GRIPS Discussion Paper 12-18

Localized knowledge spillovers and patent citations: A distance-based approach (revised version)

By

Yasusada Murata
Ryo Nakajima
Ryosuke Okamoto
Ryuichi Tamura

January 2013

National Graduate Institute for Policy Studies
7-22-1 Roppongi, Minato-ku,
Tokyo, Japan 106-8677

Localized knowledge spillovers and patent citations: A distance-based approach (revised version)*

Yasusada Murata† Ryo Nakajima‡ Ryosuke Okamoto§ Ryuichi Tamura¶

January 29, 2013

Abstract

We develop a new distance-based test of localized knowledge spillovers that embeds the concept of control patents. Using microgeographic data, we identify localization distance for each technology class while allowing for spillovers across geographic units. We revisit the debate by Thompson and Fox-Kean (2005a,b) and Henderson, Jaffe and Trajtenberg (2005) on the existence of localized knowledge spillovers, and find solid evidence supporting localization even when using fine-grained controls. We further relax the assumption of perfect controls, and show that our distance-based test detects localization for the majority of technology classes unless hidden biases induced by imperfect controls are extremely large.

Keywords: localized knowledge spillovers, distance-based tests, microgeographic data, K-density, patent citations, control patents, sensitivity analysis

JEL Codes: O31, R12

*We are very grateful to Gilles Duranton, Henry Overman, Peter Thompson and three referees for detailed comments and suggestions on earlier drafts of this paper. We have also benefited from comments and suggestions by Bill Kerr, Jungmin Lee, Matt Turner as well as other participants at various conferences and workshops.

†Advanced Research Institute for the Sciences and Humanities, Nihon University, E-mail address: murata.yasusada@nihon-u.ac.jp

‡Department of Economics, Keio University, E-mail address: nakajima@econ.keio.ac.jp

§National Graduate Institute for Policy Studies, E-mail address: rokamoto@grips.ac.jp

¶Center for Economic Growth Strategy, Yokohama National University, E-mail address: tamura@ynu.ac.jp

1. Introduction

Jaffe, Trajtenberg and Henderson (1993, henceforth, JTH) developed a matching rate method to test localized knowledge spillovers (LKS) as evidenced by patent citations. By controlling for the preexisting geographic distribution of technological activities, they found evidence supporting LKS at the state and metropolitan statistical area (MSA) levels. However, their finding was recently challenged by Thompson and Fox-Kean (2005*a*, henceforth, TFK). The major difference between these two studies lies in the selection of control patents. In JTH, control and citing patents share a technology class at the 3-digit level, whereas in TFK, both patents share a finer technology subclass at the 6-digit level.[1] The latter authors further restricted to control patents that have any subclass code in common with originating patents, and found no evidence supporting LKS at the state and MSA levels. The existence of LKS is, thus, still inconclusive (Henderson, Jaffe and Trajtenberg, 2005).

To settle this debate, we start with the question of whether states and MSAs are relevant spatial units for testing LKS. The matching rate approach by JTH and TFK focuses on within-state or within-MSA localization while abstracting from the relative position of those spatial units. Put differently, they assume that the extent of LKS to be limited by administrative boundaries while making the distance from Boston to New Haven equivalent to that of Boston to Los Angeles.[2] To capture cross-boundary knowledge spillovers, we build on distance-based kernel-density (K-density) tests of localization developed in the context of establishment agglomeration by Duranton and Overman (2005, henceforth, DO).[3]

Our distance-based approach addresses whether knowledge spillovers, as evidenced by patent citations, are localized, and examines to what extent they are localized (if they are). In doing so, we consider which technology classes are localized, and identify the localization

[1]The case-control methods have been applied to detect LKS for almost two decades (e.g., Almeida, 1996; Agrawal, Kapur and McHale, 2008; Agrawal, Cockburn and Rosell, 2010).

[2]Note also that spatial units often differ in population and area, so that spatial aggregations tend to mix different spatial scales. For instance, localization tests at the state level involve comparisons between Rhode Island and California, whose area is more than 150 times as large. Furthermore, such aggregation often leads to spurious correlations across aggregated variables, which is known as the Modifiable Areal Unit Problem.

[3]Their basic idea is to generate the distribution of distances between pairs of establishments in an industry and to compare it with that of hypothetical industries, in which establishments are randomly allocated across existing establishment sites, in order to assess the significance of departures from randomness. The DO metric is used for the analysis of determinants of industry coagglomeration in Ellison, Glaeser and Kerr (2010).

distance that is specific to the technology class of originating patents. Our key idea is to generate the distribution of actual citation distances using microgeographic data on inventors, and to compare it with that of counterfactual citation distances. To this end, we first identify, *for each observed cited-citing relationship*, a set of control patents that could have cited the originating patent as in JTH and TFK. From that set of patents, we then randomly draw counterfactual citations as in DO. We finally detect LKS by comparing the actual and counterfactual distributions of citation distances. Thus, our novelty lies in developing the K-density tests with case-controls and applying them to observed cited-citing relationships.[4]

We obtain the following results. First, distance matters. Our distance-based tests find that, even when we use 6-digit controls, knowledge spillovers are localized significantly for about one-third of all 360 technology classes in question. This is in sharp contrast to TFK who used 6-digit controls and found no evidence supporting LKS at the state and MSA levels. In the 3-digit case, more than 70% of 384 technology classes in question exhibit localization, thus confirming the result by JTH. We further show that, in both cases, the majority of technology classes displaying localization are localized at least once within 200 km, which corresponds roughly to the distance between Boston and New Haven.

Second, heterogeneity across technology classes also matters. Our 6-digit analysis reveals that, while about one-third of technology classes exhibit localization, more than 10% of technology classes display dispersion. This, together with the 6-digit result in TFK, implies that aggregating different technology classes can offset the tendency toward localization even when a substantial number of technology classes display localization at the disaggregate level.

The biases from aggregating spatial units and technology classes are shown to be substantial. To explore the difference between the matching rate and distance-based approaches in detecting LKS, we conduct class-specific matching rate tests, and compare the number of localized classes with the corresponding number generated by our distance-based tests. It turns out that, although the numbers are roughly the same for the 3-digit case, the matching rate

[4]Kerr and Kominers (2012) apply a similar distance-based method to patent data. However, they detect localization by using pairwise distances among inventors as have been done by DO in the context of establishment agglomeration. Their K-density tests thus abstract from the concept of control patents and explicit cited-citing relationships, both of which are at the heart of our analysis.

tests *underestimate* the number of localized classes for the 6-digit case. Indeed, the matching rate tests fail to detect LKS for more than 60% of the technology classes that exhibit localization by the distance-based tests.

These results rely on the premise that both the 3- and 6-digit controls are perfect. However, TFK argue that 3-digit patent classes are too broad and noisy for the purpose of identifying control patents, whereas Henderson, Jaffe and Trajtenberg (2005) state that there is no systematic evidence supporting that the 6-digit subclass classification renders "closer" technologically matched controls. Therefore, we finally conduct Rosenbaum's (2002) sensitivity analysis, provided that neither the 3-digit control nor the 6-digit control is perfect due to unobserved heterogeneity within classes or subclasses. In doing so, we deal with the 3- and 6-digit controls simultaneously by considering that matching on subclasses implies matching on classes. This specification is general in that it encompasses the cases analyzed by JTH and TFK as limiting cases, while allowing for imperfect controls. To our knowledge, no such attempt has been made so far.

In this generalized framework, we finally show that *the majority of technology classes exhibit localization* unless hidden biases induced by imperfect controls are extremely large.[5] We further confirm that, even with imperfect controls, the matching rate tests still underestimate the percentage of localized technology classes when compared with the distance-based tests.

The rest of the paper is organized as follows. Section 2 describes data and methodology. Section 3 reports our results. Section 4 generalizes our analysis and Section 5 concludes.

2. Data and methodology

Unlike the matching rate tests at the state and MSA levels, we need to combine patent citations data and microgeographic data to conduct distance-based tests. Concerning methodology, we first identify, for each cited-citing relationship, a set of control patents that could have cited the originating patent to control for the existing geographic distribution of technological activities. From that set of patents, we then randomly draw hypothetical citations. The counterfactual

[5] In this generalized framework, where the 3- and 6-digit controls are placed on a common ground, the case analyzed by TFK constitutes a special case where hidden biases are infinitely large, as shown below.

citations thus obtained, with which we compare the actual citations to detect LKS, share common features between the matching rate and the distance-based tests. Hence, we can make a direct comparison between these two tests.

