y. (2007). The wealth of networks: How social production transforms markets and freedom. New Haven, CT: Yale University Press.

Bruns, A. (2008). Blogs, Wikipedia, Second Life, and beyond: From production to produsage. New York: Peter Lang.

Craig, W.J., and S.A. Elwood. 1998. How and why community groups use maps and geographic information. Cartography and Geographic Information Systems 25 (2): 95–104.

Crampton, J. (2008). Will peasants map? Hyperlinks, map mashups and the future of information. In The hyperlinked society: Questioning connections in a digital age, ed. J. Turow and L. Tsui, 206–26. Ann Arbor: University of Michigan Press.

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology: Applied*, *17*(4), 432–432.

Dodge, M., & Zook, M. (2009). Internet-based measurement. The International Encyclopedia of Human Geography. (Eds. Kitchin, R. and Thrift, N.), Elsevier, pp. 569–579.

Elwood, S. 2006. Critical issues in participatory GIS: Deconstructions, reconstructions, and new research directions. Transactions in GIS 10 (5): 693–708.

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Science*, *112*(33), E4512–21.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. Proceedings of the National Academy of Sciences, 107(41), 17486–17490. doi:10.1073/pnas.1005962107

Goldman, E. (2008). Search engine bias and the demise of search engine utopianism. In A. Spink & M. Zimmer (Eds.), Web Search. Information Science and Knowledge Management, vol. 14 (pp. 121–133). Berlin: Springer.

Goldman, E. (2008). Search engine bias and the demise of search engine utopianism. In A. Spink & M. Zimmer (Eds.), Web Search. Information Science and Knowledge Management, vol. 14 (pp. 121–133). Berlin: Springer.

Graham, M., Straumann, R., Hogan, B. 2016. Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia. Annals of the Association of American Geographers. 105(6) 1158-1178. doi:10.1080/00045608.2015.1072791.

Graham, M., De Sabbata, S., Zook, M. 2015. Towards a study of information geographies:(im)mutable augmentations and a mapping of the geographies of information Geo: Geography and Environment.2(1)

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. Annals of the Association of American Geographers, 104(4), 746–764. doi:10.1080/00045608.2014.910087

Graham, M., Schroeder, R., & Taylor, G. (2014). Re: Search. New Media & Society, 16(2), 187–194. doi:10.1177/1461444814523872

Graham, M. 2013. The Virtual Dimension. In Global City Challenges: debating a concept, improving the practice. eds. M. Acuto and W. Steele. London: Palgrave. 117-139.

Graham, M., S.A. Hale, and M. Stephens. 2011. Geographies of the world's knowledge. London: Convoco! Edition.

Graham, M. (2010). Neogeography and the Palimpsests of Place. Tijdschrift voor Economische en Sociale Geografie. 101(4), 422-436.

Gramsci, A. 1971. Prison notebooks. New York: International Publishers.

Grimmelmann, B. J. (2010). Some Skepticism About Search Neutrality. In B. Szoka & A. Marcus (Eds.), The Next Digital Decade: Essays on the Future of the Internet (pp. 435–459). Washington, DC: TechFreedom.

Grimmelmann, J. (2013). What to Do About Google? Communications of the ACM, 56(9), 28–30. doi:10.1145/2500129

Halavais, A. (2010). Search Engine Society. Cambridge, UK: Polity. doi:10.1080/15205431003650494

Hargittai, E. (2007). The social, political, economic, and cultural dimensions of search engines: An introduction. Journal of Computer-Mediated Communication, 12(3), 769-777.

Harley, J. B. (1989). Deconstructing the map. Cartographica: The International Journal for Geographic Information and Geovisualization, 26(2), 1-20.

Hillis, K., Petit, M., & Jarrett, K. (2013). Google and the Culture of Search. New York: Routledge.

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the Politics of Search Engines Matters. The Information Society, 16(3), 169–185. doi:10.1080/01972240050133634

Jenkins, H. (2006). Convergence culture: Where old and new media collide. New York: New York University Press.

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement. Vaccine, 30(25), 3778–3789.

Keane, B. M. T., Brien, M. O., & Smyth, B. (2008). Are people biased in their use of search engines? Communications of the ACM, 51(2), 49–52.

Kitchin, Rob, and Martin Dodge. (2011) Code/Space: Software and Everyday Life. Cambridge MA: MIT Press.

König, R., & Rasch, M. (Eds.). (2014). Society of the Query Reader: Reflections on Web Search. Amsterdam: Institute for Network Cultures.

Leszczynski, Agnieszka. (2012) "Situating the Geoweb in Political Economy." Progress in Human Geography 36, no. 1: 72–89. doi:10.1177/0309132511411231.

Lessig, L. 2003. An information society: Free or feudal. Lecture given at the World Summit on the Information Society, Geneva, Switzerland. . http://www.itu.int/wsis/docs/pc2/visionaries/lessig.pdf (last accessed 14 April 2016).

Levine, J. M., & HilleRisLambers, J. (2009). The importance of niches for the maintenance of species diversity. Nature, 461(7261), 254-257.

Long, J.S. (1997). Regression Models for Categorical and Limited Dependent Variables. Sage Publishing.

Mager, A. (2012). Algorithmic Ideology: How capitalist society shapes search engines. Information, Communication & Society, 15(5), 769–787. http://doi.org/10.1080/1369118X.2012.676056

Nielsen, J. (2005). The Power of Defaults. Retrieved from http://www.nngroup.com/articles/the-power-of-defaults

Ostry, M. J. D., Berg, M. A., & Tsangarides, M. C. G. (2014). Redistribution, inequality, and growth. Washington, DC: International Monetary Fund.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google We Trust: Users' Decisions on Rank, Position, and Relevance. Journal of Computer-Mediated Communication, 12(3), 801–823. doi:10.1111/j.1083-6101.2007.00351.x

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. *Penguin*. New York: Penguin. http://doi.org/10.1353/pla.2011.0036

Pickles, J., ed. 1995. Ground truth: The social implications of geographic information systems. New York: Guilford.

Pickles, J. (2004). A history of spaces: Cartographic reason, mapping, and the geo-coded world. Hove, UK: Psychology Press.

