Geographic Biases are 'Born, not Made': Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap

Jacob Thebault-Spieker
University of Minnesota
thebault@cs.umn.edu

Brent Hecht
Northwestern University
bhecht@northwestern.edu

Loren Terveen
University of Minnesota
terveen@cs.umn.edu

ABSTRACT

The evolution of contributor behavior in peer production communities over time has been a subject of substantial interest in the social computing community. In this paper, we extend this literature to the geographic domain, exploring contribution behavior in OpenStreetMap using a *spatiotemporal* lens. In doing so, we observe a geographic version of a "born, not made" phenomenon: throughout their lifespans, contributors are relatively consistent in the places and types of places that they edit. We show how these "born, not made" trends may help explain the urban and socioeconomic coverage biases that have been observed in OpenStreetMap. We also discuss how our findings can help point towards solutions to these biases.

Author Keywords

Peer production; coverage biases; OpenStreetMap; geographic human-computer interaction.

ACM CLASSIFICATION KEYWORDS

H.5.3. Group and Organization Interfaces: Computer-supported cooperative work;

INTRODUCTION

Peer production has been very successful as a model of content production [3]. For instance, the peer produced Wikipedia is consistently the fifth most-visited website globally [1] and OpenStreetMap – the "Wikipedia of maps" [8] – provides map data to Craigslist, Apple Maps, and others [43]. Large peer produced datasets even play an essential role as training data for many AI systems (e.g. [11,27]).

Contributor choice is a fundamental characteristic of peer production: contributor choice differentiates peer production from other forms of crowdwork [3] and may even be necessary for the success of the peer production content generation model [3]. Indeed, the ethos of contributor autonomy is so foundational in peer production that, for instance, the introductory documentation of OpenStreetMap, states that "anybody can enter anything she wishes" [29].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

GROUP '18, January 7–10, 2018, Sanibel Island, FL, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5562-9/18/01…$15.00
https://doi.org/10.1145/3148330.3148350

Because of the importance of peer produced content and the role of contributor autonomy in producing that content, researchers have long sought to understand and model contributor focus in various peer production contexts (e.g. [6,14,31,33]). One common thread in this research involves studying how contributor focus evolves over the lifespan of a contributor (e.g. [2,31]). In other words, this research examines contributor focus through a *temporal* lens.

While a temporal lens is sufficient to understand contributor evolution in many peer production contexts, in *geographic peer production* – e.g. contributing to OpenStreetMap and editing geotagged Wikipedia articles – a purely temporal lens cannot detect another critical type of potential focus evolution: that which unfolds spatially. For instance, while it is useful to know that an OpenStreetMap contributor is increasing her/his contribution rate, it is also important to understand *where* and in *which types of places* the user is contributing, and *how this changes over time*. Among other applications, such knowledge can provide critical insight into the troubling coverage biases that have been observed in peer produced geographic datasets (e.g. on socioeconomic and urban/rural lines [13,20,37]).

In this research, we extend the literature on temporal focus evolution to geographic peer production with an exploratory analysis that examines contributor focus with a *spatiotemporal* lens. Our work uses OpenStreetMap – the world's largest peer produced geographic dataset – as a case study and centers around two basic research questions adapted from the temporal literature [31]. First, we ask:

(RQ1) How does contributors' geographic focus change over time?

To address this question, we operationalize four geographic contribution metrics and explore if and how they change over time. Overall, our results suggest that contributors are broadly consistent in their geographic editing behavior over the course of their contribution lifespan, although there are some deviations from this trend. Further, the consistency is of a particular nature: people tend to consistently edit in relatively specific geographic areas.

These results recall the findings of one well-known GROUP paper that examined contributor focus with a temporal lens, finding that Wikipedia power editors have different editing behavior than other users from day one of their editing career, i.e. that power editors are "born, not made" [31]. In our study, we observed this "born, not made" dynamic in a very different peer production context: the geographic editing behavior of OpenStreetMap editors (although we observe a somewhat softer version of the dynamic).

Our spatiotemporal approach also advances understanding of mechanisms behind a second (and concerning) trend that has been observed in the literature: geographic biases in peer production. In the face of peer production's immense success – which is predicated on the idea that "anyone can enter anything she wishes" – recent research shows that urban and wealthy areas receive better geographic coverage than rural and less wealthy areas [20,26]. While prior work characterizes these biases, few have studied their root causes. Thus, our second research question asks:

(RQ2) Can the spatiotemporal evolution of contributors' focus help to explain systemic coverage biases?

Our exploratory results suggest that most contributors are "born" urban-focused and wealthier-focused and stay that way. In other words, for most editors, the proportions of edits in rural and poor areas are consistent and consistently low across contribution lifespans. We also find that the few editors who do consistently focus in rural and poorer regions tend to have lower survival rates, exiting OpenStreetMap sooner than their urban- and wealthier-focused counterparts

Our study makes four primary contributions:

- We explore the geographic contribution behavior of OpenStreetMap editors over time and observe that most editors exhibit similar behavior across their entire contribution lifespans. Thus, for many people, we find evidence that *geographic editing behavior is "born, not made"*.
- We show how this consistent contribution behavior applies also to the types of regions people edit. In other words, we find some evidence that *geographic biases also are "born, not made"*.
- These *focus biases are amplified by a survival bias* – people who focus in rural and high-poverty areas tend to contribute for shorter periods of time.
- While we did observe a *small group of people who focus primarily in rural or high-poverty areas*, they *produce only a small portion of OpenStreetMap content*.

RELATED WORK

Our work here builds primarily on prior work in three areas: (1) peer production contributors' geographic contribution behavior, (2) temporal evolution of contributor behavior, and (3) geographic biases in peer production. Below we situate our work relative to each of these areas.

Contributors' Geographic Patterns

The literature examining contributor geographic patterns falls broadly into two categories: *where* contributors focus and the *geographic ranges* of contributors' work. Our research extends these two categories of prior work by considering the evolution of these types of geographic trends *over time*. Below, we describe each category in more detail and put each in the context of our work.

Where Contributors Focus

Several different studies have sought to understand and characterize the geographic focus of contributors to peer production platforms. For instance, Panciera et al. [30] examined geographic trends in the Cyclopath platform, an early bicycling-centered community. In particular, they found that "Cyclopaths" (defined as the top 5% of contributors) had geographically constrained contribution regions, even within the relatively small area in which Cyclopath operated. Zielstra et al. [40] described the geographic extents of 13 OpenStreetMap contributors and show a method of characterizing which contributions are a part of a contributors' 'home location', and which are not. They found that the contribution ranges of these 13 people do not generally exceed approximately 50 square kilometers. Lieberman et al. [25] conducted a similar study, exploring the geographic extent of Wikipedia editors' contributions.