2.1. Patents and patent citations

Our data are based on the NBER U.S. Patent Citations Data File, which is described in detail by Hall, Jaffe and Trajtenberg (2001). This data set covers all patent applications between 1963 and 1999 and those granted by 1999, as well as cited-citing relationships for patents granted between 1975 and 1999.[6] For each patent, the list of inventors, the addresses of inventors, and the technological category are recorded, along with other information such as the year of application, assignees, and the type of assignees. The detailed information of patent application month and *patent class* (3-digit) and *subclass* (6-digit) codes is supplemented with the United States Patent and Trademark Office (USPTO) Patent BIB database.[7]

We begin with 142,245 U.S. nongovernmental patents that were granted between January 1975 and December 1979. The sampling period is chosen to be comparable to those of previous studies. We identify patents as "U.S." if the country of the assignee is the United States. We observe that 115,905 (81.5%) of them were cited at least once by other U.S. patents, and we call them the *originating patents*. We then identify the *citing patents* that cited the originating patents by examining all patents that were granted between January 1975 and December 1999.

As in JTH and TFK, we consider both *intra*- and *inter*-class knowledge spillovers. Thus, citing patents need not belong to the same technology class as their originating patent. In our data intra-class knowledge spillovers account for 49.0% while the remaining 51.0% being captured by inter-class knowledge spillovers.[8]

Finally, we exclude "self-citations" by focusing on knowledge flows between different inventors of different assignees. Accordingly, a citing patent is classified as self-citing (i) if it

[6]Cited-citing relationships need not represent narrowly defined knowledge spillovers as citations might also capture knowledge flows driven by priced market transactions. However, as in the existing literature, we follow the convention that citations are proxies for knowledge spillovers throughout this paper.

[7]We use the patent classification as of December 31, 1999.

[8]In that sense, our approach is similar to that of Duranton and Overman (2008), where they use distances between establishments in vertical-linked industries. Yet, they rely on distances between arbitrary pairs of establishments in different industries, instead of distances obtained from actual cited-citing relationships.

had the same assignee as the originating patent that it cited; or (ii) if it was invented by the same inventor as the originating patent that it cited.[9] To distinguish unique inventors, we use the computerized matching procedure (CMP) proposed by Trajtenberg, Shiff and Melamed (2006).[10] The CMP uses the name of inventors, patent citations, and inventors' addresses, while allowing for possible errors in names. We find that 15.0% of citing patents are classified as self-citations. After excluding self-citations, we obtain 647,983 citing patents.

2.2. Geographic information

Our distance-based approach requires microgeographic data, namely the locations at which inventions were created. We identify the location of each invention at the census place level. The U.S. Census Bureau defines a place as a concentration of population. There are 23,789 places in the 1990 census, which we use below.[11] To be more specific, restricting patent inventors who reside in the contiguous U.S. area, we first match the address of each inventor to its 1990 census place by name. If the name match fails, we locate it via the populated place provided by the U.S. Geographic Names Information System (GNIS). We match the inventor's address with the GNIS populated place, which is more finely delineated than the census place, and then find the census place that is nearest to the identified GNIS populated place by using their spatial coordination information. This procedure allows us to identify the 18,139 census places for 97.0% of all inventors in the sample. The average of within-area distances for census places is 1.7 km, which is far smaller than those for counties (22.6 km), Consolidated Metropolitan Statistical Areas (CMSAs) (59.9 km), and states (197.9 km).[12]

[9] JTH and TFK regard only (i) as self-citations. Criterion (ii) rules out spurious knowledge spillovers associated with inventor mobility. Furthermore, in response to Henderson, Jaffe and Trajtenberg (2005), we exclude control patents that share the same inventors or the same assignees with the originating patents. Note that one assignee can be a subsidiary of the other. We identify a citation between parent and subsidiary companies as a self-citation by supplementing our data with name matching results of parent and subsidiary firms that are available from Bronwyn Hall's website. We also use SDC Platinum, the Worldwide Mergers and Acquisitions Database. Among all M&As reported therein, we focus on the cases in which the acquiring company obtains all of the stock of the target company, and then regard those pairs of two companies as to be in parent-subsidiary relationships.

[10] See Nakajima, Tamura and Hanaki (2010) for the implementation detail of the CMP.

[11] Census places are much more finely delineated than counties (there are 3,141 counties), but not as small as zip code areas (there are 29,470 zip code areas). We could use zip code areas. However, the NBER U.S. Patent Citations Data File reports zip codes for only 15.4% of all U.S. patent records.

[12] The distances are computed by the formula in Kendall and Moran (1963), which is presented in Section 2.5.

2.3. Control patents and counterfactuals

To test LKS, we must control for the existing spatial distribution of technological activities, regardless of whether or not citations come from the same technology class as their originating patent.[13] To this end, we start with *control patents*, proposed by JTH and TFK, which satisfy the following two conditions. First, control patents should belong to the same technological area as the citing patent under consideration. JTH select a control patent at the 3-digit level, whereas TFK construct a finer control at the 6-digit level.[14] In what follows, we refer to the former as a *3-digit control*, and call the latter a *6-digit control*. Second, a control patent should be in the same cohort as the citing patent. JTH choose a control patent whose application date is within a one-month window on either side of the citing patent's application date. Similarly, TFK set the application date of a control patent within plus-or-minus six month around that of the citing patent. Following these studies, we use one-month and six-month windows for the 3-digit and 6-digit controls, respectively.[15]

Insert Table 1

Table 1 presents the sample sizes. The first column shows the total numbers of the originating and citing patents. These numbers include patents with and without controls. In the second and third columns, the numbers of originating and citing patents having at least one control are reported. It should be noted that citing patents do not always have controls, and, even if they do, the control is not necessarily unique for each citing patent. As shown, 60.2% of the citing patents have 3-digit controls. The rate of the citing patents having 6-digit controls is lower, at 18.7%. The citing patents with no controls assigned (and their originating patents) are dropped out of the samples.[16] As a result, 92.6% of the originating patents remain "in-sample" for the 3-digit controls, and the corresponding number is 51.0% for the 6-digit

[13]For instance, if 20% of citations for an originating patent in patent class A are from the same class and 10% are from class B, etc., the idea is to construct a control group, where patent class A accounts for 20%, class B does for 10%, etc.

[14]The latter also claim that a control should match the originating patent.

[15]One minor difference is that we use a fixed application date window within which control patents are searched, while TFK enlarge it in incremental steps from a one-month window, then a three-month window, and, if necessary, a six-month window until the control patent is found for each citing patent.

[16]We also drop technology classes in which originating patents are distributed across less than 10 census places because we estimate the density of distances for each technology class, and a sufficient number of location points are needed to obtain well-behaved estimated density functions.

controls. In the analysis that follows, we use these in-sample patents.

Once the relevant control patents are identified, we can construct the counterfactual citations, with which we compare the actual citations, as follows. For each observed cited-citing relationship, we define an *admissible patent set* that consists of the citing and control patents at the 3- or 6-digit level, i.e., the patents that either actually cited or could have cited the originating patent.[17] We then allocate a counterfactual citation between the originating patent and a patent that is randomly drawn from the corresponding admissible patent set (see the Appendix for an example).

2.4. The matching rate approach

The idea of the matching rate approach is to compare the geographic matching rate of the actual citations with that of the counterfactual citations. Following JTH and TFK, we define the matching rate of the actual citations as the proportion of the citing patents whose geographic units such as states and CMSAs are matched with those of the originating patents. We analogously define the matching rate of the counterfactual citations by matching geographic units between an originating patent and a patent that is randomly drawn from the corresponding admissible patent set. We resample patents many times and generate a simulated distribution of the counterfactual matching rates.[18] We now describe the procedure of our matching rate test in detail.

Let p^c and p^r be the population probability that a citing patent is in the same geographic unit as the originating patent, and the corresponding probability for a randomly drawn patent from the admissible patent set. We test the null hypothesis $H_0 : p^c = p^r$ (no LKS) against the alternative hypothesis $H_1 : p^c > p^r$ (significant LKS). Let $\hat{p}^c$ be the matching rate of the actual citations that we observe in the data. Under the null hypothesis, it is not statistically different from a realization of the counterfactual matching rate, which we denote by $\hat{p}^r$. We

[17]It should be noted that, in the 6-digit case, we use the admissible patent set that consists only of the citing and control patents sharing a common technology class with the corresponding originating patent. This is a logical consequence of the additional restriction in the 6-digit case that originating-citing-control triads of patents must share at least one patent subclass in common.

[18]TFK propose a similar random sampling method to construct the matching rate of the counterfactual citations. They randomly select a patent from the admissible patent set once for each actual citation.

thus reject the null hypothesis of no LKS if the p-value, $\text{Prob}(\hat{p}^c \leq \hat{p}^r)$, is less than 5%.