Reilly, P. (2008). "Googling" Terrorists: Are Northern Irish Terrorists Visible on Internet Search Engines? In A. Spink & M. Zimmer (Eds.), Web Search. Information Science and Knowledge Management, vol. 14 (pp. 151–175). Berlin: Springer.

Said, E. (1978). *Orientalism*. Pantheon Books, New York.

Sen, S., Ford, H., Musicant, D., Graham, M., Keyes, O., & Hecht, B. (2015). Barriers to the Localness of Volunteered Geographic Information. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, 197-206.

Shirky, C. 2011. Cognitive surplus: Creativity and generativity in a connected age. London: Penguin.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.

Tapscott, D., and A.D. Williams. 2006. Wikinomics: How mass collaboration changes everything. New York: Penguin.

Vaidhyanathan, S. (2011). The Googlization of everything (and why we should worry). Berkeley, CA: University of California Press.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, *40*(4), 693–707.

Zook, M., & Graham, M. (2007). Mapping DigiPlace: geocoded Internet data and the representation of place. Environment and Planning B: Planning and Design, 34(3), 466–482. doi:10.1068/b3311

Tables and figures

Country	Engine version	Lang	Query text	Local Google	Top three URLs
Egypt	google.com	en	"Cairo"	False	1. en.wikipedia.org/wiki/Cairo 2. www.lonelyplanet.com/egypt/cairo 3. en.egypt.travel/city/index/cairo
Uganda	google.co.ug	en	"Kampala"	True	1. en.wikipedia.org/wiki/Kampala 2. www.lonelyplanet.com/uganda/kampala 3. www.kcca.go.ug
Vietnam	google.com.vn	fr	"Hanoï"	True	1. fr.wikipedia.org/wiki/Hanoï 2. tripadvisor.fr/Tourism-g293924... 3. routard.com/guide_voyage_lieu/...

Table 1: Sample of three Google queries and results.

Dataset characteristic	Value
Countries officially supported by Google	188
Languages included	103
Dates when URLs were captured	2015-03-24, 2015-04-15, 2015-04-22, 2015-05-09
Google queries	545 URLs collected four times
Query parameters	Engine version; language of results; query text
Results collected per query	All results on the first page
Collected URLs	32,327 (11,091 US; 21,236 local)
Languages for local Google	English (27.4%), French (9.8%), Spanish (6.5%), Arabic (5.1%), Russian (3.5%), Portuguese (2.3%), German (2.2%), others (44.4%)

Table 2: Overview of Google capitals dataset.

Google US Top Domain	%	Local Google Top Domain	%
wikipedia.org	10.11	wikipedia.org	10.17
wikitravel.org	5.18	lonelyplanet.com	2.56
tripadvisor.com	5.10	facebook.com	2.11
lonelyplanet.com	4.65	usembassy.gov	1.51
facebook.com	2.25	youtube.com	1.38
youtube.com	1.98	wikitravel.org	0.76
timeanddate.com	1.20	localtimes.info	0.71
nationsonline.org	0.82	tripadvisor.com	0.64
google.com	0.79	accuweather.com	0.58
britannica.com	0.73	booking.com	0.44
wikivoyage.org	0.67	timeanddate.com	0.29
theguardian.com	0.53	diplo.de	0.28
booking.com	0.51	hilton.com	0.23
usembassy.gov	0.41	gismeteo.ru	0.21
nytimes.com	0.41	24timezones.com	0.20
Other domains	*64.7*	*Other domains*	*77.7*

Table 3: Most visible 15 domains in the search results for US Google (11,091 URLs) and Local Google (21,236 URLs).

Dataset	N	Median L	Skewness L	Kurtosis L	Median PL	Outliers
Google US	144	0.22	1.70	7.15	0.90	US, UK, Australia, Canada
Local Google	297	0.37	0.29	1.91	0.85	None

Table 4: Descriptive statistics of the dataset.

Localness	L range	N	Countries
Very low	[0, 0.2)	27	Albania, Angola, Bahrain, Benin, Botswana, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Congo, Democratic Republic of the Congo, Ethiopia, Haiti, Kuwait, Lesotho, Libya, Mozambique, Palestine, Panama, Papua New Guinea, Puerto Rico, Somalia, Timor-Leste, Trinidad and Tobago, Turkmenistan, Vietnam
Low	[0.2, 0.4)	37	Afghanistan, Algeria, Azerbaijan, Côte d'Ivoire, Cyprus, Dominican Republic, Egypt, Gabon, Georgia, Ghana, India, Indonesia, Iraq, Jamaica, Kyrgyzstan, Laos, Lebanon, Malawi, Malaysia, Mali, Morocco, Namibia, Nepal, Nicaragua, Niger, Nigeria, Philippines, Qatar, Romania, Rwanda, Saudi Arabia, Sierra Leone, Sri Lanka, The Gambia, Togo, Zambia, Zimbabwe
Medium	[0.4, 0.6)	32	Armenia, Bangladesh, Belarus, Bolivia, Burkina Faso, Costa Rica, Cuba, Denmark, Estonia, Greece, Honduras, Hong Kong, Ireland, Israel, Jordan, Kazakhstan, Kenya, Latvia, Mauritius, Moldova, Myanmar/Burma, Oman, Pakistan, Peru, Senegal, Singapore, Tanzania, Thailand, Uganda, United Arab Emirates, Uzbekistan, Venezuela
High	[0.6, 0.8)	30	Argentina, Australia, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Chile, Colombia, Croatia, Ecuador, El Salvador, France, Guatemala, Hungary, Japan, Macedonia, Madagascar, Mexico, Mongolia, Netherlands, Paraguay, Poland, Portugal, Serbia, South Africa, Spain, Tajikistan, Tunisia, Ukraine, United Kingdom
Very high	[0.8, 1]	18	Austria, Canada, Czech Republic, Finland, Germany, Italy, Lithuania, New Zealand, Norway, Russia, Slovakia, Slovenia, South Korea, Sweden, Switzerland, Turkey, United States, Uruguay

Table 5: Overview of Local Google localness, aggregated by country (N=144).