Geographic Ranges of Contribution

Hecht and Gergle [16] compared different 'spatial content production models' for generating volunteered geographic information [9] and found that Flickr contributions tended to be much closer to a contributors' 'home location' than was the case with Wikipedia. Hardy et al. [15] considered geographic contribution as a *spatial interaction* process, using an exponential distance decay model for each language edition. They found that anonymous edits to geotagged Wikipedia articles decay exponentially as the contribution location gets further form a contributor's 'home'. We return to this idea of *spatial interaction* in the Discussion section.

Temporal Evolution of Contributor Behavior

Whereas the work described above focused on geographic behavior, others have focused on the evolution of non-geographic peer production contributor behavior *over time*. In one of the seminal studies in this space, Priedhorsky et al. [33] took a temporal approach to understanding how value is created in Wikipedia and by whom. Panciera et al. [31] built on this paper with a study of 'Wikipedian' lifecycles and found that 'Wikipedians' (the term they use to describe those who contribute most of the Wikipedia content) *begin* contributing at a high level and maintain this trend over time, resulting in distinctive differences in contribution behavior between different classes of users. In other words, "Wikipedians are born, not made" [31]. As noted above, this work strongly informs our study. One of the key takeaways of our work is that this finding, which describes temporal contribution levels in Wikipedia, also applies to spatiotemporal contribution behavior in OpenStreetMap. Panciera's work also inspired the methodologies in this paper: as described below, the spatiotemporal contributor class-specific analyses are a direct analogue to the temporal analyses done in Panciera et al.

Other work uses temporal evolution as a way to characterize the status of a geographic region (versus focusing on contributors and their behavior). One example of such a study is work by Gröchenig et al [12], who computationally estimated the 'completeness' of twelve urban areas, based on identifying three temporal stages ('start', 'growth', and 'saturation'), and modeling the development of a region through these stages.

More recently, others have begun to explore what roles contributors play in peer production communities, and how that changes over time. Arazy et al. [2] described 'career paths' of Wikipedia editors. Rehrl et al. [36] took a similar approach, and considered the different roles that people have in OpenStreetMap. Dittus et al. [6] explored the activation of newcomers and reactivation of previously dormant contributors during disaster events on Humanitarian OpenStreetMap (HOT).

Our study here is deeply informed by the work of Panciera et al. [31], and the studies mentioned in the subsection above. Whereas prior work has focused on understanding geographic behavior *or* the temporal evolution of behavior, our study sits at the intersection. A spatiotemporal lens helps inform our understanding how contributors' geographic behavior evolves, and how this may impact the geographic variations seen in OpenStreetMap.

Geographic Biases in Peer Production

Geographic coverage biases in peer produced datasets have become a subject of relatively substantial research interest in recent years. For instance, Sen et al. [37] found that most content in some parts of the world (e.g. sub-Saharan Africa) is not produced by people from those parts of the world, but instead by Westerners. Other work shows that these biases manifest along two important human geography variables: the urban/rural divide, and socioeconomic status variation. As one example, Johnson et al. [20] found that the quality of Wikipedia and OpenStreetMap content is much greater in urban areas than in rural areas, a result that informs key analyses below. Haklay [13] found a similar result when considering socioeconomic status as well – the quality of OpenStreetMap data is much better in wealthier regions. Informed by these (and other 'geographic HCI' [18,21,26]) studies, we focus one of our research questions on these two specific dimensions (we discuss this in more detail below).

Prior work in this area has quantified and shown the existence of these geographic biases in peer produced datasets, but little work has been done to understand the mechanisms behind these biases. As mentioned above, our work takes a spatiotemporal approach, at the intersection between studies of temporal contributor behavior and those characterizing the geographic behavior of contributors. For this reason, our work is well-situated to shed light on how the temporal evolution of geographic behavior may (or may not) facilitate the geographic biases that others have found.

METHODS

To study the spatiotemporal evolution of contributors in OpenStreetMap, we needed to (1) develop our OpenStreetMap dataset, (2) define geographic variables of interest (i.e. the '*spatio*' in spatiotemporal), and (3) characterize these variables of interest over time (i.e. the '*temporal*'). We first provide a brief introduction to how contributions occur in OpenStreetMap and then discuss each of these three steps.

Introduction to Contribution in OpenStreetMap

Where Wikipedia editors help create articles, OpenStreetMap contributors help create a worldwide map (or, more formally, a worldwide spatial database). OSM contributions either add or annotate geographic entities, e.g. bus stops, roads, buildings or even logical collections of buildings like a university. Nodes (points) are the simplest geometric unit in OSM, and they may stand alone (e.g. a bus stop), or they may comprise other types of geometries, namely 'ways' (e.g. roads or buildings) and 'relations' (e.g. a university). Early in the life of OpenStreetMap, contributions depended heavily on "GPS traces" recorded as contributors moved about the world. However, it is now much more common to trace new entities from satellite imagery using a web-based tool [44].

Similar to Wikipedia, OpenStreetMap records a "version history" for each map entity. For instance, when the node for a bus stop is first created, it will be version 1. If the location is adjusted later, the version will be incremented to 2. If the bus stop is then annotated with the available bus lines, the version would be incremented again.

Dataset

Our dataset focuses on OpenStreetMap nodes (points) and consists of the full, versioned history of OpenStreetMap, through February 2014. Because ways and relations are made up of nodes, nodes define the underlying geometry of contributions. For this reason, we limit our analysis to OSM nodes (we discuss implications for ways and relations later).

We limit our study site to the continental United States so that we can take advantage of readily-available government census data published by the U.S. Census – a common practice in geographic human-computer interaction studies (e.g. [17,18,20,21,26,39]). Because a key contribution of this work is developing an understanding of urban-rural and socioeconomic biases, it was necessary to ensure that there would be "urbanness" and socioeconomic census variables for our study site. We discuss how our work may extend to other geographic contexts in our Discussion section below.

From the broad OSM dataset, we first extracted all nodes in the continental United States, including every version of every node. We then excluded nodes created in an automated manner (e.g. large imported road datasets and bot-created geometries) using the technique in Johnson et al. [20]. Since we were interested in spatiotemporal trends, we excluded nodes created by people with fewer than five contributions out of sparsity concerns that we discuss in more detail below. Finally, we used a standard reverse geocoding approach to associate each node with the United States county that contains it. In total, we considered more than 28 million (28,021,802) contributions by 23,329 contributors.

Because contribution rates are so skewed in peer produced datasets (i.e. power-law dynamics [31,33]) and informed by Panciera et al. [31], we organize our analysis around three classes of contributors, defined by the number of edits they made:

- *1%ers*: The 1% of contributors that produce the most content. In total, "*1%ers*" contribute 68% of all OpenStreetMap nodes.
- *9%ers:* The "middle" 9% of contributors, i.e. those between the *1%ers* and the *90%ers*. "*9%ers*" produce 27% of OpenStreetMap content;
- *90%ers:* The bottom 90% of contributors. They produce only 11% of OpenStreetMap content.

Note that the percentages above refer to statistics once contributors with fewer than five edits have been removed (these contributors made only 0.07% of edits in total).