We construct the observed matching rate $\hat{p}^c$ as follows. First, for each cited-citing relationship c^{ij}, we consider location match m^{ij} between originating patent i and citing patent j, where $m^{ij} = 1$ if i and j fall into the same geographic unit and $m^{ij} = 0$ otherwise. Second, the total number of citations is given by $N = \sum_{i=1}^{n^o} n_i^c$, where n^o is the number of originating patents and n_i^c is the number of patents that cite originating patent i. Finally, the observed matching rate is given by the total number of location matches divided by the total number of citations as

$$\hat{p}^c = \frac{1}{N} \sum_{i=1}^{n^o} \sum_{j=1}^{n_i^c} m^{ij}. \tag{1}$$

We then construct the distribution of the counterfactual matching rate $\hat{p}^r$ as follows. For each cited-citing relationship c^{ij}, we identify the admissible patent set that consists of the citing patent itself and the associated control patents (see the Appendix for an example). From the admissible patent set thus defined for each cited-citing relationship c^{ij}, we randomly draw a hypothetical patent to construct a counterfactual citation relationship r^{ij}. We then calculate the counterfactual matching rate, using a formula similar to (1) for the same N. After running 1000 Monte Carlo simulations, we finally obtain the simulated distribution of the matching rate $\{\hat{p}_k^r\}_{k=1}^{1000}$ and compute the p-value of the matching rate test by using the standard percentile method.

One should be careful about multiple inventors per patent. To determine whether or not a pair of cited and citing patents falls into the same geographic unit, we use the following two matching methods. Consider, for each cited-citing relationship, all possible pairs of an inventor of the cited patent and an inventor of the citing patent. The locations of the cited and citing patents are then matched (i) if the majority of all possible inventor pairs fall into the same geographic unit (median matching); or (ii) if at least one pair of inventors falls into the same geographic unit (minimum matching). These matching methods are in accordance with those used in previous studies. For example, JTH employ a similar method as our median matching. TFK mention the minimum matching as an alternative to their random matching.

2.5. The K-density approach

The matching rate approach in the previous subsection focuses on within-state or within-CMSA localization. However, the extent of LKS is unlikely to be limited by administrative boundaries. To capture cross-boundary localization, we now develop distance-based K-density tests. We address whether knowledge spillovers are localized, and examine to what extent they are localized. As before, we allocate a counterfactual citation between the originating patent and a patent drawn randomly from the corresponding admissible patent set. Yet, unlike the matching rate approach, we compare the distribution of distances between the originating and citing patents with the counterfactual distribution generated by the randomization. We then consider the deviation from randomness as evidence of LKS. Our distance-based test uses the same counterfactuals as the matching rate test, so that we can make a direct comparison between these two tests for localization.

Such an attempt, however, poses two main difficulties. First, patents can have multiple addresses because their inventors are not necessarily unique. We thus compute, for each cited-citing relationship, all possible distances between the inventors of the originating patent and those of the citing patent, and focus on their median or minimum distance. The distance computation is in line with the median or minimum matching method of the matching rate tests, respectively, as presented above. We do the same for the counterfactual citation relationship.

Second, because of the data limitation, the location of each inventor is identified at the census place level. However, census places are not spatial points. This poses a "zero distance" problem, i.e., even when the actual distance between the originating and citing inventors is not zero, it is measured to be zero if they happen to live in the same census place. To address this problem, we consider spatial interaction between the two inventors within the same census place. Assuming that each census place is a circle, we use the distance between the two randomly chosen points in census place ℓ with area S_ℓ, which is given by $[128/(45\pi)]\sqrt{S_\ell/\pi}$ (Kendall and Moran, 1963).

It is also noted that, unlike the previous studies on patent citations, we analyze the local-

ization distance that is specific to each patent class.[19] We thus classify all originating patents into different patent classes by their primary class. The citing patents that cite each originating patent may or may not belong to the same class as that originating patent. Taking this into account, we examine whether each patent class – to which originating patents belong – displays localization.[20]

We now describe the detailed procedure of our distance-based test for originating technology class A. First, for each cited-citing relationship c^{ij}, we compute the great-circle distance d^{ij} between originating patent i and citing patent j, where we consider the minimum or median distance as mentioned above. Note that citing patents need not belong to technology class A because we allow for both intra- and inter-class knowledge spillovers as in JTH and TFK. Second, the total number of citations that the originating patents in technology class A receive is given by $N_A = \sum_{i=1}^{n_A^o} n_i^c$, where n_A^o is the number of originating patents in technology class A and n_i^c is the number of patents that cite originating patent i. Finally, following DO, the K-density estimator of citation distance for technology class A at any point d is

$$\widehat{K}_A(d) = \frac{1}{2hN_A} \sum_{i=1}^{n_A^o} \sum_{j=1}^{n_i^c} f\left(\frac{d - d^{ij}}{h}\right), \tag{2}$$

where f is a Gaussian kernel function and h is the bandwidth set as in Silverman (1986).[21] Interestingly, expression (2) is a natural extension of the matching rate (1), and implies that, unlike DO, we consider unidirectional relationships from the inventors of originating patents to those of citing patents.[22]

The construction of counterfactuals is the same as that of the matching rate test. For

[19]Since the degree of localization tends to differ across industries (e.g., Ellison and Glaeser, 1997; Duranton and Overman, 2005), it seems natural to expect that the extent of LKS can also differ across patent classes.

[20]This procedure is common regardless of whether we use the 3- or 6-digit controls. In the latter case, we could examine whether each patent subclass exhibits localization. However, the number of subclasses is about $150,000$, which significantly reduces the number of location points where originating patents in each patent subclass are distributed. In such a case, we would not obtain well-behaved estimated density functions.

[21]As in DO, we adopt the reflection method in Silverman (1986) to deal with boundary problems associated with the fact that distances cannot be negative.

[22]Kerr and Kominers (2012) recently take a regression approach to cited-citing relationships. In contrast, we take the K-density approach to cited-citing relationships because expression (2) can be readily comparable to the matching rate (1). Note that both expressions can be used for the actual and counterfactual citations.

each cited-citing relationship c^{ij} that is used in (2), we identify the admissible patent set that consists of the citing patent itself and the associated control patents (see the Appendix for an example). From the admissible patent set thus defined for each cited-citing relationship c^{ij}, we randomly draw a hypothetical patent to construct a counterfactual citation relationship r^{ij}. We then estimate the K-density for the distribution of counterfactual citation distances, using a formula similar to (2). After running 1000 Monte Carlo simulations, we finally rank the counterfactual densities at each 10 km in ascending order and select the 5-th and the 95-th percentiles to obtain a lower 5% and an upper 5% confidence interval that we denote $\overline{K}_A(d)$ and $\underline{K}_A(d)$, respectively.[23]

Detecting localization based on $\overline{K}_A(d)$ and $\underline{K}_A(d)$, however, only allows us to make local statements at a given distance. We thus finally define the global confidence bands that we use to detect LKS. Let $\bar{d}_A$ be the maximum distance for technology class A.[24] We look for the identical upper and lower local confidence intervals such that, when we consider them across all distances between 0 and $\bar{d}_A$ km, only 5% of our randomly generated K-densities hit them. Let $\overline{\overline{K}}_A(d)$ be the upper global confidence band of technology class A. When $\widehat{K}_A(d) > \overline{\overline{K}}_A(d)$ for at least one $d \in [0, \bar{d}_A]$, this technology class is said to exhibit *global localization* at a 5% confidence level. Conversely, the lower global confidence band of technology class A, $\underline{\underline{K}}_A(d)$, is such that it is hit by 5% of the randomly generated K-densities that are not localized. A technology class is then said to exhibit *global dispersion* at a 5% confidence level when $\widehat{K}_A(d) < \underline{\underline{K}}_A(d)$ for at least one $d \in [0, \bar{d}_A]$ *and* the technology class does not exhibit global localization. The definition of global dispersion requires no global localization because otherwise dispersion at large distances could be a consequence of localization at smaller distances, given that our densities must sum to one. Hence, we define an index of global localization as $\Gamma_A(d) \equiv \max\{\widehat{K}_A(d) - \overline{\overline{K}}_A(d), 0\}$, and an index of global dispersion as $\Psi_A(d) \equiv \max\{\underline{\underline{K}}_A(d) - \widehat{K}_A(d), 0\}$ if $\sum_d \Gamma_A(d) = 0$ and $\Psi_A(d) \equiv 0$ otherwise.

[23] We also repeated our simulations 2000, 5000, and 10,000 times for several technology classes, and obtained very similar results.

[24] Following DO, we define the maximum distance as the median of all distances of all possible counterfactual citations for technology class A.

3. Results

The purpose of this section is threefold. Using the matching rate tests at the aggregate level, we first replicate the same qualitative features as those of JTH and TFK. We then turn to our K-density tests, and show that a substantial number of technology classes display localization, even when control patents are selected at the 6-digit level. We finally explore in details why the discrepancy arises between these two tests by comparing our class-specific distance-based tests with the matching rate tests at the disaggregate level.

