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

Abhishek Nagaraj*

August 19, 2019

Abstract

The wild success of a few online communities (like 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 over 350 million contributions made by over 577,000 contributors to OpenStreetMap, a crowd-sourced map-making community seeded with data from the US Census. I estimate the effect of seeding using a natural experiment in which an oversight caused about 60% of US counties to be seeded with a complete Census map, while the rest were seeded with less complete versions. While 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.

*UC Berkeley-Haas. Email: nagaraj@berkeley.edu.
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 (Greenstein and Nagle, 2014; Brynjolfsson and Oh, 2012; 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 (Shah, 2006; Lakhani and Wolf, 2003; Lerner and Tirole, 2002).\footnote{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.”\footnote{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 (Kane and Ransbotham (2016); Boudreau and Lakhani (2015)). Wikipedia, for instance, was initially seeded with short articles on more than 30,000 US cities from the US Census Bureau\footnote{https://en.wikipedia.org/wiki/Wikipedia:History_of_Wikipedia_bots} 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.\footnote{http://venturebeat.com/2012/06/22/reddit-fake-users/}}

Despite theoretical arguments for the value of information seeding, we understand little about the extent to which information seeding fosters the growth of online communities in practice. First, as some theoretical models claim, it is possible that initial conditions “have little to do with the long-run success of the project,”

---

\footnote{See Shaw and Hill (2014) for one exception.}
affecting only the path to the eventual equilibrium (Athey and Ellison (2014), p. 296). Those projects destined to succeed (perhaps because they solve a real need), do succeed, while others fail and initial conditions do not matter. More interestingly, the theoretical arguments for the positive impact of information access on follow-on knowledge production might not generalize to all types and stages of online communities. Even Lerner and Tirole (2002) argue that, 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 non-pecuniary motivations driving knowledge production in these settings (Lakhani and Wolf, 2003; Belenzon and Schankerman, 2008; Franke and Shah, 2003; Shah, 2006). In particular, as I will argue, when communities are geared towards 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 while higher-levels of information seeding might crowd out these incentives. There might be therefore 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 US, the fledgling community decided to bootstrap their efforts by seeding its map with the US Census “TIGER”5 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, i.e. 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 US 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. In this way, about 60% of US counties received a higher level of information seeding than the rest. While 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 updation 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.

5Topologically Integrated Geographic Encoding and Referencing
I leverage micro-data on more than 350 million contributions to OpenStreetMap in the US 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 non-parametric 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 whole 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 updation 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 to 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 ten 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.

While a few different mechanisms could be driving these patterns, 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. While 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 will not be negatively affected by the seeding effort as compared to 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 baseline 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; Hinnosaar et al., 2019; Zhu et al., 2018). Aaltonen and Seiler (2015) argues that “content begets content”: that the provision of information encourages follow-on contributions, a finding backed up by Zhu et al. (2018). 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. While 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, while 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 these contexts, 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 (Boudreau et al., 2011; Von Hippel, 2005; Lerner and Tirole, 2002; Faraj et al., 2011; Dahlander and Piezunka, 2014). While theoretical work differs about the relative importance of initial conditions (Athey and Ellison, 2014; Lerner and Tirole, 2002; Raymond, 1999) 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, 2016), 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 (Nagle, 2018; Franke and Shah, 2003; Shah, 2006; Lakhani and Wolf, 2003).

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 over 5.3 million registered contributors 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. Although OpenStreetMap has global coverage, I will concentrate on the OpenStreetMap project in the United States. OpenStreetMap is different from other commercial providers in that the mapping data is 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, Apple Maps (Coast, 2015) as well as in self-driving cars and mobile games.

A potential contributor must register and then make a contribution to OpenStreetMap using 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 etc. 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 to contributors. After a contribution has been saved, the username of the contributor is recorded allowing her 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), which helps the analyst track contributions and contributor activity in the map over time.

2.2 The TIGER Experiment

A. The US 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 US Census Topologically Integrated Geographic Encoding and Referencing (TIGER) Map into their system. TIGER is a computer-readable map

---

6http://www.openstreetmap.org/stats/data_stats.html, accessed March 2018
7Wikipedia Executive Director, Katherine Maher’s keynote address at OpenStreetMap’s State of the Map conference 2016, https://www.youtube.com/watch?v=ywGuz1586M0
8Other related work focuses on the international dimension (Nagaraj and Piezunka, 2017).
10https://blog.openstreetmap.org/2016/12/30/tips-pokemon-go/
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.

While 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 US Census prioritized “topographical integrity” rather than absolute completeness as metrics of quality (Zandbergen et al., 2011). Further, these maps were not updated frequently due to which many new neighborhoods and corresponding street information were completely missed. While 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 US Census (Personal Communication, December 2018). The GPPB is responsible for collecting source data (including from the Census’ 12 regional offices all over the country) and concurrently making a determination of the order in this data from specific counties should be updated. Once this order was decided, the GPPB sent the data for each county “southbound” i.e. to Harris corporation where it would be updated and fixed. Once the fixes were made, Harris would send the data back “northbound,” after which they would be incorporated into the TIGER dataset released to the general public after another review by the GPPB. TIGER releases for the public were largely issued on an annual basis while northbound updates arrived as they were done. 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, while counties scheduled to be updated in 2006 missed the cut. For counties that had yet to be updated, older, less accurate data was 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, i.e. six years after its launch, that the MTAIP program was completed and all US counties were updated in the TIGER database.\footnote{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 (Krmenc, 2005).} \footnote{These updated data were never scheduled to be included within OpenStreetMap, the 2006 version was meant to be “one time updated.”}
In order 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.

B. 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.” 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, 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. *Dave Hansen, interview by Steve Coast, June 20, 2009.*

Driven by the logic that seeding OpenStreetMap with TIGER maps would fuel its development, OpenStreetMap began the process to incorporate TIGER information from the US 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. Jan-March 2008), and absent before then. Note that TIGER information was incorporated for 3,107 counties within the US; the state of Massachusetts was excluded because better quality information was available from the state government. 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 thing”.

---


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 to community
provided information.

C. Variation in TIGER Seeding: While 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
2.2A) 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 while 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 while TIGER
seemed complete and high quality in some places, it was incomplete in others. For example, contributor,
Matthew P notes:

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. [talk-us mailing list February 23, 2008]16

The fact that only about 60% of the counties in the US 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 while 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. For example, one prolific user and OpenStreetMap board member we interviewed
noticed the differences in TIGER completeness, but did not mention MTAIP, instead saying that:

I very quickly discovered that the quality of the TIGER data varied from county to county ... I
found later on that the census would get data from the counties or you know, from the states.
So it was sort of a compilation of different forces ... at the time, I just kind of took it as a given.

Personal Communication, April 2019

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

This unintentional and lesser known variation in information seeding (see Nagaraj (2014); Fischer (2013) 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.

A. 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 US 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 2004-2007, 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 twelve 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 US. For example, it does not seem to be the case that all of the major urban centers in the US 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 US Census because if the US Census had prioritized all “important” or highly populated areas at the start, there would have been difficulties with project implementation:

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—

\footnote{https://wiki.openstreetmap.org/wiki/TIGER}
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.

B. Quantitative Comparison: While 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 ($\Delta$IncomePerCapita in Table 1) than control counties. While this systematic difference between the two categories might appear problematic, it is useful to note that 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 above, 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 Appendix Figure D.1 that 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 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.

C. Alternate Samples: Finally, while 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 within counties. 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 (Personal Communication, December 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, i.e. 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 (1096 treatment and 724 control).

Second, I was able to obtain proprietary data from the US Census Bureau officials on order in which counties were scheduled to be updated as per the Geographic Program and Planning Branch’s (GPPB) 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, i.e. counties scheduled to be updated in the calendar years 2005 or 2006 to be a part of the sample. This exercise leaves us with 2218 counties (1228 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 non-parametric 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 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 different 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 US 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 American Community Survey for additional controls. This section describes these data in more detail.
A. OpenStreetMap Data: The OpenStreetMap project stores all past versions of the map in the form of a history file. 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 US for map objects relevant to the analysis, including streets, street tags and distant contributions such as a parks, buildings etc. This process creates the source data for this project, which includes almost 839.2 million contributions totaling about 100G of data made by more than 577,000 unique contributors. I then drop all contributions made by a small set of automated scripts and bots. Most importantly, I deleted all edits from the username DaveHansonTiger, which was the unique username created to make the updates from the 2006 US 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 point 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, i.e. between the years 2008-14, 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 different 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), while 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).