Geographic Variables of Interest

We operationalize four geographic variables using our historical dataset of human-generated nodes in the United States. These variables were selected because they had one of two properties: (1) they (or close variants) had been employed in non-temporal characterizations of geographic contributor focus, or (2) they are metrics related to observed geographic biases in peer produced geographic data. Our first two variables meet the first property and describe the geometric characteristics of contributors: (1) their geographic ranges [40] and (2) where they focus [25]. Our second two variables meet the second property and capture the (1) urbanness [5,18,20,21,34] and (2) socioeconomic status [5,13,35,39] of where people contribute. Below, we detail each of our four variables in turn.

Geometric Variables

std_dist: *Standard distance* is a common point-pattern analysis metric of geographic dispersion. *std_dist* is analogous to a standard deviation; it represents the geometric spread of a set of points relative to the geometric center of the set. Specifically, a *std_dist* describes the radius of a circle around the mean center point. Like a standard deviation, 68% of the points fall within this circle.

For our analysis, we computed the *std_dist* for each contributor simply by finding the mean center point of their contributions and then computing their dispersion. Prior to making this calculation, we projected all data points into a 2D reference system using the Albers' Equal Area Conic projection.

plurality_focus: While our *std_dist* variable describes the spatial distributions of people's contributions, our *plurality_focus* variable describes the actual locations where people focus. Each contributor's *plurality_focus* county is simply the county in which a plurality of their contributions were made (i.e., the mode). Prior work in geographic HCI [22] often uses this approach to attribute the "home region" of a contributor, but here we interpret "plurality county" more conservatively: we just take it as the region where a contributor has focused their contributions.

Human Geography Variables

Our next two variables focus on human geography and describe the *kinds of places* people contribute. In other words, while our first two variables describe the locations and geographic spread of contributions, the next two describe characteristics of the people who live in the contribution locations. Specifically, we define variables that describe the biases shown in prior literature: ruralness and poverty. Based on the county associated with each node, we label each contribution with: (1) a county urbanness class (from the National Center for Health Statistics' Urban-Rural Classification Scheme [28]), and (2) the percent of the county's population that is in poverty (from the US Census' American Community Survey [4]).

With these labels in place, we compute two variables for each contributor:

pct_rural: This variable describes the percent of a person's contributions that occurred in counties with urbanness classes 5 and 6 (the two nonmetropolitan classes in the classification scheme mentioned above). In Florida, for example, Miami-Dade County (where the city of Miami is located) is a 1 on this urbanness scale, whereas Monroe and Hamilton Counties (near the border with the state of Georgia, approximately halfway between the cities of Jacksonville and Tallahassee) are urbanness classes 5 and 6.

pct_high_poverty: This variable describes the percent of a person's contributions that occurred in 'high-poverty' counties, where at least 20% of the population is in poverty. We base this variable on the definition of 'high-poverty' provided by the United States Census American Community Survey [4]. For example, Webb County in Texas is a high-poverty county. Webb County is home to Laredo, Texas – one of the largest cities on the United States-Mexico border – and has an average per-capita income of approximately $10,000 (approximately $2,000 below the US poverty line in 2015).

Temporal Units of Analysis

Each of our four variables are a descriptive summary of the *geography* of contributors' focus, but they are not temporal. To understand how these geographic summaries *change*, we temporally group each person's contributions into quarters (Jan. 1 - Mar. 31st, April 1 - June 30, July 1 - Sept. 30, and Oct. 1 - Dec. 31). We selected three-month periods to ensure that (a) there would be sufficient data in each period, and (b) the temporal periods were granular enough to analyze the evolution of contributors' behaviors over time. For each *contributor-quarter*, we computed our four geographic variables. As we noted above, we excluded contributors with fewer than five contributions to avoid drawing conclusions from excessively small samples.

Figure 1 shows a histogram of the number of quarters that people participate in OpenStreetMap. Most people (71%)

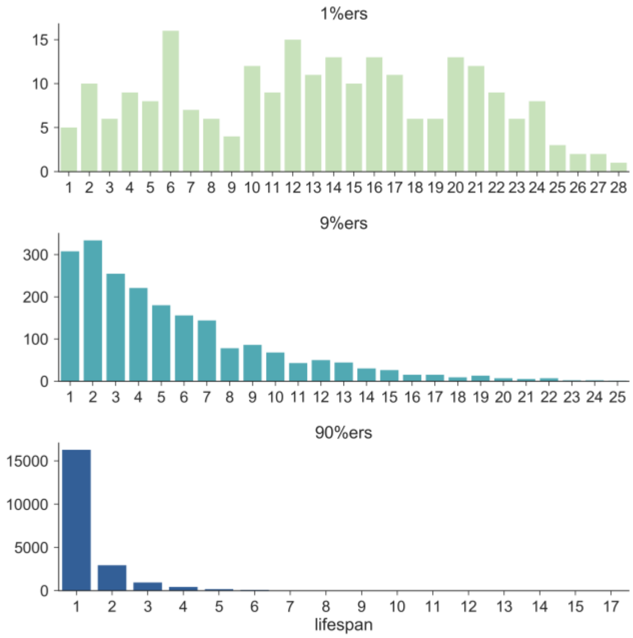

Figure 1: A histogram of contributors who participate for each number of quarters.

participate in only one quarter. These contributors are (a) predominantly *90%ers*, and (b) account for only about 4% of the total edits in our dataset. *1%ers* participate for a median of thirteen quarters, *9%ers* for a median of five quarters, and *90%ers* for a median of two quarters. We discuss the implications of these medians below.

RESULTS

We use our two main research questions to frame the presentation of our results. As we previewed, we generally find that most people are quite consistent throughout their contribution lifespans – contributors' geographic behavior tends to be 'born, not made'. Since this is exploratory work, we approach both research questions by identifying and characterizing the general trends in the data. We also highlight important deviations from those trends. We now discuss the results for each of our research questions in turn.

RQ1: How does contributors' geographic focus change over time?

The spatiotemporal trends in our *std_dist* and *plurality_focus* variables tell a relatively clear story: most contributors and contribution groups tend to have consistent geographic ranges and focus areas. In other words, most (though not all) contributors' geographic focus behavior is 'born, not made'. We now unpack these findings in more detail.

std_dist: Figure 2, which visualizes contributors' quarterly geographic ranges over time as defined by *std_dist*, shows a relatively clear trend: contributor groups have meaningfully distinct standard distances, and these distinctions are mostly consistent over time. Along the y-axis in Figure 2 – following the method used by Panciera et al. [31] – we plot the mean and 95% confidence interval in each quarter. We find that *1%ers*' and *9%ers*' average standard distances do not meaningfully vary over time. At first glance, Figure 2 may suggest that *1%ers* and *9%ers* increase their average *std_dist* over their lifespan. However, a closer inspection of the quarterly confidence intervals shows that these changes in means are not meaningfully different from one quarter to the next; the confidence intervals are highly overlapping. By contrast, we do see a meaningful uptick in *90%ers* standard distances as their lifespan increases. Note that this figure does not show quarters that exceed the 90th percentile of participation length, because the number of contributors becomes very small.