3.1. The matching rate tests

Table 2 reports the results of the matching rate tests for the state, CMSA and county levels.[25] Following JTH and TFK, the matching rate tests are implemented at the aggregate level encompassing all technology classes. Using the 3- and 6-digit controls, we compare the observed matching rate with the average of the counterfactual matching rates for each spatial scale. The standard errors of the counterfactual matching rates are computed by simulation with 1000 replications. In the 3-digit case, the observed matching rates are significantly higher than the counterfactual ones for all spatial scales. We thus reject the null hypothesis of no LKS at a 5% significance level, and find solid evidence of LKS. Yet, the null hypothesis is not rejected for the 6-digit controls. These results replicate the qualitative features in JTH and TFK.

Insert Table 2

3.2. The K-density tests

We now describe the results of the K-density tests. Let $\mathscr{A}$ be the set of all technology classes. For technology class $A \in \mathscr{A}$, knowledge spillovers are said to exhibit *localization at distance* d if $\Gamma_A(d) > 0$, whereas they are said to exhibit *dispersion at distance* d if $\Psi_A(d) > 0$. We define a technology class A as having LKS if $\Gamma_A \equiv \sum_d \Gamma_A(d) > 0$, and as having dispersed knowledge spillovers if $\Psi_A \equiv \sum_d \Psi_A(d) > 0$. Finally, we use $L^1 = \{A \in \mathscr{A} | \Gamma_A > 0\}$ and $D^1 = \{A \in \mathscr{A} | \Psi_A > 0\}$ to denote the sets of technology classes displaying localized and dispersed knowledge spillovers, respectively. Table 3 presents the results. First, concerning

[25] As in TFK, we use 16 CMSAs as defined in 1981 by excluding Puerto Rico.

the 3-digit case, we find LKS for the majority of technology classes, with about 70% being localized for both the median and minimum distances. These results are in line with those obtained by JTH. Turning to the 6-digit controls, more than 30% of technology classes exhibit LKS regardless of whether we use the median or minimum distance. Although fewer classes exhibit localization in the 6-digit case, we obtain solid evidence for LKS. This is surprising given that TFK find no evidence supporting localization at the state and CMSA levels.

Insert Table 3

To investigate more closely the scope of LKS, let $L^1(d) = \{A \in \mathscr{A} | \Gamma_A(d) > 0\}$ be the set of technology classes that exhibit localization at distance d. Figure 1 illustrates the distributions of $|L^1(d)|$ for the 3- and 6-digit controls. In each case, there is no substantial difference between the median (solid) and the minimum (dotted) distance methods. The number of localized technology classes is greater at smaller distances for both the 3- and 6-digit controls. The degree of localization decreases as the distance from the originating patents increases, thus suggesting that knowledge spillovers decay with distance. This result is consistent with the assumption in the recent theory of spatial development (Desmet and Rossi-Hansberg, 2009).[26]

Insert Figures 1 and 2

We can delineate a boundary within which knowledge spillovers are localized. Figure 2 shows the percentages of technology classes displaying localization at least once within distance d. There are substantial differences between the 3- and 6-digit cases. However, no matter which control is used, more than half of the technology classes displaying LKS are localized at least once within about 200 km, which corresponds roughly to the distance between Boston and New Haven. We can also consider 1200 km as the widest extent of LKS because more than 95% of all localized classes are localized by this distance, regardless of which controls are used.

Finally, we examine heterogeneity in the patterns of knowledge spillovers across technology classes (see the working paper version, Murata et al., 2011, for further discussion). Figure 3 illustrates the distributions of Γ_A and Ψ_A for the median distance case.[27] Interestingly, for

[26]There is no clear pattern for dispersed knowledge spillovers, although we observe some significant dispersion across various distances. Such dispersion of citing inventors may arise, for instance, when the benefits of their pooling is dominated by the costs of their poaching from firms' perspectives (Combes and Duranton, 2006).

[27]The results are fairly robust regardless of the choice between the median and the minimum distances. The

the 3-digit controls, the fraction of localized technology classes outweighs substantially that of dispersed technology classes. By contrast, in the 6-digit case, the corresponding difference between the localized and dispersed technology classes is not so large.[28]

Insert Figure 3

3.3. Comparison

We have shown that, unlike the matching rate tests, the K-density tests provide solid evidence for LKS, even for the 6-digit controls. We now highlight the differences between these two approaches. We argue that the matching rate tests using the 6-digit controls underestimate localization of knowledge spillovers due to the following two "aggregation" problems.

The first problem is "technological aggregation". As shown above, the K-density tests reveal considerable heterogeneity across technology classes in whether knowledge spillovers are localized or dispersed. This is particularly so, in the 6-digit case, where the distributions of Γ_A and Ψ_A are roughly similar. Accordingly, if these heterogeneous classes are pooled, as in the conventional matching rate tests, both localization and dispersion can be cancelled out with each other, and, thus, may leave no evidence of localization at the aggregate level.

To confirm this idea, we implement class-specific matching rate tests that are analogous to class-specific distance-based tests. Specifically, we test the hypothesis of no LKS at the 5% significance level for each technology class. Let $L_1 = \{A \in \mathscr{A} | p_A^c > p_A^r\}$ denote the set of technology classes that exhibit localization by the class-specific matching rate tests, where p^c and p^r depend on technology class A. Table 4 shows that, for the 3-digit controls, LKS are detected for 270 or 266 technology classes, depending on whether the spatial units are states or CMSAs. Since these numbers are fairly close to the 275 localized classes, obtained from the K-density tests in Table 3, we conclude that the matching rate and the K-density tests

latter results are available upon request from the authors.

[28]We can further explore heterogeneity in technology classes. Since originating technology classes have different ratios of inter-class to intra-class spillovers, we can sort those classes in descending order of the fraction of inter-class spillovers, and divide them into two – the top 50% and bottom 50% groups of the distribution. We then find that the distance-based tests with the 6-digit controls are less likely to detect localization for technology classes with greater inter-class spillovers when compared to those with the 3-digit controls. The same applies when focusing on the top and bottom 25% technology classes with respect to the fraction of inter-class spillovers. Both results are available upon request from the authors.

detect roughly the same number of localized technology classes for the 3-digit controls.[29]

However, for the 6-digit controls, the class-specific matching rate tests detect a substantially smaller number of localized technology classes than the K-density tests. More concretely, only 47 to 69 technology classes display localization in the former, depending on the spatial units, whereas more than 100 technology classes are localized in the latter tests. Yet, even in the class-specific matching rate tests, the percentages of technology classes with LKS remain in the range between 13% and 20%. Hence, we find evidence that knowledge spillovers are localized for nonnegligible technology classes even in the 6-digit case.

Insert Table 4

The second problem of the matching rate tests is "geographic aggregation". The matching rate tests allocate inventors to spatial units such as states and CMSAs. As DO pointed out, this aggregation deals with administrative units symmetrically, so that inventors in neighboring spatial units are treated in exactly the same way as inventors at the opposite ends of a country. This creates a downward bias when dealing with cross-border LKS. The distance-based tests have an advantage in that they do not overlook such knowledge flows.[30]

To investigate this possibility, we focus on the discrepancy between the matching rate and the K-density tests for the 6-digit controls. We first implement the matching rate tests for the two groups of technology classes, that is, the set of localized technology classes by the K-density tests, $L^1 = \{A \in \mathscr{A} | \Gamma_A > 0\}$, and the set of nonlocalized technology classes, $L^0 = \{A \in \mathscr{A} | \Gamma_A = 0\}$. We then define $L_0^1 = \{A \in \mathscr{A} | p_A^c = p_A^r \text{ and } \Gamma_A > 0\}$ as the set of technology classes where the K-density tests detect significant localization, while the matching rate tests do not. Thus, $L_0^1 \subseteq L^1$. Similarly, we define $L_1^0 = \{A \in \mathscr{A} | p_A^c > p_A^r \text{ and } \Gamma_A = 0\} \subseteq L^0$.

Table 5 provides the results. First, looking at the results of $|L_0^1|$ in the first and second rows, a large number of technology classes that are detected as localized by the K-density tests are not identified as localized by the matching rate tests. We thus find that the matching rate

[29]Table 4 shows the results for the median matching case. The results for the minimum matching case are qualitatively similar, and, thus, are omitted. They are available upon request from the authors.

[30]In this respect, the regression approach proposed in the recent discussion paper by Singh and Marx (2012) is similar to our approach. In particular, they have found that knowledge spills over across administrative boundaries even when country and state borders are controlled for. We thank a referee for bringing our attention to this related paper.

tests underestimate LKS. The number of underestimated technology classes ranges from 67 to 89, depending on the spatial units. These biases are substantial since the percentage of underestimated classes is as high as 61% to 62% at the state and CMSA levels, respectively, and it amounts to 81% at the county level. Moving to the results of $|L_1^0|$ in the third and fourth rows, a number of technology classes that are not detected as localized by the K-density tests are identified as localized by the matching rate tests. Thus, the matching rate tests can also overestimate LKS. Yet, the numbers of underestimated localized classes, $|L_0^1|$, much outweigh those of overestimated localized classes, $|L_1^0|$. The difference ranges from 40 to 62, which explains the difference between $|L^1|$ in Table 3 and $|L_1|$ in Table 4 for the 6-digit controls.