Explanatory variables		Pearson's *r* with localness *L*		Spearman's *rho* with localness *L*	
Data source	**Country variable**	**US Google**	**Local Google**	**US Google**	**Local Google**
WBD 2011	GDP	0.30***	0.30***	0.46**	0.44***
WBD 2011	GDPPC	0.42***	0.42***	0.40***	0.40***
WBD 2011	Internet users	0.54***	0.53***	0.49***	0.49***
WBD 2011	Internet servers	0.42***	0.39***	0.43***	0.45***
WBD 2011	Population	0.11	-0.01	0.19*	0.15
WBD 2011	Tourism revenue	0.41***	0.37***	0.49***	0.49***
WBD 2011	Tourism visitors	0.27***	0.36***	0.41***	0.43***
SciMago	Documents	0.42***	0.32***	**0.54*** **	0.55***
SciMago	Citable documents	0.41***	0.32***	**0.54*** **	0.54***
SciMago	Citations	0.42***	0.31***	0.53***	0.55***
SciMago	Self citations	0.38***	0.20**	**0.54*** **	**0.56*** **
SciMago	H index	**0.56*** **	**0.56*** **	0.53***	**0.56*** **

Table 6: Bi-variate correlations between localness *L* with of US Google (N=171) and local Google versions (N=340) (Pearson corr. coefficient). WBD stands for World Bank Data (2011). Significance: *p* value (*) < .05 (**) < .01 (***) < .001.

Dataset: US Google (N=171)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.11E-01	1.16E-01	2.675	0.0082**
H index	1.79E-04	6.31E-05	2.846	0.005 **
Region: North America	0	0	0	N/A
Region: East Asia & Pacific	-2.73E-01	1.14E-01	-2.39	0.018 *
Region: Europe & Central Asia	-2.35E-01	1.14E-01	-2.07	0.040 *
Region: Latin America & Caribbean	-3.56E-01	1.15E-01	-3.10	0.002 **
Region: Middle East & North Africa	-3.27E-01	1.17E-01	-2.79	0.006 **
Region: South Asia	-1.56E-01	1.20E-01	-1.30	0.195
Region: Sub-Saharan Africa	-2.86E-01	1.16E-01	-2.46	0.015 **
Internet users	1.76E-03	4.97E-04	3.53	0.0005***
English spoken locally	9.24E-02	3.13E-02	2.96	0.004 **

Table 7: General linear model on US Google with model M1.
Significance: *p* value (*) < .05 (**) < .01 (***) < .001.

Dataset: US Google	Df	Deviance	Resid. Df	Resid. Dev
NULL			170	3.89
H index	1	1.214	169	2.68
Region	6	0.541	163	2.14
Internet users	1	0.202	162	1.94
English spoken locally	1	0.100	161	1.84

Table 8: ANOVA test on US Google (N=171) in Table 6.

Dataset: Local Google (N=340)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.12E-01	1.36E-01	0.82	0.41
H index	4.72E-04	8.83E-05	5.35	1.65E-07***
Region: North America	0	0	0	N/A
Region: East Asia & Pacific	1.48E-02	1.34E-01	0.11	0.91
Region: Europe & Central Asia	1.65E-01	1.31E-01	1.26	0.21
Region: Latin America & Caribbean	6.19E-01	1.36E-01	0.45	0.65
Region: Middle East & North Africa	-9.66E-03	1.36E-01	-0.07	0.94
Region: South Asia	6.60E-02	1.39E-01	0.47	0.63
Region: Sub-Saharan Africa	6.00E-02	1.34E-01	0.45	0.66
Internet users	1.90E-03	6.50E-04	2.92	0.004 **
Language is local	1.44E-01	2.37E-02	6.08	3.29E-09***

Table 9: General linear model on local Google.
Significance: *p* value (*) < .05 (**) < .01 (***) < .001.

Dataset: local Google	Df	Deviance	Resid. Df	Resid. Dev
NULL			339	24.34
H index	1	7.93	338	17.41
Language is local	1	2.33	337	15.08
Region	6	1.78	331	13.30
Internet users	1	0.33	330	12.97

Table 10: ANOVA test on local Google (N=340) in Table 7.