Although the 95% confidence interval ranges in Figure 2 look small and stable over time, we wanted to ensure that individual contributors do not substantially vary their *std_dist*s over time within their group ranges. The potential for this outcome is most salient for *1%ers* for two primary reasons: (1) *1%ers* contribute most of the content in OpenStreetMap so their geographic behavior has a substantial impact, and (2) in Figure 2, *1%ers* show the largest confidence interval ranges, conceivably allowing for more individual variation.

To address this question, we did a targeted analysis of *1%ers* to evaluate their consistency over time, the results of which are visible in Figure 3. The figure plots each individual *1%ers*' *std_dist* distribution, showing the median and interquartile range (IQR) of their *std_dist* in each quarter. The IQR is the distance between the 25th and 75th percentiles of a distribution, or the width of the middle 50% of *std_dist* values here. Individuals are ranked by IQR in increasing order along the x-axis. Critically, shorter lines (smaller IQRs) indicate a higher degree of 'born, not made' behavior with regard to standard distances

The large number of small green bars in Figure 3 confirms that most 1%ers exhibit 'born, not made' *std_dist* patterns, i.e. their geographic ranges are largely consistent in every quarter. Figure 3 also reveals that the higher variance we see in Figure 2 is primarily the result of a minority of 1%ers who do not display 'born, not made' *std_dist* patterns. This non-

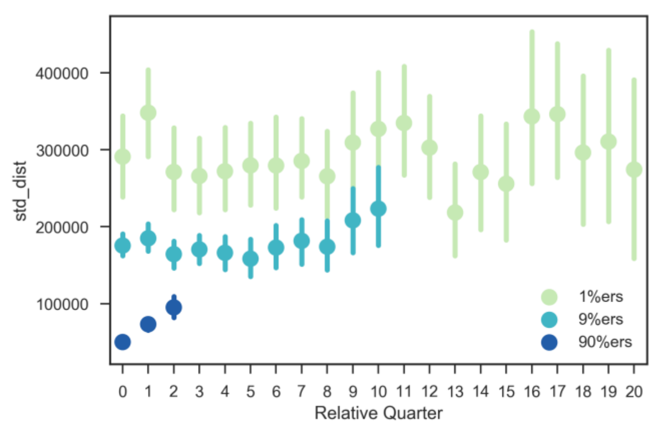

Figure 2: Mean *std_dist* over time, by user class. Error bars show 95% confidence intervals.

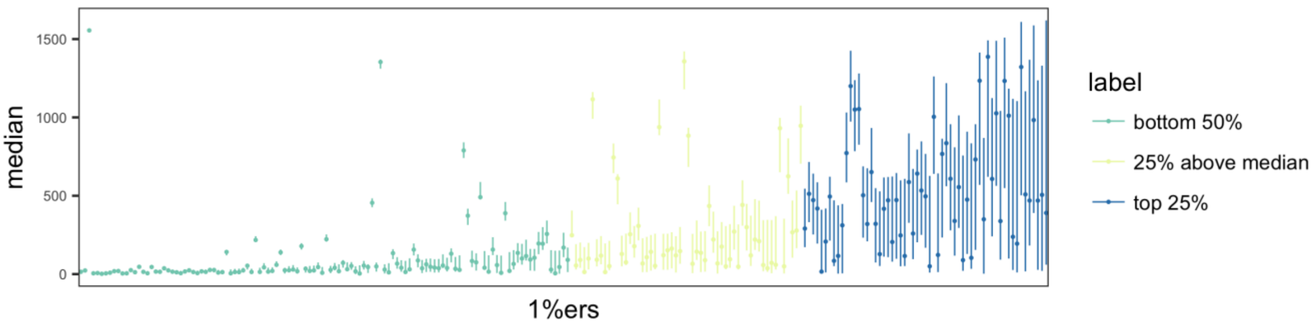

Figure 3: Distributions of each individual 1%ers' *std_dist*. Dots indicate medians, and lines indicate IQR (interquartile range).

trivial minority exhibits different geographic range patterns across quarters.

It is important to note that the IQR values in Figure 3 do not appear to be strongly driven by the number of quarters in which a contributor participates. For instance, a 1%er's *std_dist* IQR and the number of quarters they participate are only weakly correlated (Pearson's r=0.2).

plurality_focus: While *std_dist* characterizes the geographic dispersion of contributor edits, it does not capture *where* contributors focus. For this, we use *plurality_focus*.

Figure 4 plots the median number of unique *plurality_focus* counties over time. Each solid line represents a user class, truncated at the 90th percentile of participation length. The dashed line shows what would occur if the median contributor had a new *plurality_focus* county every quarter.

Figure 4 makes one trend clear: while the median contributor does increase the number of counties in which they focus over time, this increase is gradual and substantially less than

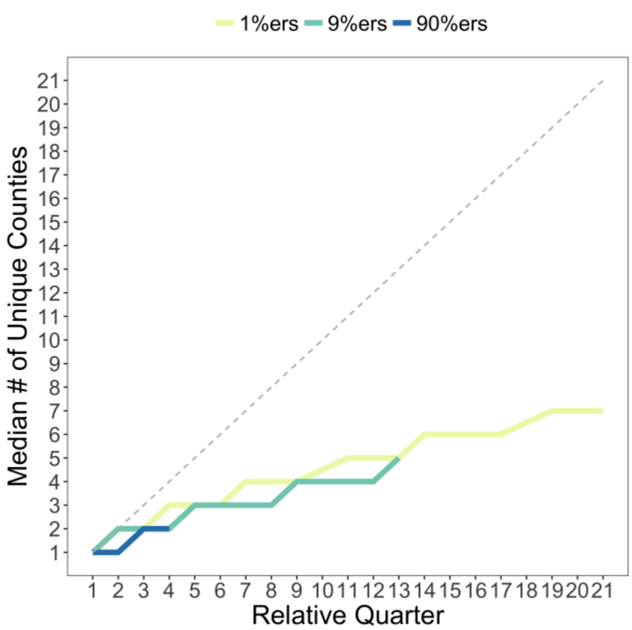

Figure 4: Plots the growth of unique *plurality_focus* counties over time. Each color is a different user class, and the dashed line represents a new *plurality_focus* county every quarter.

would be the case if the median contributor focused in new areas each quarter. Intuitively, the median contributor tends to be fairly consistent in where they focus, returning to the same few counties over time. For instance, the median 90%er participates for two quarters, but has a single *plurality_focus* county on average. The median *9%er* participates for five quarters, and this contributor has only three unique *plurality_focus* counties on average. Strikingly, the median 1%er participates for 13 quarters (more than 3 years), and on average has five unique *plurality_focus* counties.

RQ2: Can the spatiotemporal evolution of contributors' focus facilitate systemic coverage biases?