Second, I divide the total number of contributors who are active into two groups, New Contributors and Old Contributors. New Contributors are those who are making an edit in a given county for the first time, while 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 to those who make edits over a wider surface area. In practice, if most of

---


19 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.

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

21 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.
the edits of a given contributor are within a box of width 0.1 degree latitude and 0.1 degree longitude, she is classified as having a high-level of ownership, while if most edits are made in a more diffuse fashion outside of a narrow 0.1x0.1 lat/lon area, then the editor is classified as having a low-level of ownership.\footnote{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.}
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.\footnote{Thanks to a reviewer for this suggestion.}

B. 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 US Census had finished correcting the data by the end of 2006 (Fusaro and Avnayim, 2006). 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 US 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.

C. 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 were affected by the TIGER experiment. As noted before, 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 relative difference in length between a route proposed by an OpenStreetMap-based routing program compared to 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 acknowledge of Google Maps’ quality in this regard and following the OpenStreetMap literature (Zielstra et al., 2013; Goodchild and Li, 2012).

I compute the error score as follows. First, using the OpenAddresses database,\footnote{Available on https://openaddresses.io/} I collect the list of all known addresses in 2312 counties (1325 treatment and 987 control) in the US 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)\footnote{http://project-osrm.org/} as well as the Google Maps API to provide me with the shortest possible route between the two
addresses. 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 additional distance one would travel if one used OSRM in place of Google Maps. The key outcome is the average logged error at the county level, \( \log(\text{Error Score}) \) 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 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.

**D. Controls:** Finally, in addition to the dependent and independent variables mentioned above, I also collect information on nine demographic and income variables at the county level to serve as controls. Using the American Community Survey (ACS) 5-year estimates from 2009-2014, 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 IT industry, and number of highly educated (those with a master’s degree, a PhD, or a professional degree).

**E. 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 (2867.80 for contributions, 529.68 for follow-on contributions, and 7.68 for contributors) are significantly larger than the median.

In Appendix Table D.1, 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, while 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

---

26OSRM 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. While 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.
outcomes (Distant/Incremental contributions, New/Old Users, Low/High Ownership edits, Ownership 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 2007q4 (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 till 2009, after which both the stock and the flow of contributions rises dramatically. OpenStreetMap as a platform grew in popularity only starting 2009, in the US, 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 OpenStreetMap had significant traction and a sufficiently large contributor community, i.e. 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 compared to control counties.

To add to this graphical analysis, Table 1 plots 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, while 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 percent 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”, i.e. 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 is 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 to 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

A. Baseline Estimates: For the panel analysis, I estimate regressions of the following form using the county-quarter panel: 

$$Y_{it} = \alpha + \beta_1 \times Post_t \times Treat_i + \gamma_i + \delta_t + X_{it} + \epsilon_{it}$$

where \(\gamma_i\) and \(\delta_t\) 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 and \(Treat_i\) equals one for all counties that benefited from the TIGER improvement program on OpenStreetMap, i.e., they had been significantly improved through the MTAIP project in the 2006 version of the TIGER database. \(X_{it}\) 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 \(\beta_1\) should be positive and significant, but if information seeding has a negative effect on these variables, then the estimate of \(\beta_1\) 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 (Log-OLS) models because of the highly skewed distribution of the dependent variables.

Table 3 presents estimates from this regression for total contributions (cols 1-2), follow-on contributions (cols 3-4) and contributors (cols 5-6). All models include county and quarter fixed-effects, while 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. While the result of a reduction in total contributions by about 47.2% (col 2) is perhaps not surprising, even when considering only follow-on contributions the gap hovers at about 15.2% (Column 4) and about 5.6% (Column 6) when considering contributors. These effects are statistically sig-
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.

**B. 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 completely zero. Accordingly, I now turn to estimating the time-varying impact of the TIGER experiment using the following specification:

$$Y_{it} = \alpha + \sum \beta(Treat)_{i} \times 1(z) + \gamma_i + \delta_t + \epsilon_{it}$$

where $\gamma_i$ and $\delta_t$ 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 $\beta$. This specification is estimated using Log-OLS 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-percent 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 platforms overall levels of popularity discussed in Section 3.1. Specifically, even though seeding happens in late 2007, OpenStreetMap does not gain popularity till 2009-10, which is when the differences between treatment and control counties seem to become more apparent.

### 3.3 Cross-Sectional Estimates

**A. 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 “pre-trend” making it difficult to compare across treatment and control units. Therefore, while 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 variables for which treatment and control counties may differ. In particular, as shown in Figure D.1, 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 non-parametric fashion. Specifically, I estimate the effects of information seeding using the following Log-OLS specification: 

\[ \ln(Y_i + 1) = \alpha + \beta_1 \times \text{Treat}_i + \beta \times X_i + \gamma_1 + \gamma_2 + \gamma_3 + \epsilon_i \]

where \( \gamma_1, \gamma_2, \) and \( \gamma_3 \) correspond to population density, household income, and per-capita income fixed effects respectively and \( X_i \) indicates a set of ten 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 post-seeding, i.e. 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 to 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 2008-14, all follow-on contributions reduce by about 19% (as compared to 15% in panel models) and contributor activity reduces by 12.8% (as compared to 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.

**B. Yearly Graphical Estimates:** Figure 5 presents another way of looking at these results. Here, I estimate \( \beta_1 \) using ten different regressions for each calendar year from 2005 to 2014 separately, and 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 for information seeding 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 pre-trends challenging, the cross-sectional results provide a simpler and perhaps more convincing way of estimating the effect of seeding on contribution activity.

**C. Stock of Knowledge (IV 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? Appendix Table

---

27 We thank a referee for this suggestion.
D.2 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 DaveHansenTiger account which was responsible for the seeding of OpenStreetMap. Both measures, while not perfect, provide two useful ways of conceptualizing the amount of information seeded beyond a simple treatment/control dichotomy across counties.

Table D.2 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-stat. Further, following the main results, the second stage is negative, i.e. 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 percent 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, while 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 for more details.

3.4 Robustness Checks

A. Alternate Samples: I examine whether the baseline specifications remain robust when considering the boundary and timing samples described in Section 2C. Table 5 provides these estimates that 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, while 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.

28The two sets of estimates are not necessarily comparable given that a 1 percent increase in Pct. Seeded is much smaller change than a 1 percent increase in seeding contributions.

29The one exception is the estimates for contributor activity in the boundary sample which goes to 8% and 20% from 12% and 28% for the 2008-14 and the 2014 statistic respectively.
**B. Additional Tests:** Apart from the additional samples, the panel model suggests a few additional robustness tests. Since we have limited pre-trends data on the key outcome variable, more robust controls for differential pre-trends across counties might be appropriate. For example, imagine urban areas (those with a higher population density) are seeing the migration of a highly educated population 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, shown 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 as a result of the seeding experiment, but are driven by differences between the popularity of Google Maps in different regions in the US. Accordingly, I collected data on county-level popularity of Google Maps from the Google Trends database and control for this directly in Appendix Table D.3 (Columns 2 and 3). 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 changesets contributed (rather than individual contributions), and estimate the baseline specification, as shown in Table D.3 (Column 1). The co-efficient 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 \( \log(\text{Error-Score}) \), 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. Estimates from this specification are presented in Table 6. Column 1 estimates the specification on the full sample, while 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 ten 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 12-30% higher in treatment counties. 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 longer routes on OpenStreetMap than on Google Maps.

While 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 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 Appendix B for more details.

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

Finally, while 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.

---

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.
First, counties differ dramatically in terms of their population density. Most counties in the US 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 will 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 to 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 to experts like contributor B. In fact, only those who will go on to be in the top 1 percentile of contributors are unaffected by the seeding effort, while the rest are deterred from contributing in places with a high level of seeding.

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

Normally, a lower cost of information access should increase follow-on knowledge production, either because it might make the overall platform more attractive (Athey and Ellison, 2014; Zhu et al., 2018) and by lowering 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 poses a puzzle to the broader literature on information access as well as work specific to online communities. 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 Antlier and Rulles, where I made most of my contributions.” 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.”. This 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.” 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.”