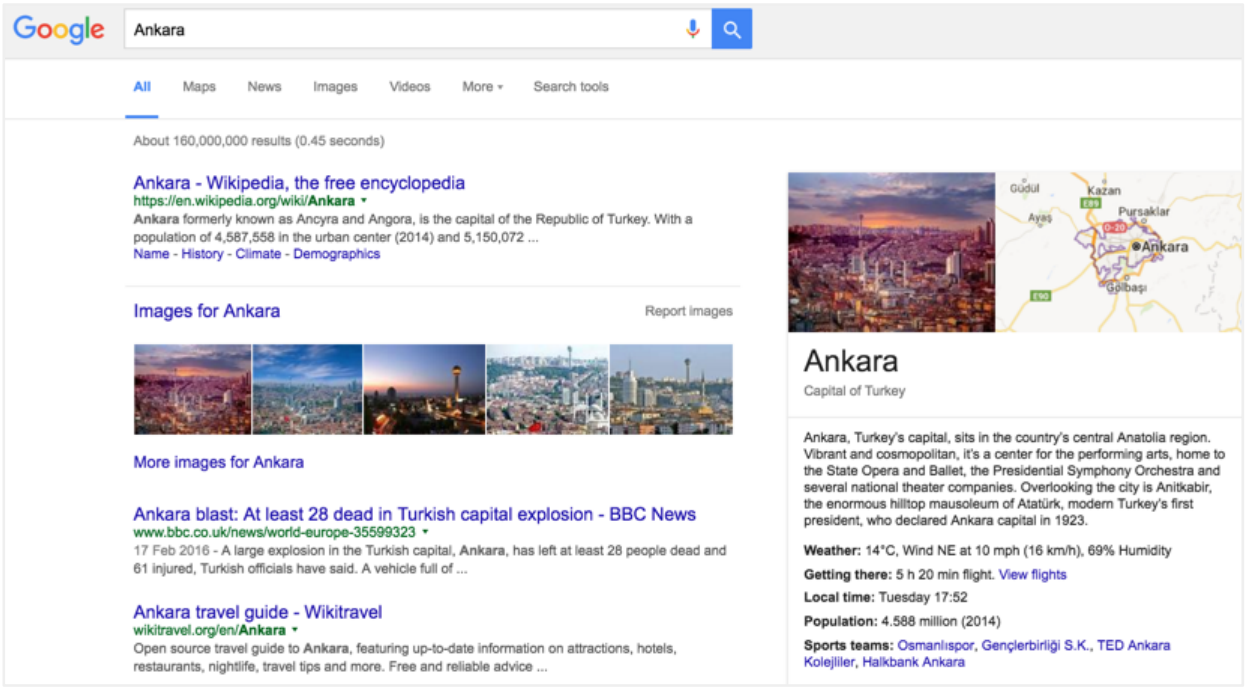

Figure 1: An example of Google search results for "Ankara" on May 10, 2016

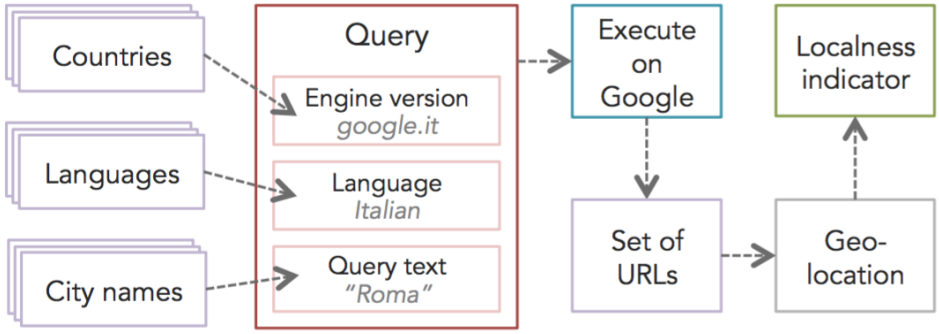

Figure 2: Overview of the study of the localness of Google search results

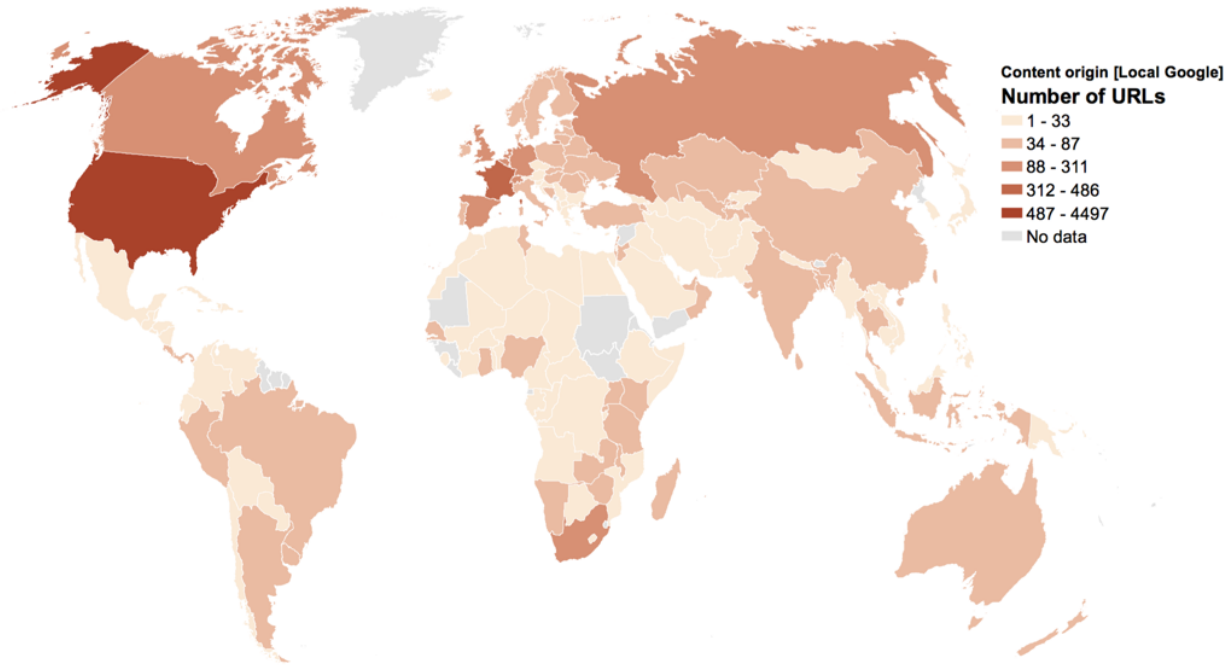

Figure 3: Number of URLs generated from each country, grouped by natural breaks. The top group contains only the US.

