Distance and Attraction: Gravity Models for Geographic Content Production

Jacob Thebault-Spieker[1,3], Aaron Halfaker[2], Loren G. Terveen[3], Brent Hecht[4]
[1]Virginia Tech, [2]Wikimedia Foundation, [3]University of Minnesota, [4]Northwestern University
{thebault, terveen}@cs.umn.edu, ahalfaker@wikimedia.org, bhecht@northwestern.edu

ABSTRACT
Volunteered Geographic Information (VGI), such as contributions to OpenStreetMap and geotagged Wikipedia articles, is often assumed to be produced locally. However, recent work has found that peer-produced VGI is frequently contributed by non-locals. We evaluate this approach across hundreds of content types from Wikipedia, OpenStreetMap, and eBird, and show that these models can describe more than 90% of "VGI flows" for some content types. Our findings advance geographic HCI theory, suggesting some spatial mechanisms underpinning VGI production. We also discuss design implications that can help (a) human and algorithmic consumers of VGI evaluate the perspectives it contains and (b) address geographic coverage variations in these platforms (e.g. via more effective volunteer recruitment strategies).

Author Keywords
Volunteered geographic information (VGI), gravity models, OpenStreetMap, Wikipedia, eBird, spatial interaction models, geographic HCI.

ACM Classification Keywords
H.5.3. Information interfaces and presentation (e.g., HCI): Group and Organization Interfaces;

INTRODUCTION
Geotagged Wikipedia articles, OpenStreetMap contributions, bird sightings submitted to eBird, and other types of peer-produced volunteered geographic information (VGI) represent critical information resources. For instance, geotagged Wikipedia articles are among the most-visited articles on Wikipedia [22]. OpenStreetMap underpins many consumer maps (e.g. Mapbox, Craigslist, and Apple Maps, among others [28]). eBird is the largest biodiversity dataset of its kind [61]. VGI also directly enables other important endeavors: it helps in disaster relief [8], can aid in epidemiology [14] and earthquake prediction [50], and may even influence regional economic growth [27].

Because VGI pervades many aspects of computing and beyond, factors that influence its use – e.g. quality and completeness – have significant impact. Recent work has begun to show that locally-produced VGI contributions are higher quality [10] and reflect richer [28] or more diverse [60] information. Additionally, local perspectives are known to have innate value for certain use cases (e.g. [52]). Thus, how and where VGI is produced has important implications for its use.

VGI often has been assumed to be largely local. This idea can be traced back to when Michael Goodchild coined the term "volunteered geographic information". Goodchild conceptualized a network of "humans as sensors" [18] wherein people mostly contribute information that is nearby. Goodchild even suggested that "*the most important value of VGI may lie in what it can tell about local activities in various geographic locations"*. The intuition behind Goodchild's conception of VGI is easy to understand. After all, it is probably easier to contribute nearby because people are more likely to be knowledgeable about their home area.

However, recent work has problematized this "localness assumption" [30] for peer-produced VGI. For instance, Hecht and Gergle [24] found that up to 93% of peer-produced VGI content is non-local. This disconnect between the localness assumption and the reality of peer-produced VGI – and the impact this disconnect has on the value of VGI content – calls for an alternative model of how VGI is produced.

In this study, we propose and evaluate one such alternative model. Our approach is based on *spatial interaction models*, long used as a means of understanding geographic interaction patterns in the social sciences. For instance, spatial interaction models are commonly leveraged to understand the transportation of goods between countries [32,35,36].

Here, we apply *gravity models* – a sub-class of spatial interaction models – to understand how VGI is produced. We compare gravity models against two baselines that evaluate opposing perspectives on VGI production. The first (our *local production* baseline) implements a version of the localness assumption, i.e. that where people contribute is determined by distance alone. The second (our *distance is dead* baseline) represents the complete inverse of the localness assumption, i.e. that people merely contribute based on the attractiveness of the contribution target, and distance has absolutely no effect. Conceptually, gravity models merge the ideas from both of these two baselines –

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.
ACM ISBN 978-1-4503-5620-6/18/04.
https://doi.org/10.1145/3173574.3173722

distance impacts where contributions occur, but attractiveness of the location also informs where contributions occur, and may counteract the effect of distance. As a fundamental part of this evaluation, we follow recent calls for simultaneously considering multiple online communities rather than focusing on a single community (e.g. [1,49]). Specifically, we examine three different VGI platforms: Wikipedia, OpenStreetMap, and eBird.

Our gravity models yield an important theoretical insight: we find that gravity models perform meaningfully better than either baseline and describe more than 90% of the 'VGI flows' from one region to another in some cases. Overall, this suggests that understanding peer production VGI as a form of traditional spatial interaction is more effective than understanding peer production VGI as a type of local production (as in the localness assumption).

Our results also have implications for multiple types of VGI stakeholders and suggest important areas of future work. In particular, we discuss how variations in the effects of distance problematize some approaches to VGI "editathons", might suggest mechanisms for understanding demographically-linked coverage biases in VGI, and help to define where local perspectives may be present in VGI and where they may be absent.

RELATED WORK

Our research is mainly informed by four threads of prior work: (1) applications of gravity models in the social sciences, (2) studies that explore local contributions in VGI, (3) research into geographic variations in VGI content, and (4) overall applications of VGI. Below, we describe each of these areas and how they informed our work.

Spatial Interaction Models in the Social Sciences

Modeling spatial interactions between regions has a long history in the social sciences, particularly in the field known as economic geography. *Gravity models*, which date back to 1948 [54], are the most common approach to spatial interaction modeling. Gravity models aim to capture the interaction between two regions based on the 'gravitational' attractiveness of each region and the friction of distance between the two regions. In the almost 70 years of their existence, gravity models have been used to effectively understand a wide variety of spatial phenomena, primarily in two domains: (1) *transportation* of goods and people (e.g. international wheat transactions [35], inter-state gun trades [32], international meat trades [36]) and (2) *communication* patterns (e.g. inter-city [37] and international phone calls [9]).

Our research is directly motivated by the effectiveness of gravity models in explaining these phenomena. As we describe below, we hypothesized that a contribution to one region by a VGI contributor based in another region could be modeled similarly to a product (e.g. meat, wheat) being exported from one region and imported to another. In other words, even with the complex dynamics associated knowledge production in online communities, we believed the spatial dynamics of VGI contribution can be thought of as transfers of units of information from one region to another. Our results indicate that this hypothesis was supported.

While gravity models have traditionally been a social science method, a few studies in HCI and related fields have utilized gravity models to model transportation and communication phenomena. For instance, Smith et al. [53] used a gravity model to model public transit flow and evaluate its potential to predict urban socioeconomic status, García-Gavilanes et al. [13] used a gravity model to model the flow of tweets between countries, Scellato et al. [51] discussed the relationship of gravity models to their approaches to examining the socio-spatial properties of location-based social networks, and techniques related to gravity models have been used in other projects as well (e.g. [31,33]).