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 in the first place. 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, 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 while 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.

32https://www.openstreetmap.org/user/escada/diary/41779
33https://www.openstreetmap.org/user/lewis_pusey/diary/1106
34https://www.openstreetmap.org/user/petphi/diary/20458#comment24694
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. For example, Wikipedians often consider themselves to be “parents” of certain pages they have contributed to (Nagaraj et al., 2009). Creative workers in off-line settings can develop “work-product attachment” (Ranganathan, 2017) over the output of their labor and make personal sacrifices for it. More broadly, the idea of greater attachment to information created by oneself than to knowledge from an external source is related to the “endowment effect” in behavioral economics (Kahneman et al., 1991; Knetsch, 1989), which argues that individuals value products more highly after they own them rather than before.

4.2 Predictions

Building on the qualitative findings, I develop a set of predictions which follow logically from the theory and test these predictions using observational data. While 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), while she is more likely to be motivated to add more distant information (such as a park or restaurant). Accordingly, I predict that:

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

Second, I can exploit information on the timing of when contributors entered the OpenStreetMap community to further analyze the ownership mechanism. Specifically, if a contributor 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, while contributors who began editing after the TIGER experiment have less opportunity to develop ownership. Accordingly: The negative effect of information seeding should be larger for new contributors than for old (i.e. pre-2008) contributors. (Prediction 2).

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 to those who didn’t. Among the older
contributors, information seeding should not affect those who have already developed some ownership over their knowledge, and should hurt those who haven’t. 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, while this prediction focuses on the contribution activity of existing users. Accordingly, I argue: The negative effect of information seeding should be larger for older contributors with a low-level of pre-existing ownership as compared to those with a high-level of ownership. (Prediction 3).

Finally, one can directly measure the number of times a contributors makes a follow-on contribution to an object that she created from scratch, indicating a sense of ownership over it. Examples of such “ownership contributions” could be when the user Petphi in the quote above, modified a street that he had originally created when he has more data about it in the future. I am able to trace the history of all contributions at the object-level on OpenStreetMap and therefore could count the number of ownership contributions in a given county-quarter. These should be higher in control counties than in treatment counties, providing perhaps the most direct test of the theory. Accordingly: Information seeding should lead to a lower number of ownership contributions in treatment counties as compared to control counties (Prediction 4).

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 2.4A.

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, while 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 in fact seems to encourage them. In terms of magnitude, it seems that information seeding decreases follow-on incremental contributions by 22.7% while 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, while 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) as compared to 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 that a contributor makes 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 analysis so far measures ownership by tracking the number of improvements to pre-existing objects and summing them over a county-quarter. For example, if four streets were modified once by their owner in a given county-quarter, the number of ownership contributions would be four in the specification testing Prediction 4. A parallel approach would focus on objects themselves (like the street segments in the example before) and ask whether seeding crowds out ownership contributions when comparing objects in treated and control counties. This approach provides perhaps an even more direct examination of the ownership channel.

Accordingly, in Appendix Table D.4, I provide an alternate cross-sectional analysis that relies on an object-level analysis. Since these data are large, I focus on the state of Florida, given its relative balance between treatment and control counties. For the 85,292 objects created from scratch in the state, I measure the total number of follow-on edits and the total number of ownership edits, i.e. when the original contributor makes a follow-on contribution on the same object. 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.4. Objects in treatment counties are less likely to see follow-on contributions by about 11-32 percentage points, and ownership edits are likely to reduce by about 3.9-8.7 percentage points. The relatively large magnitude of the drop on ownership contributions (4 percentage points) as compared to 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 which would attract more members to the
community and establish a virtuous cycle (Zhang and Zhu, 2011). While plausible, appendix Table D.5 tests this idea and finds that seeding does not affect the formation of location 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; i.e. if a contributor adds basic information, she is more likely to attract others to add follow-on information who are recognizing her efforts or want to create a community around her. As show in Table 7 and Table D.4, 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.

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 it, 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 (Athey and Ellison, 2014; Lerner and Tirole, 2002). 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. While past evidence does clearly suggest 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 which 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 In order to explain the generaliz-

35 Thank to you to a referee for this suggestion.
ability of our findings to this domain, in 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 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 set 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. Appendix B provides a more detailed discussion of all of these points.

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 seeded with no information, some with moderate information, and some with high information more insight could be gained. While 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, while 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. While we believe this explanation could have some (albeit limited) merit, we do not separate the effects of accuracy from those of completeness in our research. It would interesting to examine the counter-intuitive 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 of future study.

References


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

Panel B: Example of Differences in Information Seeding between Treatment and Control Counties

Note: This figure provides an overview of the variation in information seeding. Panel A highlights the counties with higher level of information seeding in green solid color, while counties that received a lower amount of information seeding are presented using an orange dotted pattern. Note that counties in the state of Massachusetts have been excluded because they were not seeded with TIGER information. Panel B provides 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.
Figure 2. Alternate Samples

Panel A: Boundary Sample (Excluding Contiguous Treatment/Control Counties)

Panel B: Timing Sample (Excluding Early/Late Updated Counties)

Note: This figure describes the alternate samples used to test the baseline hypothesis. Panel A shows the subset of 1820 counties (of 3107) that are included in boundary sample, i.e. the sample when only TIGER counties that adjoin a control county (or vice versa) are included. Panel B, the timing sample, is based on data on 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 and counties scheduled to be updated before or after this period are excluded. This sample includes 2218 (of 3107) counties.
Figure 3. **Mean Outcomes for Treatment and Control Counties**

**Panel A: Total Contributions**

*i. Logged*  

\[ \begin{array}{c}
0 & 2 & 4 & 6 & 8 \\
\end{array} \]

- Treatment Counties
- Control Counties

**ii. Cumulative**

\[ \begin{array}{c}
0 & 100000 & 200000 & 300000 & 400000 \\
\end{array} \]

- Treatment Counties
- Control Counties

**Panel B: Total Follow-on Contributions**

*i. Logged*  

\[ \begin{array}{c}
0 & 1 & 2 & 3 & 4 \\
\end{array} \]

- Treatment Counties
- Control Counties

**ii. Cumulative**

\[ \begin{array}{c}
0 & 5000 & 10000 & 15000 & 20000 & 25000 \\
\end{array} \]

- Treatment Counties
- Control Counties

**Note:** This figure compares the average number of contributions (Panel A) and follow-on contributions (Panel B) between treatment and control counties per quarter. In both panels, figure (i) represents logged version of the outcome variable, while figure (ii) represents a cumulative sum of contributions up until a given quarter. The vertical line represents the quarter when TIGER data was imported into OpenStreetMap. The blue dashed line represents average values for outcome variables in control counties, while the red solid line represents outcomes in treatment counties.
Figure 4. Time-varying Impacts of the TIGER Experiment on Contributions

Note: This figure plots estimates (and 95 percent confidence intervals) of \( \beta_t \) from the event study specification specified below. On the x axis is calendar year. This figure is based on county-year observations, the coefficients are estimates from Log-OLS models, the sample includes all county-year observations in the sample and the standard errors are clustered at the county level. The specification is \( Y_{it} = \alpha + \sum \beta_T(\text{Treat}_i) \times 1(z) + \gamma_t + \delta_i + \epsilon_{it} \) where \( \gamma_t \) and \( \delta_i \) represent county and year fixed effects respectively for block \( i \) and year \( t \). \( z \) represents the lag, or the number of years after TIGER information was first included in the map.