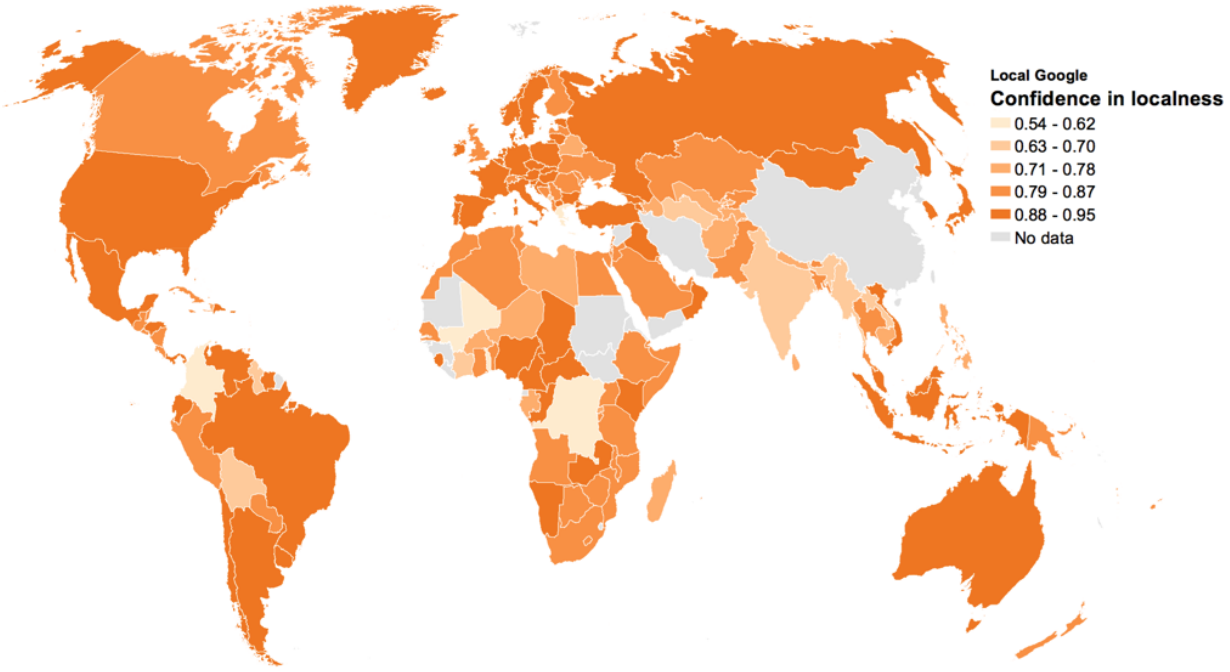

Figure 4: Confidence in localness L, as the mean probability of correct classification of content origin.

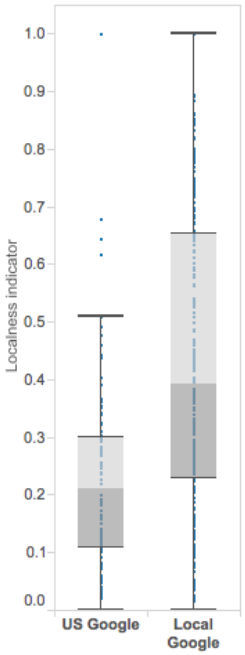

Figure 5: Distribution localness indicator *L* in US Google (N=144) and local Google (N=297).

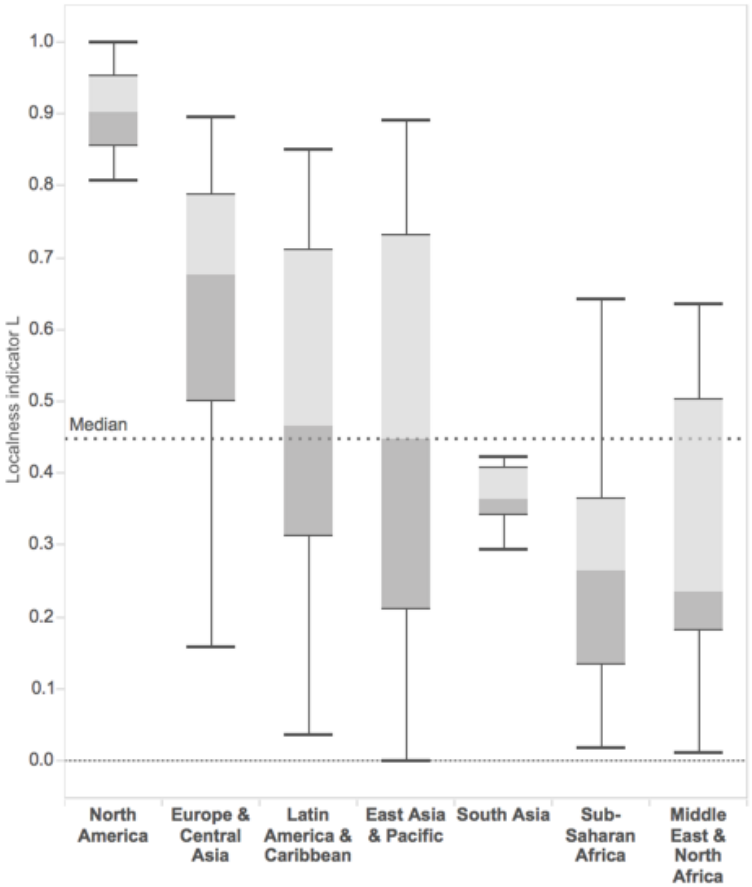

Figure 6: Distribution localness indicator *L* for local Google in 144 countries, grouped by World Bank regions, with global median (N=144)