Insert Table 5

We can investigate where we observe the downward biases of the matching rate tests using the 6-digit controls in detecting LKS. Figure 4 plots $|L_0^1(d)|$ for each distance d, where $L_0^1(d) = \{A \in \mathscr{A} | p_A^c = p_A^r \text{ and } \Gamma_A(d) > 0\}$. The downward biases tend to be most substantial around 200 km or 500 km, depending on whether we focus on counties or on CMSAs and states. For example, the county-level matching rate tests fail to detect localization for about 40 technology classes at 200 km. This underestimation is inherent in their construction as the matching rate tests cannot discern knowledge spillovers that travel longer than their predetermined administrative boundaries. For example, given that the average of within-area distances for the U.S. states is 197.9 km, LKS whose scope significantly exceeds that distance are unlikely to be captured by the state-level matching rate test. In this light, the matching rate tests with smaller spatial units, which have the smaller average of within-area distances, tend to more severely underestimate LKS that can be detected by the K-density tests.

Insert Figure 4

In order to further elucidate the biases of the matching rate tests from omitting cross-boundary knowledge spillovers, we conduct augmented matching rate tests in which a pair of patents are counted as geographically matched if they are either in the same geographic unit, or in the adjacent units. The results are presented in Figure 5, where we plot in the solid line the number of technology classes, $|L_0^1(d)|$, for which the K-density tests detect localization at

each distance d while the augmented matching rate tests do not. For comparison, we draw the results for the original matching rate tests in the dotted line.[31]

The downward biases of the original matching rate tests are alleviated as the augmented matching rate tests partially capture localization across administrative boundaries. The bias reduction is most significant around 400 km at the state level, and it is around 200 km at the county level. Yet, the augmented matching rate tests still fail to detect a significant fraction of localized technology classes. This provides a rationale for employing the distance-based tests, rather than the matching rate tests, no matter whether to include neighboring units or not, in order to detect LKS.

Insert Figure 5

In summary, the existing matching rate tests systematically understate LKS. We explain this by two aggregation problems, namely, technological and geographic aggregations. If we control for heterogeneity in localization and dispersion by disaggregating technology classes, the matching rate tests provide evidence of LKS for a fraction of technology classes. Yet, they still fail to identify a substantial number of localized technology classes that are detected by the distance-based K-density tests. Our analysis also suggests that the matching rate tests with smaller administrative units tend to exacerbate the underestimation problem. In view of this, the geographic aggregation problem with the matching rate tests cannot be resolved, even when taking smaller administrative units such as counties. Rather, in that case, the downward biases become more substantial.

4. Sensitivity analysis

We have so far constructed counterfactual citations by drawing patents randomly from the admissible patent set. This amounts to assuming that citing and control patents are equally likely to cite the originating patent (see the Appendix for an example of citation probabilities). This assumption relies on the premise that the control patents perfectly mimic the citing patents, except that the former do not cite the originating patents while the latter do.

[31]The augmented matching rate tests are performed at the state and county levels, but not at the CMSA level, because most CMSAs are not adjacent with each other.

However, TFK argue that the 3-digit patent classes are too broad and noisy for the purpose of identifying control patents, whereas Henderson, Jaffe and Trajtenberg (2005) state that there is no systematic evidence supporting that the 6-digit subclass classification renders "closer" technologically matched controls.

This section generalizes our analysis, provided that neither the 3-digit controls nor the 6-digit controls are perfect. As we will illustrate in Figure 6 below, the generalized framework relies on the unit simplex regarding citation probabilities. The simplex includes the previous JTH and TFK cases with perfect controls as a single point (i.e., the point denoted by either JTH or TFK). Furthermore, we will show that more general cases with imperfect controls can be depicted as a hexagon, instead of a single point, on the simplex. The goal of this section is to conduct the distance-based tests for each vertex of the hexagon and to create the bounds of the percentage of localized technology classes under imperfect controls. We are particularly interested in whether the lower bound exceeds 50%, i.e., whether the majority of technology classes are localized.

The assumption that the 3- and 6-digit controls are imperfect is relevant because, as argued in Henderson, Jaffe and Trajtenberg (2005), it is exceedingly difficult to perfectly identify anything akin to well-circumscribed technologies by using the USPTO patent classification system only.[32] Thus, the patent classification system, no matter how strict the criteria, can be used as just a proxy for the true technological environment. Hence, there may exist unobserved factors in matching between the citing and control patents.

To address how sensitive our localization results are to various magnitudes of unobserved factors, we perform Rosenbaum's (2002) sensitivity analysis.[33] More specifically, we reconstruct counterfactual citations in the presence of imperfect controls, and show that citing and control patents need not be drawn with equal probability. Using these generalized counter-

[32]Henderson, Jaffe and Trajtenberg (2005) state that "the patent classification system has been morphing and growing over time in response to the evolving needs of patent examiners faced with fast-changing technologies ... the subclass classification layer has changed quite rapidly, and it consists by now of about 150,000 patent subclasses". Moreover, the USPTO patent classification system is not the unique system for classifying a myriad of patents. The International Patent Classification (IPC) provides a different classification system.

[33]See Imbens (2003), Altonji, Elder and Taber (2005) or Ichino, Mealli and Nannicini (2008) for recent applications of Rosenbaum's sensitivity analysis to program evaluations.

factual citations, we conduct both the matching rate and distance-based tests, where we use the median matching and median distance, respectively. In doing so, we deal with the 3- and 6-digit controls simultaneously by considering that matching on subclasses implies matching on classes. This approach encompasses our previous analysis as limiting cases, and provides some robust bounds of localization results. In particular, we obtain the lowest possible percentage of localized technology classes for a given magnitude of hidden biases, and show that the majority of technology classes exhibit localization unless the magnitude of hidden factors is extremely large. We further confirm that, even with imperfect controls, the matching rate tests systematically underestimate the percentage of localized technology classes when compared with the distance-based tests.

To see this, we first restate the tests of LKS in terms of matching estimators.[34] Let m be a dummy variable indicating whether a pair of patents match at the same geographic unit or not. Denote by t a treatment assignment dummy that takes one if there is a citation link between a pair of patents. Then, the matching rate test measures the mean difference of the match variable m between a treatment group ($t = 1$) and a non-treatment group ($t = 0$), conditional on the propensity score. That is, we compare $\mathrm{E}(m|t = 1, p(x))$ with $\mathrm{E}(m|t = 0, p(x))$, where x is a vector of technology class dummies, and $p(x)$ is the propensity score defined as the probability that the patent with technology class x receives treatment. Similarly, letting d be the geographic distance between a pair of patents, the distance-based test detects any significant difference in the density at distance d between treatment and non-treatment groups, conditional on the propensity score. That is, we compare $K(d|t = 1, p(x))$ with $K(d|t = 0, p(x))$, where K is a conditional density function of citation distance d.

The basic premise of these localization tests is the conditional independence assumption, i.e., the outcomes, m and d, are independent of treatment assignment t, conditional on the technology class x. If this assumption holds, then the potential outcome is independent of treatment, conditional on the propensity score $p(x)$ (see, e.g., Angrist and Pischke, 2009; Wooldridge, 2010). However, if patent classes fail to control technological activities, the treat-

[34] A similar idea can be found in Thompson and Fox-Kean (2005*b*).

ment assignment is influenced by hidden factors. Then, a pair of patents having the same technology class x have different probabilities $p(x)$ of receiving treatments. Accordingly, the outcomes between the treatment and non-treatment groups are not comparable, and the localization tests will be biased (see, e.g., Imbens, 2004).

4.1. The case with a single control group

Consider an admissible patent set that consists of the citing and control patents that share the same 3-digit patent class. In general, each citing patent has multiple control patents, but, for the moment, we assume that the control patent is unique, so that the set is given by $\{b, c\}$, where b denotes the 3-digit control patent corresponding to citing patent c. Following Rosenbaum (2002), the treatment assignment probability of a patent in the admissible patent set conditional on technology class x_r is given by

$$p_r = \text{Prob}(t_r = 1|x_r) = F(\kappa(x_r) + \lambda u_r), \tag{3}$$

where κ is an unknown function of technology class, $u_r \in [0, 1]$ is an unobserved factor, λ is the effect of u_r on the citation probability, and F is the logistic distribution function. As the control patent b and the citing patent c share the same technology class, $x_b = x_c = x$ must hold. Hence, the assignment probability p_r is nothing but the propensity score $p(x)$.

If there is no hidden bias ($\lambda = 0$), the treatment assignment probabilities are the same between citing and control patents, $p_b = p_c = F(\kappa(x))$, because $x_b = x_c = x$. This provides a rationale for why we draw a hypothetical patent randomly from the admissible patent set with equal chances. However, if hidden bias exists ($\lambda \neq 0$), the difference in unobservables, $u_b \neq u_c$, implies different assignment probabilities for citing and control patents, $p_b \neq p_c$. We take this into account in the modified simulation process by drawing citing and control patents from the admissible patent set with different probabilities, reflecting the magnitudes of hidden biases.