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

Note: This figure plots estimates (and 95 percent 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 for a host of county-level control variables and fixed effects. I estimate \( Y_i = \alpha + \beta_1 \times \text{Treat}_i + X_i + \gamma^1 + \gamma^2 + \gamma^3 + \epsilon_i \) where \( \gamma^1, \gamma^2 \) and \( \gamma^3 \) represent population density, household income and per-capita income fixed effects respectively and \( X_i \) indicates a set of ten controls, including a county’s population, unemployment, population density, the percent of males etc. Heteroskedasticity robust standard errors are estimated.
Figure 6. **Heterogeneous Effects of Seeding on OpenStreetMap**

**A. By Population Density**

Follow-on Contrib. | Contributors
---|---

![Graph A](image)

**B. By Contributor Type**

![Graph B](image)

**Note:** This figure plots the heterogeneous effects of the seeding intervention on different dimensions of activity on OpenStreetMap. 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 while 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.
Table 1. **Cross-Sectional Comparison**

**County Level (N=3107)**

<table>
<thead>
<tr>
<th>Contribution Outcomes</th>
<th>(1) Treatment $\bar{y}$</th>
<th>(2) Control $\bar{y}$</th>
<th>(3) Diff</th>
<th>(4) p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributions</td>
<td>2375.5</td>
<td>3593.3</td>
<td>-1217.7</td>
<td>0.00</td>
</tr>
<tr>
<td>Follow-on Contribs.</td>
<td>457.9</td>
<td>635.5</td>
<td>-177.6</td>
<td>0.05</td>
</tr>
<tr>
<td>Community Outcomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contributors</td>
<td>7.482</td>
<td>7.966</td>
<td>-0.484</td>
<td>0.26</td>
</tr>
<tr>
<td>Quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error-Score</td>
<td>1830.2</td>
<td>2046.7</td>
<td>-216.5</td>
<td>0.09</td>
</tr>
<tr>
<td>Mechanism</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owner-Contribs</td>
<td>149.6</td>
<td>203.6</td>
<td>-54.08</td>
<td>0.21</td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ Population</td>
<td>7324.3</td>
<td>6756.5</td>
<td>567.8</td>
<td>0.62</td>
</tr>
<tr>
<td>$\Delta$ Households</td>
<td>2186.4</td>
<td>1899.8</td>
<td>286.6</td>
<td>0.42</td>
</tr>
<tr>
<td>$\Delta$ Unemployed Pop.</td>
<td>-646.1</td>
<td>-543.6</td>
<td>-102.5</td>
<td>0.27</td>
</tr>
<tr>
<td>$\Delta$ Educ. Population</td>
<td>1995.9</td>
<td>1884.8</td>
<td>111.0</td>
<td>0.68</td>
</tr>
<tr>
<td>$\Delta$ Male Population(18 – 45)</td>
<td>555.3</td>
<td>410.5</td>
<td>144.8</td>
<td>0.42</td>
</tr>
<tr>
<td>$\Delta$ Income per Capita</td>
<td>3152.7</td>
<td>2773.7</td>
<td>379.0</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*Note:* This table provides summary statistics that help to evaluate the impact of the TIGER seeding experiment on OpenStreetMap outcomes, as well as evaluate the selection of counties into the treatment and control groups. Mean values of the variables are presented in column (1) for 1851 treatment counties and in Column (2) for 1242 Control counties. Column (3) presents estimates for the difference in the mean values while Column (4) presents the p-value for this estimate. For the control variables, $\Delta$ represents the difference in the variable between 2014 and 2005 for a given county.
Table 2. **Summary Statistics**

**County-Quarter Level (N=121,173)**

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Contributions</td>
<td>2867.80</td>
<td>28510.38</td>
<td>69.00</td>
<td>0.00</td>
<td>3625156.00</td>
</tr>
<tr>
<td>Total Follow-on Contrs</td>
<td>529.68</td>
<td>13416.84</td>
<td>11.00</td>
<td>0.00</td>
<td>3625156.00</td>
</tr>
<tr>
<td>Contributors</td>
<td>7.68</td>
<td>46.12</td>
<td>4.00</td>
<td>0.00</td>
<td>8535.00</td>
</tr>
<tr>
<td>Distant Follow-on</td>
<td>120.21</td>
<td>3612.47</td>
<td>0.00</td>
<td>0.00</td>
<td>422200.00</td>
</tr>
<tr>
<td>Incremental Follow-On</td>
<td>366.00</td>
<td>12551.49</td>
<td>2.00</td>
<td>0.00</td>
<td>3625156.00</td>
</tr>
<tr>
<td>New Contributors</td>
<td>4.77</td>
<td>45.36</td>
<td>2.00</td>
<td>0.00</td>
<td>8532.00</td>
</tr>
<tr>
<td>Old Contributors</td>
<td>0.10</td>
<td>1.14</td>
<td>0.00</td>
<td>0.00</td>
<td>293.00</td>
</tr>
<tr>
<td>Low Ownership Contrs</td>
<td>2.00</td>
<td>87.47</td>
<td>0.00</td>
<td>0.00</td>
<td>20595.00</td>
</tr>
<tr>
<td>High Ownership Contrib</td>
<td>123.52</td>
<td>6411.82</td>
<td>0.00</td>
<td>0.00</td>
<td>1268382.00</td>
</tr>
<tr>
<td>Owner-Contribs</td>
<td>171.42</td>
<td>5729.41</td>
<td>1.00</td>
<td>0.00</td>
<td>1695215.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Timing Vars</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>0.60</td>
<td>0.49</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Post</td>
<td>0.72</td>
<td>0.45</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Year</td>
<td>2009.62</td>
<td>2.82</td>
<td>2010.00</td>
<td>2005.00</td>
<td>2014.00</td>
</tr>
<tr>
<td>Quarter</td>
<td>200.00</td>
<td>11.25</td>
<td>200.00</td>
<td>181.00</td>
<td>219.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Select Controls</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>100029.21</td>
<td>320590.11</td>
<td>25905.00</td>
<td>41.00</td>
<td>10230943.00</td>
</tr>
<tr>
<td>Households</td>
<td>37143.93</td>
<td>112520.17</td>
<td>9880.00</td>
<td>22.00</td>
<td>3325103.00</td>
</tr>
<tr>
<td>Pop. Density</td>
<td>265.36</td>
<td>1771.88</td>
<td>45.36</td>
<td>0.06</td>
<td>72839.69</td>
</tr>
<tr>
<td>Per-Capita Income</td>
<td>23765.17</td>
<td>5854.17</td>
<td>22966.00</td>
<td>1601.00</td>
<td>82817.00</td>
</tr>
</tbody>
</table>

*Note:* This table presents summary statistics for the main sample used to estimate the baseline specification. The data are a balanced panel for 3107 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.
### Table 3. The Effects of Information Seeding: Difference-in-Difference Estimates

<table>
<thead>
<tr>
<th></th>
<th>Total-Contrib.</th>
<th>Follow-on-Contrib.</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Post X Treat</td>
<td>-0.471***</td>
<td>-0.472***</td>
<td>-0.141***</td>
</tr>
<tr>
<td></td>
<td>(0.0466)</td>
<td>(0.0403)</td>
<td>(0.0484)</td>
</tr>
<tr>
<td>Quarter FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>County FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>N</td>
<td>121173</td>
<td>121173</td>
<td>121173</td>
</tr>
</tbody>
</table>

*Note:* This table estimates the impact of the TIGER experiment on contribution activity on OpenStreetMap. I use the following specification:  
\[ Y_{it} = \alpha + \beta_1 \times Post_t \times Treat_i + \gamma_i + \delta_t + \epsilon_{it} \]
where \( \gamma_i \) and \( \delta_t \) represent county and quarter fixed effects respectively for county \( i \) in quarter \( t \). \( Post_t \) equals one for all quarters after December 2007 and \( Treat_i \) equals one for all counties for received the higher level of information seeding. All specifications are estimated using Log-OLS models, with one added to the dependent variable for all zero values.