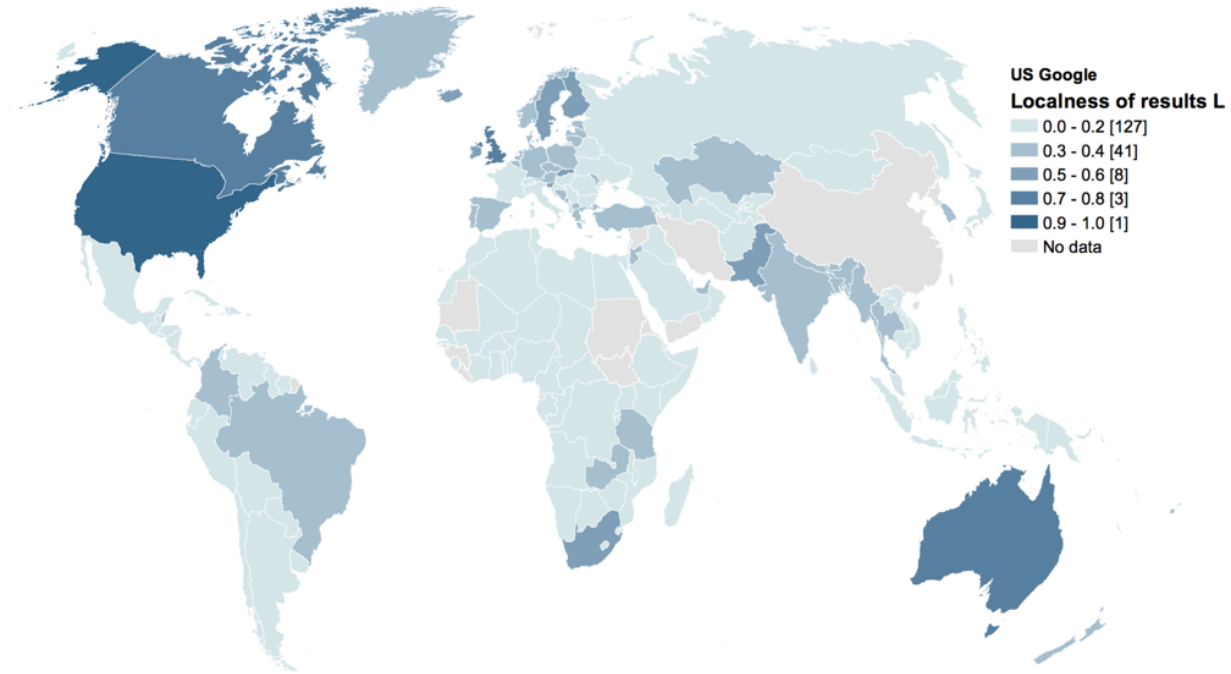

Figure 7: Localness indicator per country for the US version of Google and queries in English.

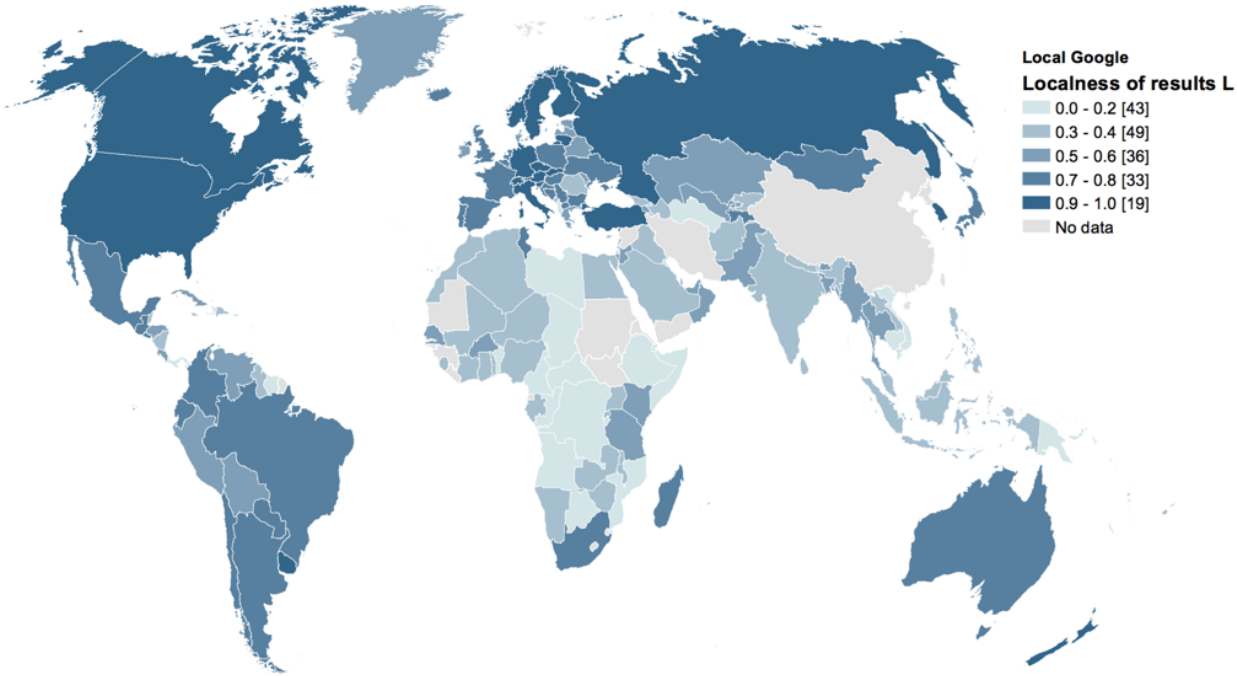

Figure 8: Localness indicator per country for the local versions of Google and queries in local languages.

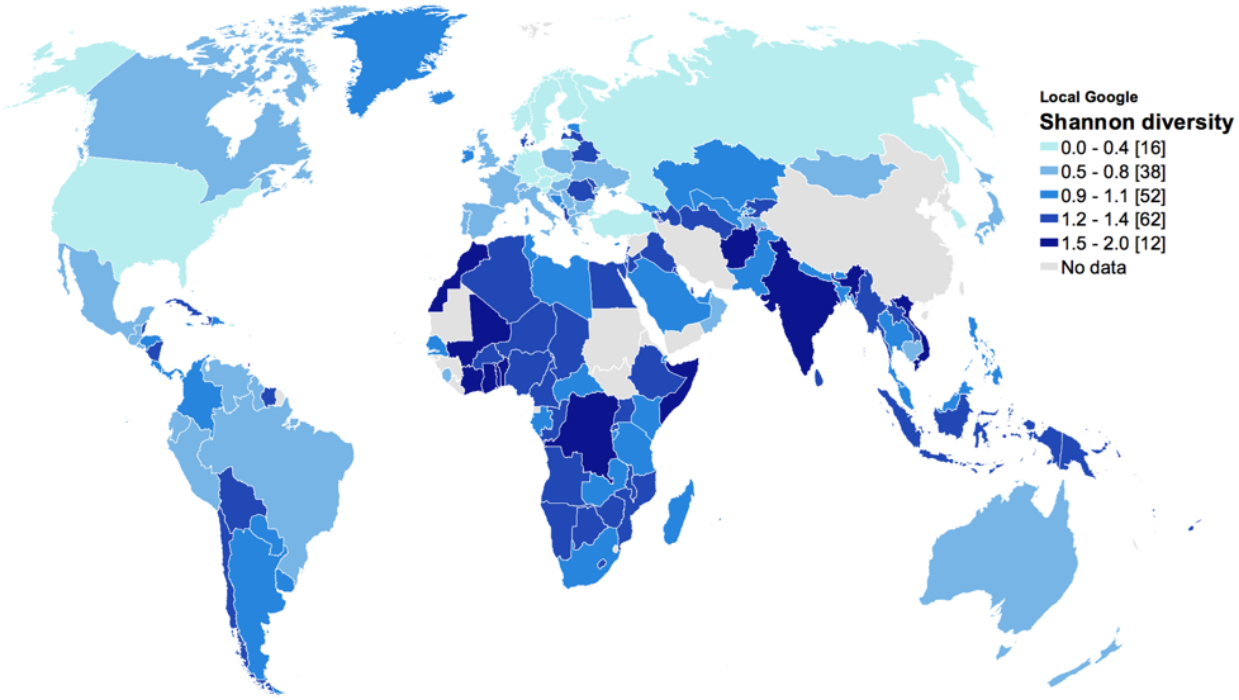

Figure 9: Diversity of results for local versions of Google. Low values indicate results from fewer countries, while high values indicate results from many countries.