We now turn to our second research question, which. uses the *pct_rural* and *pct_high_poverty* variables to investigate patterns in geographic behavior concerning *kinds of places* (e.g. poor vs. rich) rather than specific places (i.e. individual counties). We highlight the general trends in these variables as well as impotant deviations from the trends.

Overall Trends

Figure 5 (*pct_rural*) shows the mean rate of contributions in counties classified as 5 or 6 on the National Center for Health Statistics urbanness scale. Figure 6 (*pct_high_poverty*) shows the mean rate of contributions in counties designated as 'high-poverty', according to the US Census. As before, these plots show the 90th percentile number of participation quarters. In both cases, the means of these distributions remain consistent across time for all three user classes, suggesting that most people consistently contribute a relatively small proportion of their edits in rural and poor counties. Even *1%ers*, who have the largest standard distance (and thus contribute across larger distances) make less than one fifth of their contributions in rural areas on average, and even fewer in high-poverty areas (and do so consistently across their lifespans).

As before, while the community-level trends are consistent over time, we also wanted to check whether these trends hold at the idividual level. We again focused on *1%ers*, who have the widest confidence intervals in Figures 5 and 6 and who contribute the most edits. Figures 7 and 8 confirm that the majority of *1%ers* tend to be quite individually consistent, having persistently low individual median *pct_rural* and *pct_high_poverty* values. The median *pct_rural* IQR is 0.11

and the median *pct_high_poverty* IQR is 0.02, both of which are quite small (on a scale from 0 to 1)[1]. Moreover, the small variation is centered on mostly urban and mostly-non-poor regions, as can be seen by the tendency of the green lines in Figures 7 and 8 to be at the bottom of the y-axis.

The results in Figures 7 and 8 indicate that there is a strong 'born, not made' signal in our *pct_high_poverty* and *pct_rural* variables. In other words, *geographic biases may be "born, not made'.* If contributors start by contributing the large mjaority of their content in urban areas, this trend typically will persist for their entire time in OpenStreetMap. Our *pct_high_poverty* variable shows the same result – most contributors (a) do not contribute much content in high-poverty areas, and (b) maintain this trend over time.

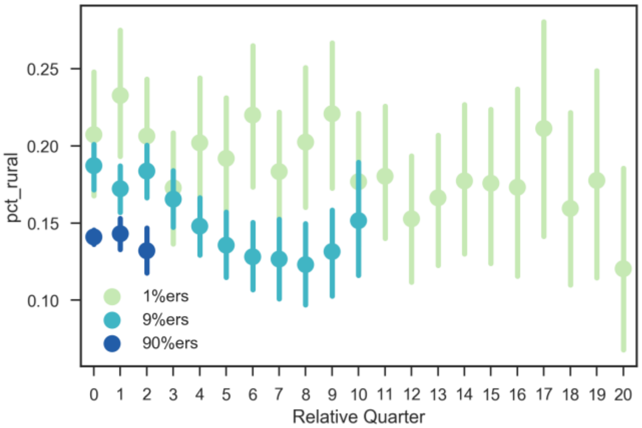

Figure 5: Mean *pct_rural* over time, by user class. Error bars show 95% confidence intervals.

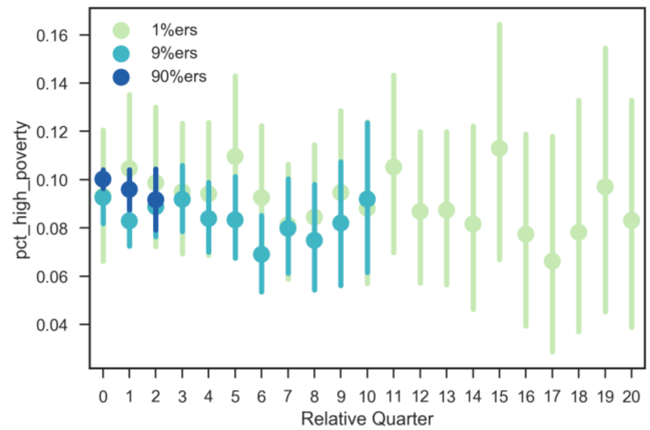

Figure 6: Mean *pct_high_poverty* over time, by user class. Error bars show 95% confidence intervals.

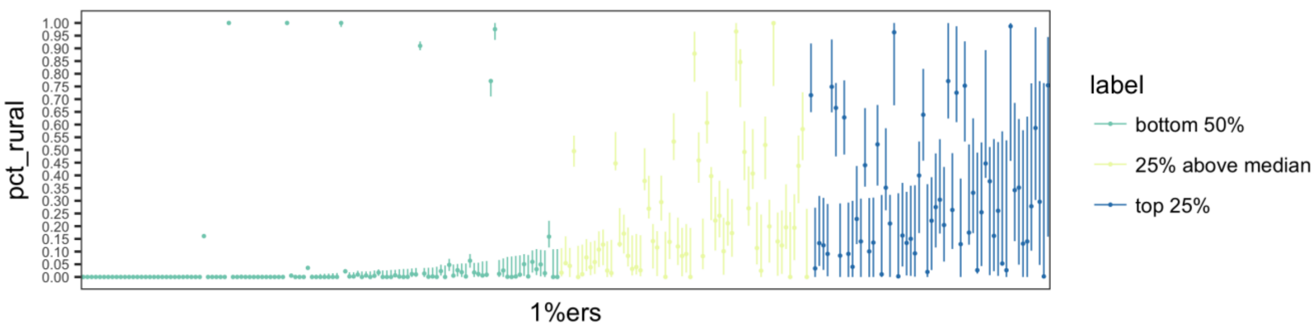

Figure 7: Distributions of each individual 1%ers' *pct_rural* values. The dot indicates the median, and the line indicates their interquartile range.

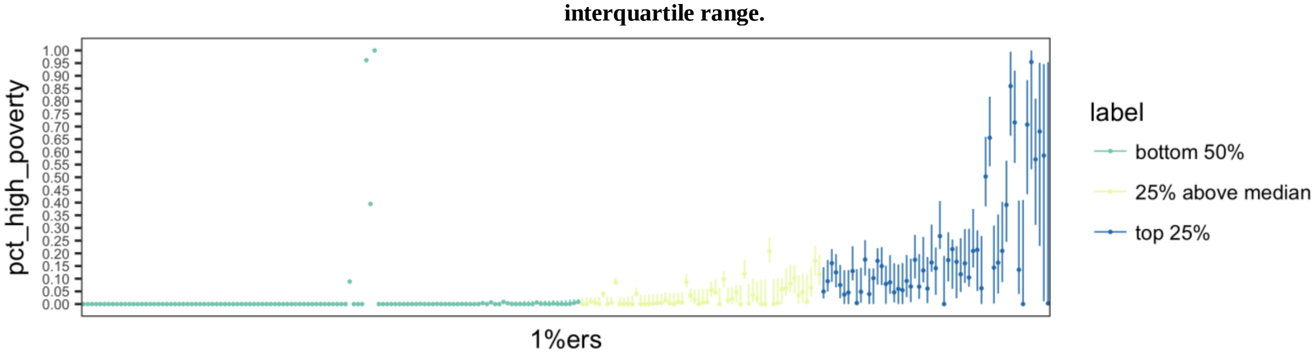

Figure 8: Distributions of each individual 1%ers' *pct_high_poverty* values. The dot indicates the median, and the line indicates their interquartile range.

