Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Information Seeding and Knowledge Production in Online Communities: Evidence from OpenStreetMap

Abhishek Nagaraj

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–27
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Information Seeding and Knowledge Production in Online Communities: Evidence from OpenStreetMap

Abhishek Nagaraj[a]

[a] Haas School of Business, University of California, Berkeley, Berkeley, California 94720-1900

Contact: nagaraj@berkeley.edu, https://orcid.org/0000-0001-9049-0522 (AN)

Received: September 27, 2017
Revised: August 19, 2019; March 9, 2020
Accepted: May 8, 2020
Published Online in Articles in Advance: January 14, 2021

https://doi.org/10.1287/mnsc.2020.3764

Abstract. The wild success of a few online communities (such as Wikipedia) has obscured the fact that most attempts at forming such communities fail. This study evaluates *information seeding*, an early-stage intervention to bootstrap online communities that enables contributors to build on externally sourced information rather than have them start from scratch. I analyze the effects of information seeding on follow-on contributions using data on more than 350 million contributions made by more than 577,000 contributors to OpenStreetMap, a crowd-sourced map-making community seeded with data from the U.S. Census. I estimate the effect of seeding using a natural experiment in which an oversight caused about 60% of U.S. counties to be seeded with a complete census map, while the rest were seeded with less complete versions. Although access to basic knowledge generally encourages downstream knowledge production, I find that a higher level of information seeding significantly lowered follow-on contributions and contributor activity on OpenStreetMap, and was associated with lower levels of long-term quality. However, seeding did benefit densely populated urban areas and did not discourage more committed users. To explain these patterns, I argue that information seeding can crowd out contributors' ability to develop ownership over baseline knowledge and thereby disincentivize follow-on contributions.

History: Accepted by Chris Forman, information systems.
Funding: Funding from the Massachusetts Institute of Technology Initiative on the Digital Economy and the Berkley-Haas Fisher Center for Business Analytics is acknowledged.
Supplemental Material: The online appendices are available at https://doi.org/10.1287/mnsc.2020.3764. Code and data to replicate the analysis in the paper can be found at https://osf.io/a47e9/.

Keywords: online communities • knowledge production • crowdsourcing • innovation • digitization

1. Introduction

More and more, knowledge production is being performed outside of traditional organizations and in online communities, such as those behind Wikipedia and open source software (including projects such as Linux and Apache). Public goods produced by such communities are important drivers of economic growth and productivity at the national level (Brynjolfsson and Oh 2012, Greenstein and Nagle 2014, Harhoff and Lakhani 2016). Despite their importance and extensive research in this area, the fact remains that the success of online communities is far from certain. Most projects fail, rarely attracting more than one contributor (Healy and Schussman 2003, Hill 2013). Existing research is limited in its ability to offer guidance on increasing the success of communities as most research has focused on understanding or improving motivation within communities that are already successful (Lerner and Tirole 2002, Lakhani and Wolf 2003, Shah 2006).[1] When it comes to the question of how to build new online communities, "support for most of the design claims . . . come not from empirical evidence but from anecdotes and theoretical arguments," a gap that this paper begins to address (Resnick et al. 2011, p. 232).

The central challenge in building new online communities is the chicken-and-egg problem: without any existing information, community members are hard to attract, and without any community members, new information is hard to accumulate (Athey and Ellison 2014). *Information seeding* is a prominent early-stage intervention designed to crack this problem by enabling potential contributors to improve and build on externally sourced information, rather than starting from scratch. Forms of information seeding are emphasized in early case studies of online communities and open source software. In his famous essay, Raymond (1999) argues that "when you start community-building, what you need to be able to present is a plausible promise" (p. 37), often through seeding a piece of useful code. Lerner and Tirole (2002) state that a project must "assemble a critical mass of code . . . to show that the project is doable and has merit" (p. 220). Theoretically, this practice is based on the "cumulative growth effect" (Aaltonen and Seiler 2015) or the notion that "content begets content."[2] According to this principle, contributors are more likely to be attracted to a project and add follow-on knowledge if the project already has pre-existing information to build upon (Boudreau and Lakhani 2015, Kane and Ransbotham 2016). Wikipedia, for instance, was initially seeded with short articles on more than 30,000 U.S. cities from the U.S. Census Bureau[3] and Reddit's founders used a "fake it till you make it" strategy in which they seeded the website with content from fake user accounts to attract additional follow-on contributions.[4]

Despite theoretical arguments and anecdotal evidence for the value of information seeding, we understand little about how and to what extent information seeding fosters the growth of online communities in practice. In particular, as I will argue, when communities are geared toward lower-level information-provision tasks (such as mapping cities or tagging images), which offer little scope for career progression or skills development (Franzoni and Sauermann 2014, Lyons and Zhang 2018), contributors might be driven by a sense of ownership over the knowledge that they create, motivating follow-on contributions. Allowing contributors to create new knowledge from scratch might foster a greater sense of ownership, whereas higher levels of information seeding might crowd out these incentives. As Lerner and Tirole (2002) argue, in some cases, "it may be important that the leader does not perform too much of the job on his own" (p. 220) given the strong nonpecuniary motivations driving knowledge production in these settings (Franke and Shah 2003, Lakhani and Wolf 2003, Shah 2006, Belenzon and Schankerman 2008). Therefore, there might be important limits to the benefits of information seeding. I test this possibility that a high level of information seeding is ultimately harmful for follow-on knowledge.

My empirical design exploits a rare natural experiment in which an unintentional error caused variation in the level of information seeding in an online crowdsourcing platform. Specifically, I analyze OpenStreetMap, a Wikipedia-style open source GIS community (Maurer and Scotchmer 2006) that leverages user contributions to build a digital map similar to Google Maps. In the fourth quarter of 2007, about two years after OpenStreetMap was launched in the United States, the fledgling community decided to bootstrap their efforts by seeding its map with the U.S. Census Topologically Integrated Geographic Encoding and Referencing (TIGER) map, which provides bare bones information about streets and their names. Rather than start from scratch, the idea was for the community to build on this information by adding follow-on contributions, that is, additional information such as road tags (speed limits, one ways, etc.) as well as information on local businesses and points of interest.

Unbeknownst to OpenStreetMap contributors, the U.S. Census was itself in the process of updating and correcting a mostly outdated and incomplete TIGER map in preparation for the 2010 census. Consequently, the 2006 version of the TIGER map that was used by OpenStreetMap contained accurate and complete information for only about 60% of the approximately 3,100 counties in the United States. Information for the remaining 40% provided largely out of date and incomplete information.[5] In this way, about 60% of U.S. counties received a higher level of information seeding than the rest. Although the counties that were updated earlier in the program were not explicitly chosen at random, the Bureau was not selectively choosing the most interesting or important counties for early update either. As I will explore in significant detail through qualitative and quantitative analysis (see Section 2.3), high and low-information seeding counties were largely comparable along a number of dimensions, providing a unique opportunity to analyze the long-run effects of information seeding.

I leverage microdata on more than 350 million contributions to OpenStreetMap in the United States between 2005 and 2014, matched to either a *treatment* county (that received the higher-quality TIGER map) or a *control* county. Armed with these data, I rely on two types of specifications. First, I compare treatment and control counties over time in a difference-in-difference framework. This strategy is quite robust because it allows me to include nonparametric county and time fixed effects. However, since information seeding interventions, by definition, happen early in a community's lifespan, there are not many contributions in either treatment or control counties before the seeding took place, making it challenging to evaluate the parallel trends assumption. Therefore, I also estimate cross-sectional specifications that compare treatment and control counties along with a host of relatively flexible controls and fixed effects. Further, to address lingering concerns that treatment and control counties are not comparable, I test both specifications on two refined subsets of treatment and control counties. The boundary sample includes only those treatment counties that share a border with at least one control county (dropping treatment and control counties that are clumped together). And the second timing sample exploits novel data on the scheduled timing of county updates from the Census Bureau to drop counties that were scheduled very early or very late in the update process, including only those that were scheduled relatively close but differed in terms of their treatment or control status.

Perhaps surprisingly, both the difference-in-difference and cross-sectional results suggest that a high degree of information seeding hurts, rather than helps, the long-term development of OpenStreetMap. Despite having a higher quantity of baseline information, treatment counties see about 4–5.5% fewer contributors and receive about 10–15% fewer follow-on knowledge contributions compared with control counties, depending on the specification. These differences are striking because they represent an apples-to-apples comparison in follow-on map layers, such as street tags or points of interest, that must be added from scratch in both treatment and control counties. Importantly, differences in follow-on knowledge have significant long-run effects on quality: despite their early lead, treatment counties have an error rate that is about 10% higher than that of control counties over a 10-year period. Further, I find that information seeding is not uniformly harmful for community outcomes. In relatively dense urban counties, seeding helps rather than hurts follow-on contributions and contributor activity, and the negative effects of seeding do not apply to users who are relatively more committed to the platform to begin with.

As a potential mechanism driving these results, I investigate the ownership channel, whereby contributors are more likely to make follow-on contributions to the knowledge they contributed rather than information seeded from an external source. I provide a simple sketch of the theoretical idea and some qualitative validation from interviews with OpenStreetMap contributors. Although I do not have a direct measure of this psychological concept, I develop and test three indirect predictions that follow from this theory, including the idea that contributors who demonstrate a high sense of ownership prior to seeding are not discouraged by the seeding effort as compared with those who do not. Further, a fourth prediction using the direct count of the total number of times an owner makes a follow-on contribution to their initial contribution combined with an object-level analysis that traces out the sequence of every contribution to specific elements on the map (e.g., a building) provide a direct illustration of this mechanism in action. Overall, the ownership mechanism seems to be an important channel for the negative effects of information seeding on follow-on contributions in this setting.

This work is closely related to recent papers that examine the relationship between existing knowledge and the propensity of contributors to contribute new information (Aaltonen and Seiler 2015, Kane and Ransbotham 2016, Zhu et al. 2019, Hinnosaar et al. 2019). Aaltonen and Seiler (2015) argue that "content begets content": that the provision of information encourages follow-on contributions, a finding backed up by Zhu et al. (2019). In contrast, Kane and Ransbotham (2016) argue that while pre-existing information might foster contributions at an initial stage, over the long term, this relationship might break down when pre-existing knowledge becomes relatively complete. Finally, when relying on a field experiment that adds content randomly to Wikipedia pages, Hinnosaar et al. (2019) find no significant effect of pre-existing content on driving follow-on contributions. Although these papers make important advances, they do not analyze the impact of information seeding at an early stage in the project's life on long-term dynamics. Further, although many of these studies analyze communities with problem-solving and open-ended tasks (such as Wikipedia or open source), I analyze information seeding in a context that is largely about low-level information provision. In this context, it seems like the limits of information seeding might be particularly salient.

More broadly, the present study contributes to the literature on knowledge production in user and open innovation communities (Lerner and Tirole 2002, Von Hippel 2005, Boudreau et al. 2011, Faraj et al. 2011, Dahlander and Piezunka 2014). Although theoretical work differs about the relative importance of initial conditions (Raymond 1999, Lerner and Tirole 2002, Athey and Ellison 2014), this work provides the first empirical evidence to suggest that initial conditions can shape the long-term evolution of online communities. Further, I evaluate a new practice, information seeding, which complements past work that investigate how online communities are shaped by factors such as the disclosure of intermediate results (Boudreau and Lakhani 2015), intellectual property (Nagaraj 2017), awards (Gallus 2017), incentives (Lyons and Zhang 2018), demand shocks (Kummer 2013), audience size (Zhang and Zhu 2011, Piezunka and Dahlander 2015), and competition (Nagaraj and Piezunka 2017). Finally, the idea that contributors might be motivated by a sense of ownership over their contributions adds to the wide-ranging literature focused on the question of why contributors exert costly effort for no financial compensation (Franke and Shah 2003, Lakhani and Wolf 2003, Shah 2006, Nagle 2018).

The rest of the paper proceeds as follows. Section 2 describes the setting, research design, and data. Section 3 provides the empirical estimates on contribution and contributor activity, long-term quality, as well as the heterogeneous effects of information seeding on OpenStreetMap. Section 4 explores the ownership mechanism in detail and Section 5 concludes.

2. Setting, Research Design, and Data

2.1. Setting: OpenStreetMap

OpenStreetMap is an online, collaborative project to create a free, editable map of the world (Haklay and Weber 2008). It was inspired by Wikipedia and was launched in the United Kingdom in 2004 when other popular online mapping tools such as Google Maps were not yet available. OpenStreetMap has grown to more than 5.3 million registered contributors[6] and is one of the largest community-based knowledge production platforms on the web, with about half the number of active contributors as Wikipedia has.[7] Although OpenStreetMap has global coverage, I will concentrate on the OpenStreetMap project in the United States.[8] OpenStreetMap is different from other commercial providers in that the mapping data are sourced from volunteers and is available under a relatively open license. It is therefore reused freely, including in popular Internet services such as Craigslist, Foursquare, Uber, Snapchat,[9] Apple Maps (Coast 2015), as well as in self-driving cars and mobile games.[10]

To contribute to OpenStreetMap, a potential contributor must register and then use a specialized editor or a browser (Neis et al. 2011). In places with a blank map, contributors can make a basic contribution, such as the geometry of a street and its name, using first-person surveys with GPS devices or tracing satellite imagery. In places where baseline information is already present, contributors can improve the map by adding follow-on contributions, including more incremental information like speed limits and turn restrictions and more distant information like buildings, parks, restaurants, and so on. It is not uncommon, in some places, for experienced OpenStreetMap contributors to seed baseline information from external sources such as government or city mapping databases (copying information from copyrighted sources, including Google Maps, is not permitted), leaving largely follow-on editing for the community. After a contribution has been saved, the username of the contributor is recorded allowing the contributor to feel a sense of ownership over the object (for example, a street or a restaurant). OpenStreetMap stores the entire history of contributions to the map, including the date, time, and contributor of each edit (anonymous edits are not permitted), thereby tracking contributions and contributor activity in the map over time.

2.2. The TIGER Experiment

2.2.1. The U.S. Census TIGER Map. When OpenStreetMap was launched in the United States, rather than start the map from scratch, the fledgling community decided to import the U.S. Census TIGER map into their system. TIGER is a computer-readable map that was developed in cooperation with the U.S. Geological Survey (USGS) in response to problems in the 1980 census (Marx 1986). It provides basic street information, the location of populated areas (including cities and towns), and administrative boundaries for all regions in the United States. By design, TIGER does not include any follow-on information such as speed limits or lane information that are relevant for GPS routing, nor does it contain information about buildings, parks, or local points of interest.