Beyond the critical implicit value of better understanding VGI production processes, our work also highlights that gravity models may be useful in computing domains further afield from their typical applications in transportation and communication. In this case, we identify that gravity models are surprisingly effective at capturing a knowledge production relationship between a person and a place as mediated by complex online community dynamics. We return to this point in the Discussion section.

Localness and VGI

As noted above, VGI has been considered a largely local phenomenon since the term "volunteered geographic information" was coined a decade ago [18]. As a result, the extensive and interdisciplinary VGI literature tends to presume that VGI is contributed by locals. Studies and systems that make a VGI "localness assumption" [30] range from studies of VGI contributors' "spatial footprints" (e.g. [39,45]) to epidemiological analyses (e.g. [14]) to applications of sentiment analysis algorithms (e.g. [43]). Johnson et al. [30] provides a summary of applications of the localness assumption in VGI research and practice.

Recent work, however, has begun to call the localness assumption into question. This work – the direct inspiration for our research – has found that a substantial proportion of VGI contributions are in fact, non-local. For instance, Hecht and Gergle [24] observed that between 75 and 93% of edits to geotagged Wikipedia articles by anonymous (non-registered) editors were non-local, depending on the language edition (77% for English). Similar findings were observed by Hardy et al. [21], who modeled the relationship between IP-geolocated anonymous Wikipedia editors and the locations of their geotagged contributions (we used this approach to inform our *local production* baselines). Hecht and Gergle also observed that geotagged Flickr photos tended to be more local – although far from exclusively so – with around 50% of photos being taken by people outside their home region (100km). Sen et al. [52] found that geotagged Wikipedia articles about certain areas are

significantly more local than others, with articles about sub-Saharan Africa being written almost entirely by foreigners in most language editions. Thebault-Spieker et al. [56] found that the 1% of contributors who produce the most OpenStreetMap content also have the largest geographic contribution ranges. Finally, with respect to social media VGI (e.g. geotagged tweets), Johnson et al. [30] also observed a substantial degree of non-local contribution, averaging roughly around 25% depending on the definition of local.

The above work robustly establishes that the localness assumption in peer production VGI is problematic. This raises an important question: If peer production VGI is largely contributed by non-locals, (1) where are these non-locals and (2) how do they make their contribution decisions? We begin to address both aspects of this question in this paper. More generally, being able to effectively model peer production VGI contributions can help fill the theoretical and practical gap between the localness assumption and the reality of peer-produced VGI.

Geographic Variations in VGI Content

A growing body of work has shown that demographic factors are often associated with geographic variations in the quantity and quality of VGI contributions (e.g. [19,28,38]). Two demographic factors that are particularly linked to VGI content variations are socioeconomic status and the rural/urban divide. In short, low-SES and rural areas have been found to have fewer and lower-quality VGI contributions than wealthier and more urban areas [19,28,38]. For instance, in OpenStreetMap, Haklay [19] found less and lower-quality content in low-SES regions of London. Similarly, in Wikipedia, Johnson et al. [28] reported a similar trend concerning the rural/urban divide, observing that Wikipedia content about rural areas is often little more than bot-written template articles. In social media VGI, Li and Goodchild [38] found fewer tweets and photos submitted from low-SES regions of California.

The above work shows that VGI repositories tend to advantage urban and wealthier areas (among trends in other demographic dimensions). However, when taken together with work that suggests the traditional localness assumption does not hold in peer-produced VGI, it becomes clear that very little is understood about the spatial mechanisms behind VGI production. In this paper, we show that by formulating peer-produced VGI contributions as a type of spatial interaction, we can begin to gain a better understanding of these mechanisms.

Applications of VGI

There are three main types of applications of VGI: (1) direct consumption by readers/users, (2) scientific studies, and (3) intelligent technologies and other systems. The success of these applications tends to be closely tied to the coverage and quality of their underlying VGI datasets. As we discuss below, our work here suggests spatial production mechanisms that may underlie variations in VGI coverage and quality.

With regard to direct consumption, geotagged articles are some of the most persistently popular articles on Wikipedia [22] and OpenStreetMap powers many prominent mobile maps applications like Apple Maps [28]. In this case, VGI coverage and quality have a direct impact that is highly visible to the public. Scientific applications of VGI that rely on the coverage and quality of VGI include the effects of tourism on water quality [34], detecting the epicenter of earthquakes [50], and others discussed in more detail by Venerandi et al. [57] and Wood et al. [59]. VGI has also become a key input to many intelligent technologies, like geolocation inference techniques (e.g. [6,31]), among others [5,12]. Indeed, geolocation inference (e.g. of people and documents), is a domain in which coverage and quality has verified importance [29]. Further still, there is some evidence that VGI coverage and quality can impact economic growth [27].

METHODS

Datasets

One of the key findings in previous work is that the spatial production dynamics in VGI may differ based on the community. Therefore, to more robustly evaluate the role of spatial interaction dynamics in VGI production, we examined three VGI platforms: Wikipedia, OpenStreetMap, and eBird.

Further, contribution in each of these communities is a heterogeneous process; that is, some types of content in a given community may support different types of spatial interactions than other types of content. For example, editing Wikipedia articles about national parks (which are globally known) may have a different spatial interaction profile than editing Wikipedia articles about elementary schools (for which information is more locally concentrated). A similar dynamic may exist in OSM with respect to, for example, encoding state borders versus tracing and labeling ("tagging") specific electrical infrastructure.

Therefore, within each of our three platforms, we examine contributions at the level of the *content type*. We analyze the effect of spatial interaction for each content type individually, as well as at the overall platform level. Example content types include articles about schools for Wikipedia (as defined by WikiProjects), electrical towers for OSM (as defined by tags), and bald eagles (*Haliaeetus leucocephalus*) for eBird (as defined by species). In total, we examined 561 different content types, with 101 content types in Wikipedia, 192 content types in OpenStreetMap, and 268 content types in eBird.

As is common in VGI research (e.g. [29,38,42,48]), we focus on data from a single study area: the continental United States. We explore how our research can be expanded to other study areas in our discussion of future work below.

We next describe in more detail the datasets we developed for each of our three VGI platforms.

Wikipedia

Our Wikipedia dataset focused on contributions to geotagged Wikipedia articles. A contribution can be anything from creating new article text to fixing a typo. We queried the English Wikipedia public database [55] for all contributions by registered users to geotagged articles that were saved in the year between Oct 2015 and Oct 2016 (resulting in 3.5 million edits). We then limited the data to articles located within the continental United States, leaving 644,480 total contributions.

For each edit, we used its associated *WikiProject* as the content type (approximately 4% of the contributions had no WikiProject assigned and were excluded). A *WikiProject* is a self-organized group of people working to improve Wikipedia content on a certain topic. For instance, WikiProject Schools is a group of contributors who work to curate school-related content in Wikipedia. We excluded the smallest WikiProjects (with fewer than 1,000 contributions) in order to ensure sufficient data to fit a model, resulting in *101 Wikipedia content types*.