4.2. The case with multiple control groups

So far, we have illustrated the effect of hidden biases on the localization tests in the case of a single control group (either the 3- or 6-digit control). We now turn to a general class of Rosenbaum's sensitivity analysis that encompasses multiple control groups (both the 3- and 6-digit controls). Let $\mathbf{b}_3$ be the set of 3-digit controls that match the citing patent c at the 3-digit level. We allow the number of controls to be multiple, $n_3 = |\mathbf{b}_3|$. Note that a pair of patents that match at the 6-digit level also match at the 3-digit level by construction of the 3- and 6-digit codes. We thus have $\mathbf{b}_3 = \mathbf{b}_6 \cup \mathbf{b}_{3\backslash 6}$, where $\mathbf{b}_6$ is the set of 6-digit controls and $\mathbf{b}_{3\backslash 6}$ is the set of controls that match the citing patent at the 3-digit level but not at the 6-digit level.[35] Let $n_6 = |\mathbf{b}_6|$ and $n_{3\backslash 6} = |\mathbf{b}_{3\backslash 6}|$ with $n_3 = n_6 + n_{3\backslash 6}$, $n_6 \geq 1$ and $n_{3\backslash 6} \geq 1$. Then, the admissible patent set at the 3-digit level is given by $\{\mathbf{b}_6, \mathbf{b}_{3\backslash 6}, c\}$.

Let p_6, $p_{3\backslash 6}$, and p_c be the treatment assignment probabilities for $\mathbf{b}_6$, $\mathbf{b}_{3\backslash 6}$, and citing patent c, respectively. Since the originating patent could have been cited by any patent in the admissible patent set, the treatment assignment probabilities must satisfy the restriction:

$$n_6 p_6 + n_{3\backslash 6} p_{3\backslash 6} + p_c = 1. \tag{4}$$

When the 3-digit control is perfect, $p_6 = p_{3\backslash 6} = p_c$ holds, whereas we have $p_6 = p_c$ but $p_{3\backslash 6} \neq p_c$ when the 6-digit control is perfect. Each control patent is thus comparable to the citing patent in some ways but need not be in other ways. Rosenbaum (2002, Ch. 7) calls this property "partial comparability".

Following Rosenbaum (2002) we express partial comparability as a restriction on hidden factors in the treatment assignment probabilities (3). Let x_6 and $x_{3\backslash 6}$ be 3-digit technology class dummies for $\mathbf{b}_6$ and $\mathbf{b}_{3\backslash 6}$, respectively. Since any patent in the admissible patent set shares the same 3-digit code, the observed factors are perfectly comparable, i.e., $x_6 = x_{3\backslash 6} = x_c$.

[35] In this sensitivity analysis with multiple control groups, we remove the restriction, which is applicable only to the 6-digit controls, that control patents must share any subclass in common with originating patents. This allows us to analyze both the 3- and 6-digit controls on a common ground. Alternatively, one could impose the restriction that the 3-digit controls must also share any subclass in common with originating patents. Yet, this restriction makes $\mathbf{b_3}$ for the sensitivity analysis very different from the set of JTH's controls. Indeed, $\mathbf{b}_3$ contains only about one percent of the original 3-digit controls that we have used in the previous sections.

In contrast, the unobserved terms are partially comparable. As in Rosenbaum (2002), u_r is given by a weighted sum of unobserved factors, $v_r \in [0,1]$ and $w_r \in [0,1]$, as $u_r = (1-\phi)v_r + \phi w_r$, where $\phi \in [0,1]$. We impose the restriction $w_6 = w_c$, while allowing for $w_{3\backslash 6} \neq w_c$. In words, the 6-digit controls and the citing patent share some unobserved similarities that are not shared by the 3-digit controls.

The partial comparability parameter ϕ plays a role in reducing uncertainty in hidden factors. To see this, letting $q_r \equiv p_r/(1-p_r)$ and using (3), we compute the odds ratios: $q_6/q_c = \exp[\lambda(1-\phi)(v_6 - v_c)]$; $q_{3\backslash 6}/q_c = \exp[\lambda(u_{3\backslash 6} - u_c)]$; and $q_6/q_{3\backslash 6} = \exp[\lambda(u_6 - u_{3\backslash 6})]$. Since $0 \leq u, v \leq 1$, the bounds of the odds ratios are given by

$$\Lambda^{\phi-1} \leq \frac{q_6}{q_c} \leq \Lambda^{1-\phi} \tag{5}$$

$$\Lambda^{-1} \leq \frac{q_{3\backslash 6}}{q_c} \leq \Lambda \tag{6}$$

$$\Lambda^{-1} \leq \frac{q_{3\backslash 6}}{q_6} \leq \Lambda, \tag{7}$$

where $\Lambda = \exp(\lambda)$. Since $\Lambda^{1-\phi} \leq \Lambda$ for $0 \leq \phi \leq 1$, the bounds of q_6/q_c are narrower than the others due to the restriction $w_6 = w_c$.

Figure 6 depicts feasible probability distributions $(p_{3\backslash 6}, p_6, p_c)$ implied by the bounds of the odds ratios (5)–(7) on the simplices for different values of parameters (Λ, ϕ), where we set $n_6 = n_{3\backslash 6} = 1$ for illustrative purposes. When $\Lambda = 1$, only $p_{3\backslash 6} = p_6 = p_c = 1/3$ — the centroid of the equilateral triangle — is feasible regardless of the value of ϕ. As denoted by JTH in Figure 6 (a), this point corresponds to the JTH case, where the 3-digit control and citing patent are equally likely to cite the originating patent. In contrast, when $\phi = 1$ and $\Lambda = \infty$, the feasible probability set is given by the line segment such that $\{(p_{3\backslash 6}, p_6, p_c) | p_6 = p_c\}$, i.e., the 6-digit control and citing patent cite the originating patent with equal likelihood. Indeed, TFK explore the admissible patent set corresponding to one of the end points of the segment. As denoted by TFK in Figure 6 (d), this point implies $p_{3\backslash 6} = 0$ and $p_6 = p_c = 1/2$.

Insert Figure 6

We consider a more general case where $1 \leq \Lambda \leq \infty$ and $0 \leq \phi \leq 1$ to examine how sensitive

our results of LKS are to various values of parameters (Λ, ϕ). Then, as seen in Figures 6 (b) and (c), the set of feasible probability distributions can be depicted as a hexagon with six vertices, each of which is characterized by a pair of bounds given by (5)–(7). For each vertex, we can obtain the treatment assignment probabilities by noting that (4) can be rewritten as

$$n_6 \left(\frac{q_6}{1+q_6} \right) + n_{3\backslash 6} \left(\frac{q_{3\backslash 6}}{1+q_{3\backslash 6}} \right) + \frac{q_c}{1+q_c} = 1. \tag{8}$$

First, to obtain vertex ① in Figures 6 (b), consider the upper bounds of (5) and (6), i.e., $q_6/q_c = \Lambda^{1-\phi}$ and $q_{3\backslash 6}/q_c = \Lambda$. Plugging these expressions into (8) and rearranging the terms yield the cubic equation for q_c: $A_3(q_c)^3 + A_2(q_c)^2 + A_1(q_c) + A_0 = 0$, where the coefficients are given by: $A_3 > 0$; $A_2 > 0$; $A_1 > 0$; and $A_0 < 0$. We can show that the equation has the unique solution for $q_c \geq 0$. Given the solution q_c, we find $q_6 = \Lambda q_c$ and $q_{3\backslash 6} = \Lambda^{1-\phi} q_c$. The assignment probability p_r is then computed by $p_r = q_r/(1+q_r)$. The assignment probabilities for the other five vertices are analogously obtained.[36] To obtain the bounds of the percentage of localized technology classes, we finally conduct the matching rate and distance-based tests of localization for each set of assignment probabilities associated with each vertex.

4.3. Results

Figure 7 presents the sensitivity analysis for the K-density tests. Each panel illustrates, for a fixed value of Λ, the estimated percentages of localized technology classes with different values of ϕ. The six lines in each panel correspond to the vertices of the hexagon in Figure 6. As Λ increases, the difference between the upper and lower bounds of localized technology classes gets larger, reflecting increasing uncertainty in the admissible patent set. If there were no hidden bias ($\Lambda = 1$), localization would be observed for about 70% of technology classes (not graphed), which is comparable to the previous localization result for the 3-digit controls.