Although the TIGER map offers many benefits, notably its national coverage and lack of copyright, critics have pointed to serious problems with its completeness. TIGER maps were designed to guide census officers in matching census units with their geographical location and were not designed for use in web-mapping applications. Therefore, the U.S. Census prioritized topographical integrity rather than absolute completeness as metrics of quality (Zandbergen et al. 2011). Further, these maps were not updated frequently, and consequently many new neighborhoods and corresponding street information were completely missed. Although it is difficult to quantify the exact extent of this missing information in different locations, as of 2002, there was general consensus that the TIGER map needed to be updated and improved for the 2010 census.

The census undertook a large and ambitious project to improve the TIGER map and make it available in time for the 2010 census (Broome and Godwin 2003). This project, the MAF/TIGER Accuracy Improvement Project (MTAIP), was executed through a $200 million contract awarded to the Harris Corporation in June 2002 (Harris 2002). The program was organized through the Geographic Program and Planning Branch (GPPB) of the U.S. Census (Liadis 2018). The GPPB is responsible for collecting source data (including from the census' 12 regional offices throughout the country) and concurrently determining the order in which this data from specific counties should be updated. Once this order was decided, the GPPB sent the data for each county "southbound," that is, to Harris Corporation where it would be updated and fixed.[11] Once the fixes were made, Harris would send the data back "northbound," after which they would be incorporated into the TIGER data set released to the general public after another review by the GPPB. Although northbound updates arrived as they were completed, TIGER releases for the public were largely issued on an annual basis. This fact led to the situation that updated information for a county was included in a public release of TIGER without delay if it arrived in time for the next TIGER release. For example, counties scheduled to be updated in 2005 were updated and largely included in the 2006 TIGER release used in OpenStreetMap, whereas counties scheduled to be updated in 2006 missed the cut. For counties that had yet to be updated, older, less accurate data were included in the TIGER release, although the underlying difference in the status of updated and yet-to-be updated counties was not made salient to the end user. This important detail ultimately led to the natural experiment that I exploit in this paper. It was not until the 2008 TIGER release, that is, six years after its launch, that the MTAIP program was completed and all U.S. counties were updated in the TIGER database.[12]

To validate my research design, it is important to investigate whether the GPPB systematically selected counties of a certain type to be fast-tracked, leaving other types of counties for later. The Harris Corporation was simply updating counties in the order in which they were received, and so understanding the logic by which the GPPB ordered counties is relevant here. This concern, as well as more information on the process through which counties were ordered, is discussed in detail in Section 2.3. However, I note here that my interviews with census officials and some archival sources give the impression that such systematic selection is not a central concern, even though the ordering of counties was not explicitly random.

2.2.2. Seeding OpenStreetMap with TIGER. When the OpenStreetMap community began to gain momentum in the United States, the possibility of using the TIGER map as a baseline for follow-on contributions seemed attractive to the community. The idea was that rather than having a blank map that people needed to fill in, it would be better to have "a skeleton to build off on."[13] Dave Hansen, a key OpenStreetMap community member involved with the seeding process, wrote the programs that would convert TIGER information to OpenStreetMap format and then import it into the database. In an interview Hansen gave in 2009,[13] he notes:

> The great thing about TIGER was that it may not have been the perfect data, but gave people a place to start. . . . Instead of my street being a completely blank area on the map, there is at least something there that looks like my street [that I can fix]. . . . Having this framework makes it a lot more approachable.

Driven by the logic that seeding OpenStreetMap with TIGER maps would fuel its development, OpenStreetMap began the process to incorporate TIGER information from the U.S. Census in late 2007 and completed the process by January 2008 (Zielstra et al. 2013). Since 2002, the TIGER map was issued annually with updates from the MTAIP program in the previous year and OpenStreetMap imported the 2006 version. In this paper, I assume that the full 2006 TIGER map was present in OpenStreetMap beginning in the first quarter of 2008 (i.e., January–March 2008), and absent before then. Note that TIGER information was incorporated for 3,107 counties within the United States; the state of Massachusetts was excluded because better quality information was available from the state government.[14] I will restrict my analysis to these 3,107 counties. Finally, it is important to note that while there was a small but fledgling community of OpenStreetMap contributors before the TIGER experiment, most of the United States was relatively empty. In places where information existed from previous contributors, the process of information seeding tried to preserve these contributions, although in some cases, contributors agreed to have TIGER overwrite pre-existing information. Further, contributors usually start editing from the map view, where it is difficult to tell the provenance of the data. Even when the contributor has entered the editing window, discovering that the data came from Dave Hansen's TIGER account is relatively difficult under the web interface. It is only through detailed analysis (such as the one I undertake in this paper) or related methods that a contributor could learn the provenance of TIGER data. Finally, even if contributors discovered the source of the TIGER data, none of our interviews suggested that contributors had increased trust in TIGER data as compared with community provided information.

2.2.3. Variation in TIGER Seeding. Although all counties were incorporated in OpenStreetMap using a similar computer program from the 2006 TIGER map, the partial completion of the MTAIP program (see Section 2.2.1) by this date meant that there was wide variation in the level of information that was seeded. As of 2006, only 1,851 of the 3,107 counties that OpenStreetMap included had been updated by the MTAIP initiative. The remaining 1,256 were slated to be completed by 2008, and although this goal was achieved on schedule, the fully complete basemap was never used within OpenStreetMap. Consequently, once the 2006 TIGER map was fully incorporated within OpenStreetMap, many contributors noted that although TIGER seemed complete and high quality in some places, it was incomplete in others. For example, contributor Matthew Perry notes[15]:

> Some TIGER data I've seen suffers from horrible spatial accuracy. . . . Areas are missing crucial data (entire sections of long-established highways). . . . On the other hand, many areas of TIGER are beautifully accurate.

The fact that only about 60% of the counties in the United States were seeded with a complete basemap seems to have been missed entirely by the OpenStreetMap community, perhaps because the census did not prominently advertise this fact. In numerous online discussions of the TIGER import that I examined, I was not able to find a single mention of the MTAIP project, suggesting that OpenStreetMap contributors were unaware of this project and its implications for the quality of the TIGER data. Further, we conducted interviews with a few contributors to OpenStreetMap during this time, and although each one talked about the varying quality of the TIGER map, they did not mention the MTAIP program and were unaware of the systematic differences that I highlight here.

Further, there was no expectation that the TIGER import process would ever be repeated and that incomplete county information would be updated, partially because it was difficult to merge existing data with a new import of this scale. For example, the OpenStreetMap help pages document that "It is unlikely that the TIGER data ever will be imported again."[16] This unintentional and lesser known variation in information seeding (see Fischer 2013, Nagaraj 2014 for some related commentary) during the U.S. OpenStreetMap community's formative years forms the basis of the natural experiment that I exploit in this paper.

2.3. Research Design and Validity of the TIGER Experiment

In theory, the variation introduced by the introduction of the TIGER map into OpenStreetMap can be used to estimate the impact of information seeding on follow-on contributions, if the counties affected by the MTAIP update were comparable to those that were not. This section provides qualitative background on the process through which counties were ordered and some quantitative comparisons between treated and control counties. Note that I will account for any possible differences through county fixed effects and a variety of controls (depending on the specification) but understanding the process through which certain counties were treated provides further confidence in the research design, and suggests specific robustness tests.

2.3.1. Qualitative Background. First, to qualitatively understand the order in which data for the over three thousand U.S. counties were updated, I contacted multiple senior officials at the U.S. Census Bureau who were familiar with the procedural details and organization of the MTAIP project. The officials clarified that the Census Bureau was responsible for setting the order in which counties were to be updated; Harris was responsible only for making the updates. Approximately 700 counties were scheduled to be updated every year from 2004 to 2007, with the remaining to be completed in 2008, although this schedule was not followed exactly. In determining the order in which counties were to be updated, the Census Bureau relied on input from its 12 regional offices (in Atlanta, Boston, Charlotte, Chicago, Dallas, Denver, Detroit, Kansas City, Los Angeles, New York, Philadelphia, and Seattle). Each county was under the purview of one of these offices, and each office tried to obtain updated data for its region. This countrywide distribution ensured that treatment counties did not disproportionately represent any one part of the country, but rather came from all regions of the United States. These source data were then collated by the GPPB and then sent to Harris in an order that the GPPB determined.

Figure 1 depicts the relatively even distribution of treatment and control counties across the United States and their relative balance across all the different regions in the United States. For example, it does not seem to be the case that all of the major urban centers in the United States were in the treatment group. In fact, in my interviews I also found that the decision to balance the updating of rural and urban counties was also intentional on the part of the U.S. Census because if the U.S. Census had prioritized all important or highly populated areas at the start, there would have been difficulties with project implementation Ratcliffe 2014

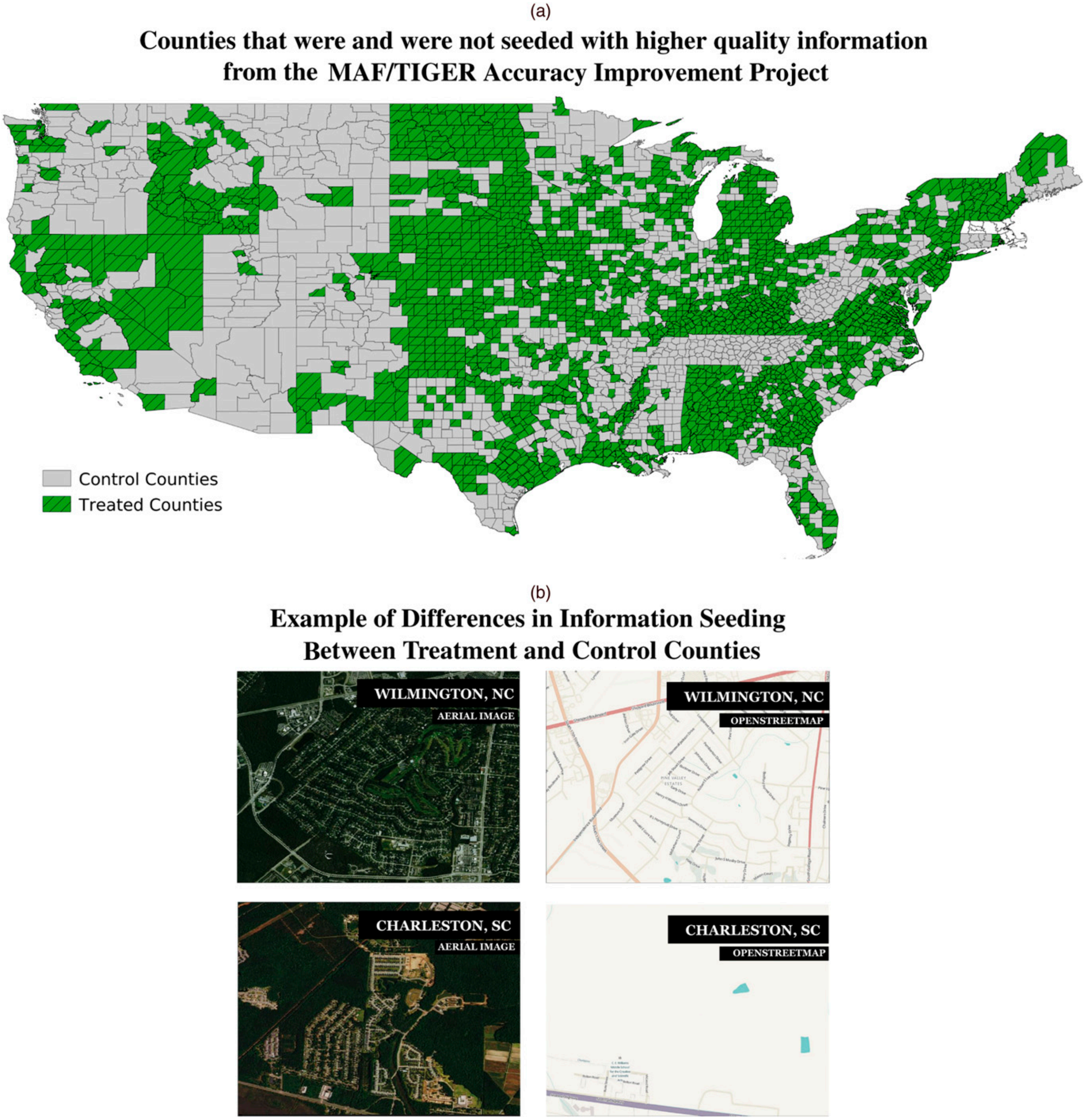
> We had about 12 regional offices around the country and wanted to make sure that we had a distributed work load. . . . We did not think, OK we're going to start with highest population—large land areas and dense areas are all difficult to process at the same time!

Overall, the qualitative evidence suggests that the variation in the timing of MTAIP updates could be used to estimate the effect of seeding on follow-on OpenStreetMap contributions.

2.3.2. Quantitative Comparison. Although the qualitative background from my interview is reassuring, it is important to quantitatively evaluate the assumption that treatment and control counties exhibit similar rates of change over time. To formally assess this conjecture, I collect information on income, population, population density, and demographics from the American Community Survey (ACS) at the county level. These data are useful with the cross-sectional specifications since they help control for any systematic differences between counties in a granular way. Further, for the panel models that include county fixed effects, one can use these data to examine the differences in trends between treatment and control counties. For example, given that young, highly educated males are more likely to contribute to open source projects than other demographic groups (Glott et al. 2010), if this demographic is growing at a higher rate in treatment counties than in control counties, the validity of the natural experiment might be called into question.