OpenStreetMap

Our OpenStreetMap dataset focused on node (point) contributions. An OpenStreetMap node may be a tree, a traffic circle, or a label point for an electrical tower. In OpenStreetMap, tags are used to describe different types of nodes. A tag consists of a key-value pair, with only one value allowed per key. For example, a 'natural=tree' tag on a node denotes that this node represents a tree and 'junction=roundabout' denotes a traffic circle node. Entities like buildings or roads are normally represented by 'ways', logical groups of nodes. However, attributes of the way (e.g. height of the building, street name) are applied to the way, not the with individual nodes, and therefore are not included in our dataset. We used the full history of OpenStreetMap nodes in the continental USA through February 2014. We excluded nodes that did not have one of the 1,000 most-popular tags (e.g. to eschew typos). From this set of nodes, we than randomly sampled 2,000,000 nodes for analysis.

We used the tags of a node to define its content type(s). As such, all tree contributions were defined as one content type, all traffic circles as another, and so on. As noted above, we excluded the smallest content types (with fewer than 1,000 contributions) in order to successfully fit our models, resulting in *192 total OSM content types*.

eBird

eBird is an observational citizen science project in which a contribution is a bird sighting. As opposed to Wikipedia and OSM, in which one does not need to be physically present in order to contribute, eBird *contributors need to be at or near the location of their contributions.* This geographic proximity requirement makes eBird an interesting comparison point to Wikipedia and OpenStreetMap. As is shown in the Results section, this comparison point will prove to be a valuable reference for understanding contributions in OSM and Wikipedia.

To gather an eBird dataset, we began with the full history of eBird observations through April 2015. We then randomly sampled 2,000,000 observations from this data set, and again limited this data to the continental United States, resulting in 1,573,798 total observations. To understand spatial interaction by content type, we defined content type by sightings of a particular bird species. Again, we excluded the smallest species (with fewer than 1,000 observations) to ensure successful model fitting, resulting in *268 eBird content types*.

Defining the Geographic Origin of Contributions

Prior to modeling spatial interaction processes in peer-produced VGI, we first had to verify our three datasets actually are largely non-local. To do so, we needed to define two properties for each contribution: (1) the local (home) region of its contributor (*i*) and (2) the region in which the contribution was made (*j*). We also had to determine the spatial scale at which a region would be defined. For this, we used the scale of U.S. counties, a common choice in VGI analyses [2,3,26,28,42].

Determining the county in which a contribution is made (*j)* is straightforward: we use the geotag attached to each contribution and perform a reverse geocoding operation. Determining the home region of a contributor (*i)*, on the other hand, is significantly more complex. Unlike social media user profiles, contributors to our VGI repositories have no widely-used means by which they state their home location. Although some contributors do so voluntarily in venues like Wikipedia user pages, participation is low and available only in certain repositories. Similarly, prior work has used IP address geolocation [21,24] when studying Wikipedia, but contributor IP is not available in all of our repositories (and would likely suffer accuracy problems at the county scale [47]). Moreover, even within Wikipedia, IP addresses are only available for anonymous editors [52].

As such, it was necessary to do *home location inference* to determine the county *i* of each contribution. Fortunately, this is a common task, and numerous solutions exist [30]. We adopted the home location inference technique known as *plurality*, which defines a contributor's home region (county) as the region (county) in which s/he has made the plurality of their contributions; this technique has been used in a number of VGI and VGI-related studies (e.g. [26,30,44]). We excluded contributors with fewer than 5 contributions in order to be confident in the inferred county. Following recent calls for researchers to validate home location results across multiple inference techniques [30], we also calculated the home location of each contributor using the *geographic median* approach [6,30,31]. We found that well over 90% of identified home counties were identical across the two approaches, giving us high confidence that both approaches would lead to very similar results in a spatial interaction

model. Therefore, we used the plurality approach in our analysis.

To verify that our datasets violate the assumption of being local, we examined the percentage of contributions in which the contributor's home county *i* is not equal to the contribution county *j*. The results of this simple analysis made clear that the large degree of non-local contributions identified in prior work is replicated in our datasets: only 26% of Wikipedia contributions, 23% of OSM contributions, and 57% of eBird contributions occurred in the plurality-defined home county of their contributor.

These findings justified our further exploration of spatial interaction as an alternative model of VGI production. Below, we describe how we performed these analyses using gravity models.

Gravity Modeling

Intuition

Spatial interaction models seek to explain the relationship between two locations (*i* and *j*) using the distance between them and their individual attributes. More formally, they ask the following: *how does location* i *interact with location* j*, based on the attributes of* i*, the attributes of* j*, and the distance between* i *and* j? Gravity models specifically assume that these relationships can be modeled through an analogy to the basic formula for gravity in the physical world [54]:

$$F_{ij} = \frac{M_i M_j}{D_{ij}^2}$$

When considering the physical gravitational pull two objects have on one another, the *mass* of each object describes their attraction to one another, which is moderated by the *distance* between them. The gravity model takes this intuition, and applies it to interaction between geographic regions, rather than, for example, planets in outer space. The amount of interaction – the dependent variable – is commonly represented as F_{ij} (or 'flow between regions'). The 'mass' variables (M_i for region *i*, and M_j for region *j*) are typically the population of the area, GDP of the area, or other 'attraction' attributes (e.g. [4,32,35,36]). Because gravity models are intended to help understand *interaction*, it is critical that the mass variables incorporate both *potential outflow* (leaving *i*) and *potential inflow* (entering *j*). For instance, using GDP for both mass variables (M_i and M_j) is common for physical processes like international meat trading [36], because it accounts for both exports (potential outflow from *i*) and imports (potential inflow to *j*). The final variable in a gravity model, distance (D_{ij}), is often operationalized as geodesic (straight-line) distance between two regions. Note that D_{ij} has an exponent of 2. In this traditional formulation of the gravity model, 2 is the *friction of distance* – the rate at which interactions between *i* and *j* decay as distance increases.

Airline travel is a common intuitive example for understanding how these variables relate to one another. Consider the case of three cities: New York City, Los Angeles, and Bangor, Maine (a city of about 33,000 residents), with the mass variables set to the population of each city. In this case, population operationalizes both the *potential outflow* from a city and the *potential inflow* to a city (more people usually means more business and personal travel, etc.). New York City and Los Angeles are on opposite coasts of the United States, and thus have a large D_{ij}. However, many people fly back and forth between New York City and Los Angeles due to the large 'attraction' (i.e. large product of masses) between the two cities, which overcomes the large distance (large D_{ij}). On the other hand, despite the much smaller D_{ij} between Bangor and New York City, the tiny mass of Bangor counteracts the shorter distance, and many fewer people fly between Bangor and New York City.