Insert Figures 7 and 8

Figure 8 presents the sensitivity analysis for the matching rate tests at the state level. The overall patterns are roughly similar to those for the K-density tests. However, for a given

[36] See the working paper version of our paper, Murata et al. (2011).

set of parameter values, (Λ, ϕ), the matching rate tests yield lower percentages of localized technology classes than the K-density tests. In particular, we find that the underestimation is more noticeable for larger hidden biases. For example, when $\Lambda = 16$, the lower bound for the matching rate tests is 50%, whereas that for the K-density tests is 56%.[37] This confirms our previous finding that, with the 6-digit controls, the matching rate tests understate LKS. Our underestimation result thus remains true, even with a more general choice of control patents.

Figures 7 and 8 show that the lower bound of the percentage of localized technology classes decreases as the magnitude of hidden biases, Λ, increases. Figure 9 further investigates this relationship. For a given value of Λ, the worst-case scenario bound is computed as the lowest percentage of localized technology classes within the range of $\phi \in [0, 1]$.[38] As shown, the worst-case scenario bound for the matching rate tests is uniformly lower than that for the K-density tests. Again, the matching rate tests understate LKS. Focusing on the K-density tests, the worst-case scenario bound exceeds 50% even at $\Lambda = 25$. Thus, even if we allow for significant unobserved factors that make the odds of receiving a citation differ by a factor of 25 between the actual citing patents and the control patents – an extreme departure from no hidden factor – LKS remain dominant. In this light, the K-density tests with the 6-digit controls, which show that only about 30% of technology classes are localized, are rather extreme because they constitute a limiting case of the worst-case scenario bound when $\Lambda \to \infty$. In a nutshell, our sensitivity analysis provides solid evidence of localization unless hidden biases are extremely large.

Insert Figure 9

5. Conclusion

We have proposed a distance-based approach to LKS and revisited the recent debate by Thompson and Fox-Kean (2005*a*,*b*) and Henderson, Jaffe and Trajtenberg (2005) on the existence of LKS. Our concern has been two aggregation problems, namely technological and

[37] We also conduct the sensitivity analysis at the CMSA and county levels. The results are qualitatively similar to those at the state level, although the percentages of localized technology classes are somewhat smaller for more disaggregated geographic units: 47% at the CMSA level; and 46% at the county level. The more detailed results are available from the authors upon request.

[38] Our worst-case scenario bound is related to the bounding approach proposed by Manski (2007).

geographic aggregations, both of which are ignored in that literature. Overcoming these two problems, our distance-based tests have found solid evidence supporting LKS for a substantial number of technology classes, even when the 6-digit controls are used. At the same time, nonnegligible technology classes exhibit dispersion, thus implying considerable heterogeneity across classes. We show that the class-specific matching rate tests for the 6-digit controls understate the number of localized technology classes that are detected by the distance-based tests. These aggregation biases may thus explain why the matching rate tests, implemented by TFK, could not find any significant evidence for intranational knowledge spillovers.

To compare our distance-based tests with the conventional matching rate tests by JTH and TFK, we have relied on typical case-control methods by specifying the technology level at which control patents are selected. However, as discussed by Thompson and Fox-Kean (2005*a*,*b*) and Henderson, Jaffe and Trajtenberg (2005), neither the 3-digit control nor the 6-digit control is perfect due to unobserved heterogeneity within classes or subclasses. Therefore, we have developed a new framework to detect localization even when these controls are imperfect. It is worth emphasizing that, even with imperfect controls, our sensitivity analysis shows that the majority of technology classes exhibit localization. Since our approach does not require additional data such as the information on examiner added citations, it can be readily used to settle the debate over the existence of LKS between JTH and TFK who rely on the 1975-1999 data for which that information is not available.[39]

Finally, following JTH and TFK, we have abstracted from underlying forces that generate LKS. Kerr and Kominers (2012) have recently made an important attempt to provide a micro-foundation capturing benefits and costs of interactions that determine the extent of knowledge flows and the resulting shapes of agglomeration clusters. Developing theoretical frameworks that can account for LKS is left for future research.

[39]To cope with imperfect controls, Thompson (2006) develops an alternative way that does not involve case controls. However, this requires more recent data that can distinguish citations added by inventors from those added by examiners. Although Thompson (2006) shows that inventor citations are more likely to match the state or CMSA of their originating patents than examiner citations, this result may be biased as well, given our result that the matching rate tests are subject to the two aggregation problems.

References

Agrawal, A., Cockburn, I., and Rosell, C. (2010), "Not Invented Here? Innovation in company towns", *Journal of Urban Economics*, 67: 78–89.

Agrawal, A., Kapur, D., and McHale, J. (2008), "How do spatial and social proximity influence knowledge flows? Evidence from patent data", *Journal of Urban Economics*, 64: 258–268.

Almeida, P. (1996), "Knowledge sourcing by foreign multinationals: Patent citation analysis in the U.S. semiconductor industry", *Strategic Management Journal*, 17: 155–165.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005), "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools", *Journal of Political Economy*, 113: 151–184.

Angrist, J. D., and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton: Princeton University Press).

Combes, P.-P., and Duranton, G. (2006), "Labour pooling, labour poaching, and spatial clustering", *Regional Science and Urban Economics*, 36: 1–28.

Desmet, K., and Rossi-Hansberg, E. (2009), "Spatial development", Working Paper 2009-18, Instituto Madrileño de Estudios Avanzados.

Duranton, G., and Overman, H. (2005), "Testing for localization using micro-geographic data", *Review of Economic Studies*, 72: 1077–1106.

Duranton, G., and Overman, H. (2008), "Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data", *Journal of Regional Science*, 48: 213–243.

Ellison, G. D., and Glaeser, E. L. (1997), "Geographic concentration of in US manufacturing industries: A dartboard approach", *Journal of Political Economy*, 105: 889–927.

Ellison, G. D., Glaeser, E. L., and Kerr, W. R. (2010), "What causes industry agglomeration? Evidence from coagglomeration patterns", *American Economic Review*, 100: 1195–1213.

Hall, B., Jaffe, A., and Trajtenberg, M. (2001), "The NBER patent citation data file: Lessons, insights and methodological tools", NBER Working Paper #8498.

Henderson, R., Jaffe, A., and Trajtenberg, M. (2005), "Patent citations and the geography of knowledge spillovers: A reassessment: comment", *American Economic Review*, 95: 461–464.

Ichino, A., Mealli, F., and Nannicini, T. (2008), "From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?", *Journal of Applied Econometrics*, 23: 305–327.

Imbens, G. W. (2003), "Sensitivity to exogeneity assumptions in program evaluation", *American Economic Review*, 93: 126–132.

Imbens, G. W. (2004), "Nonparametric estimation of average treatment effects under exogeneity: A review", *Review of Economics and Statistics*, 86: 4–29.

Jaffe, A., Trajtenberg, M., and Henderson, R. (1993), "Geographic localization of knowledge spillovers as evidenced by patent citations", *Quarterly Journal of Economics*, 108: 577–598.

Kendall, M., and Moran, P. (1963), *Geometrical Probability* (London: Charles Griffin & Company Limited).

Kerr, W. R., and Kominers, S. D. (2012), "Agglomerative forces and cluster shapes", Working Papers 12-09, Center for Economic Studies, U.S. Census Bureau.

Manski, C. F. (2007), *Identification for Prediction and Decision* (Cambridge: Harvard University Press).

Murata, Y., Nakajima, R., Okamoto, R., and Tamura, R. (2011), "Localized knowledge spillovers and patent citations: A distance-based approach", GRIPS Discussion Papers 11-11.

Nakajima, R., Tamura, R., and Hanaki, N. (2010), "The effect of collaboration network on inventors' job match, productivity and tenure", *Labour Economics*, 17: 723–734.

Rosenbaum, P. R. (2002), *Observational Studies, Second Edition* (New York: Springer-Verlag).

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis* (New York: Chapman and Hall).

Singh, J., and Marx, M. (2012), "Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity", Mimeographed.

Thompson, P. (2006), "Patent citations and the geography of knowledge spillovers: Evidence from inventor- and examiner-added citation", *Review of Economics and Statistics*, 88: 383–388.

Thompson, P., and Fox-Kean, M. (2005*a*), "Patent citations and the geography of knowledge spillovers: A reassessment", *American Economic Review*, 95: 450–460.

Thompson, P., and Fox-Kean, M. (2005*b*), "Patent citations and the geography of knowledge spillovers: A reassessment: Reply", *American Economic Review*, 95: 465–466.

Trajtenberg, M., Shiff, G., and Melamed, R. (2006), "The NAMES GAME: Harnessing inventors' patent data for economic research", NBER Working Paper #12479.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data, Second Edition* (Cambridge: MIT Press).

Appendix. The admissible patent set: An example

Consider originating patent 4164057 (Food casing stuffing sizing control method) applied in October 1977 for technology class 452 (Butchering) and subclass 38 (Sizing ring). It received a citation from patent 4649602 (Stuffing method, apparatus and article for use therewith) applied in July 1986 for technology class 452 and subclass 38.