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

<table>
<thead>
<tr>
<th></th>
<th>Total-Contrib.</th>
<th>Follow-on Contrib.</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>-0.609***</td>
<td>-0.934***</td>
<td>-0.193***</td>
</tr>
<tr>
<td></td>
<td>(0.0442)</td>
<td>(0.0521)</td>
<td>(0.0604)</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Quantile FEs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>3107</td>
<td>3107</td>
<td>3107</td>
</tr>
</tbody>
</table>

*Note:* This table estimates the impact of the TIGER seeding on total contributions, follow-on contributions and total number of users within OpenStreetMap at the county-level. For each variable, 2008-14 indicates the cumulative total of contributions or users over the post-seeding period, while 2014 indicates that number of contributions or users in the last year, i.e. 2014. The specification employed is:  
\[ Y_i = \alpha + \beta_1 \times Treat_i + X_i + \gamma_1 + \gamma_2 + \gamma_3 + \epsilon_i \]
where \( \gamma_1 \), \( \gamma_2 \) and \( \gamma_3 \) represent fixed effects for population density, household income and per-capita income quartiles respectively and \( X_i \) indicates a set of ten controls, including a county’s population, unemployment, population density, the percent of males etc. Heteroskedasticity robust standard errors are estimated. See text for more details.
Table 5. **Alternate Samples and Additional Robustness**

**Panel A – Difference in Difference Analysis**

<table>
<thead>
<tr>
<th></th>
<th>Diff-Time-Trends</th>
<th>Placebo</th>
<th>Boundary-Sample</th>
<th>Timing-Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Follow-on</td>
<td>Contributors</td>
<td>Follow-on</td>
<td>Contributors</td>
</tr>
<tr>
<td>Post X Treat</td>
<td>-0.187***</td>
<td>-0.0917***</td>
<td>0.00783</td>
<td>-0.135**</td>
</tr>
<tr>
<td></td>
<td>(0.0469)</td>
<td>(0.0144)</td>
<td>(0.0483)</td>
<td>(0.0174)</td>
</tr>
<tr>
<td>Quarter FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>County FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>121173</td>
<td>121173</td>
<td>121173</td>
<td>86502</td>
</tr>
</tbody>
</table>

**Panel B – Cross-Sectional Specification**

<table>
<thead>
<tr>
<th></th>
<th>Boundary Sample (Follow-on)</th>
<th>Boundary Sample (Users)</th>
<th>Timing Sample (Follow-on)</th>
<th>Timing Sample (Users)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>-0.183*** (0.0704)</td>
<td>-0.412*** (0.0761)</td>
<td>-0.265*** (0.0780)</td>
<td>-0.121*** (0.0216)</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Quantile FEs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>2218</td>
<td>2218</td>
<td>1820</td>
<td>1820</td>
</tr>
</tbody>
</table>

*Note:* This table presents a number of robustness checks for the difference-in-difference (Panel A) and cross-sectional (Panel B) estimates for the main results. 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 3107 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.
Table 6. **Impact of Information Seeding on Long-Term Quality**

<table>
<thead>
<tr>
<th></th>
<th>Log/Error-Score</th>
<th>Log/Error-Score</th>
<th>Log/Error-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>0.126* (0.0708)</td>
<td>0.149* (0.0800)</td>
<td>0.297*** (0.0915)</td>
</tr>
<tr>
<td>Distance</td>
<td>0.0340*** (0.00262)</td>
<td>0.0339*** (0.00292)</td>
<td>0.0376*** (0.00320)</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Quartile FEs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sample</td>
<td>Full</td>
<td>Boundary</td>
<td>Timing</td>
</tr>
<tr>
<td>N</td>
<td>81007</td>
<td>58715</td>
<td>49095</td>
</tr>
</tbody>
</table>

*Note:* This table estimates the impact of TIGER seeding on a measure of long-term quality of the OpenStreetMap database. 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 to Google Maps). All specifications are estimated using linear models. See text for more details and for more complete descriptions of the outcome variable.

Table 7. **Testing the Ownership Mechanism**

<table>
<thead>
<tr>
<th>Follow-on-Contribs</th>
<th>Contributors</th>
<th>Ownership-Level</th>
<th>Owner-Contribs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distant Incremental</td>
<td>Old New</td>
<td>High Low</td>
<td>–</td>
</tr>
<tr>
<td>Post X Treat</td>
<td>-0.131*** (0.0391)</td>
<td>-0.00281 (0.00280)</td>
<td>-0.00847 (0.00574)</td>
</tr>
<tr>
<td></td>
<td>-0.227*** (0.0497)</td>
<td>-0.0586*** (0.0127)</td>
<td>-0.0357** (0.0173)</td>
</tr>
<tr>
<td></td>
<td>-0.00847 (0.00574)</td>
<td>-0.0357** (0.0173)</td>
<td>-0.161*** (0.0369)</td>
</tr>
<tr>
<td>County FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Quarter FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>121173</td>
<td>121173</td>
<td>121173</td>
</tr>
</tbody>
</table>

*Note:* This table presents estimates evaluating predictions from the ownership mechanism. 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, while Old Contributors are those who have made at least one edit before the information seeding took place. The Ownership-Level section picks apart 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-Contribs 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.
Appendices

Appendix A: External Validity: Hadoop and Tensorflow Case Study

Introduction

In this section, we present a small comparative case study to examine the external validity of our core finding that information seeding might hurt rather than help the long term development of communities. In particular, OpenStreetMap represents a case where a community performs an information gathering task. In contrast, one can think of traditional open source software as being a community-based knowledge good that is more interested in problem solving. This appendix tries to shed light on what aspect of our results might generalize to such problem-based communities and what aspects might not? It must be noted that this appendix is based on very preliminary qualitative work and is therefore meant to be an inspiration for future work, rather than any systematic form of evidence in favor of one theory over another.

The exercise we undertake is as follows. We profile two open-source projects; Hadoop, a big-data processing framework from the Apache Foundation, and TensorFlow a machine-learning framework sponsored by Google. These projects were seeded by contributors inside Apache and Google respectively and were then open-sourced. In contrast to the findings of the main study, these projects seem to have gained significant traction and success, despite being based on extensive seeding from an organization. Some facts about these two cases are summarized in Table A.1. Our case studies help to provide further detail on this seeming contradiction, and point out the extent to which our theory might generalize to different types of community-based knowledge goods. The overall conclusion from our exercise is that a closer reading of these cases supports rather than contradicts the information seeding theory. However, it also helps to point out other mechanisms that help to sustain online communities even in the face of extensive information seeding, and therefore provides for ideas for future work that could build on and modify our theory.

The Role of Information Seeding

First, we conducted a review of the contributions to these projects and their history as summarized in the table below. Our main conclusion is that even though both projects were seeded by a private organization – the extent of seeding differed considerably between them. TensorFlow was released earlier in its evolution, while Hadoop was more fully developed before its release. TensorFlow has many more active contributors, and these contributors are more likely to be from outside of Google and contribute on a voluntary basis. Hadoop on the other hand has fewer external contributors and a smaller proportion of volunteer contributors, which suggests that it has been less successful in attracting unpaid, active contributors. While these projects are dramatically different (including the complexity of the problem they are solving, and the years in which they were released), it is possible that a lower level of information seeding helped TensorFlow attract more follow-on contributions as compared to Hadoop. At the very least, this comparative exercise suggests that an “optimal” level of information seeding might be relevant for problem-oriented open source software and not just for information-oriented projects like OpenStreetMap.

Novel Mechanisms Parallel to Information Seeding

While the differential information seeding between TensorFlow and Hadoop could explain the divergence in their engagement among volunteer contributors, the broader question of why users might contribute code to these projects seeded by an external organization remains. Unlike information-oriented projects, problem oriented projects like open source software attract skilled contributors and therefore some mechanisms unrelated to ownership seem at work and serve to mute the demotivating effect of information seeding. Table A.2 provides case studies of contributors to TensorFlow and Hadoop. Based on these case studies we derived
the following list of alternate mechanisms that could explain why contributors might contribute even when incentives for ownership might seem limited.

1. **Contributions as evidence of a specialised skill:** As noted in Table A.2, Tensorflow contributor @facaiy showed an inclination to pursue tasks that will help him gain specialised skills he has outlined for himself, like learning C++, an advanced programming language. This learning motivation is likely to arise the more specialized the skills needed to contribute to a project.

2. **Seeking career opportunities with the organisation owning an open source project:** In the Tensorflow project we also notice instances of several users signing their contributions with personal email addresses and subsequently signing their contributions with a work email address, at Google, after a few months into the project; this could be indicative of absorption into the Google workforce. This is a special version of the career concerns mechanism that applies specifically to the project to which a user contributes.

3. **Achieving recognition within and outside an open source organisation:** @terrytangyuan, an external contributor who has been contributing to Tensorflow from a few weeks after its first public release, published the first book to teach Tensorflow in the Chinese language, and regularly makes large additions to the project documentation. For the same, he has received three awards, Outstanding China Mainland Books Copyright Exported to Taiwan (The Publishers Association of China), Outstanding Author (Beijing Publishing House of Electronics Industry) and the Open Source Peer Bonus Award (Google Inc.). This has enabled him to use his Tensorflow connection to extend his influence to other communities.