Applying Gravity Models to Our Datasets

In the traditional formulation of the gravity model (above) the friction of distance is defined as -2, and the weights of M_i and M_j are held constant, predefining the degree to which they affected F_{ij}. Because of this, the traditional form was generalized and transformed to a log-linear OLS model (below) [11]. The friction of distance was no longer held constant (at -2), and M_i, M_j, and D_{ij} all became independent variables, predicting the dependent variable F_{ij}.

$$F_{ij} = \frac{M_i^{\beta_1} M_i^{\beta_2}{}_i}{D_{ij}^{\beta_3}}$$

$$\ln(F_{ij}) = \beta_0 + \beta_1 * \ln(M_i) + \beta_2 * \ln(M_j) - \beta_3 * \ln(D_{ij})$$

This straightforward approach, however, causes problems when the variables contain zeroes. After all, the natural log of zero is undefined. Further, the common practice of adding a small constant (such that there are no zeroes) produces biased estimates [11]. To address this problem, we employ one of the most common solutions (recommended by [11]): fitting a Poisson linear regression, which does not risk biased estimates in the scenario mentioned above. It is common to take the natural log of all independent variables, so we use this strategy in our models.

The first step in operationalizing our gravity models is defining *i* and *j*. We use the same definitions as before: *i* is the 'home' region of a contributor, and *j* is the region in which a contribution is made. Every inter-county interaction is thus modeled as someone based in county *i* contributing information about county *j*, aggregated over all contributions in a content type. In other words, if *i* = Wayne County, Michigan and *j* = Baltimore County, Maryland, the goal of the models is to accurately predict the number of contributions about places in Baltimore County made by people whose home county is Wayne County (F_{ij}). To make these predictions, we define M_i to be the number of contributors from county *i* (e.g., Wayne County) that make contributions elsewhere (potential outflow), and M_j to be the number of contributors from anywhere that make

contributions (potential inflow) into county j, (e.g. Baltimore County. We follow common practice, and consider D_{ij} to be the geodesic distance between i and j. We make these predictions separately for each content type in each repository. In other words, we run a separate gravity model for each of our 561 (101 Wikipedia + 192 OSM + 268 eBird) content types.

Traditionally, predictions for intra-regional flow are excluded when constructing gravity models, for two primary reasons. First, intra-regional flows are not intuitive for many physical processes, e.g. we generally do not speak of a country trading wheat with itself. Second, it is not intuitive what the distance from a region to itself ought to be, and using zero can be problematic for reasons mentioned above.

However, for our purposes, these reasons do not apply. First, in our data, non-trivial quantities of VGI content is locally produced (intra-regional) – as much as 43% in the case of eBird. Second, because we follow more recent common practice and implement our gravity models as Poisson regressions, defining a small intra-regional distance will not cause biased estimates. Therefore, we adopt two approaches for intra-regional distances that have been used in the literature [16,20]:

1. constant 1 km for every region, and
2. $\frac{1}{2} * \sqrt{regional\_area}$.

We compare results from both these approaches, which we term *constant-distance* and *regional-distance*, respectively.

Because we build so many models, statistical intuition suggests that a small portion of any significant results would be due to chance. However, as we will see below, our overall results are sufficiently widespread that they are quite robust against the occasional Type I error.

Contextualizing Gravity Models

To provide context for our evaluation of how well gravity models describe VGI production, we also construct two baselines against which to compare our gravity models. To make sure our baselines and our gravity models are directly comparable, we construct both baseline approaches with a Poisson regression. As is the case in our gravity models, all variables we include in these models are log-scaled.

Our first baseline is a set of *distance is dead* models. These represent an interpretation of VGI production in which the distance between i and j is irrelevant. After all, in Wikipedia and OpenStreetMap there are no technical limitations preventing a contributor from contributing about anywhere in the world. This model seeks to predict F_{ij} with only M_i and M_j as independent variables. If indeed these models perform better than our gravity models, it will indicate that distance is unimportant in VGI production.

We also have our *local production* baseline, which represents an interpretation of VGI production in which the localness assumption holds. More formally, this baseline postulates that distance between i and j is the only meaningful factor in why contributions flow between i and j. This model seeks to predict F_{ij} using only D_{ij} as an independent variable, for each content type. Because of the two different approaches to intra-regional distances we discuss above, this baseline is composed of two different sets of models, one for each approach. If these models perform better than our gravity models, it will indicate that attributes of i and j have no bearing on VGI production, and would provide support for the localness assumption.

To properly evaluate if gravity models are even effective at characterizing VGI production, and because the literature suggests two alternatives for incorporating intra-zonal predictions into gravity models, we construct five separate models, across hundreds of different content types. Specifically, we construct one *distance is dead* baseline, two instances of our *local production* baselines, and two instances of gravity models. We then compare all five, and evaluate which are most successful at describing peer-produced VGI.

Summary of Methods

To summarize:

- M_i is the number of contributors from county i, M_j is the number of contributors who contribute in county j, and D_{ij} is geodesic distance between i and j.
- We construct all models as Poisson regressions, following the recommendations of [11].
- Because some VGI contributions occur in the same county where their contributor lives, we evaluate two approaches for defining D_{ij} when i and j are the same county: 1 km, and $\frac{1}{2} * \sqrt{regional\_area}$. This is true for both our *local production* baseline, and our gravity models.
- We construct five different models for each of the 561 different types of content.

RESULTS

We now turn to our results, first evaluating if our gravity models are even a reasonable approach to understanding VGI production, and then engaging in a deeper exploration into the effect of distance in spatial interaction models.

Evaluating Model Fit

Informed by their long history and theoretical underpinnings, we believed that gravity models would likely be effective descriptors of VGI production, though this was by no means guaranteed. Therefore, our first task was to evaluate this conjecture. We did so by comparing the pseudo-R^2 values from each content type, across our two *local production* baselines, our *distance is dead* baseline, and both instances of our gravity models.

Figure 1 shows the distributions of pseudo-R^2 values along the x-axis (as measured by the pseudo-R^2 metric suggested in [7]). The y-axis lists each of our five models, and each

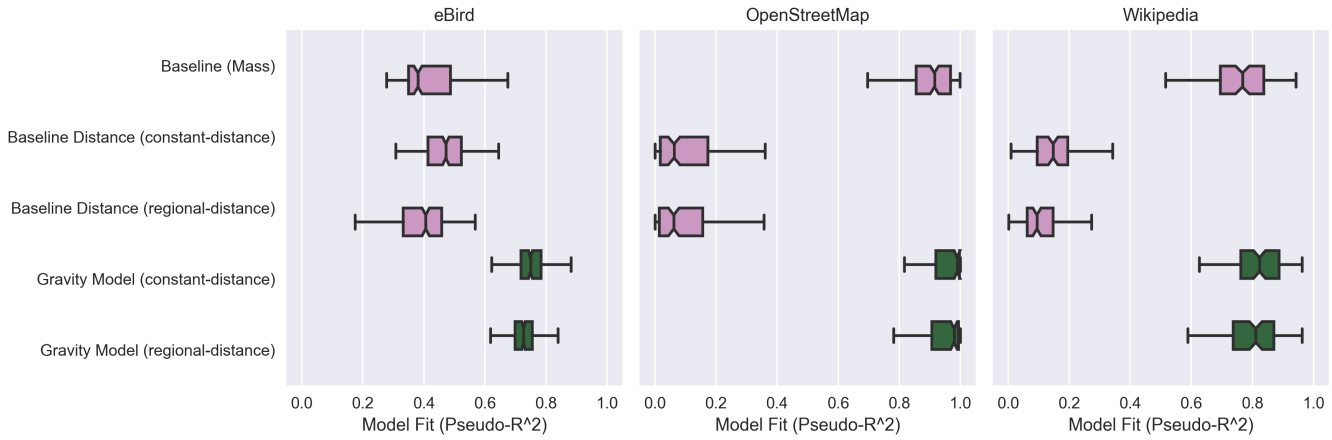

Figure 1: On the *x*-axis, each plot shows the model fit (pseudo-R^2). The *y*-axis shows each of the five models we are evaluating. Each distribution excludes outliers. The top three are baseline models, and the bottom two are gravity models.

chart indicates a different platform. The top three models are baselines, and the bottom two show our two different instantiations of gravity models.