Using these data, the controls section in Table 1 presents the rate of change of six of these primary control variables between 2014 and 2005 (after and before the TIGER experiment) for treatment and control counties. These data make it clear that the two sets of counties are largely similar along five of the six dimensions I compare. The one exception is the fact that treatment counties appear to have a slightly greater increase in per-capita income (*Δ Income Per Capita* in Table 1) than control counties. Although this systematic difference between the two categories might appear problematic, it is useful to note that

Figure 1. (Color online) Research Design



(a)

Counties that were and were not seeded with higher quality information from the MAF/TIGER Accuracy Improvement Project

(b)

Example of Differences in Information Seeding Between Treatment and Control Counties

Notes. Panel (a) highlights the counties with higher level of information seeding in dark grey hatched pattern, whereas counties that received a lower amount of information seeding are presented using light grey solid pattern. Note that counties in the state of Massachusetts have been excluded because they were not seeded with TIGER information. Panel (b) gives an example of the research design for two neighboring towns, a control county, Charleston, SC (bottom) and a treated county, Wilmington, NC (top). The satellite image provides "ground truth" for the two towns, and the map on the right shows their status on OpenStreetMap after TIGER seeding was completed.

richer counties are more likely to contribute to projects like OpenStreetMap, and this difference makes it less likely that we will find a negative effect of seeding on contributions, the main hypothesis of this paper. Having said that, in addition to the comparison checks presented earlier, I include time-varying controls for income (and all of the other control variables) in the regression analysis. These patterns can also be seen in Figure D.1 in the online appendix, which plots these variables at the annual level between treatment and control counties. Note here the level differences between some of these variables that it will be important to

Table 1. Cross-Sectional Comparison (County Level (N = 3,107))

Variables	(1) Treatment (N = 1,851) ȳ	(2) Control (N = 1,242) ȳ	(3) Diff.	(4) *p*-value
Contribution outcomes				
Contributions	2,375.5	3,593.3	–1,217.7	0.00
Follow-on Contributions	457.9	635.5	–177.6	0.05
Community outcomes				
Contributors	7.482	7.966	–0.484	0.26
Quality				
Error-Score	1,830.2	2,046.7	–216.5	0.09
Mechanism				
Owner-Contributions	149.6	203.6	–54.08	0.21
Controls				
Δ *Population*	7,324.3	6,756.5	567.8	0.62
Δ *Households*	2,186.4	1,899.8	286.6	0.42
Δ *Unemployed Pop.*	–646.1	–543.6	–102.5	0.27
Δ *Educ. Population*	1,995.9	1,884.8	111.0	0.68
Δ *Male Population (18–45)*	555.3	410.5	144.8	0.42
Δ *Income per Capita*	3,152.7	2,773.7	379.0	0.01

Notes. The summary statistics in this table help to evaluate the impact of the TIGER seeding experiment on OpenStreetMap outcomes, as well as the selection of counties into the treatment and control groups. For the control variables, Δ represents the difference in the variable between 2014 and 2005 for a given county.

control for in the cross-sectional specifications, as well as the similar trends among treatment and control counties that is reassuring for the difference-in-difference specification. In other words, even if the difference-in-difference specification has limited "pre" data on the key outcome variable, finding balance in trends across these covariates serves as a useful proxy.

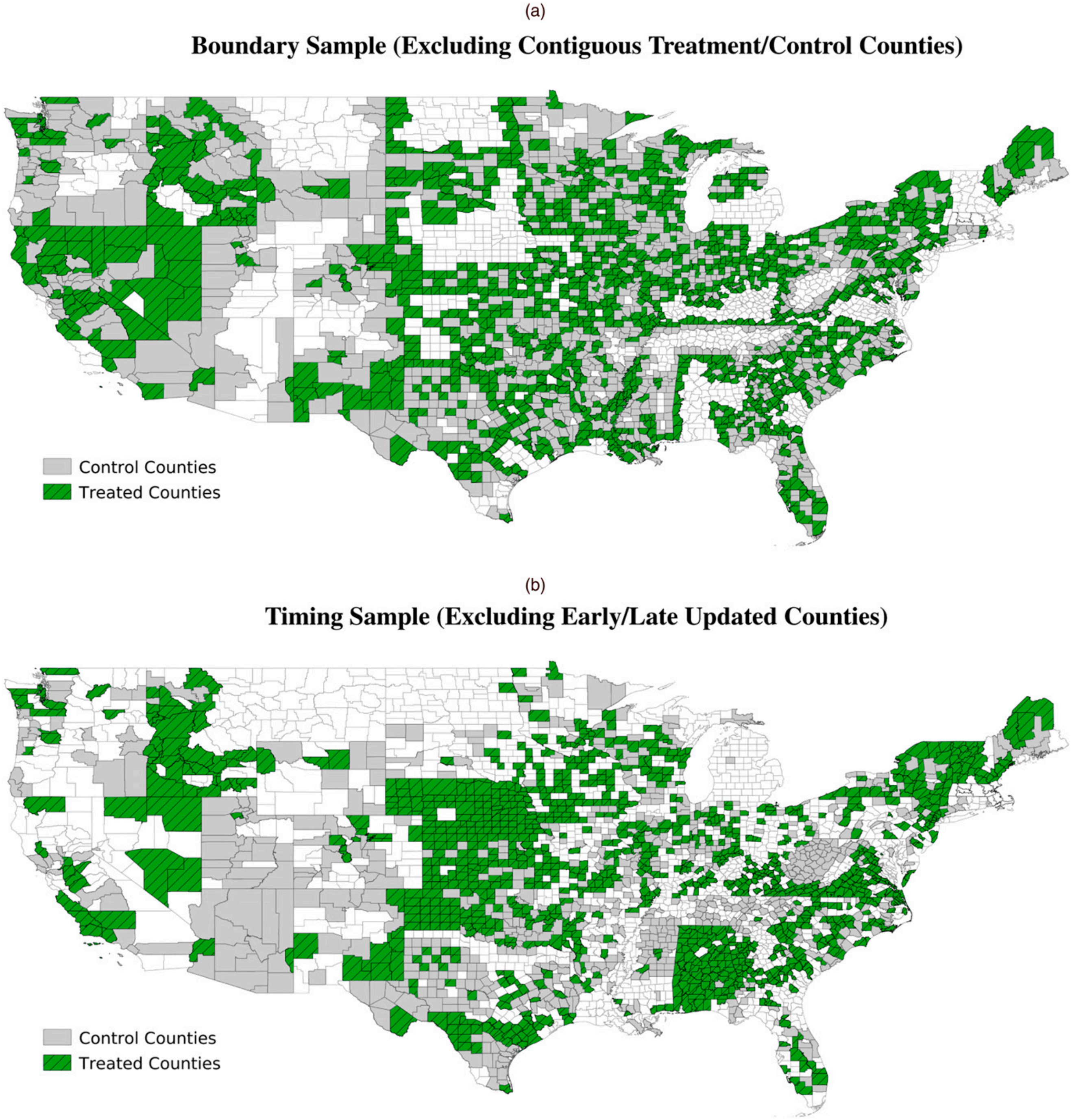
2.3.3. Alternate Samples. Finally, although the qualitative and quantitative analysis strongly suggests that treatment and control counties are comparable, I conducted additional interviews with Census Bureau officials to see if there were any additional sources of selection that I had not accounted for. In these interviews, Census Bureau officials admitted to me that the ordering of counties was driven by practical considerations leading to two sources of selection that might have crept in. First, conditional on a county being chosen to be updated in a given month, the logistics made it quite likely that a cluster of nearby counties would be chosen as well. This created a pattern where the ordering of county clusters was determined in no systematic order, although there was significant clustering across counties within a given region. Second, the census tried to hold out a small group of fast-growing regions for the last year of the updating so that TIGER maps would not go out of date by the 2010 census. North Dakota, in particular, was growing fast given the fracking boom at the time and was therefore scheduled to be updated near the end of the program (Liadis 2018).

Fortunately, it is possible to design alternate samples to tackle both of these challenges. First, inspired by some recent work in the labor literature that uses counties on either side of a minimum wage threshold (Dube et al. 2010), I construct a sample of counties that neighbor another county of the opposite treatment status. In other words, using a GIS algorithm, I identify counties that are completely surrounded by others of only one type, that is, treatment or control counties, and drop them from the sample. The resulting sample is shown in Figure 2, panel (a). This exercise helps get rid of clustering among treatment and control counties and provides an alternate sample to establish the robustness of the baseline results. In total, this boundary sample includes 1,820 counties (1,096 treatment and 724 control).

Second, I was able to obtain proprietary data from the U.S. Census Bureau officials on order in which counties were scheduled to be updated as per the GPPB's instructions. Most of the counties that were updated in the 2006 TIGER edition (and which form the treatment sample) were scheduled to be sent to the Harris Corporation by the end of 2005. I therefore include only those counties scheduled to be updated one year either side of this date, that is, counties scheduled to be updated in the calendar years 2005 or 2006 to be a part of the sample. This exercise leaves us with 2,218 counties (1,228 treatment and 990 control) and forms the timing sample, as show in Figure 2, panel (b).

Combined, the qualitative and quantitative tests in this section, coupled with the granular cross-sectional controls and fixed effects (in the cross-sectional specification) and the nonparametric county fixed effects combined with time-varying demographic, population, and income controls (in panel models), help to establish the robustness of the research design and the validity of

Figure 2. (Color online) Alternate Samples


(a)
Boundary Sample (Excluding Contiguous Treatment/Control Counties)
Control Counties
Treated Counties
(b)
Timing Sample (Excluding Early/Late Updated Counties)
Control Counties
Treated Counties

Notes. Panel (a) shows the subset of 1,820 counties (of 3,107) that are included in boundary sample. Panel (b), the timing sample, is based on a schedule of counties to be updated by the MTAIP program by month and by TIGER/control status. This map shows only those counties scheduled to be updated in the years 2005 and 2006 with counties scheduled to be updated before or after this period excluded. This sample includes 2,218 (of 3,107) counties.

the TIGER experiment. The boundary sample and the timing sample provide the chance to further test the robustness of the main hypotheses.

2.4. Data

I rely on four sources of data. First, I employ the complete history of OpenStreetMap to measure contribution and contributor activity. Second, I collect data on the implementation of the TIGER program from internal U.S. Census Bureau information, including data on the locations of the treatment and control counties and the scheduled timing of county updates. Third, I build measures of quality of OpenStreetMap county maps by comparing routing distance between OpenStreetMap and commercial alternatives. And finally, I leverage data from the ACS for additional controls. This section describes these data in more detail.

2.4.1. OpenStreetMap Data. The OpenStreetMap project stores all past versions of the map in the form of a history file.[17] I use a version of this history file that contains data for the North American continent from the start of the project in 2005 to the end of 2014. From this file, I employ scripts to extract data for each county in the United States for map objects relevant to the analysis, including streets, street tags, and distant contributions such as a parks, buildings, and so on.[18] This process creates the source data for this project, which includes almost 839.2 million contributions totaling about 100 gigabytes of data made by more than 577,000 unique contributors.[19] I then drop all contributions made by a small set of automated scripts and bots. Most importantly, I deleted all edits from the username DaveHansenTiger, which was the unique username created to make the updates from the 2006 U.S. Census TIGER map.

I then calculate the primary outcome variables, *Contributions*, *Follow-on Contributions*, and *Contributors*. *Contributions* measures the total number of edits including basic information (such as street geometry and street names) as well as follow-on information. *Follow-on Contributions* includes the sum total of contributions that either (a) added a tag to existing streets with information about one ways, speed limits, or access type; (b) created or modified a building; or (c) created or modified amenities, such as a restaurant, park, or other points of interest. Finally, *Contributors* measures the total number of unique user IDs making an edit. I calculate three versions for all three variables. For the cross-sectional specification, I consider the cumulative total of these variables post-TIGER, that is, between the years 2008 and 2014, as well as the total for the year 2014, to measure long-run effects. For the panel models, these measures are calculated at the county-quarter level from 2005 to 2014.

Next, I calculate variables that investigate the mechanisms through which information seeding affects knowledge production. First, I measure two types of follow-on information: *Distant Follow-on* information contains knowledge that is related to buildings (and is unrelated to street information provided by TIGER), whereas *Incremental Follow-on* information contains information that adds to street information that may have been provided from TIGER (including data on speed limits, one ways, and access type).[20]

Second, I divide the total number of contributors who are active into two groups: *New Contributors* are those who are making an edit in a given county for the first time, whereas *Old Contributors* are those who have made at least one edit before the seeding effort. Third, within the set of old contributors, I classify those with a high versus low level of ownership. Conceptually, I label those contributors who make contributions within a narrowly bounded geographic region as having a high level of ownership as compared with those who make edits over a wider surface area. In practice, if most of the edits of a given contributor are within a box of width 0.1 degree latitude and 0.1 degree longitude, the contributor is classified as having a high level of ownership, whereas if most edits are made in a more diffuse fashion outside of a narrow 0.1×0.1 latitude/longitude area, the editor is classified as having a low level of ownership.[21] Finally, I also directly measure the number of owner contributions. These are the total number of follow-on contributions to a given object by a contributor who created the object from scratch in the first place. For example, if contributor A adds the basic information, this measure includes the total number of times the same contributor A adds follow-on information to the same object.[22]

2.4.2. MTAIP Implementation Data. The main independent variable, the treatment status of a county, is derived from internal documents charting the progress on the MTAIP implementation. In particular, I rely on a map that records the counties for which the U.S. Census had finished correcting the data by the end of 2006.[23] Using this map, I designate counties updated by the MTAIP program by 2006 using an indicator variable, *Treatment*, where this variable equals one if the data for a county had been corrected before its use within OpenStreetMap. Further, I complement this treatment assignment data with newly obtained data from the U.S. Census Bureau on the schedule for the updates for the MTAIP program, which includes the date on which a particular county was scheduled to be sent to Harris for updating.

2.4.3. Quality Data. To examine the impact of seeding on the quality of OpenStreetMap, I build a measure of quality that can pick up differences in the type of follow-on information that was affected by the TIGER experiment. As noted earlier, differences in street information (one type of follow-on information affected by TIGER) are significant for the quality of automobile routing. Therefore, as an indicator of quality, I measure the error score as the absolute value of the difference in length between a route proposed by an OpenStreetMap-based routing program compared with one suggested by a routing program from a reliable and well-regarded third-party source. I rely on comparisons between the OpenStreetMap-based routing program and Google Maps, given the widespread acknowledgment of Google Maps' quality in this regard and following the OpenStreetMap literature (Goodchild and Li 2012, Zielstra et al. 2013).

I compute the error score as follows. First, using the OpenAddresses database,[24] I collect the list of all known addresses in 2,312 counties (1,325 treatment and 987 control) in the United States for which such information is available. For each county, I randomly chose a set of 50 address pairs (a starting address and a destination address), netting a total of 98,711 address pairs, including some counties for which I am not able to obtain the full quota of 50 addresses. For each address pair, I queried the OpenSource Routing Machine (OSRM)[25] as well as the Google Maps application programming interface (API) to provide me with the shortest possible route between the two addresses.[26] I then compute the difference in routing distance offered by the two services, logging it to normalize outliers. This measure can be easily interpreted as the logged value of the difference in distance one would travel (either longer or shorter) if one used OSRM in place of Google Maps. The key outcome is the average logged error at the county level, Log(ErrorScore), which is the key time-invariant outcome variable that measures the quality of OpenStreetMap information as of 2017.