4. **Fulfilling specific tasks pre-determined by their place of work:** In the Hadoop project, we notice strong evidence of contribution segmentation by organisation, in addition to contributors using a work email address from their first contribution, indicating their organisation’s involvement with the project.

We conclude that these mechanisms help explain some reasons why contributors might contribute to seeded projects even when the ownership incentive has been crowded out. We further note that unlike in information oriented projects, where contributor skill can be picked up in the course of a couple of days and is not a determining factor for contributions, problem oriented projects may discourage certain contributors simply because they require specialised skills that may take a potential contributor weeks or months to pick up. This paper does not delve into some of these factors, and instead presents them as a possibility for future work. With growing commercial interest from companies like Mapbox, Facebook, Apple and Grab in the OpenStreetMap project, data from these additional factors may present interesting results and analyses that help understand the influence of these factors on open knowledge projects as well.

Overall, these case studies help to validate that information seeding might be relevant across broad types of community-based knowledge projects, but also call for additional work that analyzes the impact of seeding on follow-on contributions through related, but distinct, mechanisms.
Table A.1. Comparing Hadoop and Tensorflow Information Seeding

<table>
<thead>
<tr>
<th>Factor</th>
<th>Hadoop (seeded more)</th>
<th>TensorFlow (seeded less)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genesis</td>
<td>The first stable release of the Hadoop project was made under the aegis of Yahoo and the Apache foundation in April, 2006, and was built on top of the Apache Nutch project, which was started in 2004. The initial code that was factored out of Nutch consisted of about 5,000 lines of code for HDFS and about 6,000 lines of code for MapReduce.</td>
<td>Work on Tensorflow began from scratch in the year 2014, a year before the project was open sourced in November 2015. First external contributor started contributing a few weeks after the project was open sourced.</td>
</tr>
<tr>
<td>Number of contributors till date</td>
<td>267 (not unique) contributors who have contributed code and 203 (not unique) contributors use apache/hortonworks email addresses while adding contributions.</td>
<td>2053 (not unique) contributors who have contributed code and 403 (not unique) contributors use google email addresses while adding contributions, but many more are likely google employees.</td>
</tr>
<tr>
<td>Contribution Patterns</td>
<td>Contributions are striated well into clusters of contributions from specific organisations like Hortonworks, Apache, Intel, ZTE, Cloudera (Acquired by HortonWorks in Jan 2019).</td>
<td>Several contributions from Google, but contributions are from a mix of individual contributors and organisations interested in machine learning.</td>
</tr>
</tbody>
</table>

Appendix B: Measuring Quality Through Amenity Names

In the main manuscript, we measured quality by comparing OpenStreetMap and Google Maps in terms of routing error. Counties with a higher level of information seeding seem to demonstrate a higher level of errors almost a decade after the map was first seeded. This is likely due to a lack of information about one or two-way streets on OpenStreetMap in seeded counties. While lack of information is one definition of quality, an alternate definition might consider quality, conditional on information being added. In particular, one might ask, conditional on information being added – how likely is it that the information is accurate?

While information accuracy is harder to evaluate when considering road geometry, in this section, I evaluate it by comparing the accuracy of information labels for local amenities, specifically, full service restaurant names. In particular, we licensed data from a private provider of local business records, InfoUSA on the exact location and names of all full service restaurants in the United States. We then took a sample of about 100,000 restaurants balanced across counties. Since we know the latitude and longitude for every restaurant, we query the OpenStreetMap database as of 2018 and find the closest matching restaurant in a 100m radius. Through this exercise we identify about 12,000 restaurants that match the source data. Further for every match, we compute the Levenshtein distance between the restaurant’s name in our database (which we assume to be more accurate) and the name for the amenity on OpenStreetMap. This distance is used to assign a “match score” to every restaurant name in the OpenStreetMap data. These scores vary between 60 and 100, where a 100 signifies the most accurate record, while a 60 is a much poorer score. For example, the restaurant “Cici’s Pizza” in Foley, AL in our database matches exactly with the corresponding restaurant on OpenStreetMap, and this match gets a score of 100. In contrast, “Joe and Aggie’s Cafe” in Holbrook, AZ is named “Joe and Aggies” in OpenStreetMap and gets a score of 73. It is this dimension of “quality” that we are measuring in this exercise. In addition to being a good measure for name accuracy, we also think this score is correlated with other measures of quality such as exact location, other description (restaurant opening hours, website etc) and is therefore a good way to measure differences in information accuracy.
Table A.2. Case Studies of External Contributors to Tensorflow and Hadoop

<table>
<thead>
<tr>
<th>Tensorflow:</th>
</tr>
</thead>
<tbody>
<tr>
<td>We consider the case study of a user @facaiy. @facaiy has been contributing to the project since 2017, 2 years after the inception of the project, and is a senior engineer at Alibaba. He works out of Beijing. @facaiy contributes to the project in three predominant ways.</td>
</tr>
<tr>
<td>• Creates notes to outline technical issues with the project which he then proceeds to fix himself.</td>
</tr>
<tr>
<td>• Fixing similar issues proposed by others as long as they involve simple syntactical fixes to the relevant files.</td>
</tr>
<tr>
<td>• Migrating features in the project from the Python programming language to the C++ programming language.</td>
</tr>
<tr>
<td>@facaiy’s notes on the Tensorflow GitHub repository provide some insight into his motivations. He asks to be assigned to certain tasks since they would provide him the opportunity to learn the C++ programming language or to explore parts of the code that intrigue him, like one issue which concerns code related to GPUs. Sometimes, he also selects simple fixes that would allow him to interact with issues opened by newer contributors. Most of @facaiy’s fixes are large, involve the addition of new features, and touch multiple files. There are also fixes and issues that evince ownership, since @facaiy chooses similar types of problems (outlined above) and interacts with files scoped to a specific directory to fix these problems. He does not engage in grunt work, does not participate in conversations regarding programming styles or review the work of other contributors for accuracy. For most of @facaiy’s contributions, these aspects are taken care of by Googlers like @drpngx, @suharshs, @sriramkb, @yifeif, @skye.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hadoop:</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Hadoop project was hosted on SVN, an older version control system, and moved to the GitHub version control system several years after the inception of the project. To keep the newer system up to date, each released version was moved in its entirety into GitHub. As a result, we do only have a high-level view of the project. We also note that contributors can be clustered by their place of work owing to the strong interest exhibited by several companies in Hadoop. This interest leads to a segmenting of the project, wherein contributors only touch parts of the project relevant to their company. This segmentation again shows an evidence of ownership.</td>
</tr>
<tr>
<td>We take a look at contributions from bobhansen and Thomas Marquardt, two users who we selected, specifically because of their non-affiliation with Apache/Hortwornworks. Bob works at Hewlett Packard Enterprise and all of his contributions are related to a single library, libhdfs++. Each fix is large and touches multiple files, but is a feature that adds to libhdfs++. Thomas works at Microsoft, and his contributions are related to Windows Azure, and other Microsoft related dependencies. Each contribution is again a new feature fix and scoped to parts of the code most relevant to Microsoft.</td>
</tr>
</tbody>
</table>
Table B.1. **Alternate Measure of Quality: Accuracy of Restaurant Names**

<table>
<thead>
<tr>
<th></th>
<th>Match Score</th>
<th>Match Score</th>
<th>Match Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treat</strong></td>
<td>-0.00853**</td>
<td>-0.00944**</td>
<td>-0.0150***</td>
</tr>
<tr>
<td></td>
<td>(0.00345)</td>
<td>(0.00382)</td>
<td>(0.00443)</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Quartile FE</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td>Full</td>
<td>Boundary</td>
<td>Timing</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>11315</td>
<td>8300</td>
<td>6592</td>
</tr>
</tbody>
</table>

**Note:** This regression uses a similar specification as the baseline cross-sectional regression including ten demographic controls and fixed-effects for quartiles of population density, household income, and per-capita income. The main outcome variable is a score indicating name similarity between OpenStreetMap and InfoUSA in terms of restaurant name. The first column uses the full sample, while the next two use the boundary adn timing samples respectively. See main text of Appendix B for more details.

between tiger and control counties.