To evaluate differences between distributions, we use a notched boxplot. In a notched boxplot, the medians of two distributions can be considered significantly different when the boxplot notches (the indentation in the middle) do not overlap [41]. However, because we are testing the significance of differences between groups of model output, the distributional assumptions are unclear and significance should be interpreted with some caution. Effect sizes, on the other hand, do not have this issue.

Figure 1 shows that all six gravity models perform better than the *distance is dead* or *local production* baselines (five of them significantly so). Put another way: the gravity model medians are larger, and in most cases the notches in our gravity model boxplots do not overlap with the notches in our baseline boxplots. Examining the medians of each distribution in more detail, the general trend is clear: spatial interaction models are very successful at describing VGI contributions with median pseudo-R^2s as high as 0.99 in some cases. eBird has the lowest median pseudo-R^2s, at 0.74 and 0.72 for the *constant-distance* gravity models and *regional-distance* gravity models respectively. Wikipedia content types show better fits than eBird content types, with median pseudo-R^2s of 0.82 and 0.91 for the *constant-distance* and *regional-distance* gravity models. OpenStreetMap content types tend to show the highest median pseudo-R^2s, at 0.99 and 0.98 (for *constant-distance* and *regional-distance*, respectively).

Focusing on the baseline model distributions in more detail, we noticed some striking differences between platforms. In eBird, a platform where contributors *must* travel to make contributions, the *distance is dead* and *local production* baselines tend to be much more similar in terms of model fit. This is in contrast to OpenStreetMap and Wikipedia, where the *distance is dead* models fit substantially better than the *local production* models.

Comparing the pseudo-R^2s of the baselines to those of the gravity models provides additional insight into how the mass and distance affect the model fit of our gravity models. Our gravity models fit quite well, and in OpenStreetMap and Wikipedia, the *distance is dead* baseline models also fit quite well. This suggests that in the 'wiki' platforms (Wikipedia and OpenStreetMap) where "armchair editing" is possible, the mass variables drive a substantial portion of the gravity model fit. Put another way: in OpenStreetMap and Wikipedia, the content type and its attraction dynamics between regions matters much more than geographic distance for how contributions flow between regions. However, in eBird both distance and content type matter for where people contribute. Someone may contribute to Wikipedia about famous golf courses in Florida, regardless of where they live. Conversely, while some bird species may be rare or more interesting, a bird-watcher still must travel to contribute about these rare birds.

The high-level conclusions of our results at this stage are clear. These findings suggest that VGI production in these platforms indeed can be understood as a gravity model spatial interaction process. Even in OpenStreetMap and Wikipedia – where contribution is not an obviously physical process – most of our gravity models still have more explanatory power than our *distance is dead* baselines. Further still, while the *distance is dead* baseline indicates that the mass variables play a substantial role in model fit for OpenStreetMap and Wikipedia, the addition of a distance variable does improve model performance. This means in spite of fact that a contributor does not need to move at all to edit Wikipedia or OpenStreetMap, there is a degree to which contributions in these platforms are impacted by how far away they are from a contributor.

Interpreting Our Models

Our results in the previous subsection indicate that our gravity models perform better than our baselines, and effectively explain a large portion of the spatial contribution decisions of VGI contributors. Therefore, we now limit our

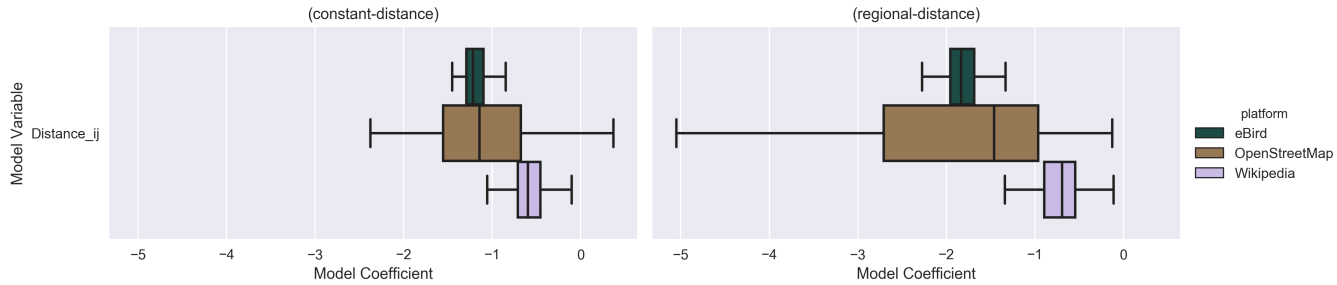

Figure 2: These plots show our D_{ij} variable coefficients, for each platform. On the left are our *constant-distance* models, and the right shows our *regional-distance* models. Both exclude outliers.

discussion of results to our gravity models. While we present both our *constant-distance* and *regional-distance* models, we will focus our discussion of results around *constant-distance* gravity models, because all three perform significantly better than our *distance is dead* baseline (whereas only two *regional-distance* models outperform the *distance is dead* baseline). We exclude 22 types of content with variables that are not significant (predominantly from OSM), to ensure all distributions are comparable.

We focus specifically on the D_{ij} (friction) coefficient in order to shed light on the degree to which contributions are likely to be local. Recall that the more negative a friction coefficient is (further left in Figure 2), the stronger friction effect exists.

Figure 2 shows the distributions of our D_{ij} coefficients. On the left are the coefficients from our *constant-distance* instance of a gravity model, and on the right are the coefficients from our *regional-distance* models. Each boxplot represents a different platform. eBird is on top, OpenStreetMap is in the middle, and Wikipedia on the bottom.

Immediately visible in Figure 2 is that there are clear differences in the friction coefficients between eBird and Wikipedia. Content types from both Wikipedia and eBird are quite clustered together, and the platforms themselves center around different points on the friction of distance spectrum. In some cases, eBird has species that have similar friction coefficients to some Wikipedia content types, but the overlap between these distributions is small. Surprisingly, content types from OpenStreetMap have a much wider overall distribution – some are much closer to eBird content types, and others are much closer to Wikipedia content types. Put simply, distant contributions are much more expensive in eBird than Wikipedia, and OpenStreetMap contains some types of content that have similar friction coefficients to fundamentally physical processes like eBird.