---

[1] As was the case above with std_dist, we see very weak correlation between the number of quarters a 1%er spends in OpenStreetMap and their IQR (Pearson's r = 0.09 and 0.06 for *pct_rural* and *pct_high_poverty*, respectively).

Contextualizing pct_rural and pct_high_poverty values

To put our *pct_rural* and *pct_high_poverty* results into context, we now consider three dimensions against which to compare these results. Specifically we ask if the *pct_rural* and *pct_high_poverty* findings in Figures 5-8 are proportional to what would be expected given (1) the population of these counties, (2) the number of rural or high-poverty counties themselves, or (3) the number of contributors focusing in rural or high-poverty areas.

With regard to county population, according to the United States Census [28], nearly 15% of the US population lives in rural areas, and approximately 14% live in high-poverty areas. Comparing these numbers against Figures 5 and 6 suggests that the average rate of rural contribution is actually proportional to the population rate in these counties. However, this is not true for our *pct_high_poverty variable*. The average rate of contribution in high-poverty areas is approximately 10%, indicating that high-poverty counties are underrepresented across the board.

Another option to consider is whether these *pct_rural* or *pct_high_poverty* rates are proportional to the number of counties that are rural or high-poverty counties, i.e. maybe there are just fewer of these counties. 63% of counties are rural (have urbanness classes 5 or 6), and 24% of counties are high-poverty (at least 20% of their population is in poverty). Comparing these numbers to the median *pct_rural* or *pct_high_poverty* rates shown in Figures 5 and 6, the conclusion is clear: in terms of the number of counties, OSM contributors in all user classes are undercovering rural and high-poverty counties. While there may be fewer people in many of these counties, these counties still have road networks, natural features like lakes and rivers, and many other entities that are not directly correlated with population [19].and that typically are mapped in OpenStreetMap.

A third consideration is whether the number of contributors focusing in rural or high-poverty counties is proportoinal to the population of these regions. One important reason to consider this dimension is the effect it may have on content quality. Prior work has shown that people who focus near where they live produce more diverse [40], richer [20], and higher quality [7] content. Unfortunately, Figures 7 and 8 suggest concerning trends here too. As noted above, 15% of the US population live in rural areas, and 14% live in high-poverty areas. However, Figures 7 and suggest substantially fewer *1%ers* focus in rural or high-poverty areas – very few have medians near the top of the y-axis.

Thus far, our results suggest that most contributors – across all user classes – are consistent across time, and contribute in consistently urban and wealthier areas. Further still, rural and high-poverty areas are disproportionately undercovered in comparison to (a) the number of rural and high-poverty counties, and (b) the number of contributors who focus in these areas. Taken together, our results suggest that (a) where contributors focus, (b) the kinds of places they focus in, and (c) the consistency with which this occurs all contribute to the geographic coverage biases shown in prior literature.

Additional Mechanisms of Bias

We noted above that *1%ers* participate for the longest period of time, which creates a secondary mechanism facilitating bias – longevity bias. Specifically, people who participate longer *contribute* longer and because of 'born, not made' trends, *contribute in the same places (and kinds of places) longer*.

While this trend is intuitive when comparing *1%ers* and *90%ers* (after all, *1%ers* produce most of the content), we wanted to understand how a longevity bias might facilitate socioeconomic and urbanness focus biases. Therefore, we split contributors into two groups, those who tend to be rural-focused (have a median *pct_rural* of at least 50%), and those who tend to be urban-focused (have a median *pct_rural* below 50%). We computed how long each contributor participated, and compared the urban-focused and rural-focused groups. Examining the means of these groups (urban-focused: 1.9 quarters, rural-focused: 1.65 quarters) suggests that urban-focused contributors participate longer, on average. Due to a skewed distribution, we conducted a Wilcoxon Rank-Sum Test which found significant differences between the two groups ($z=2.67$, $p < 0.01$). We ran the same analysis for our *pct_high_poverty* contributors. Again, the means (non-high-poverty focused: 1.9 quarters, high-poverty focused: 1.55 quarters) suggest that high-poverty focused contributors participate longer, on average. A Wilcoxon Rank-Sum Test also found significant differences between the two groups ($z=4.81$, $p < 0.001$).

While these findings are not causal – and future work should examine predictors of retention in OSM – they do potentially have implications for the evolution of bias in OSM. Specifically, these results suggest that the bias in where people focus is perpetuated by who remains a contributor. Most people, across all user classes, consistently contribute small amounts of content in rural and high-poverty areas over the course of their time in OSM. People who do focus in rural and high-poverty areas stop contributing earlier than people who focus in more urban, or wealthier areas. This finding potentially has important implications for improving coverage in rural and high-poverty areas, something to which we return in the Discussion section.

Deviations from Trend

Faced with results that suggest that most people consistently contribute in urban and non-high-poverty areas, we sought to better understand contributors who *do* primarily focus in rural and/or high-poverty areas and the contributions that they make. What we found aligns strongly with what is shown in Figures 7 and 8. The majority of rural and high-poverty content is not contributed by consistently rural or consistently high-poverty contributors. Figures 7 and 8 indicate that relatively few *1%ers* have high median *pct_rural* and *pct_high_poverty* values, and that many of those who do also tend to have wider IQRs, indicating that they are less consistent over time in terms of the types of places they edit than the median *1%er*.

To understand these rural and high-poverty focused contributors in more detail, we use the same metric as above: if a contributors' median *pct_rural* and median *pct_high_poverty* are at least 50%, we consider them rural-focused and high-poverty-focused, respectively.

Beginning with rural-focused contributors, we found that 3,126 people tend to contribute in rural areas, and as a group contribute less than 40% of content in rural areas. There are 27 rural-focused *1%ers* (those nearer the top of the Y axis in Figure 7), 315 rural-focused *9%ers*, and the rest (2,748) are *90%ers*. They account for 25%, 11%, and 2% of rural content, respectively (totaling 38% of rural content). Because *1%ers* contribute most of the content in OpenStreetMap, we have mapped the plurality focus counties for the seven most prolific rural-focused *1%ers* in Figure 9. We selected only the seven most prolific to aid in map legibility [41].

There are two primary trends in Figure 9: (1) people who contribute in national parks (and national forests), and (2) people who contribute regionally. With respect to the national parks, (a) prior studies have shown that vacation destinations are common locations for VGI contribution [32], and (b) very few people live in counties with national parks. What this suggests is that some of the participants who we termed rural-focused may instead be 'national park-focused', with national parks serving huge numbers of urban visitors. The second pattern in Figure 9 involves regional contributors. To take one example, consider the person contributing in northern Maine (in the upper northeast corner of Figure 9). This area is very sparsely populated, and yet a single, consistently rural 1%er contributes most their content, over multiple quarters, in those counties. Both groups have implications for recruitment in peer production communities, which we discuss further below.