The error score captures quality in the sense that when OpenStreetMap has more complete information it is said to have higher quality. If one is interested in purely the quality of information, not its mere presence, in Online Appendix B, I develop an additional measure of quality that focuses on the accuracy of restaurant names as a complement to the error score metric developed here, and find that the basic results are largely similar.

2.4.4. Controls. Finally, in addition to the dependent and independent variables mentioned earlier, I also collect information on nine demographic and income variables at the county level to serve as controls. Using the ACS five-year estimates from 2009 to 2014, I extract the following nine variables at the county level: area, population, number of housing units, earnings, median age, number of males between 18 and 44, number of college educated individuals, number of workers in the information technology (IT) industry, and number of highly educated (those with a master's degree, a Ph.D., or a professional degree).

2.4.5. Summary Statistics. Table 2 provides a list of the main variables used in the quantitative analysis and summary statistics for the sample at the county-quarter level. The sample contains information for 3,107 counties over 39 quarters from the second quarter of 2005 to the last quarter of 2014, a total of 120,627 observations. The primary outcome variables of interest are *Total Contributions*, *Total Follow-on Contributions*, and *Total Contributors*. The median county-quarter experiences about 69 contributions, of which 11 seem to be follow-on contributions, made by about four contributors. These data display extremely skewed distributions and the means for these variables (2,867.80 for contributions, 529.68 for follow-on contributions, and 7.68 for contributors) are significantly larger than the median.

In Table D.1 in the online appendix, I provide summary statistics separately at the county and quarter level to investigate this skew. The medians are significantly smaller than the mean in the county-level panel A, rather than the quarter-level panel B, implying that most of the skewness is driven by cross-county differences, rather than those over time. This is not surprising given that interest in OpenStreetMap over time grows in a stable and consistent way, whereas contributions vary dramatically across counties depending on factors such as county size, population, and demographics.

In addition to the key outcome variables, Table 2 also presents some summary statistics for the related outcomes (distant/incremental follow-on, new/old contributors, low/high ownership contributions, owner-contributions), and the timing and control variables. For example, as is clear from this table, of the mean of 529 follow-on contributions, almost 171 are owners editing objects they created, showing the relatively important role that ownership could play in driving contributions in this setting.

3. Results: Does Information Seeding Hurt Follow-on Contributions and Contributor Activity?

3.1. Simple Differences in Descriptive Statistics

I first begin by exploring differences between treatment and control counties in the raw data. Figure 3 plots the logged and cumulative number of quarterly contributions (panel (a)) and follow-on contributions (panel (b)) in treatment and control counties over time. Contribution activity is quite low before the TIGER map was seeded in 2007, quarter 4 (as indicated by the vertical line) in both treatment and control counties. After the TIGER information is seeded, treatment and control counties start to diverge, with control counties receiving significantly more contributions than treatment counties. Note the secular trends in both groups: overall activity is quite low until 2009, after which both the stock and the flow of contributions rises dramatically. OpenStreetMap as a platform grew in popularity starting 2009 in the United States, and this trend is reflected in these data. If seeding affects how potential contributors contribute and stay engaged in the platform, after they've discovered it, large differences between treatment and control counties should show up only after

Table 2. Summary Statistics

Variables	Mean	SD	Median	Min	Max
Outcomes					
Total Contributions	2,867.80	28,510.38	69.00	0.00	3,625,156.00
Total Follow-on Contributions	529.68	13,416.84	11.00	0.00	3,625,156.00
Contributors	7.68	46.12	4.00	0.00	8,535.00
Distant Follow-on	120.21	3,612.47	0.00	0.00	422,200.00
Incremental Follow-On	366.00	12,551.49	2.00	0.00	3,625,156.00
New Contributors	4.77	45.36	2.00	0.00	8,532.00
Old Contributors	0.10	1.14	0.00	0.00	293.00
Low Ownership Contributions	2.00	87.47	0.00	0.00	20,595.00
High Ownership Contributions	123.52	6,411.82	0.00	0.00	1,268,382.00
Owner-Contributions	171.42	5,729.41	1.00	0.00	1,695,215.00
Timing variables					
Treat	0.60	0.49	1.00	0.00	1.00
Post	0.72	0.45	1.00	0.00	1.00
Year	2,009.62	2.82	2,010.00	2,005.00	2,014.00
Quarter	200.00	11.25	200.00	181.00	219.00
Select controls					
Population	100,029.21	320,590.11	25,905.00	41.00	10,230,943.00
Households	37,143.93	112,520.17	9,880.00	22.00	3,325,103.00
Population Density	265.36	1,771.88	45.36	0.06	72,839.69
Per-Capita Income	23,765.17	5,854.17	22,966.00	1,601.00	82,817.00

Notes. The data are a balanced panel for 3,107 counties (excluding Massachusetts) and 39 quarters from the second quarter of 2005 to the last quarter of 2014 for a total of 121,173 observations. See text for data and variable descriptions.

OpenStreetMap had significant traction and a sufficiently large contributor community, that is, about a couple of years after the TIGER seeding. This is what we see in these data. Note that the patterns in panel (a) could be explained by the possibility that there are more opportunities for contributions in control counties. However, this gap persists even when comparing only follow-on contributions (such as speed limits or points of interest) for both treatment and control counties (panel (b)). This apples-to-apples comparison shows that the difference between treatment and control counties could be linked to the higher level of information seeding in treatment counties than in control counties.

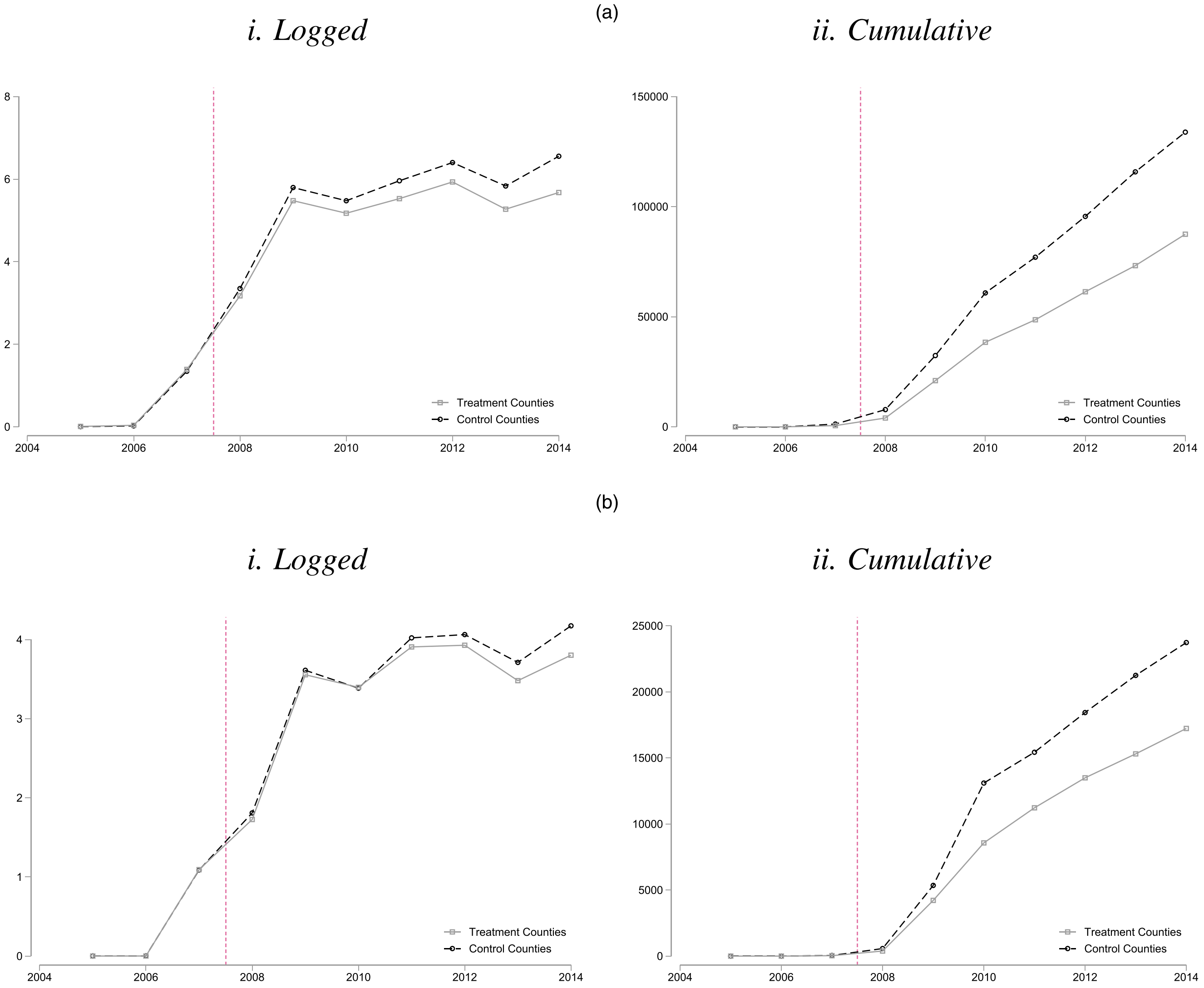
To add to this graphical analysis, Table 1 describes estimated differences in means for the key outcome variables. According to these data, despite receiving a lower level of seeding from the TIGER maps, control counties received a larger number of total contributions, and more importantly, a larger number of follow-on contributions. In particular, treatment counties received about 457.9 follow-on contributions on average in a quarter, whereas control counties received 635.5, a difference of almost 177.6 contributions or about 38.8%. The difference in the number of active contributors is smaller, with control counties having about 0.484 more active contributors, a difference of about 5%, which is not statistically significant at the 95% level. Regarding the long-term quality of the map, Table 1 suggests that treatment counties have slightly lower error scores, although in regression analysis I will establish that treatment counties exhibit a moderately higher level of errors than control counties when some basic controls are added to this cross-sectional comparison. Finally, when considering the number of owner contributions, that is, the number of times the creator of an object makes a follow-on contribution on the same object, there is also a difference in the mean values (albeit not significant), suggesting a potential channel driving the overall differences in follow-on contributions.

Figure 3 and Table 1 together suggest that when the raw data are evaluated, there do seem to be some negative consequences of the higher level of information seeding on follow-on contributions and contributor activity in treatment counties as compared with control counties. Having explored the raw data, I now turn to formally testing the main hypotheses, first via difference-in-difference models and second, cross-sectional specifications. Both methods provide alternate and complementary approaches to identify the effects of seeding.

3.2. Difference-in-Difference Estimates

3.2.1. Baseline Estimates. For the panel analysis, I estimate regressions of the following form using the county-quarter panel: $Y_{it} = \alpha + \beta_1 \times Post_t \times Treat_i + \gamma_i + \delta_t + \beta \cdot X_{it} + \varepsilon_{it}$, where γ_i and δ_t represent county and quarter fixed effects, respectively, for county i in quarter t; $Post_t$ equals one for all quarters after December 2007 when the TIGER implementation was completed on OpenStreetMap; $Treat_i$ equals one for all counties that benefited from the TIGER improvement program on OpenStreetMap, that is, they had been

Figure 3. (Color online) Mean Outcomes for Treatment and Control Counties


(a)

i. Logged

ii. Cumulative

(b)

i. Logged

ii. Cumulative

Notes. In both panels (a) and (b), (i) represents logged versions of the outcome variable, whereas (ii) represents a cumulative sum of contributions up until a given quarter. The vertical line represents the quarter when TIGER data were imported into OpenStreetMap. The dark grey dashed line represents average values for outcome variables in control counties, while the light grey solid line represents outcomes in treatment counties.

significantly improved through the MTAIP project in the 2006 version of the TIGER database; and X_{it} denotes a series of county-quarter control variables.

This specification compares the difference between treatment and control counties in a differences-in-differences framework. If a higher level of information seeding encourages follow-on contributions and contributor activity within OpenStreetMap, then the coefficient on β_1 should be positive and significant, but if information seeding has a negative effect on these variables, then the estimate of β_1 should be less than zero. The presence of county and time fixed effects is quite powerful because they control for time-invariant differences in the underlying proclivity of each county to contribute to OpenStreetMap. Further, these controls help to account for trends in the popularity of the OpenStreetMap platform over time and technology trends such as the rise of smartphones, which increased interest in mapping technology. I estimate this model using log-linear models because of the highly skewed distribution of the dependent variables.

Table 3 presents estimates from this regression for total contributions (columns (1) and (2)), follow-on contributions (columns (3) and (4)), and contributors (columns (5) and (6)). All models include county and quarter fixed effects, and columns (2), (4), and (6) also include time-variant county-level controls for population, demographic, and income characteristics. The estimates suggest a negative impact of seeding the OpenStreetMap platform with TIGER in terms of all three outcomes. Although the result of a reduction in total contributions by about 47.2% (column (2)) is perhaps not surprising, even when considering only follow-on contributions, the gap is about 15.2% (column (4)) and about 5.6% (column (6)) when considering contributors. These effects are statistically significant and economically meaningful for OpenStreetMap. Further, the coefficients do not change much after the inclusion of quarterly demographic and related controls, suggesting that the counties that had been updated in TIGER through the MTAIP project by 2006 were reasonably comparable to counties that had not yet been updated. These baseline results, therefore, support the conclusion that instead of spurring the development of communities and growing the contributor community, seeding the OpenStreetMap platform with higher levels of information crowded out follow-on contributions and discouraged active contributors on OpenStreetMap. Table D.3 in the online appendix presents similar results but shows the estimated coefficients for the control variables, which are suppressed in the Table 3, as well results from models without any county fixed effects for comparability with the cross-sectional specification presented later.

3.2.2.Time-Varying Estimates. Next, I evaluate the parallel trends assumption, which is the key assumption underlying the difference-in-difference specification. Specifically, this test verifies that the primary outcome variables evolve in a similar way in both treatment and control counties prior to the implementation of information seeding. Note that this test has a key limitation in that engagement on the platform is low before the seeding effort. However, it is still useful to run this important check given that contributions are not zero. Accordingly, I now turn to estimating the time-varying impact of the TIGER experiment using the following specification: $Y_{it} = \alpha + \Sigma_z \beta_t (Treat)_i \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ_i and δ_t represent county and year fixed effects for county i in year t, and z accounts for the number of years after the TIGER information was first included on the map. Note that this specification is estimated using a county-year sample (rather than a county-quarter sample) for simplicity and because this setup provides more precise values for β_t. This specification is estimated using log-linear models as before. The results are presented in Figure 4, which plots the difference in follow-on contributions and contributors between treated and control counties for every quarter before and after 2007. The dotted lines represent 95% confidence intervals.