To explore these friction coefficients in more detail, we now discuss some examples from each platform, moving from left to right, and from top to bottom (highest friction coefficient to lowest, eBird to Wikipedia).

eBird

Near the high-friction end of the spectrum is the purple finch (*Haemorhous purpureus*) sighting content type, with a friction coefficient of -1.45. Succinctly, many contributions of this bird would be highly 'local', or nearby the contributors' home. One hypothesis for why this might be the case is that the purple finch has a very large range, spanning most of the eastern United States. This means that even though eBird contributors might upload reports of purple finch sightings on an everyday basis, while traveling it perhaps might be somewhat boring to continue to upload sightings of the same species when there are novel species available.

On the low end of the friction spectrum for eBird we see the ladder-backed woodpecker (*Picoides scalaris*) with a friction coefficient of -0.85. The ladder-backed woodpecker has a range that contains very popular tourist areas in the United States (e.g. Las Vegas, the Grand Canyon). As such, one hypothesis is that this bird is often reported while eBird users are on vacation in these areas, thereby making these reports distinctly non-local (i.e. having a small friction of distance).

OpenStreetMap

The OpenStreetMap content types are much more heterogeneous, and have a much wider distribution than either of the other two platforms. Near the left-hand side of the OSM distribution is 'addr:city=San Diego', which has a friction coefficient of -2.58. Because San Diego is a city of 1.5 million people (has a large mass), it is likely that this friction coefficient reflects a highly-local bulk import done by a resident of San Diego. This would cause a large number of highly local contributions, and thus a high friction of distance. Another interesting example is 'power=tower' with a friction coefficient of -0.66. This tag might have a weaker friction of distance for a simple reason: using satellite imagery, it is fairly straightforward to identify large steel structures intended for holding electricity lines. A contributor would only need to know where electricity is commonly run, but would not require any local knowledge or context.

Wikipedia

What is initially clear is that the Wikipedia friction coefficients tend to be quite similar to one another. Starting from the left-hand side of the distribution is WikiProject Politics, with a friction coefficient of -1.05. Contributions to

WikiProject politics decrease nearly linearly as the places they contribute about get further away. Intuitively, it seems likely that people are less interested or less aware of the details of politics that are further away from them – as the common saying goes: "all politics is local". On the other end of the spectrum is WikiProject Golf, with a friction coefficient of -0.1. This friction coefficient is quite low. One reason this may be the case is the topic itself – to participate in WikiProject Golf, a contributor would likely be highly motivated by Golf as a topic, and may treat golf courses as vacation destinations as well. The combination of being highly motivated and traveling to play Golf would lead to a quite low friction coefficient.

Summary and Generalizable Conclusions

To summarize our results, we found that in all our content types, gravity models are very effective at describing VGI production. Additionally, we found that contributions in Wikipedia and OpenStreetMap are largely driven by attraction between regions, whereas distance is much more important when describing eBird contribution trends. Further, in two of our platforms (eBird and Wikipedia), the friction coefficients are quite consistent, indicating that some platforms facilitate a specific 'style' of spatial interaction. In contrast, in OpenStreetMap the content types span a large range of friction coefficients.

DISCUSSION

Our results have implications for a several constituencies and research areas. Below, we detail these implications.

Implications for VGI Contributors and Platform Managers

Our model fits align well with an idea implicit in the editing ethos of some large VGI communities – the belief that distance has relatively minimal impact on VGI contribution. For instance, Wikipedia states "anyone can edit almost every page" [58], and OpenStreetMap's introductory documentation says "You can map from your armchair" [46]. From a purely technical perspective, it is just as easy for a person who lives in e.g. Montreal to log into Wikipedia or OSM and contribute information about McGill University as it is for that person to contribute information about, for instance, Nazarbayev University in Kazakhstan.

This raises a key question: what limits some content types from being advantaged by the affordances to map anywhere or write articles about anywhere from "armchairs"? A number of factors likely are responsible, and may help explain reasons behind the attraction processes shown in our models. For example, physical world processes are still highly correlated with knowledge about a region, and knowledge about a region can help one more easily write a Wikipedia article, do OSM mapping, or see and recognize a specific bird. Regional boosterism may also be at play, causing people to build up information about known locations. However, future work should seek to examine the presence and strength of these and other factors. One approach might be a qualitative study to understand where people choose to contribute, and why. This exploratory approach would help shed light on some of the mechanisms that underpin the large attraction processes we see in our results.

Our work has several implications for the design of VGI communities and platforms. Our results present challenges for a particularly common means by which VGI communities attempt to address coverage issues: "editathons". Editathons are usually in-person events and are typically held in urban areas where many potential new contributors can attend. Our results show that – although the various aspects of gravity models can have complex interactions – for high-friction content types these types of in-person contribution drives are unlikely to affect the variations in coverage. To do so requires localized contributors, and it is unlikely that editathons occur in places where contributors are needed most. This is especially troubling as editathons are often funded by the cash-strapped organizations that operate VGI platforms (e.g. the Wikimedia Foundation). Our results suggest that organizations like the OpenStreetMap Foundation may want to redirect some of their resources towards efforts that work towards these goals.

Implications for Coverage Biases

More generally, our work may help to reveal mechanisms for the coverage biases linked to socioeconomic status, the rural/urban spectrum, and other demographics. One hypothesis as to the mechanisms for this coverage variation is that "self-focus bias" is playing a role [23]. That is, people are contributing about places near where they have lived, and, given the demographics of VGI contributors (e.g. [17]), it is likely that they will have lived in higher-SES areas and urban areas. Our results provide a direct means of testing this hypothesis: If this is true, then content types for which the friction of distance is high may exhibit more coverage bias then content types for which armchair mapping is more common. Evaluating this hypothesis is an immediate opportunity for future work.

Our results also highlight a hypothesis for a potential second cause of these biases: preferential attachment. It may be that high-SES areas and urban areas were some of the first areas to be covered in these datasets, thereby making them more "attractive". Because of this attraction – and the importance of attraction shown in our baseline models more generally – these areas' early leads in coverage became effectively permanent. More generally, our baseline models suggest that, at least for OpenStreetMap and Wikipedia, preferential attachment may be a particularly potent force. Testing this "geographic preferential attachment" hypothesis is also an excellent direction of future work.

Implications for Algorithms

Hecht and Gergle showed that AI systems that use VGI for world knowledge can adopt the perspectives of their underlying VGI datasets [15]. Since our results suggest that certain VGI content types will *innately* contain more local perspectives than others, this suggests that VGI-based AI

systems that rely on certain types of data may innately be biased towards perspectives that are more or less local.

Implications for Human Consumers

The exact same biases that may affect algorithms with respect to non-local and local perspectives will also affect human consumers of VGI. For instance, our results suggest that Wikipedia content about golf courses will be less local than its content about politics. This highlights a number directions of future work. Two of the most interesting might be (1) building tools that can surface the fact that local perspectives may not be present for certain content types and (2) using this surfacing to perhaps incentivize more contributions from the local area (e.g. using a prompt like "This article about your local golf course was written entirely by non-locals. Do you have any local expertise to add?")