Armed with matchscores for about 12,000 establishments all across the US, we estimate regressions similar to the ones in Section 3.5 with a few modifications. Specifically for establishment $i$ in county $j$, we estimate,

$$
Matchscore_{ij} = \alpha + \beta_1 \times Treat_i + controls_{ij} + control_i + \epsilon_{ij}
$$

where, $controls_{ij}$ include controls for county-level variables such as area, population, num of households, income per capital, education level and number of young males. In addition we separate all counties into 4 groups by population density and include fixed effects for each group – this allows us to control for urban and rural areas quite clearly. Finally, establishment level controls include sales and size (in terms of number of employees) derived from InfoUSA data. Given these controls and the relatively random allocation of seeding between treatment and control counties, we expect $\beta_1$ to estimate the difference in matchscore between treatment and control counties. In some specifications we estimate this regression using the boundary and timing samples described in the main text.

Table B.1 includes results from this analysis. As is clear, treatment counties have a lower matchscore as compared to control counties across all specifications. Having said that, the magnitude of this difference is more modest than the difference in quality scores reported in the main analysis. As shown in the paper, this is likely due to the fact that distant follow on edits (such as restaurant names) are less affected by seeding. Nevertheless, there seems to be some evidence to suggest that even along an alternate measure of “quality”, i.e. accuracy in naming, treatment counties do worse as compared to control counties even a decade after the seeding experiment in late 2007.
Appendix C: The Ownership Mechanism–Theory and Simulations

I provide a simple theoretical illustration of how information seeding could reduce total follow-on contributions through the ownership channel.

Simple Framework

The Map: To fix ideas, assume that a region $R^i$ is made up of $N \times N$ unique objects that need to be mapped. One can think of these objects as a grid of streets, restaurants or other points of interest. Region $i$ is represented by a map, $M^i$ which is a matrix of dimension $N \times N$, where item $M^i[x,y]$ represents the level at which object $[x,y]$ in region $i$ has been mapped. $M^i[x,y] \in \{0,1,2\}$, where state 0 represents that the object has not been mapped, state 1 represents that it has been mapped but follow-on information (such as its name, or one-way status) has not been added, while state 2 represents that the object has been mapped and follow-on information has been added. When a region is seeded, a subset of objects are mapped automatically at level 1, i.e. for object $[x,y]$, $M^i[x,y] = 1$.

User Behavior: Users may now start making contributions, both by mapping objects that were not seeded, and by adding follow-on information to seeded and unseeded objects. Assume that for each region $i$, a unique user $U^i_{xy}$ is allocated to a given object $[x,y]$ and she can only add information for that object. One can think of this as the user living close to a given street or restaurant. The user must exert costly effort $c^i_{xy}$ to change the value of the map element $M^i[x,y]$ from 0 to 1, and from 1 to 2. We call the first contribution a “basic” contribution, and the second a “follow-on” contribution. For a basic contribution, the user receives a benefit $B$, and for a follow-on contribution a benefit $B/k$, where $k > 1$. One can think of this benefit as the “warm glow” of having contributed to a public good, and this benefit is likely to be smaller for follow-on contribution than for a basic contribution. However, and this is where the ownership mechanism kicks in, the follow-on benefit $B/k$ increases to $B/\tilde{k}$, where $\tilde{k} < k$ if the user also made the basic contribution for the given object.

To be concrete, if a user is assigned to an object that has been seeded she has the option to make a follow-on contribution and gain benefit $B/k$ at a cost $c^i_{xy}$. However, if the user is assigned to an object that has not been seeded, then she must first make a basic contribution at cost $c^i_{xy}$ and benefit $B$, and also have the choice to make a follow-on contribution if her cost of making a contribution is lower than $B/k$.

Seeding: Now consider two regions $R^1$ and $R^2$, where $R^1$ receives a higher level of seeding as compared to $R^2$. Formally, $M^i[x,y] = 1$ for all $x < X_1$ and $y < Y_1$, and where $X_1 > X_2$ and $Y_1 > Y_2$. Therefore, Region 1 has $(X_1 - X_2) \times (Y_2 - Y_1)$ more objects that are seeded.

The key question of the paper is: which region will have a higher level of objects set to 2 (i.e. with follow-on contributions)? Note that, both regions have the potential for $N^2$ follow-on contributions, since none of the objects have $M^i[x,y]$ set to 2. Let $Follow\_on^1 = \Sigma(M^i[x,y] = 2)$, i.e. the number of objects that are mapped to the level 2 for region $i$, which is the same as the total number of follow-on contributions.

The main prediction is that even though Region 1 has a head-start, Region 2 will have a more complete map, more follow-on contributions and more objects set to the value 2, i.e. $Follow\_on^2 > Follow\_on^1$. The key mechanism is that in $R^1$, there are fewer opportunities to gain the higher benefit level $B/\tilde{k}$ from ownership, while in $R^2$ there are more such opportunities, and users have a higher incentive to make follow-on contributions driven by the ownership effect.

Simple Example To see this, consider a simple example. Let, $N = 3$, i.e. a region has $3 \times 3 = 9$ objects. Let $X_1 = Y_1 = 2$ and $X_2 = Y_2 = 1$, i.e. four objects $(1,1),(2,1),(1,2)$ and $(2,2)$ are seeded in region 1, and only one object $(1,1)$ is seeded in region 2. Let $B = k = 2$, $\tilde{k} = 1$, i.e. users get a benefit of 2 from a basic contribution, a benefit of 1 from a follow-on contribution, which increases to a benefit of 2 if they have ownership over...
the object. Now, assume that all users have cost 1.5 of making a contribution. Users assigned to seeded objects can choose to make a follow-on contribution at cost, but will gain a benefit of only 1 unit, since they are not the owners of the object. These objects will therefore be left unedited. On the contrary, users assigned to unseeded objects will make the basic contribution and gain a benefit of 2 units at a cost of 1.5 units. Further, since these users are motivated by the ownership effect, they stand to gain another 2 units from follow-on contributions and make those edits as well. Unseeded objects therefore receive both basic and follow-on contributions. Note that this pattern applies to all objects in both $R^1$ and $R^2$. However, since $R^2$ has 8 unseeded objects as compared to only 4 for $R^1$, it receives 8 follow-on contributions as compared to only 4 in $R^2$.

Simulation

In order to make the simple example above more realistic, I also present results from a simulation model. Here, $N = 4$, $B = 75$, $k = 2$ and $\tilde{k} = 1$. Rather than a constant cost across users, the simulation allows a more realistic assumption that user costs are drawn from a uniform distribution $U[0,100]$. 12 of the 16 objects in $R^1$ are seeded, while only 4 objects in $R^2$ are seeded. The simulation is then run over a 1000 different cost distributions and the total difference in the number of follow-on contributions between $R^2$ and $R^1$ is shown in the following histogram (Figure C.1).

The simulation shows that $R^1$ gets on average about 2.9 more follow-on contributions than $R^1$. Mean follow-on contributions for $R^1$ are 7.4 edits as compared to 10.3 for $R^2$. Note here that when user costs are low, follow-on contributions will be made even for seeded nodes, while under high costs, even ownership benefits are not enough to incentivize follow-on edits. Despite these more general patterns, we find robust evidence under this theory that the ownership mechanism will lead to fewer follow-on contributions in seeded regions.