Figure 4 shows that before information was seeded into OpenStreetMap, treatment and control counties were following a parallel trajectory in terms of both follow-on contributions and contributors. Note that it seems like it takes about two to three years for the adverse impact of the TIGER seeding to become apparent. Rather than represent some general delay that one might see with seeding in all platforms, this is likely because of the trends in the platform's overall levels of popularity discussed in Section 3.1. Specifically, even though seeding happens in late 2007, OpenStreetMap does not gain popularity until 2009–2010, which is when the differences between treatment and control counties seem to become more apparent.

3.3. Cross-Sectional Estimates

3.3.1. Baseline Estimates. Overall, the evidence presented in Figure 4 is reassuring because it provides some support for the parallel trends assumption, but should be interpreted with caution. By definition, early-stage seeding interventions do not have much of a pretrend, making it difficult to compare across treatment and control units. Therefore, though the difference-in-difference estimates are useful, it is important to complement them with cross-sectional analysis and careful controls to firmly establish that seeding did hurt long-term outcomes on OpenStreetMap.

Since the cross-sectional specification cannot include county-level fixed effects, I include controls for

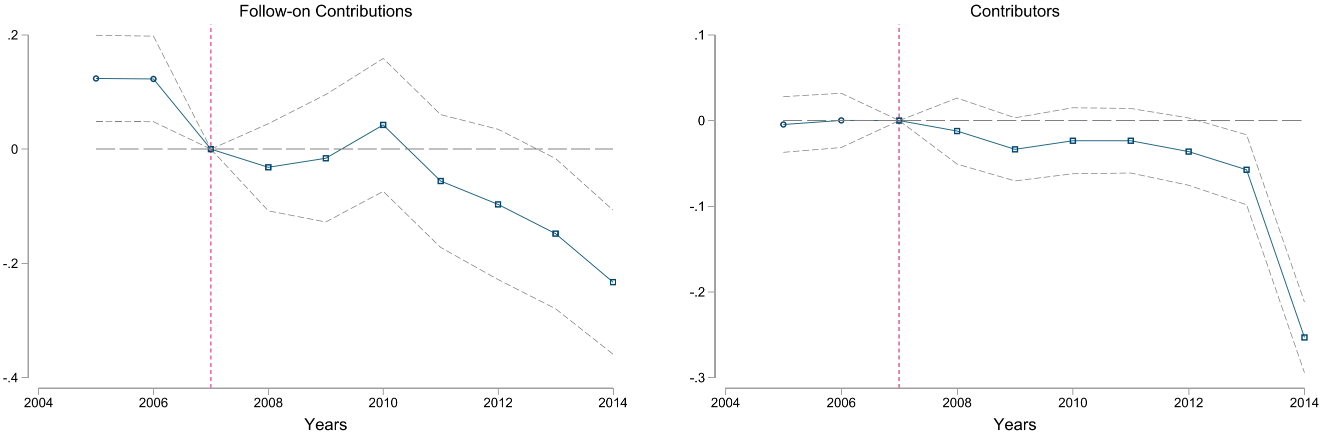
Table 3. Effects of Information Seeding: Difference-in-Difference Estimates

	Total Contributions		*Follow-on-Contributions*		*Contributors*	
Coefficients	(1)	(2)	(3)	(4)	(5)	(6)
Post × Treat	−0.471*** (0.0466)	−0.472*** (0.0403)	−0.141*** (0.0484)	−0.152*** (0.0448)	−0.0557*** (0.0171)	−0.0556*** (0.0143)
Quarter fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
County fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	121,173	121,173	121,173	121,173	121,173	121,173

Note. All specifications are estimated using Log-OLS models, with one added to the dependent variable for all zero values

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Figure 4. (Color online) Time-Varying Impacts of the TIGER Experiment on Contributions



Notes. This figure plots estimates (and 95% confidence intervals) of β_t from the event study specification. The figure is based on county-year observations, the coefficients are estimates from Log-ordinary least squares (OLS) models, the sample includes all county-year observations in the sample, and the standard errors are clustered at the county level.

variables for which treatment and control counties may differ. In particular, as shown in Figure D.1 in the online appendix, we might be especially worried about differences in population density, household income, and per-capita income between treatment and control counties. Accordingly, I divide treatment and control counties into four equal groups by their percentile rank along these three dimensions and include the fixed effects to control for these factors in a nonparametric fashion. Specifically, I estimate the effects of information seeding using the following log-linear specification: $Ln(Y_i + 1) = \alpha + \beta_1 \times Treat_i + \beta \times X_i + \gamma_i^1 + \gamma_i^2 + \gamma_i^3 + \varepsilon_i$, where γ^1, γ^2, and γ^3 correspond to population density, household income, and per-capita income fixed effects, respectively, and X_i indicates a set of 10 controls, including a county's population, unemployment, population density, and the percent of males. Results from this analysis are presented in Table 4. I use two forms of the dependent variable Y_i for each of the three main outcomes (contributions, follow-on contributions, and contributors); the sum total across all quarters postseeding, that is, from 2008 to 2014, and only the levels as of 2014. The idea is to estimate the total impact of seeding over the seven-year period as well as to examine the long-run persistent effects as of 2014.

Across the board, the estimates suggest a reduction in contribution activity on OpenStreetMap in treatment counties as compared with control counties. As before, there is a large and negative reduction in total contributions in treatment counties, but the reduction in follow-on contributions and contributor activity is also maintained. Specifically when considering the totals between 2008 and 2014, all follow-on contributions reduce by about 19% (compared with 15% in panel models) and contributor activity reduces by 12.8% (compared with 5.5% in panel models). These magnitudes increase to 40.5% and 28.3% when considering only the levels in 2014, indicating the large and persistent effect of the seeding experiment on OpenStreetMap. Table D.4 in the online appendix presents similar results, but includes the estimated coefficients for the control variables, which are suppressed in Table 4.

Table 4. Effects of Information Seeding: Cross-Sectional Specifications

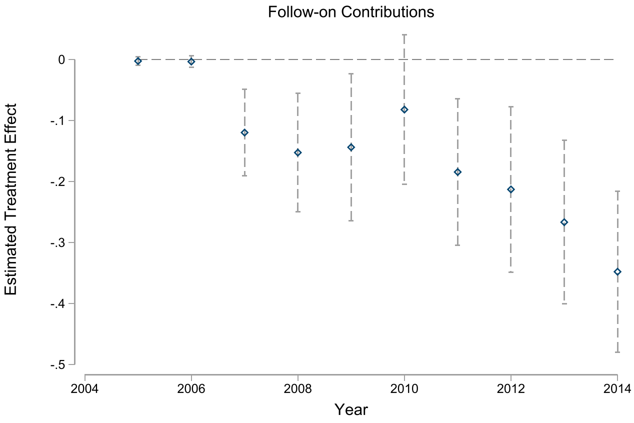
	Total Contributions		*Follow-on-Contributions*		*Contributors*	
Coefficients	2008–2014	2014	2008–2014	2014	2008–2014	2014
Treat	−0.609*** (0.0442)	−0.934*** (0.0521)	−0.193*** (0.0604)	−0.405*** (0.0657)	−0.128*** (0.0166)	−0.283*** (0.0186)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Quantile fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	3,107	3,107	3,107	3,107	3,107	3,107

Notes. For each variable, 2008–2014 indicates the cumulative total of contributions or users over the postseeding period, and 2014 indicates that number of contributions or users in the last year, that is, 2014. Heteroskedasticity robust standard errors are estimated. See text for more details.

3.3.2. Yearly Graphical Estimates. Figure 5 presents another way of looking at these results. Here, I estimate β_1 using 10 different regressions for each calendar year from 2005 to 2014 separately, which are useful to look at the timing of the impact of the seeding on contribution activity. The results are consistent with the difference-in-difference results: the coefficient on the variable indicating treatment is close to zero for the first two years (before information seeding occurs), and becomes significant around the TIGER map implementation. Given that the difference-in-difference estimates make analyzing pretrends challenging, the cross-sectional results provide a simpler and perhaps more convincing way of estimating the effect of seeding on contribution activity.

3.3.3. Stock of Knowledge (Instrumental Variable Specification). The cross-sectional specification also helps address an interpretational issue with the existing analysis. Specifically, the estimates show that a higher level of information seeding leads to about a 15–10% decrease in follow-on contributions. However, the more general question of the elasticity of the flow of follow-on knowledge, per unit of existing knowledge remains unanswered.[27] In other words, how does the existing stock of knowledge affect future knowledge creation? Table D.2 in the online appendix provides estimates that answer this question. Rather than use a binary treatment variable, I instead consider two endogenous measures of the stock of knowledge, which are then instrumented using the treatment dummy. The first, *Pct. Seeded*, is the total length of all highways added by the seeding effort divided by the total length of all highways as per the TIGER database in 2018 (which is presumed to be complete). This measure provides a useful proxy in terms of what percent of a given county was seeded in late 2007. The second, ln(*Seeding Contributions*), measures the total number of individual contributions made by the Dave-HansenTiger account that was responsible for the seeding of OpenStreetMap. This account was setup to chunk the TIGER data into small chunks (usually small segments of streets at one time) and each chunk was added as a separate contribution, allowing us to measure how much information was seeded in a granular manner. Both measures, though not perfect, provide two useful ways of conceptualizing the amount of information seeded beyond a simple treatment/control dichotomy across counties.

Figure 5. Year-by-Year Estimates From Cross-Sectional Regressions.



Notes. This figure plots estimates (and 95% confidence intervals) from multiple regressions (one each for every year 2005 to 2014), estimating the effect of treatment status in a cross-sectional specification, and after accounting population density, household income. and per-capita income fixed effects and indicates a set of 10 county-level controls. Heteroskedasticity robust standard errors are estimated.

Table D.2 in the online appendix provides the estimates from the IV specification for both follow-on contributions and total contributors. The specification is similar to the baseline cross-sectional regression (with the same number of controls), except that the endogenous variable is instrumented with the treatment dummy. First, note the strong first stage in all regressions and the large F-statistic. Further, following the main results, the second stage is negative, that is, a greater stock of knowledge is associated with lower follow-on activity. A one percentage point higher level of seeding (*Pct. Seeded* variable) implies a 2.9% drop in follow-on contributions and a 2% drop in contributors. Similarly, a 1% increase in the number of seeding contributions (which is a large change) is associated with a 27% drop in follow-on contributions and an 18% drop in contributors.[28] These results, although specific to OpenStreetMap and the TIGER experiment, provide more interpretable estimates of the impact of the amount of seeding (in terms of the stock of knowledge) on follow-on contributions and contributor activity. See the footnote for Table D.2 in the online appendix for more details.

3.4. Robustness Checks

3.4.1. Alternate Samples. I examine whether the baseline specifications remain robust when considering the boundary and timing samples described in Section 2.3.3. Table 5 provides these estimates, which are obtained by estimating the baseline specification on a more restricted set of treatment and control counties. Estimates for the difference-in-difference model are in panel A, whereas those for the cross-sectional model are in panel B. When considering panel models, the estimates remain negative and significant, and largely increase in magnitude (except the estimate in the timing sample for follow-on contributions). In the cross-sectional specification, too, the estimated size of the effects are largely similar and larger in magnitude.[29] Note that the set of counties under consideration is quite different in these samples, so the estimates are not necessarily comparable. However, it is reassuring to see that the baseline findings remain robust across a range of dependent variables in both the panel and the cross-sectional specifications.

3.4.2. Additional Tests. Apart from the additional samples, the panel model suggests a few additional robustness tests. Since we have limited pretrends data on the key outcome variable, more robust controls for differential pretrends across counties might be appropriate. For example, imagine that urban areas (those with a higher population density) are seeing the migration of a highly educated population that is more likely to contribute to OpenStreetMap. If these urban areas are disproportionately represented in the control group of counties, my estimates could be the result of this differential time trend between urban and less urban regions. To address this concern, I now divide the 3,107 counties in the analysis into four equal groups based on their population density and including 156 effects at the county-group/quarter level (39 quarters × four county-groups) rather than one fixed effect for each of the 39 quarters in the analysis. The estimates from this specification, show, in Table 5, are robust, and somewhat larger than the baseline analysis.

Second, Google Maps was launched at a similar time as OpenStreetMap in 2005, and it is possible that some of the differences between treatment and control counties are not a result of the seeding experiment, but are driven by differences between the popularity of Google Maps in different regions in the United States. Accordingly, I collected data on county-level popularity of Google Maps from the Google Trends database and control for this directly in Table D.5 (columns (2) and (3)) in the online appendix. The results are, again, slightly larger and significant, suggesting that Google Maps popularity cannot explain the basic patterns here.

Third, it is important to evaluate the concern that the results are driven purely by the research design or that the dependent variables are mechanically related to the independent variables in some way. To address this concern, I evaluate a placebo version of the baseline specification where the primary independent variable, assignment to the treatment county-group, is assigned randomly rather than according to the actual value of this categorization. Reassuringly, estimates from this placebo specification, presented in Table 5, disappear when treatment status is randomly assigned. Finally, there are often cases when a single contributor makes several consecutive edits to one county within a short time interval (for example, adding, saving, and then deleting information). OpenStreetMap codes a set of edits made during one session as a changeset. I replace the main dependent variable to be the total number of

Table 5. Alternate Samples and Additional Robustness

Panel A. Difference-in-Difference analysis

	Diff-Time-Trends		Placebo		Boundary sample		Timing sample	
Coefficients	*Follow-on*	*Contributors*	*Follow-on*	*Contributors*	*Follow-on*	*Contributors*	*Follow-on*	*Contributors*
Post × Treat	−0.187***	−0.0917***	0.00783	−0.0218	−0.135**	−0.282***	−0.0335*	−0.0795***
	(0.0469)	(0.0144)	(0.0483)	(0.0174)	(0.0563)	(0.0634)	(0.0194)	(0.0227)
Quarter fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	121,173	121,173	121,173	121,173	86,502	70,980	86,502	70,980

Panel B. Cross-Sectional specification

	Boundary sample (*follow-on*)		Boundary sample (*users*)		Timing sample (*follow-on*)		Timing sample (*users*)	
	2008–2014	2014	2008–2014	2014	2008–2014	2014	2008–2014	2014
Treat	−0.183***	−0.412***	−0.0815***	−0.200***	−0.265***	−0.535***	−0.121***	−0.291***
	(0.0704)	(0.0761)	(0.0183)	(0.0210)	(0.0780)	(0.0849)	(0.0216)	(0.0245)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quantile fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	2,218	2,218	2,218	2,218	1,820	1,820	1,820	1,820

Notes. In panel A, the specification is similar to that in Table 3. Diff-Time-Trends columns present estimates where the time fixed effects are replaced by county-specific time trends. Specifically, all the 3,107 counties in the study are divided into four equal groups depending on their population density, and county-group specific time trends are included. The placebo column estimates a version of the baseline specification where counties are assigned to the treatment group randomly, rather than based on their true classification. Finally, the boundary sample and timing sample columns estimate the difference-in-difference specification on the subsamples defined in Figure 2. Estimates from panel B are presented using the same specification as in Table 4, except that the sample is limited to the boundary sample (columns (1)–(4)) or the timing sample (columns (5)–(8)). These subsamples are as shown in Figure 2.

changesets contributed (rather than individual contributions), and estimate the baseline specification, as shown in Table D.5 (column (1)) in the online appendix. The coefficient remains negative and significant.