Implications for Gravity Models and Social Computing

As discussed above, the predominant use of gravity models in HCI and social computing contexts have tended to be in the traditional gravity model domains of transportation and communication (using datasets of interest to the HCI and social computing communities). Our results suggest that gravity models are also quite useful for understanding processes further afield from transportation and communication. At the very least, this work suggests that researchers who are examining the role of distance in a geographic HCI [25] process should consider utilizing gravity model techniques. The primary challenge of moving beyond simple distance involves operationalizing the mass variables, and our discussion of our implementation of mass can provide a reference point along these lines.

One particularly interesting future application builds directly on another prior application of gravity models. Gravity models are commonly used in planning [40] to identify where to place, e.g., Coca-Cola distribution centers. By considering geographic attraction and distance, planners maximize the region a distribution center serves, while minimizing the number (and cost) of distribution centers. By analogy, future work could use gravity models to help allocate volunteer resources within peer production platforms. These adapted models could be used to predict where contributions are likely to go and which places will never receive contributions. These models can also predict where to focus recruitment to maximize the region contributors serve and perhaps mitigate the geographic biases shown in prior literature.

FUTURE WORK AND LIMITATIONS

In constructing our gravity models, we made a number of decisions that open up opportunities for future work. First, we used a plurality inference method to infer home regions and verified that a geographic median approach was unlikely to change our results. Recent work [56] has used a plurality inference method to identify *where contributors focus* instead of where they live. Operationalizing plurality assignments as focus regions provides a different view on our results, one in which home regions become regions of substantial knowledge or expertise. In this view, we believe our conclusions would largely remain unchanged, e.g. different content types would still have different degrees of local expertise associated with them. This alternative conception of plurality does, however, re-emphasize a suggestion we made above: future work should develop a deeper understanding of the attraction processes – and their relationship to local expertise – that seem to drive a substantial amount of contribution. Future work should also consider additional home location inference approaches as a way to account for contributors potentially being 'local experts' in multiple regions.

Second, we did not distinguish between human contributors and bots in Wikipedia, or bulk imports in OpenStreetMap. After all, bulk imports are a key part of these ecosystems, especially for OpenStreetMap. In a quick analysis removing all bulk imports from our OpenStreetMap data, we do not see substantial changes in the distributions of either model fit, or friction of distance. Future work should explore additional content types, with an eye towards similarities and differences between human behavior and bots or bulk imports (e.g. [28]).

Finally, it is standard practice in the VGI literature to focus on a single study site, as we did here (e.g. [29,38,48]). However, future work in this space should likely seek to select study sites in different human geographic regions than those considered here in order to determine whether the distance relationships change in different human geographic contexts.

CONCLUSION

In this paper, we showed that VGI contributions can be modeled effectively using spatial interaction techniques, and gravity models in particular. We also explored the implications of these findings for our understanding of VGI, for stakeholders currently managing large VGI communities, and for the development of future VGI platforms.

ACKNOWLEDGEMENTS

This research was supported by NSF grants IIS-1707319, IIS-1707296, IIS-0964695, IIS-1111201, and IIS-1218826. This research was also supported by the Wikimedia Foundation.

REFERENCES

1. Judd Antin, Ed H. Chi, James Howison, Sharoda Paul, Aaron Shaw, and Jude Yew. 2011. Apples to Oranges?: Comparing Across Studies of Open Collaboration/Peer Production. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (WikiSym '11), 227–228. https://doi.org/10.1145/2038558.2038610
2. Michael Bailey, Ruiqing (Rachel) Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2017. *Measuring Social Connectedness*. National Bureau of Economic Research. https://doi.org/10.3386/w23608

3. Saeideh Bakhshi, Partha Kanuparthy, and Eric Gilbert. 2014. Demographics, weather and online reviews: a study of restaurant recommendations. 443–454. https://doi.org/10.1145/2566486.2568021
4. Jonathan M. Bossenbroek, Clifford E. Kraft, and Jeffrey C. Nekola. 2001. Prediction of Long-Distance Dispersal Using Gravity Models: Zebra Mussel Invasion of Inland Lakes. *Ecological Applications* 11, 6: 1778–1788. https://doi.org/10.1890/1051-0761(2001)011[1778:POLDDU]2.0.CO;2
5. Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning* (ICML '08), 160–167. https://doi.org/10.1145/1390156.1390177
6. Ryan Compton, Craig Lee, Jiejun Xu, Luis Artieda-Moncada, Tsai-Ching Lu, Lalindra De Silva, and Michael Macy. 2014. Using publicly visible social media to build detailed forecasts of civil unrest. *Security Informatics* 3, 1: 4. https://doi.org/10.1186/s13388-014-0004-6
7. Stefany Coxe, Stephen G. West, and Leona S. Aiken. 2009. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment* 91, 2: 121–136. https://doi.org/10.1080/00223890802634175
8. Martin Dittus, Giovanni Quattrone, and Licia Capra. 2017. Mass Participation During Emergency Response: Event-centric Crowdsourcing in Humanitarian Mapping. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 1290–1303. https://doi.org/10.1145/2998181.2998216
9. Stuart Carter Dodd. 1950. The Interactance Hypothesis: A Gravity Model Fitting Physical Masses and Human Groups. *American Sociological Review* 15, 2: 245–256. https://doi.org/10.2307/2086789
10. Melanie Eckle. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. Retrieved September 24, 2015 from http://iscram2015.uia.no/wp-content/uploads/2015/05/5-1.pdf
11. Robin Flowerdew and Murray Aitkin. 1982. A Method of Fitting the Gravity Model Based on the Poisson Distribution*. *Journal of Regional Science* 22, 2: 191–202. https://doi.org/10.1111/j.1467-9787.1982.tb00744.x
12. Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJcAI*, 1606–1611.
13. Ruth García-Gavilanes, Yelena Mejova, and Daniele Quercia. 2014. Twitter Ain't Without Frontiers: Economic, Social, and Cultural Boundaries in International Communication. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '14), 1511–1522. https://doi.org/10.1145/2531602.2531725
14. Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y Del Valle, and Reid Priedhorsky. 2014. Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Computational Biology* 10, 11: e1003892. https://doi.org/10.1371/journal.pcbi.1003892
15. Darren Gergle and Brent Hecht. 2010. The tower of Babel meets web 2.0. In *the 28th international conference*, 291. https://doi.org/10.1145/1753326.1753370
16. M. Gibson and M. Pullen. 1972. Retail turnover in the East Midlands: A regional application of a gravity model. *Regional Studies* 6, 2: 183–196. https://doi.org/10.1080/09595237200185161
17. Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. 2010. Wikipedia Survey - Overview of Results.
18. Michael F Goodchild. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 4: 211–221. https://doi.org/10.1007/s10708-007-9111-y
19. Mordechai Haklay. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37, 4: 682–703. https://doi.org/10.1068/b35097
20. Mark E Hanson. 1966. *Project METRAN: an integrated, evolutionary transportation system for urban areas*. Cambridge, Mass.: MIT Press.
21. Darren Hardy, James Frew, and Michael F Goodchild. 2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science* 26, 7: 1191–1212. https://doi.org/10.1080/13658816.2011.629618
22. Brent Hecht. 2013. The Mining and Application of Diverse Cultural Perspectives in User-generated Content. Northwestern University, Evanston, IL, USA.
23. Brent Hecht and Darren Gergle. 2009. Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. In *Communities and Technologies 2009: 4th International Conference on Communities and Technologies*, 11–19.
24. Brent Hecht and Darren Gergle. 2010. On the "localness" of user-generated content. 229. https://doi.org/10.1145/1718918.1718962
25. Brent Hecht, Johannes Schöning, Muki Haklay, Licia Capra, Afra J Mashhadi, Loren Terveen, and Mei-Po Kwan. 2013. Geographic human-computer interaction. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 3163. https://doi.org/10.1145/2468356.2479637
26. Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *Eighth International AAAI Conference on Weblogs and Social Media*. Retrieved February 13, 2015 from