Figure C.1. Difference in Follow-on Contributions Between Unseeded and Seeded Region
Appendix D: Additional Figures and Tables

Figure D.1. Comparing Covariate Trends Between Treatment and Control Counties

Note: This figure compares treatment and control counties along a number of different covariates. Each panel plots a chart of the dependent variable (for example, population, households etc.) by county separately for treatment counties using a red solid line and control counties in a blue, dotted line. The standard errors are also represented using light gray solid and dashed lines respectively.
Table D.1. **Summary Statistics by County/Quarter**

### Panel A – County Level (N=3107)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributions (2008-14)</td>
<td>111844.15</td>
<td>299098.30</td>
<td>29230.00</td>
<td>418.00</td>
<td>8175957.00</td>
</tr>
<tr>
<td>Follow-on Contribs (2008-14)</td>
<td>20657.63</td>
<td>97220.59</td>
<td>3105.00</td>
<td>65.00</td>
<td>4138528.00</td>
</tr>
<tr>
<td>Contributors (2008-14)</td>
<td>299.42</td>
<td>459.03</td>
<td>225.00</td>
<td>3.00</td>
<td>12212.00</td>
</tr>
<tr>
<td>Contributions (2014)</td>
<td>15761.61</td>
<td>75658.48</td>
<td>3075.00</td>
<td>0.00</td>
<td>2239903.00</td>
</tr>
<tr>
<td>Follow-on Contribs (2014)</td>
<td>2251.21</td>
<td>10832.92</td>
<td>361.00</td>
<td>0.00</td>
<td>386446.00</td>
</tr>
<tr>
<td>Contributors (2014)</td>
<td>68.51</td>
<td>62.13</td>
<td>54.00</td>
<td>0.00</td>
<td>976.00</td>
</tr>
</tbody>
</table>

### Panel B – Quarter Level (N=39)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributions</td>
<td>8910250.90</td>
<td>7315968.17</td>
<td>9786033.00</td>
<td>0.00</td>
<td>27426628.00</td>
</tr>
<tr>
<td>Follow-on Contribs</td>
<td>1645724.51</td>
<td>1942820.07</td>
<td>1525216.00</td>
<td>0.00</td>
<td>9690495.00</td>
</tr>
<tr>
<td>Contributors</td>
<td>23853.51</td>
<td>19494.03</td>
<td>25441.00</td>
<td>0.00</td>
<td>73300.00</td>
</tr>
</tbody>
</table>

*Note:* This table provides a version of the summary statistics of the key outcome variables at the county level (Panel A) and quarter level (Panel B) separately.
Table D.2. **Instrumental Variables (IV) Estimates of the Impact of the Stock of Pre-Existing Knowledge on Follow-on Outcomes**

<table>
<thead>
<tr>
<th></th>
<th>Follow-on Contribs</th>
<th>Follow-on Contribs</th>
<th>Contributors</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st Stage</td>
<td>IV</td>
<td>1st Stage</td>
<td>IV</td>
</tr>
<tr>
<td>Treat</td>
<td>6.301***</td>
<td>0.698***</td>
<td>6.301***</td>
<td>0.698***</td>
</tr>
<tr>
<td></td>
<td>(0.668)</td>
<td>(0.0549)</td>
<td>(0.668)</td>
<td>(0.0549)</td>
</tr>
<tr>
<td>Pct. Seeded</td>
<td>-0.0294***</td>
<td>-0.0201***</td>
<td>-0.0294***</td>
<td>-0.0201***</td>
</tr>
<tr>
<td></td>
<td>(0.0101)</td>
<td>(0.00337)</td>
<td>(0.0101)</td>
<td>(0.00337)</td>
</tr>
<tr>
<td>ln(Seeding Contribs)</td>
<td>-0.277***</td>
<td>-0.183***</td>
<td>-0.277***</td>
<td>-0.183***</td>
</tr>
<tr>
<td></td>
<td>(0.0883)</td>
<td>(0.0312)</td>
<td>(0.0883)</td>
<td>(0.0312)</td>
</tr>
<tr>
<td>1st Stage F-Stat</td>
<td>97.12</td>
<td>178.2</td>
<td>97.12</td>
<td>178.2</td>
</tr>
<tr>
<td>Area/Pop Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Quantile FEs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>3092</td>
<td>3092</td>
<td>3107</td>
<td>3107</td>
</tr>
</tbody>
</table>

*Note:* This table presents estimates from IV regressions estimated using 2SLS. Rather than examining the impact of treatment status on follow-on outcomes directly, this regression examines the impact of the “stock of knowledge” on follow-on outcomes, instrumented by the treatment variable (i.e. exogenous variation in TIGER seeding). For each county, two measures of the stock of knowledge are used: Pct Seeded, which 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, and ln(Seeding Contribs) which is the logged number of total contributions by the seeding effort (i.e. by the Dave Hansen TIGER account). These two variables provide two different ways of measuring how much knowledge was added via the seeding effort to the OpenStreetMap database. Similar to the baseline specifications, two follow-on outcomes are considered, the logged number of follow-on contributions and the logged number of users at the county level between 2008-2014. The 1st Stage for each model regresses the endogenous “stock of knowledge” variable on the treatment dummy, and the IV regresses the eventual outcome (i.e follow-on contribs or total users) on this instrumented stock-of-knowledge variable. All models include controls for the total land area, population etc. exactly as in the baseline specifications (see Table 1). The Cragg-Donald Wald F statistic for the first stage and robust standard errors are reported. Pct. Seeded data is missing for 15 counties, and hence total observations when using this varible is 3092 and not 3107.
### Table D.3. Additional Robustness for Panel Models

<table>
<thead>
<tr>
<th>Post X Treat</th>
<th>Total Changesets</th>
<th>Follow-on-Contrib.</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.0966***</td>
<td>-0.149***</td>
<td>-0.0623***</td>
</tr>
<tr>
<td></td>
<td>(0.0258)</td>
<td>(0.0483)</td>
<td>(0.0170)</td>
</tr>
<tr>
<td>G-Maps Popularity</td>
<td>-0.00575***</td>
<td>-0.00433***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00114)</td>
<td>(0.000435)</td>
<td></td>
</tr>
</tbody>
</table>

Quarter FE | Yes | Yes | Yes |
County FE  | Yes | Yes | Yes |
N          | 121173 | 121173 | 121173 |

**Note:** This table provides additional robustness for the baseline panel results. In order to account for multiple edits that are close together in time, the outcome variable in column 1 is Total Changesets, which aggregates edits in a single session into a single “changeset”. The next two columns estimate the baseline regression after controlling for a measure of Google Maps popularity in a given county in a given year as measured by data from Google Trends for the term “Google Maps”.

### Table D.4. Testing Ownership Effect At the Object Level

<table>
<thead>
<tr>
<th>Treat</th>
<th>Follow-on Edits</th>
<th>Follow-on Edits</th>
<th>Owner Edits</th>
<th>Owner Edits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.323***</td>
<td>-0.113***</td>
<td>-0.0869***</td>
<td>-0.0392***</td>
</tr>
<tr>
<td></td>
<td>(0.0115)</td>
<td>(0.0127)</td>
<td>(0.00734)</td>
<td>(0.00833)</td>
</tr>
</tbody>
</table>

Area/Pop Controls | No | Yes | No | Yes |
Quantile FEs      | Yes | Yes | Yes | Yes |
N                  | 85292 | 85292 | 85292 | 85292 |

**Note:** This table compares the 85,292 user-created nodes in treatment and control counties in the state of Florida. The regression specification and controls are similar to the baseline cross-sectional specification. Follow-on Edits is the number of times an object was modified after creation, and Owner Edits is the number of follow-on edits that came from the object’s owner, i.e. the user who first created the object.
### Table D.5. The Impact of Seeding on Governance

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th></th>
<th>Boundary Sample</th>
<th></th>
<th>Timing Sample</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Events</td>
<td>Groups</td>
<td>Events</td>
<td>Groups</td>
<td>Events</td>
<td>Groups</td>
</tr>
<tr>
<td>Treat</td>
<td>0.00202</td>
<td>0.000614</td>
<td>0.000650</td>
<td>0.00548</td>
<td>0.00493</td>
<td>0.000816</td>
</tr>
<tr>
<td></td>
<td>(0.00442)</td>
<td>(0.00403)</td>
<td>(0.00532)</td>
<td>(0.00471)</td>
<td>(0.00567)</td>
<td>(0.00567)</td>
</tr>
<tr>
<td>Area/Pop Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Quantile FEs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>3107</td>
<td>3107</td>
<td>2218</td>
<td>2218</td>
<td>1820</td>
<td>1820</td>
</tr>
</tbody>
</table>

Note: This table presents estimates examining the impact of information seeding on an alternate mechanism through which it could’ve affected outcomes: governance. Specifically, I examine whether information seeding is associated with increased level of organized events or formal “maptime” groups in a given county. The specification is similar to the cross-sectional specifications employed in the baseline analysis, and the key dependent variables are an indicator for whether any mapping events (“mapping parties”) were organized in a given county or whether there exists a formal “maptime” group in that county.