Turning to high-poverty contributions (Figure 10), the trend we observed for rural areas is even more severe. We found that 2,014 people consistently contribute in high-poverty areas, and as a group contribute slightly more than one-fourth of the content in high-poverty areas. There are 11 high-poverty-focused *1%ers* (those nearer the top of the Y axis in Figure 8), 126 high-poverty-focused *9%ers*, and the rest (1,877) are *90%ers*. They contribute 16%, 8%, and 2% of high-poverty content, respectively (totaling 26% of high-poverty content). We have mapped the plurality focus counties for the seven most prolific high-poverty focused *1%ers* in Figure 10. We again selected only the seven most prolific to aid in map legibility.

The contributors in Figure 10 show similar trends to those in Figure 9: many of the counties shown contain national parks and forests and a few are contributors who contribute regionally. One example of the first trend is the large teal section in the southwestern section of the map (the area surrounding the Grand Canyon). The counties that contain the Grand Canyon also contain the Navajo Indian Reservation, one of the five most impoverished reservations in the United States [42]. This lends further credence to the idea that some contributors focus in natural parks, and it is likely that these contributors *are not* contributing in the very impoverished parts of this region. However, there are some contributors who are consistently focused in high-poverty areas. For example, consider Sierra County, New Mexico (reddish), also in the southwestern corner of the map. The person primarily contributing here *is* focused on high-poverty counties. Residents of Sierra County tend to be quite poor, with a median household income of $25,583, and a per-capita income of $16,667. Another example of a high-poverty area is the more northern county in Texas (pink, central southern section of the map) – Webb County, Texas. Webb County is home to Laredo, the third largest city on the Mexico-United States border. The median household income in Webb County is $28,100, and the per-capita income is $10,179. As before, both examples suggest implications for recruitment that we discuss below.

DISCUSSION

In this section, we step up a level and discuss the implications of our findings more broadly. This section follows the same structure as the results section. Specifically, we first discuss what our findings mean for our understanding of contributor behavior in peer production systems. Second, we discuss

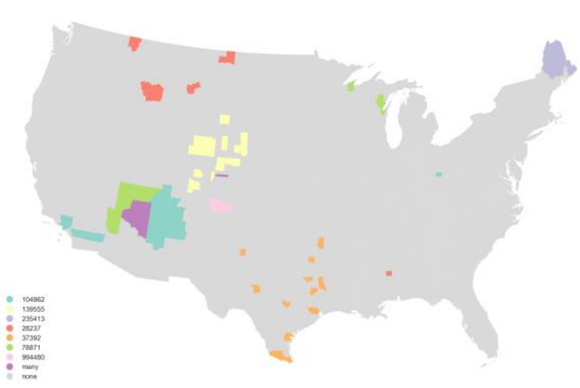

Figure 9: All counties for rural focused 1%ers.

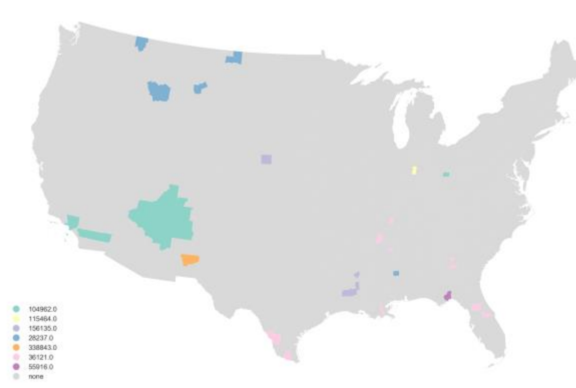

Figure 10: All counties for high poverty focused 1%ers.

what our findings suggest for the mitigation of urban and socioeconomic coverage biases in peer production systems.

Implications for Peer Production

Standard Distance and Spatial Interaction Behavior

Closely related to our *std_dist* variable is a concept from geography called *spatial interaction* [23,24,38], which is used to describe 'flow' between regions, e.g., of physical goods [23,24] or people [38]. This process often is modeled with gravity models and characterizes, e.g., the rate at which travel between regions changes as a function of distance and attributes of the regions. The 'cost of distance' aspect of these models is particularly relevant to our findings here.

We find that different classes of contributors (e.g. *1%ers* vs. *90%ers*) have consistently distinct sizes of geographic range, which presents an important opportunity for future work. Intuitively, these findings suggest that different contributor classes interact consistently differently across distance. Prior "GeoHCI" [17] work using gravity models has not accounted for contributor class, but doing so may provide for better understanding of the mechanisms behind spatial content production. This may also help support predictions about which areas would receive contributions if, for instance, a concerted recruiting effort were made in rural areas (as is discussed in more detail below).

Mitigating Coverage Biases

Our results suggest that 'born, not made' dynamics may naturally facilitate the creation of geographic coverage biases, which are in part enabled by who remains a contributor over time. We next reflect on how our results suggests mechanisms for reducing these biases.

Existing Consistently Rural or High-poverty Contributors

The first intuitive approach to mitigating biases is to examine those participants who *do* consistently focus in rural and high-poverty areas. After all, these participants are contributing a non-trivial amount of content in rural and high-poverty areas already. As noted above, our results suggest that there are two trends in where these rural-focused or high-poverty-focused *1%ers* contribute: national parks and regional areas.

National Parks: Leveraging existing contributors who focus in the counties that contain national parks and forests to address poverty or urban/rural bias is likely to be difficult. Prior work suggests that vacation destinations are common locations for geographic contributions [32], and that people tend to be more aware of the geography in places with which they are familiar [10]. Thus, it is likely that the contributors who focus in counties that contain national parks are not producing content in the rural or high-poverty sections of those counties (although investigating this hypothesis in detail is a good targeted direction of future work).

Regional Focus: The other group of rural- or high-poverty-focused *1%ers*, however, may be more promising. These contributors *already* are focusing their effort in rural or high-poverty areas. We see two implications for design here. The first is simple: find ways to keep these contributors in the community! Our results suggest that the longevity of these contributors in OpenStreetMap is less than their peers, and targeting this issue would be one immediate and effective partial solution to coverage biases. Second, our results suggest that targeted recruitment of regionally-focused *1%ers* in low-coverage areas could be effective.

LIMITATIONS AND FUTURE WORK

In this work, we took an exploratory approach to understanding contributor behavior. Our results outline important specific hypotheses that should be investigated using more formal quantitative approaches (e.g. targeted hypothesis testing). Future work might also consider deeper qualitative approaches that would help shed light on *why* contributors choose to focus in the areas they do.