http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114
27. Marit Hinnosaar, Toomas Hinnosaar, Michael Kummer, and Olga Slivko. 2017. Wikipedia Matters.
28. Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. 13–25. https://doi.org/10.1145/2858036.2858123
29. Isaac L. Johnson, Connor J McMahon, Johannes Schöning, and Brent Hecht. 2017. The Effect of Population and "Structural" Biases on Social Media-based Algorithms -- A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI 2017)*. http://dx.doi.org/10.1145/3025453.3026015
30. Isaac L. Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. The Geography and Importance of Localness in Geotagged Social Media. 515–526. https://doi.org/10.1145/2858036.2858122
31. David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*.
32. Leo H. Kahane. 2013. Understanding the Interstate Export of Crime Guns: A Gravity Model Approach. *Contemporary Economic Policy* 31, 3: 618–634. https://doi.org/10.1111/j.1465-7287.2012.00324.x
33. Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. 2013. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd international conference on World Wide Web*, 667–678.
34. Bonnie L Keeler, Spencer A Wood, Stephen Polasky, Catherine Kling, Christopher T Filstrup, and John A Downing. 2015. Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes. *Frontiers in Ecology and the Environment* 13, 2: 76–81. https://doi.org/10.1890/140124
35. Won W. Koo and David Karemera. 1991. Determinants of World Wheat Trade Flows and Policy Analysis. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* 39, 3: 439–455. https://doi.org/10.1111/j.1744-7976.1991.tb03585.x
36. Won W. Koo, David Karemera, and Richard Taylor. 1994. A gravity model analysis of meat trade policies. *Agricultural Economics* 10, 1: 81–88. https://doi.org/10.1016/0169-5150(94)90042-6
37. Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D. Blondel. 2009. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment* 2009, 07: L07003. https://doi.org/10.1088/1742-5468/2009/07/L07003
38. Linna Li, Michael Goodchild, and Bo Xu. 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science* 40, 2: 61–77. https://doi.org/10.1080/15230406.2013.777139
39. Michael D Lieberman and Jimmy Lin. 2009. You Are Where You Edit : Locating Wikipedia Contributors Through Edit Histories. *Proceedings of the Third International ICWSM Conference*: 106–113.
40. M J Hodgson. 1978. Toward More Realistic Allocation in Location—Allocation Models: An Interaction Approach. *Environment and Planning A: Economy and Space* 10, 11: 1273–1285. https://doi.org/10.1068/a101273
41. Robert McGill, John W. Tukey, and Wayne A. Larsen. 1978. Variations of Box Plots. *The American Statistician* 32, 1: 12–16. https://doi.org/10.2307/2683468
42. Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM* 11: 5th.
43. Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE* 8, 5: e64417. https://doi.org/10.1371/journal.pone.0064417
44. Mohamed Musthag and Deepak Ganesan. 2013. Labor Dynamics in a Mobile Micro-task Market. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 641–650. https://doi.org/10.1145/2470654.2470745
45. Pascal Neis and Alexander Zipf. 2012. Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information* 1, 3: 146–165. https://doi.org/10.3390/ijgi1020146
46. OpenStreetMap. Armchair Mapping - OpenStreetMap Wiki. Retrieved October 24, 2016 from http://wiki.openstreetmap.org/wiki/Armchair_mapping
47. Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. IP Geolocation Databases: Unreliable? *SIGCOMM Comput. Commun. Rev.* 41, 2: 53–56. https://doi.org/10.1145/1971162.1971171
48. Giovanni Quattrone, Davide Proserpio, Daniele Quercia, Licia Capra, and Mirco Musolesi. 2016. Who Benefits from the "Sharing" Economy of Airbnb? *arXiv:1602.02238 [physics]*. Retrieved February 25, 2016 from http://arxiv.org/abs/1602.02238
49. Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213: 1063–1064. https://doi.org/10.1126/science.346.6213.1063

50. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web* (WWW '10), 851–860. https://doi.org/10.1145/1772690.1772777
51. Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. 2011. Socio-Spatial Properties of Online Location-Based Social Networks. In *Fifth International AAAI Conference on Weblogs and Social Media*. Retrieved June 15, 2016 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2751
52. Shilad W. Sen, Heather Ford, David R. Musicant, Mark Graham, Oliver S.B. Keyes, and Brent Hecht. 2015. Barriers to the Localness of Volunteered Geographic Information. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 197–206. https://doi.org/10.1145/2702123.2702170
53. Chris Smith, Daniele Quercia, and Licia Capra. 2013. Finger on the Pulse: Identifying Deprivation Using Transit Flow Analysis. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (CSCW '13), 683–692. https://doi.org/10.1145/2441776.2441852
54. John Q. Stewart. 1948. Demographic Gravitation: Evidence and Applications. *Sociometry* 11, 1/2: 31–58. https://doi.org/10.2307/2785468
55. Jacob Thebault-Spieker, Aaron Halfaker, Brent Hecht, and Loren Terveen. 2018. enwiki.revisions_with_coords.201510-201610.csv. https://doi.org/10.6084/m9.figshare.5764626.v1
56. Jacob Thebault-Spieker, Brent Hecht, and Loren Terveen. 2018. Geographic Biases Are "Born, Not Made": Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork* (GROUP '18), 71–82. https://doi.org/10.1145/3148330.3148350
57. Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. 2015. Measuring Urban Deprivation from User Generated Content. 254–264. https://doi.org/10.1145/2675133.2675233
58. Wikipedia. 2015. Wikipedia:Introduction. *Wikipedia*. Retrieved October 22, 2016 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Introduction&oldid=680454568
59. Spencer A. Wood, Anne D. Guerry, Jessica M. Silver, and Martin Lacayo. 2013. Using social media to quantify nature-based tourism and recreation. *Scientific Reports* 3. https://doi.org/10.1038/srep02976
60. Dennis Zielstra, Hartwig H. Hochmair, Pascal Neis, and Francesco Tonini. 2014. Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap. *ISPRS International Journal of Geo-Information* 3, 4: 1211–1233. https://doi.org/10.3390/ijgi3041211
61. About eBird | eBird. Retrieved September 19, 2017 from http://ebird.org/content/ebird/about/