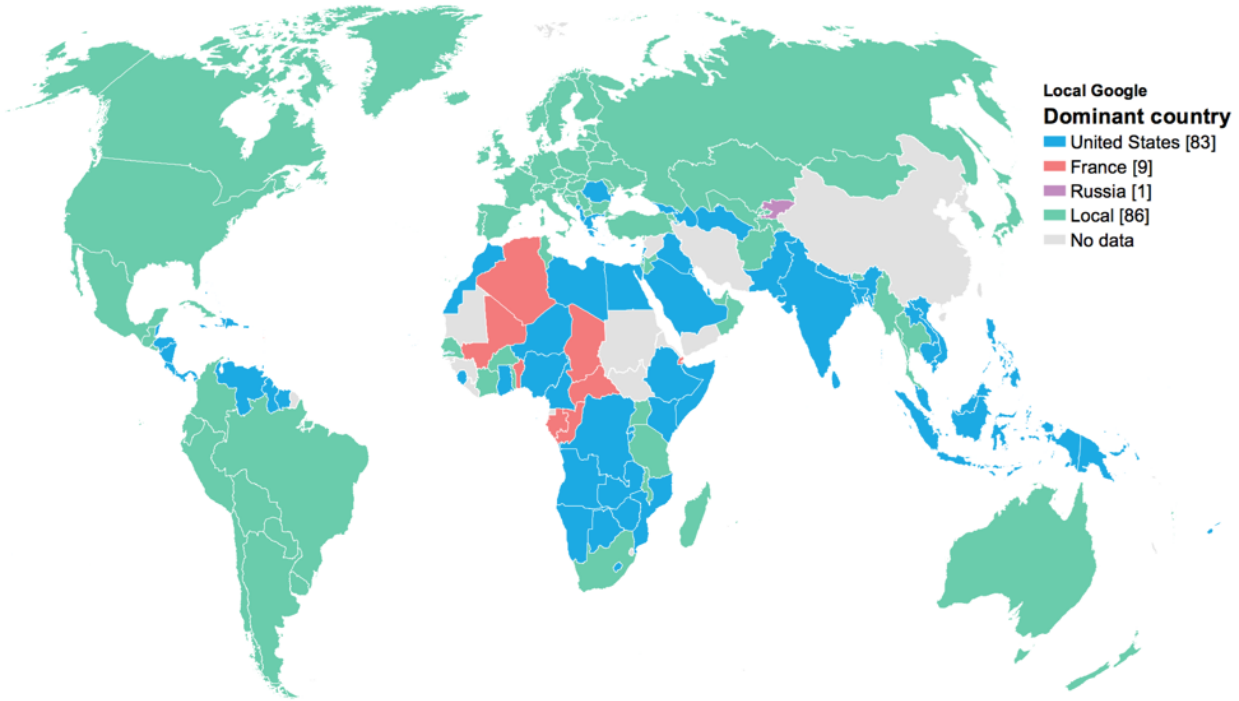

Figure 10: Countries that dominate results in local versions of Google.

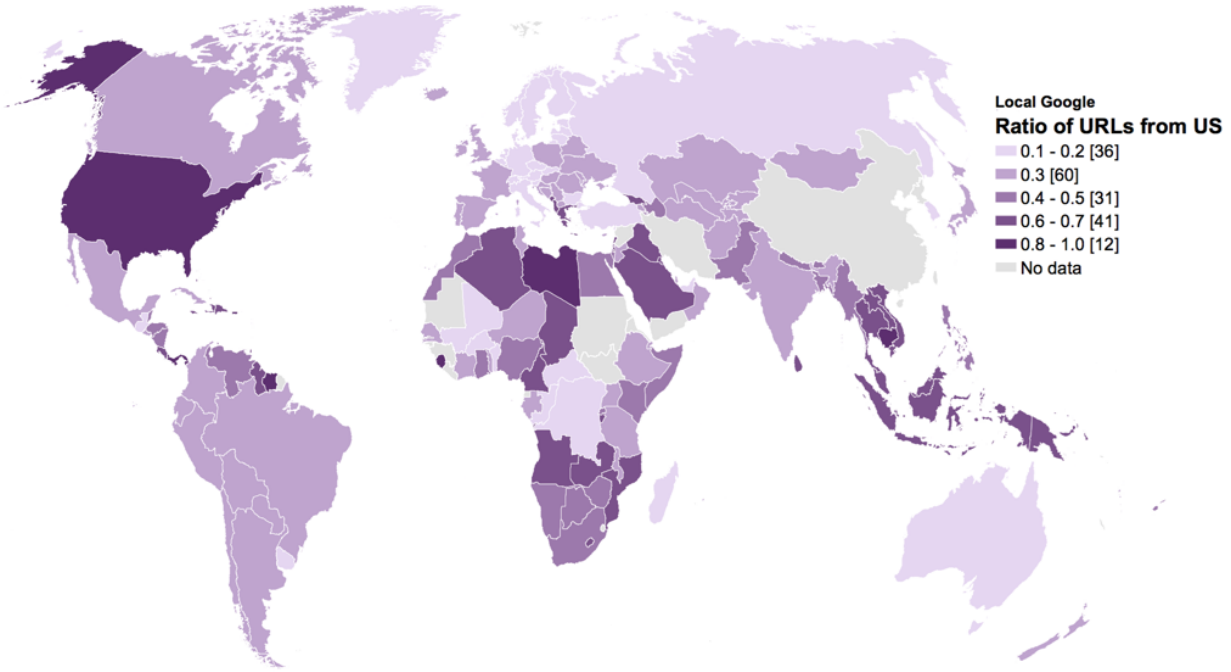

Figure 11: Proportion of URLs from the US in local versions of Google.