3.5. Effects on Long-Term OpenStreetMap Quality

I now turn to evaluate the impact of information seeding on long-term quality within OpenStreetMap. Seeding might be a worthwhile intervention if it lowers contributor activity and follow-on contributions, but ultimately increases the quality. However, if seeding not only decreased follow-on contributions and contributor activity, but through this decline also reduced the quality of the OpenStreetMap database in the long run, the welfare impact of seeding would more clearly be negative. Recollect that for each county, I collected 50 address pairs to create the quality measure Log(Error-Score), which is calculated as the log-difference in the trip distance provided by OpenStreetMap and Google Maps, as of early 2017. Given the cross-sectional nature of the data, I regress the Log(Error-Score) on variables defined by the cross-sectional specification from the baseline regressions along with an additional control, $distance_{ij}$, which is the distance between the two addresses as the crow flies.[30] Estimates from this specification are presented in Table 6. Column (1) estimates the specification on the full sample, whereas columns (2) and (3) use the boundary and the timing sample, respectively. As before, all regressions include the fixed effects for quartiles of population density, household income, and per-capita income fixed effects as well as a set of 10 controls, including a county's population, unemployment, population density, and the percent of males.

As expected, the distance between address pairs is positively correlated with a higher error score. More interestingly, the Log(Error-Score) is significantly higher in treatment counties than in control counties. Treatment counties seem to have an error score that is 12.6% higher than that of control counties for the full sample, and this number increases to 14.9% and 29.7% for the boundary and timing samples, respectively. In other words, the distance a person would drive using OpenStreetMap instead of Google Maps is different by 12–30% in treatment counties as compared with control counties.[31] This result is striking because treatment counties were mechanically provided with a higher level of quality when the improved TIGER information was seeded within OpenStreetMap. However, it seems like control counties go on to achieve significantly lower error scores in the long run. This evidence is consistent with the idea that treatment counties possess less contributor activity and lower levels of follow-on information, which in turn leads to missing information and routes that are either too long or too short on OpenStreetMap as compared with Google Maps.

Although I use the term "quality" here to indicate the long-run amount of information, one might have an alternate conception of quality. Specifically, quality could be conceived as the accuracy of the mapping information, conditional on information being present. In Online Appendix B, I present a test of this idea by comparing restaurant names on OpenStreetMap with a third-party, verified list of restaurant names and assess information quality in terms of the similarity of names across the two databases. Even when this alternate measure of accuracy is considered, I find that seeding lowers the quality of information in treated OpenStreetMap counties, although the magnitude of this difference is smaller. See Online Appendix B for more details.

Although the results here show that seeding lowered the long-term quality of OpenStreetMap in terms of routing direction (and the accuracy of restaurant names), it is important to note that these results might

Table 6. Impact of Information Seeding on Long-Term Quality

Coefficients	Log(*Error-Score*)	Log(*Error-Score*)	Log(*Error-Score*)
Treat	0.126*	0.149*	0.297***
	(0.0708)	(0.0800)	(0.0915)
Distance	0.0340***	0.0339***	0.0376***
	(0.00262)	(0.00292)	(0.00320)
Controls	Yes	Yes	Yes
Quartile fixed effects	Yes	Yes	Yes
Sample	Full	Boundary	Timing
N	81,007	58,715	49,095

Notes. The regression is estimated using a cross-sectional specification at the county-address-pair level with a similar specification as the baseline cross-sectional specification, with one additional control $distance_{ij}$, the geodesic (as the crow flies) distance between the two addresses in the address pair j. The main dependent variable $Ln(Error - Score_{ij})$ is a measure of the quality of the route between the address pairs according to the OpenStreetMap database (as compared with Google Maps). All specifications are estimated using linear models.

not generalize to other measures of completeness. As I will explore in the next section, more distant types of follow-on information, such as buildings and parks, were less likely to be present in control places, suggesting that the map was less complete on many dimensions, and potentially of lower visual quality. Routing quality is quite firmly tied to street geometry and related follow-on information, but is not directly related to broader or alternate measures of quality that might be of interest, depending on the particular use case at hand.

3.6. Heterogeneous Effects: When Might Information Seeding Improve Follow-on Contributions

Finally, although the evidence so far suggests an overall negative effect of information seeding, is seeding harmful for all regions and contributors within OpenStreetMap? I explore heterogeneity in the main results in this section.

First, counties differ dramatically in terms of their population density. Most counties in the United States are rural, with a low population density, but there are a small number of counties that are in large metropolitan and urban areas. Urban areas are likely to have a richer set of information to add beyond the basic data provided by the TIGER project. As I show in Table 7, distant contributions tend to increase as a result of seeding, and urban areas offer greater scope for distant contributions, such as restaurants and parks. Accordingly, I estimate the effects of seeding on follow-on contributions and contributor activity separately for counties based on the decile of their rank in the population density distribution. Figure 6, panel (a), plots these estimates with the bottom 10 percentile counties to the left, and the most dense counties to the right. As is clear from this chart, the negative effect of information seeding is seen for the bottom 80 percentile of counties. However, for the densest counties, in the top 20 percentile of the population density distribution, the effects of information seeding are significantly more positive. For example, counties in the top 10 percentile of the population density distribution double the number of follow-on contributions and increase the number of contributors by about 60% in treated counties as compared with control counties. This is a remarkable reversal and points to the potential benefits of information seeding where information seeding leaves significant room for follow-on activity.

Next, I examine the effects of seeding on new contributors with differing levels of commitment to the platform. Imagine two contributors, a novice A, who is not sure about contributing to the platform and will never become a heavy contributor, and an expert B, who is steeped in the open source philosophy and is eager to join OpenStreetMap. Which contributor is more deterred by the seeding effort? To answer this question, I measure the number of new contributors in a given county-quarter who will go on to meet a minimum number of contributions measured in terms of their percentile rank. Figure 6, panel (b), presents these results. As is clear from this chart, the negative effect of seeding on novices like contributor A is much larger as compared with experts like contributor B. In fact, only those who will go on to be in the top one percentile of contributors are unaffected by the seeding effort, whereas the rest are deterred from contributing in places with a high level of seeding.

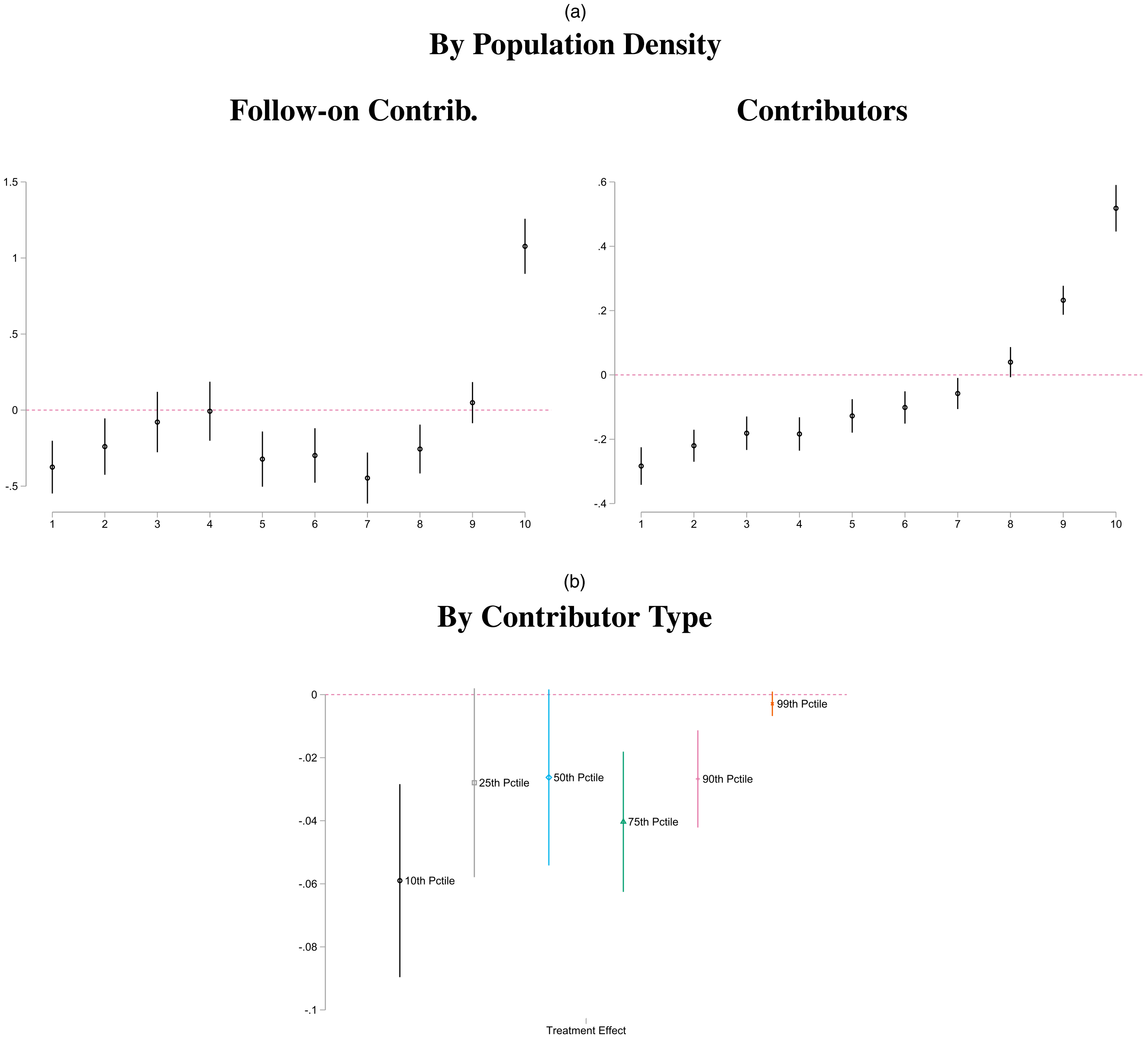
The results from Figure 6 make clear that though seeding does have some important negative impacts

Table 7. Testing the Ownership Mechanism

	Follow-on-Contributions		*Contributors*		*Ownership-Level*		*Owner-Contributions*
	Distant	*Incremental*	*Old*	*New*	*High*	*Low*	—
Post × *Treat*	0.131*** (0.0391)	−0.227*** (0.0497)	−0.00281 (0.00280)	−0.0586*** (0.0127)	−0.00847 (0.00574)	−0.0357** (0.0173)	−0.161*** (0.0369)
County fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quarter fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	121,173	121,173	121,173	121,173	121,173	121,173	121,173

Notes. *Distant Follow-on Contributions* are those that do not modify the street information related to the TIGER project, while *Incremental Follow-on Contributions* are ones that do. *New Contributors* are those who are making an edit in a given county for the first time, whereas *Old Contributors* are those who have made at least one edit before the information seeding took place. The *Ownership-Level* section categorizes contributions from old contributors into two groups: those with a high sense of ownership and those without. Those with a high sense of ownership make most of their edits in a small concentrated area, while those without make diffuse edits. Finally, *Owner-Contributions* measures the total number of follow-on contributions where the owner of an object makes a follow-on contribution. In all columns, the specification is similar to the baseline specification and is estimated using Log-OLS models.

Figure 6. Heterogeneous Effects of Seeding on OpenStreetMap


(a)

By Population Density

Follow-on Contrib.

Contributors

(b)

By Contributor Type

Notes. Panel (a) looks at the effect of seeding by counties divided by the decile of their population density. Counties on the lower end on the left are least densely (rural) populated whereas the most densely populated (urban) counties are on the right. The effect of seeding is estimated for both follow-on contributions and active contributors. Panel (b) looks at the effect of the seeding intervention on different types of new contributors, classified by the percentile of their total lifetime contributions in a given county, from the 10th percentile on the left, to the 99th percentile on the right.

on OpenStreetMap, there are important cases when these effects are muted (high-commitment contributors) or even reversed (high-density, urban counties). These estimates might be instructive for those wishing to deploy seeding in a more targeted fashion.

4. Why Did Information Seeding Lower Follow-on Contributions?

4.1. Theoretical Mechanism: Ownership

Imagine that the higher amount of information seeding increases the level of information in treatment counties as compared with control counties. As a stylized example, after seeding, treatment counties might be at a level 20, whereas control counties might be at a level 10. The question is which type of county is more likely to reach level 30 with a greater number of follow-on contributions, which are contributions between level 20 and level 30? What we have found so far is that, in the long run, control counties are more likely to reach level 30 than treatment counties, even though contributors must first make the basic contributions (from level 10 to 20) before they can start making follow-on contributions. This finding poses a puzzle to standard models where availability of baseline information should encourage follow-on contributions, either because it might make the overall platform more attractive (Athey and Ellison 2014, Zhu et al. 2019) or it lowers the cost of the marginal contribution (Aaltonen and Seiler 2015). Yet, save for the exceptions pointed out in Section 3.6, the finding that a higher degree of information seeding might decrease, rather than increase, follow-on contributions needs further investigation. This section explores an alternate theoretical mechanism, ownership, that might help resolve this puzzle.