We limited this study to the contiguous United States, but examining different study sites would be valuable. For instance, it may be the coverage biases in different parts of the world (e.g. China [20]) are connected to different spatiotemporal focus patterns.

While the atomic unit of our study – the OpenStreetMap node – captures most editing behavior in OpenStreetMap, it does not capture all editing behavior, e.g. edits that added a tag to other OSM geometries (ways or relations). We also focused on human contributor behavior here and excluded automated edits. Future work should directly examine automation behavior in OpenStreetMap, as this may provide useful insight into mitigating coverage biases.

Both our *plurality_focus* and our *std_dist* metrics build a measure of central tendency (mode and mean center, respectively). We rely on *plurality_focus* because it is a common approach for this purpose, and makes no assumptions about the 'normality' of a distribution (analogous to using the mode vs. the mean). Future work should consider similarities and differences between these metrics, and alternative approaches to characterizing where people primarily focus their contributions (e.g. [22]).

CONCLUSION

In this paper, we performed the first examination of the spatiotemporal behavior of contributors to geographic peer production communities. We observed that contributors' spatiotemporal behavior is generally consistent throughout their contribution lifespans, both with respect to the geometric structure of contributions and with respect to the types of places to which contributions are made (e.g. urban places vs. rural places). In other words, we saw evidence that there is a strong (but not omnipresent) 'born, not made' tendency in spatiotemporal peer production behavior. More generally, this work sheds light on some of the mechanisms by which the coverage (and coverage biases) of peer produced geographic datasets may occur.

ACKNOWLEDGEMENTS

This research was supported by NSF grants IIS-1218826, IIS-1707296 and IIS-1707319.

REFERENCES

1. Alexa. 2010. wikipedia.org Traffic Statistics. Retrieved January 13, 2017 from http://www.alexa.com/siteinfo/wikipedia.org
2. Ofer Arazy, Hila Liifshitz-Assaf, Oded Nov, Johannes Daxenberger, Martina Balestra, and Coye Cheshire. 2017. On the "How" and "Why" of Emergent Role Behaviors in Wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 2039–2051. https://doi.org/10.1145/2998181.2998317
3. Yochai Benkler and others. 2016. Peer production and cooperation. *Handbook on the Economics of the Internet* 91.
4. US Census Bureau. American Community Survey (ACS). Retrieved July 4, 2017 from https://www.census.gov/programs-surveys/acs
5. Ashley Colley, Jacob Thebault-Spieker, Allen Yilun Lin, Donald Degraen, Benjamin Fischman, Jonna Häkkilä, Kate Kuehl, Valentina Nisi, Nuno Jardim Nunes, Nina Wenig, and others. 2017. The Geography of Pokémon GO: Beneficial and Problematic Effects on Places and Movement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
6. 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
7. Melanie Eckle and João Porto de Albuquerque. 2015. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. In *ISCRAM*.
8. Killian Fox. 2012. OpenStreetMap: "It's the Wikipedia of maps." *The Guardian*. Retrieved October 25, 2016 from https://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals
9. 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
10. Peter Gould and Rodney White. 1986. *Mental Maps*. Routledge, London; New York.
11. Mark Graham, Matthew Zook, and Andrew Boulton. 2013. Augmented reality in urban places: contested content and the duplicity of code. *Transactions of the Institute of British Geographers* 38, 3: 464–479. https://doi.org/10.1111/j.1475-5661.2012.00539.x
12. Simon Gröchenig, Richard Brunauer, and Karl Rehrl. 2014. Estimating Completeness of VGI Datasets by Analyzing Community Activity Over Time Periods. 3–18. https://doi.org/10.1007/978-3-319-03611-3_1
13. 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
14. A Halfaker, Aaron Halfaker, J T Morgan, and J Riedl. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57, 5: 664–688. https://doi.org/10.1177/0002764212469365
15. 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
16. Brent Hecht and Darren Gergle. 2010. On the "localness" of user-generated content. 229. https://doi.org/10.1145/1718918.1718962
17. 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
18. 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 October 20, 2016 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114
19. Isaac Johnson and Brent Hecht. 2016. Structural Causes of Bias in Crowd-derived Geographic Information: Towards a Holistic Understanding. In *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.
20. 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
21. 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)*. https://doi.org/http://dx.doi.org/10.1145/3025453.3026015
22. 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
23. 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

24. 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
25. Michael D Lieberman and Jimmy Lin. 2009. You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories. In *ICWSM*.
26. 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
27. Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies* 67, 9: 716–754. https://doi.org/10.1016/j.ijhcs.2009.05.004
28. NCHS. 2010. NCHS Urban-Rural Classification Scheme For Counties. Retrieved January 13, 2017 from https://www.cdc.gov/nchs/data_access/urban_rural.htm
29. OpenStreetMap. Good practice - OpenStreetMap Wiki. Retrieved October 24, 2016 from http://wiki.openstreetmap.org/wiki/Good_practice
30. Katherine Panciera. 2011. User lifecycles in cyclopath. In *the 2011 iConference*, 741–742. https://doi.org/10.1145/1940761.1940892
31. Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made. In *Proceedings of the ACM 2009 international conference*, 51. https://doi.org/10.1145/1531674.1531682
32. Adrian Popescu and Gregory Grefenstette. 2009. Deducing Trip Related Information from Flickr. In *Proceedings of the 18th International Conference on World Wide Web* (WWW '09), 1183–1184. https://doi.org/10.1145/1526709.1526919
33. Reid Priedhorsky, Jilin Chen, Shyong Tony K Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, destroying, and restoring value in wikipedia. In *the 2007 international ACM conference*, 259. https://doi.org/10.1145/1316624.1316663
34. Giovanni Quattrone, Licia Capra, and Pasquale De Meo. 2015. There's No Such Thing As the Perfect Map: Quantifying Bias in Spatial Crowd-sourcing Datasets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 1021–1032. https://doi.org/10.1145/2675133.2675235
35. 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
36. Karl Rehrl, Simon Gröechenig, Hartwig Hochmair, Sven Leitinger, Renate Steinmann, and Andreas Wagner. 2013. A Conceptual Model for Analyzing Contribution Patterns in the Context of VGI. In *Progress in Location-Based Services*, Jukka M. Krisp (ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 373–388. Retrieved July 1, 2017 from http://link.springer.com/10.1007/978-3-642-34203-5_21
37. 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
38. 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
39. Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Towards a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. *ACM TOCHI*.
40. 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
41. 2012. Cartographic guidelines for public health. Retrieved October 6, 2017 from https://stacks.cdc.gov/view/cdc/13359
42. 2017. Reservation poverty. *Wikipedia*. Retrieved July 4, 2017 from https://en.wikipedia.org/w/index.php?title=Reservation_poverty&oldid=776694412
43. Apple Maps | OpenStreetMap Blog. Retrieved July 4, 2017 from https://blog.openstreetmap.org/2012/10/02/apple-maps/
44. OpenStreetMap Statistics. Retrieved October 6, 2017 from http://www.openstreetmap.org/stats/data_stats.html