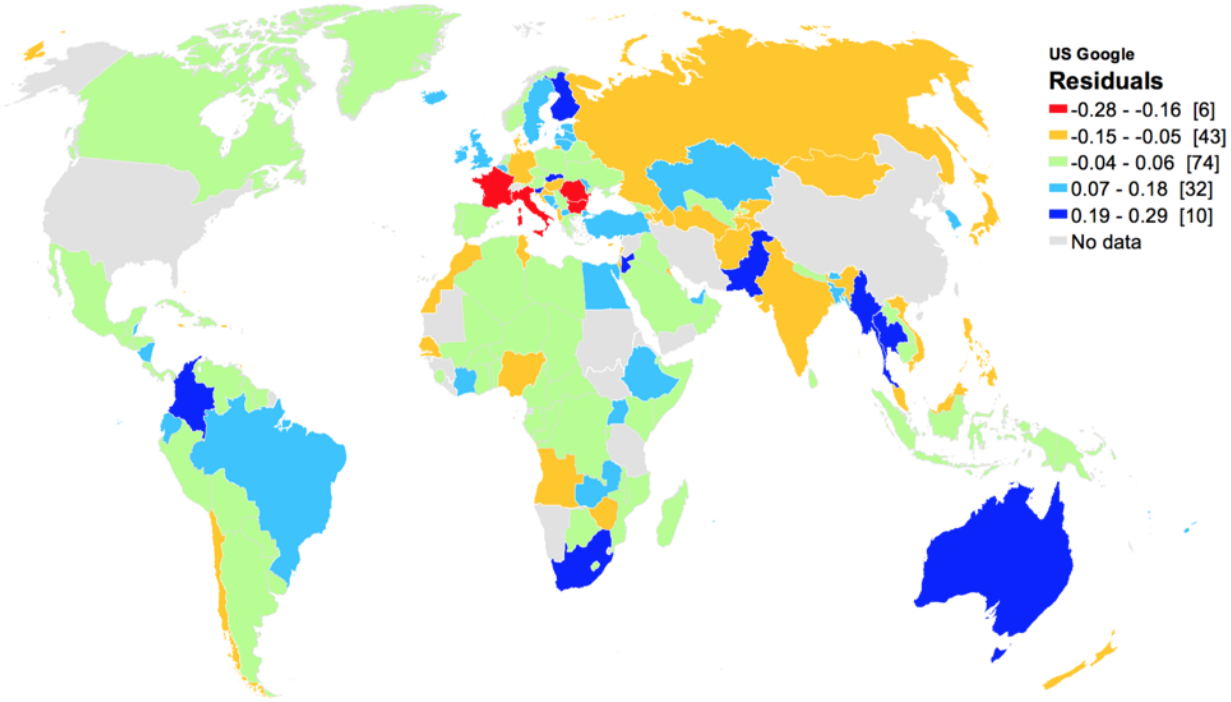

Figure 12: Residuals of regression model for US Google (model M1)

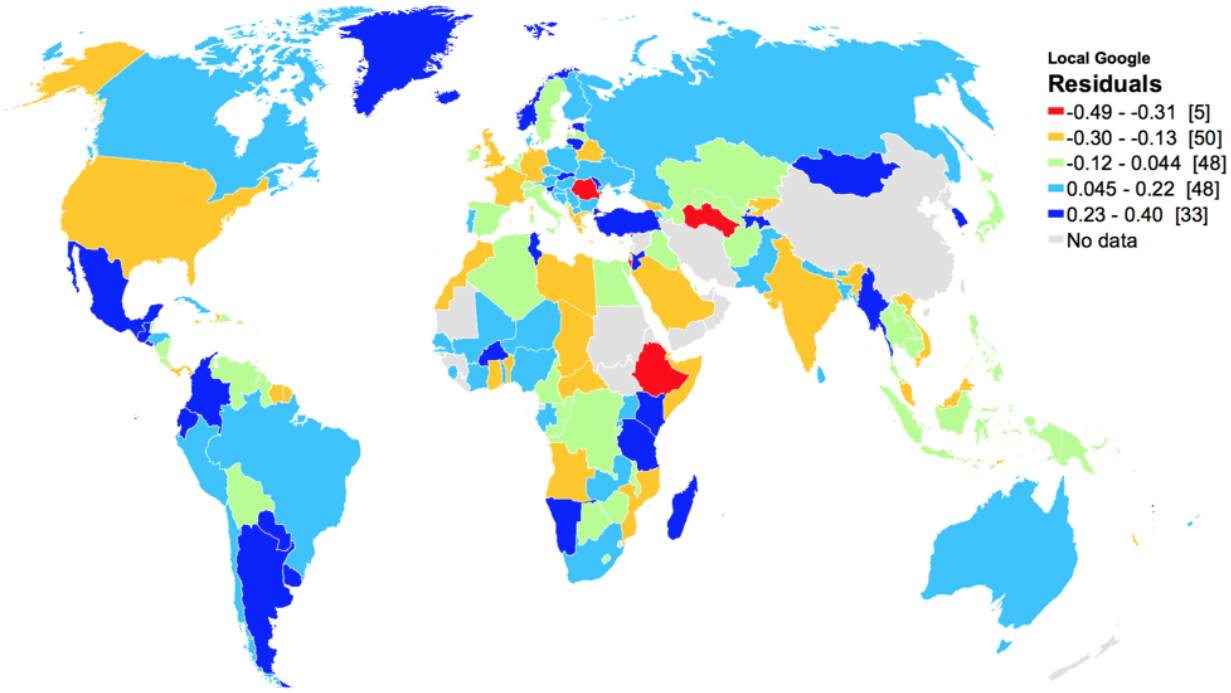

Figure 13: Residuals of regression model for Local Google (model M1)