To uncover the mechanism that might underlie the puzzling findings, I rely on qualitative data obtained from attending a number of conferences and OpenStreetMap events and my participation in the community myself, making over 150 contributions. Working with a research assistant, I also conducted a number of interviews with OpenStreetMap participants and analyzed online discussions. From these observations and qualitative data, it became clear that contributors are motivated by a sense of attachment to the local area in which they contribute and develop a sense of ownership over it. For example, when asked about what he is most proud of, user Julien Minet says, “I am proud of ‘my area’, roughly described as the Forest of Anlier and Rulles, where I made most of my contributions.”[32] and then goes on to explain how he keeps this area up-to-date to reflect changes in the real world, saying, “I am especially happy with the result, as the official IGN maps of the forests are not always up-to-date. Some paths can disappear rapidly under the vegetation and new ones are created by the exploitation of the forest and by mountain bikers.” The fact that many users have a “my area” on a map was pretty common in our observations. Users with such ownership motivations were much more likely to return to objects and make follow-on contributions. For example, user Lewis Pusey mentions how he has “done extensive reworking of ‘my area’ of the Upper Valley of the Connecticut River on the New Hampshire - Vermont Border.”[33] Another user, Petphi, reinforces this point: “now that its been a few years . . . a quick review in OSM . . . makes me realise that I have to re-edit some of the initial tracks that I drew with a single trace, now that I have better data.”[34] It is important to note that unlike other open source projects where contributors like Minet or Pusey might be awarded formal authority and special editing privileges, in OpenStreetMap, such contributors do not receive special editing privileges.

Building on these observations and qualitative data, I argue that contributors develop ownership over the product of their digital labor when they make basic contributions and this force motivates them to stay engaged and make follow-on contributions. In other words, if they contributed the original knowledge in the first place, people experience a desire to maintain and improve the information that they were responsible for providing. According to this theory, information seeding provides the baseline information at a lower cost, but crowds out the ability of the contributor to establish ownership over the piece of knowledge and thereby demotivates follow-on contributions. To further clarify this theory, Online Appendix C provides a simple model of how such a process might work. In this model, the map is simply a representation of the objects in the real world (Nagaraj and Stern 2019). A certain percent of objects are seeded whereas the rest need to be filled in. Follow-on information must be added for all objects. There are benefits and user-specific costs to making both basic and follow-on contributions. Contribution costs are constant for both types. However, the benefits from follow-on contributions are higher if the user made the basic contribution that underlies it. This critical assumption is an operationalization of the ownership effect. Using a simple example and simulation, I show how such a model could lead to more follow-on contributions in an area, even with a lower level of seeding.

This ownership theory is also supported by past work. In another example, Wikipedians often consider themselves to be parents of certain pages they have contributed to (Nagaraj et al. 2009). For example, customers who design their own products (such as t-shirts) on websites demonstrate significantly higher willingness to pay for such customized products (Franke et al. 2010). Similarly, Norton et al. (2012, p. 453) argue that labor alone can increase individuals to “overvalue their . . . creations,” a phenomenon they term the IKEA effect. As a consequence of this labor of love, even poor individuals can develop “work-product attachment” (Ranganathan 2018) over the output of their labor and make personal sacrifices for it. My qualitative observations suggest that a similar effect is likely to be in play in this setting. Seeding limits the ability of individuals to contribute their digital labor, crowding out their ability to develop such attachment and ultimately shaping their contribution activity on OpenStreetMap.

4.2. Predictions

Building on the qualitative findings, I develop a set of predictions that follow logically from the theory and test these predictions using observational data. Though imperfect, these represent useful tests to identify the ownership effect using observational data.

First, if information seeding from the TIGER project crowded out follow-on contributions, one should expect this effect to be the most concentrated for follow-on contributions that directly modify the TIGER information. If there are follow-on contributions that are not closely related to the baseline information, I expect contributors to treat this information as a novel contribution, thereby negating the detrimental impact of information seeding on follow-on contributions. This logic follows from the theory of cumulative innovation (Scotchmer 1991), which argues that more distant steps (those recombining previously less-known ideas) are perceived to be more novel (Fleming 2001). In this setting, information seeding is more likely to disincentivize a contributor from adding incremental contributions to information that first came from the TIGER map (such as the speed limits to a road that TIGER provided), but is more likely to be motivated to add more distant information (such as a park or restaurant). Accordingly, I predict the following.

Prediction 1. The negative effect of information seeding should be stronger for incremental contributions rather than for distant contributions.

Second, I can exploit information on the timing of when contributors entered the OpenStreetMap community to further analyze the ownership mechanism. Specifically, if contributors had begun making edits before the TIGER maps were uploaded, I expect them to have had a greater opportunity to develop ownership in both treatment and control counties, whereas contributors who began editing after the TIGER experiment have less opportunity to develop ownership. Accordingly, I make my second prediction.

Prediction 2. The negative effect of information seeding should be larger for new contributors than for old (i.e., pre-2008) contributors.

The next prediction focuses on the older contributors (i.e., those active before seeding) and splits them into those that demonstrated a high level of ownership as compared with those who did not. Among the older contributors, information seeding should not affect those who have already developed some ownership over their knowledge, and should hurt those who have not. For example, older contributors who repeatedly make edits in a small region over time, seemingly marking their territory, could be designated to have a high level of ownership. Such contributors should be less affected by the seeding. This is distinct from the heterogeneity around committed contributors explored in Section 3.6, which focused on the longevity of new contributors entering the platform. Instead, this prediction focuses on the contribution activity of existing users. Accordingly, I make the following prediction.

Prediction 3. The negative effect of information seeding should be larger for older contributors with a low level of pre-existing ownership as compared with those with a high level of ownership.

Finally, one can directly measure the number of times a contributor makes a follow-on contribution to an object that the contributor created from scratch, indicating a sense of ownership over it. Examples of such ownership contributions could be when the user Petphi, introduced in the earlier quote, later modified a street that he originally created. I count the number of ownership contributions at the county-quarter level. This variable measures the total number of contributions in a given county-quarter where a contributor is modifying an object that individual created. Such contributions should be higher in control counties than in treatment counties, providing perhaps the most direct test of the theory. Accordingly, I make the following prediction.

Prediction 4. Information seeding should lead to a lower number of ownership contributions in treatment counties as compared with control counties.

4.3. Empirical Estimates

Table 7 presents regression analysis testing the theoretical predictions. For brevity, I use the panel specification as before, replacing the dependent variable with the outcome relevant to each prediction. The first set of results examines the differential effects of the TIGER experiment on distant and incremental follow-on contributions, the second on old and new contributors, the third on old users with a high and low level of ownership, and finally the fourth, on the number of ownership contributions itself. These variables are defined in Section 2.4.1.

All four sets of predictions stemming from the ownership hypothesis seem to be validated according to the estimates presented in Table 7. The effect of the TIGER experiment on incremental follow-on contributions (i.e., those that are closely related to street-level information) is strongly negative, whereas the effect on distant modifications (such as new amenities, restaurants, etc.) seems to be positive and significant. This result validates Prediction 1. In other words, not only does a higher level of information seeding not discourage distant contributions, it seems to encourage them. In terms of magnitude, it seems that information seeding decreases follow-on incremental contributions by 22.7%, whereas distant follow-on contributions and modifications appear to increase by about 13.1%.

Next, Prediction 2 is validated by the tests that evaluate the differential impact of information seeding on old and new contributors. As predicted, most of the negative effects of the TIGER experiment seem to be concentrated among new contributors, who are yet to develop ownership over knowledge, whereas contributors who were active before the TIGER information was included appear to be less affected. Perhaps more interestingly, even when considering old users, I split the effects by those who demonstrate a high level of ownership (i.e., make a majority of their edits in a concentrated area) compared with those without. As predicted, information seeding does not crowd out contributions from those with a high level of ownership, but it is the low-ownership group that seems to reduce contribution activity.

Finally, the final regression looks at the number of times contributors make a follow-on contribution on an object they created. These owner contributions do drop significantly in response to information seeding, which provides perhaps the most direct evidence of the hypothesized mechanism.

4.4. Object-Level Analysis

The ownership analysis so far has proceeded by summing up the total number of contributions at the county-quarter level across treatment and control counties. For example, if four streets were modified by their owner once in a treatment county, and six streets were modified eight times in a control county, we would say that the control county received more ownership edits. Since treatment and control counties are quite comparable, and since we also control explicitly for variables such as county size and population, we expect that we are comparing the number of ownership edits against a similar base of total objects. However, rather than summing up such ownership contributions at the county-quarter level, an alternate approach would focus on the objects themselves and compare their likelihood of receiving follow-on contributions across treatment and control counties. In other words, one could compare streets and related objects created in treatment and control counties and examine their likelihood of receiving follow-on ownership contributions. This approach provides perhaps an even more direct examination of the ownership channel.

Table D.6 in the online appendix provides an alternate cross-sectional analysis along these lines. Since object-level data are significantly larger, I focus on the state of Florida, given its relative balance between treatment and control counties. There are 85,292 objects created from scratch in Florida across treatment and control counties. For each object, I measure the total number of follow-on edits and the total number of ownership edits, which are a subset of these follow-on contributions where the contributor is the owner of the object. The key question is then: Do objects in control counties receive fewer follow-on contributions and, more importantly, fewer ownership contributions?

As before, I regress these two outcomes on the treatment dummy along with a host of controls, using the same specification as the main cross-sectional regressions. Results are presented in Table D.6 in the online appendix. I find that objects in treatment counties are less likely to see follow-on contributions by about 11 to 32 percentage points, confirming the baseline result. More strikingly, ownership edits contribute 3.8–8.5% to the drop in follow-on contributions. The relatively large magnitude of the drop in ownership contributions (four percentage points) as compared with the total follow-on contributions (11 percentage points) suggests an important role for the ownership theory in driving the overall effects of the seeding experiment.

4.5. Alternate Mechanisms

Overall, the empirical results provide support for the ownership channel as a potential mechanism linking a high level of seeding with lower follow-on contributions. Note that this evidence should be seen as tentative given that I do not measure ownership directly at the contributor level. Further, it is not my intention to claim that this is the only mechanism through which the effects of information seeding play out. In particular, it is possible that the lack of information galvanizes groups of contributors to create offline and online governance structures that are known to be related to the health of online communities (Nagaraj and Piezunka 2017). These interactions could create network effects that would attract more members to the community and establish a virtuous cycle (Zhang and Zhu 2011). Although plausible, Table D.7 in the online appendix tests this idea and finds that seeding does not affect the formation of regional meeting groups, one measure of governance in this setting. However, other measures of strong governance might show different results and are worth investigating. Further, recognition or collaboration effects are also possible, That is, a contributor who adds basic information is more likely to attract others to add follow-on information who recognize the contributor's efforts or want to create a community around the contributor. As shown in Table 7 and Table D.6, seeding affects the owner's edits directly, even after excluding follow-on contributions from other members. This validates the ownership channel. However, such recognition or collaboration effects are theoretically valid channels and require future investigation.

Finally, although the heterogeneity results presented in Section 3.6 were largely exploratory, what do they imply about the ownership mechanism? These results showed that the effects are largely driven by less densely populated counties and for novice rather than expert users. Rural areas do not change much and have a lower scope for adding new information. Seeding is therefore likely to crowd out ownership more effectively in such places. Further, expert users, that is, new users who will go on to make a large number of contributions, are less influenced by the ownership mechanism, possibly because they are already quite motivated. This mechanism is most relevant for ex ante casual contributors, who might later become more dedicated given the chance to develop ownership over map elements. The heterogeneity results therefore seem aligned with the ownership mechanism. However, compared with the hypotheses and the detailed exploration presented earlier, they do not offer a strong test of the ownership mechanism, and it is possible that these patterns are a result of parallel and complementary mechanisms as well.

5. Conclusion

This study investigates the role of information seeding in shaping the long-term development of communities. The main findings is that a higher level of information seeding might be counterproductive to the goal of encouraging follow-on contributions, contributor activity, and project quality. This might be because seeding crowds out the ability of contributors to create objects from scratch and develop ownership over them, a mechanism that needs further investigation. Further, seeding is not always harmful. It encouraged follow-on contributions in dense, urban areas and did not discourage motivated heavy contributors.

These results provide the first empirical evidence that speak to theoretical arguments for and against the importance of initial conditions in shaping the long-run dynamics of communities (Lerner and Tirole 2002, Athey and Ellison 2014). Echoing some results from Boudreau and Lakhani (2014) in the context of online contests, our results suggest caution in the broad application of information seeding to encourage community development. Although past evidence clearly suggests that content is often useful to attract contributors (Aaltonen and Seiler 2015, Kane and Ransbotham 2016), it does seem that there might be diminishing, even negative, returns to a high level of information seeding in online communities. Practically, managers and communities looking to design information seeding interventions in online communities might be advised to use it in moderation and in a targeted fashion. For example, seeding might be appropriate when tasks offer plenty of scope for creativity and ownership and for motivated contributors. For OpenStreetMap contributors interested in the value of imports, this work suggests that they might be useful for encouraging distant contributions and in urban areas, but might demotivate incremental contributions such as road tags.

Despite the contributions of this work, the external validity of these results to a broader set of online communities must be considered. In particular, open source software communities provide numerous examples of projects, such as Mozilla Firefox or Eclipse, that were seeded as complete packages but have still thrived. Do results from my context, which is largely about information provision translate to these software projects, which could be seen to be more problem-solving oriented?[35] To explain the generalizability of our findings to this domain, in Online Appendix A, we present short case studies of two projects, Hadoop and Tensorflow, which seem to have thrived despite being seeded by companies (Yahoo and Google, respectively) from the point of view of seeding and follow-on contributions. We have two broad findings. First, there seems to be variation in the extent to which these projects were seeded. Tensorflow was released at a less mature stage, and this lower level of information seeding does seem to be correlated with a greater number of follow-on edits and external contributors. Second, we also discover a number of different motivations that help these communities to flourish, even when opportunities for code ownership are muted. We provide a typology of four such prominent motivations, including the desire to obtain a job at the firm sponsoring the software project. These more diverse sets of motivations offer an opportunity for future researchers to build on the results of this paper in a more problem-solving context such as open source software. Online Appendix B provides a more detailed discussion of all these points. Further, the idea of a plausible promise (Raymond 1999) might differ across information-provision and problem-solving contexts. For OpenStreetMap, even though the map lacks important information, the visible rough street network might be sufficient to establish credibility, whereas for software projects, one might need to provide code that compiles and executes basic functions. This lower bar for information-provision-type projects might make it easier for seeding to crowd out follow-on contributions.

In addition to the challenge of extrapolating the results to open source projects, another limitation must also be acknowledged. The TIGER experiment helps us to compare a moderate and a high level of information seeding. It is possible that if I constructed an experiment in which some counties were not seeded at all, some were seeded with moderate information, and some with high information, more insight could be gained. Although the contribution of this work is to examine the limits of high levels of information seeding, it is up to future research to evaluate the optimal level of information seeding. Finally, although we exploit the TIGER experiment for variation in the levels of completeness between treatment and control counties, these counties could also have differed along measures of accuracy. Although we believe this explanation could have some (albeit limited) merit, we do not separate the effects of accuracy (such as information being out of date) from those of completeness in our research. It would be interesting to examine the counterintuitive prediction that intentionally introducing errors in seeded information could spur follow-on contributions.

In conclusion, this paper contributes to our understanding of the role of early-stage design factors in promoting the long-term health and success of online communities. Future research could further elaborate on the conditions under which information seeding encourages or discourages different aspects of community development. For example, research could investigate other early-stage interventions, such as the seeding of specific contributors, the role of different leadership styles, and the role of socialization initiatives, such as welcome messages and onboarding (Narayan et al. 2017). Finally, online communities are increasingly seeing an increase in the use of bot or automated agents that add new information, similar to the script that added TIGER information. How contributors react to bots is also an exciting question deserving future study.

Acknowledgments

The author thanks Scott Stern, Heidi Williams, Pierre Azoulay, and Catherine Tucker for their support; and Aruna Ranganathan, Sameer Srivastava, Mathijs de Vaan, Ned Augenblick, Shane Greenstein, Lisa McKay, and seminar participants at the Massachusetts Institute of Technology (MIT) Economic Sociology Working Group, MIT Applied Economics Seminar, Berkeley Haas Macro Lunch, and the OpenStreetMap State of the Map Conference 2018 for comments. Several individuals from the OpenStreetMap community and John S. Liadis from the U.S. Census Bureau also provided invaluable guidance. Melissa Staha, Jiaqi Meng, and Aruna Sankaranarayan provided superb research assistance. All errors are the author's.

Endnotes

[1] See Shaw and Hill (2014) for one exception.

[2] See https://www.gsb.stanford.edu/insights/wikipedias-army-volunteer-editors-content-begets-content (accessed February 20, 2020).

[3] See https://en.wikipedia.org/wiki/Wikipedia:History_of_Wikipedia_bots (accessed February 20, 2020).

[4] See http://venturebeat.com/2012/06/22/reddit-fake-users/ (accessed February 20, 2020).

[5] Although these two types of information could presumably have different effects, we are unable to measure precisely to what extent the seeded information was out of date compared with incomplete.

[6] See http://www.openstreetmap.org/stats/data_stats.html (accessed March 2018).

[7] Wikipedia Executive Director Katherine Maher's keynote address at OpenStreetMap's State of the Map conference 2016, https://www.youtube.com/watch?v=ywGuz1586M0 (accessed February 20, 2020).

[8] Other related work focuses on the international dimension (Nagaraj and Piezunka 2017).

[9] See https://www.axios.com/startup-mapbox-is-helping-power-snapchats-new-map-feature-2445906143.html (accessed February 20, 2020).

[10] See https://blog.openstreetmap.org/2016/12/30/tips-pokemon-go/ (accessed February 20, 2020).

[11] Depending on the county, Harris Corporation would use high-quality third-party data, including satellite imagery and ground surveys to update the counties roughly in the order in which they received them (Krmenec 2005).

[12] These updated data were never scheduled to be included within OpenStreetMap, the 2006 version was meant to be a one-time thing.

[13] Dave Hansen, in an interview with Steve Coast, June 20, 2009, available at https://blog.openstreetmap.org/2009/06/20/podcast-dave-hansen/ (accessed February 20, 2020).

[14] See https://lists.openstreetmap.org/pipermail/talk/2006-June/004443.html (accessed February 20, 2020).

[15] Talk-us mailing list, February 23, 2008, https://lists.openstreetmap.org/pipermail/talk-us/2008-February/000043.html (accessed February 20, 2020).

[16] See https://wiki.openstreetmap.org/wiki/TIGER (accessed February 20, 2020).

[17] See https://wiki.openstreetmap.org/wiki/Planet.osm/full (accessed February 20, 2020).

[18] For objects such as street segments or buildings composed of more than one point, I infer the county based on the location of the centroid.

[19] Note that for the purposes of this analysis, the same username making edits in more than one county would be counted more than once.

[20] Note that these categories are not exhaustive; there might be some information unrelated to either buildings or streets that is not classified as either distant or incremental.

[21] I exclude the small number of users who make edits more diffuse than 0.5 degree since these are likely to be operating in multiple regions.

[22] Thanks to a reviewer for this suggestion.

[23] Accessed at https://i.imgur.com/WuJQLjy.png (last accessed February 20, 2020).

[24] Available at https://openaddresses.io/ (accessed February 20, 2020).

[25] See http://project-osrm.org/ (accessed February 20, 2020).

[26] OSRM is the most widely used routing engine based on OpenStreetMap data, and directions from this service serve as a useful proxy for the quality of directions provided by data from OpenStreetMap. Although automobile routing algorithms also incorporate traffic information, we are interested purely in the differences in the shortest route without consideration for traffic or time taken, as a measure of OpenStreetMap's quality. This is why we focus on differences in the shortest route ignoring considerations around estimated time required.

[27] We thank a referee for this suggestion.

[28] The two sets of estimates are not necessarily comparable given that a one percent increase in *Pct. Seeded* is much smaller change than a one percent increase in seeding contributions.

[29] The one exception is the estimates for contributor activity in the boundary sample, which go to 8% and 20% from 12% and 28% for the 2008–2014 and the 2014 statistic, respectively.

[30] Errors are likely to be mechanically larger for addresses that are further apart from each other and this variable increases the precision of the estimates.

[31] Note that this estimate accounts for the general difference in quality between OpenStreetMap and Google Maps, and isolates the additional impact of the TIGER experiment on the gap in quality between treatment and control counties.

[32] See https://www.openstreetmap.org/user/escada/diary/41779 (accessed February 20, 2020).

[33] See https://www.openstreetmap.org/user/lewis_pusey/diary/1106 (accessed February 20, 2020).

[34] See https://www.openstreetmap.org/user/petphi/diary/20458#comment24694 (accessed February 20, 2020).

[35] Thank you to a referee for this suggestion.

References

Aaltonen A, Seiler S (2015) Cumulative growth in user-generated content production: Evidence from Wikipedia. *Management Sci.* 62(7):2054–2069.

Athey S, Ellison G (2014) Dynamics of open source movements. *J. Econom. Management Strategy* 23(2):294–316.

Belenzon S, Schankerman M (2008) Motivation and sorting in open source software innovation. Working paper, The Fuqua School of Business, Durham, NC.

Boudreau K, Lakhani KR (2014) Cumulative innovation and open disclosure of intermediate results: Evidence from a policy experiment in bioinformatics. Working paper, Northeastern University, Boston.

Boudreau KJ, Lacetera N, Lakhani KR (2011) Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Sci.* 57(5):843–863.

Boudreau KJ, Lakhani KR (2015) "Open" disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. *Res. Policy* 44(1):4–19.

Broome FR, Godwin LS (2003) Partnering for the people. *Photogrammetric Engrg. Remote Sensing* 69(10):1119–1123.

Brynjolfsson E, Oh J (2012) The attention economy: Measuring the value of free digital services on the Internet. Working paper, Sloan School of Management, Cambridge, MA.

Coast S (2015) *The Book of OSM* (CreateSpace, Scotts Valley, CA).

Dahlander L, Piezunka H (2014) Open to suggestions: How organizations elicit suggestions through proactive and reactive attention. *Res. Policy* 43(5):812–827.

Dube A, Lester TW, Reich M (2010) Minimum wage effects across state borders: Estimates using contiguous counties. *Rev. Econom. Statist.* 92(4):945–964.

Faraj S, Jarvenpaa SL, Majchrzak A (2011) Knowledge collaboration in online communities. *Organ. Sci.* 22(5):1224–1239.

Fischer E (2013) What's new in TIGER 2013. *Maps for Developers* (August 23), https://blog.mapbox.com/whats-new-in-tiger-2013-6f225a7e0a17.

Fleming L (2001) Recombinant uncertainty in technological search. *Management Sci.* 47(1):117–132.

Franke N, Schreier M, Kaiser U (2010) The "I designed it myself" effect in mass customization. *Management Sci.* 56(1):125–140.

Franke N, Shah S (2003) How communities support innovative activities: An exploration of assistance and sharing among end-users. *Res. Policy* 32(1):157–178.

Franzoni C, Sauermann H (2014) Crowd science: The organization of scientific research in open collaborative projects. *Res. Policy* 43(1):1–20.

Gallus J (2017) Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia. *Management Sci.* 63(12):3999–4446.

Glott R, Schmidt P, Ghosh R (2010) Wikipedia survey: Overview of results, working paper, UNU-Merit, Maastricht, Netherlands.

Goodchild MF, Li L (2012) Assuring the quality of volunteered geographic information. *Spatial Statist.* 1:110–120.

Greenstein S, Nagle F (2014) Digital dark matter and the economic contribution of Apache. *Res. Policy* 43(4):623–631.

Haklay M, Weber P (2008) OpenStreetMap: User-generated street maps. *IEEE Pervasive Comput.* 7(4):12–18.

Harhoff D, Lakhani KR (2016) *Revolutionizing Innovation: Users, Communities, and Open Innovation* (MIT Press Cambridge, MA).

Harris C (2002) Harris Corporation awarded $200 million contract for U.S. Census Bureau's MAF/TIGER accuracy improvement project. Press release, Harris Corporation (June 25), https://www.harris.com//press-releases/2002/06/harris-corporation-awarded-200-million-contract-for-us-census-bureaus.

Healy K, Schussman A (2003) The ecology of open-source software development. Technical report, University of Arizona, Tucson, AZ.

Hill BM (2013) Almost Wikipedia: Eight early encyclopedia projects and the mechanisms of collective action. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Hinnosaar M, Hinnosaar T, Kummer M, Slivko O (2019) Externalities in knowledge production: Evidence from a randomized field experiment. Working paper, University of Nottingham, Nottingham, UK.

Kane GC, Ransbotham S (2016) Content as community regulator: The recursive relationship between consumption and contribution in open collaboration communities. *Organ. Sci.* 27(5): 1258–1274.

Krmenec G (2005) MAF/TIGER Accuracy Improvement Program (MTAIP). Presentation, Indiana GIS 2005 Conference, March 10, Indiana Geographic Information Council, Indianapolis, IN.

Kummer ME (2013). Spillovers in networks of user generated content: Evidence from 23 natural experiments on Wikipedia. Working paper, University of East Anglia, Norwich, UK.

Liadis, J (2018) Telephone interview, December 4.

Lakhani KR, Wolf RG (2003) Why hackers do what they do: Understanding motivation and effort in free/open source software projects. Working paper, Harvard Business School, Cambridge, MA.

Lerner J, Tirole J (2002) Some simple economics of open source. *J. Indust. Econom.* 50(2):197–234.

Lyons E, Zhang L (2018) Research as leisure: Experimental evidence on voluntary contributions to science. Working paper, School of Global Policy and Strategy, University of California, San Diego, CA.

Marx RW (1986) The TIGER system: Automating the geographic structure of the United States census. *Government Publ. Rev.* 13(2):181–201.

Maurer SM, Scotchmer S (2006) Open source software: The new intellectual property paradigm. Technical report, National Bureau of Economic Research, Cambridge, MA.

Nagaraj A (2014) Fixing Tiger deserts: The progress so far. OpenStreetMap (March 1), http://www.openstreetmap.org/user/dalek2point3/diary/21111.

Nagaraj A (2017) Does copyright affect reuse? Evidence from Google Books and Wikipedia. *Management Sci.* 64(7):3091–3107.

Nagaraj A, Piezunka H (2017) The impact of competition on contributions in online communities: Evidence from digital mapping platforms. Working paper, Haas School of Business, Berkeley, CA.

Nagaraj A, Stern S (2020) The economics of maps. *J. Econom Perspect.* 34(1):196–221.

Nagaraj A, Seetharaman P, Roy R, Dutta A (2009) Do wiki-pages have parents? An article-level inquiry into Wikipedia's inequalities. *Workshop Inform. Tech. Systems* (WITS), 14–15.

Nagle F (2018) Learning by contributing: Gaining competitive advantage through contribution to crowdsourced public goods. *Organ. Sci.* 29(4):569–587.

Narayan S, Orlowitz J, Morgan JT, Hill BM, Shaw AD (2017) The Wikipedia adventure: Field evaluation of an interactive tutorial for new users. *ACM Conf. Comput.-Supported Cooperative Work Soc. Comput.*, 1785–1799.

Neis P, Zielstra D, Zipf A (2011) The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet* 4(1):1–21.

Norton MI, Mochon D, Ariely D (2012) The IKEA effect: When labor leads to love. *J. Consumer Psych.* 22(3):453–460.

Piezunka H, Dahlander L (2015) Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Acad. Management J.* 58(3):856–880.

Ranganathan A (2018) The artisan and his audience: Identification with work and price-setting in a handicraft cluster in Southern India. *Admin. Sci. Quart.* 63(3):637–667.

Ratcliffe, M (2014) Telephone Interview, March 11, U.S. Census Bureau.

Raymond E (1999) The cathedral and the bazaar. *Knowledge Tech. Policy* 12(3):23–49.

Resnick P, Konstan J, Chen Y, Kraut RE (2011) Starting new online communities. Kraut RE, Resnick P, Kiesler S, eds. *Building Successful Online Communities: Evidence-Based Social Design* (MIT Press, Cambridge, MA), 231–280.

Scotchmer S (1991) Standing on the shoulders of giants: Cumulative research and the patent law. *J. Econom. Perspect.* 5(1):29–41.

Shah SK (2006) Motivation, governance, and the viability of hybrid forms in open source software development. *Management Sci.* 52(7):1000–1014.

Shaw A, Hill BM (2014) Laboratories of oligarchy? How the iron law extends to peer production. *J. Comm.* 64(2):215–238.

Von Hippel E (2005) Democratizing innovation: The evolving phenomenon of user innovation. *J. Betriebswirtschaft* 55(1): 63–78.

Zandbergen PA, Ignizio DA, Lenzer KE (2011) Positional accuracy of TIGER 2000 and 2009 road networks. *Trans. GIS* 15(4): 495–519.

Zhang X, Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *Amer. Econom. Rev.* 101(4):1601–1615.

Zhu K, Walker D, Muchnik L (2019) Content growth and attention contagion in information networks: A natural experiment on Wikipedia. Working paper, Boston University, Boston.

Zielstra D, Hochmair HH, Neis P (2013) Assessing the effect of data imports on the completeness of OpenStreetMap: A United States case study. *Trans. GIS* 17(3):315–334.