For this citing patent, we find two patents that share the same 3-digit code but do not cite the originating patent, namely, patent 4662028 (Apparatus for splitting animal heads) applied in July 1986 for technology class 452 and subclass 160 (Cutting longitudinally through body or body portion) and patent 4683617 (Disposable tension sleeve for a stuffing machine) applied in August 1986 for technology class 452 and subclass 38.

For the cited-citing relationship 4164057-4649602, the admissible patent set can be defined as follows. In the 3-digit case, the set includes citing patent 4649602 itself and control patents 4662028 and 4683617 because they share the same technology class 452. These patents in the admissible patent set $\{4649602, 4662028, 4683617\}$ are equally likely to cite the originating patent with probabilities being $(1/3, 1/3, 1/3)$.

In the 6-digit case, the admissible patent set includes citing patent 4649602 and control patent 4683617. However, the set does not include patent 4662028 because it does not belong to subclass 38. The two patents in the admissible patent set $\{4649602, 4683617\}$ are equally likely to cite the originating patent with probabilities being $(1/2, 1/2)$. In both the 3- and 6-digit cases, we draw hypothetical patents randomly from the respective admissible patent set with equal probability.

The admissible patent set for the sensitivity analysis in Section 4 consists of patents 4649602, 4662028, and 4683617 as we encompass the 3- and 6-digit cases. Unlike the previous cases, however, our sensitivity analysis allows for different probabilities across $\{4649602, 4662028, 4683617\}$ to be drawn as hypothetical patents. The probabilities depend on the parameter for the magnitude of hidden biases Λ and the partial comparability parameter ϕ as explained in Section 4.2.

Tables and figures.

Table 1: Sample Patent Sizes

	Total	3-digit	6-digit
Originatings	115,905	107,561	59,168
Percent	(100.00)	(92.64)	(51.04)
Citings	647,983	390,104	120,876
Percent	(100.00)	(60.20)	(18.65)
Controls	—	33,472,826	941,532

Notes: The first column reports the total numbers of the originating and citing patents, whereas the second and third columns report the numbers of the originating and citing patents having at least one control.

Table 2: Matching Rate Test Results

		3-digit Control		6-digit Control	
		Median	Minimum	Median	Minimum
State	Observed Rate (%)	12.53*	13.54*	13.38	14.31
	Counterfactual Rate (%)	9.33	10.16	13.45	14.49
	Std. Error	(0.04)	(0.04)	(0.07)	(0.06)
CMSA	Observed Rate (%)	9.24*	10.29*	10.12	11.18
	Counterfactual Rate (%)	6.54	7.32	10.33	11.37
	Std. Error	(0.03)	(0.03)	(0.06)	(0.06)
County	Observed Rate (%)	4.08*	5.27*	4.34	5.62
	Counterfactual Rate (%)	2.54	3.31	4.63	5.88
	Std. Error	(0.02)	(0.02)	(0.04)	(0.05)

Notes: $*$ denotes statistically significant at 5% level.

Table 3: K-density Test Results

	3-digit Control		6-digit Control	
	Median	Minimum	Median	Minimum
All Classes $\lvert\mathscr{A}\rvert$	384	384	360	360
Localized Classes $\lvert L^1\rvert$	275	273	109	109
Dispersed Classes $\lvert D^1\rvert$	39	40	41	51
$\lvert L^1\rvert/\lvert\mathscr{A}\rvert \times 100$ (percent)	(71.61%)	(71.09%)	(30.28%)	(30.28%)

Notes: $\lvert L^1\rvert$ is the number of technology classes that exhibit localized knowledge spillovers, and $\lvert D^1\rvert$ is the number of technology classes that exhibit dispersed knowledge spillovers.

Table 4: Matching Rate Test Results for Disaggregated Technology Classes

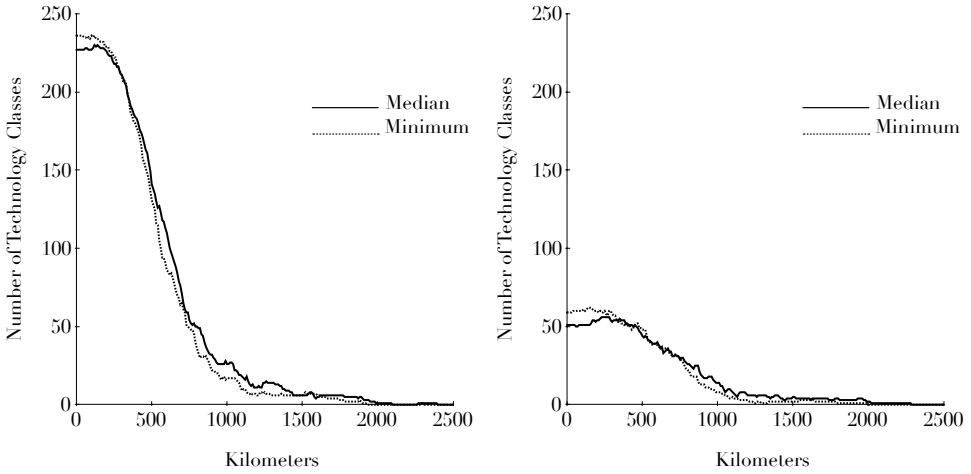
	3-digit Control			6-digit Control		
	State	CMSA	County	State	CMSA	County
All Classes $\lvert\mathscr{A}\rvert$	384	384	384	360	360	360
Localized Classes $\lvert L_1\rvert$	270	266	247	68	69	47
$\lvert L_1\rvert/\lvert\mathscr{A}\rvert \times 100$ (percent)	(70.31%)	(69.27%)	(64.32%)	(18.89%)	(19.17%)	(13.06%)

Notes: $\lvert L^1\rvert$ is the number of technology classes that exhibit localized knowledge spillovers. All the tests are based on the median distance.

Table 5: Matching Rate Tests Conditional on K-density Tests for 6-digit Controls

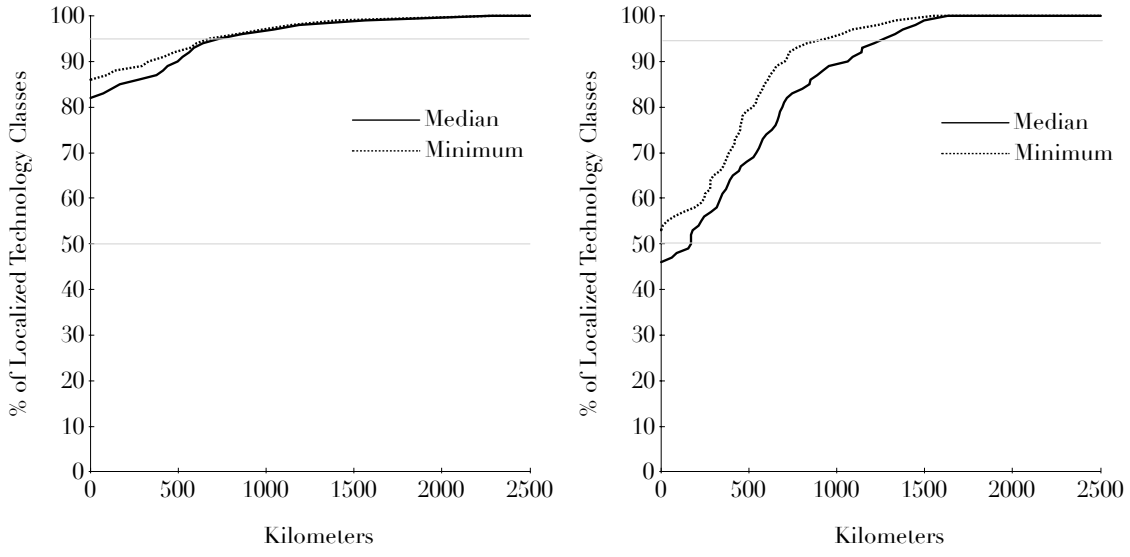
	State	CMSA	County
$\lvert L_0^1\rvert$: $p_A^c = p_A^r$ and $\Gamma_A > 0$	67	68	89
$\lvert L_0^1\rvert/\lvert L^1\rvert \times 100$ (percent)	(61.47%)	(62.39%)	(81.65%)
$\lvert L_1^0\rvert$: $p_A^c > p_A^r$ and $\Gamma_A = 0$	26	28	27
$\lvert L_1^0\rvert/\lvert L^0\rvert \times 100$ (percent)	(10.36%)	(11.16%)	(10.76%)

Notes: $\lvert L^1\rvert$ is the number of technology classes that the K-density tests detect localization while $\lvert L^0\rvert$ is the number of technology classes that the K-density tests do not detect localization. $\lvert L_0^1\rvert$ is the number of technological classes for which the K-density tests detect localization while the matching rate tests do not. $\lvert L_1^0\rvert$ is the number of technology classes for which the matching rate tests detect localization while the K-density tests do not. All the tests are based on the median distance.



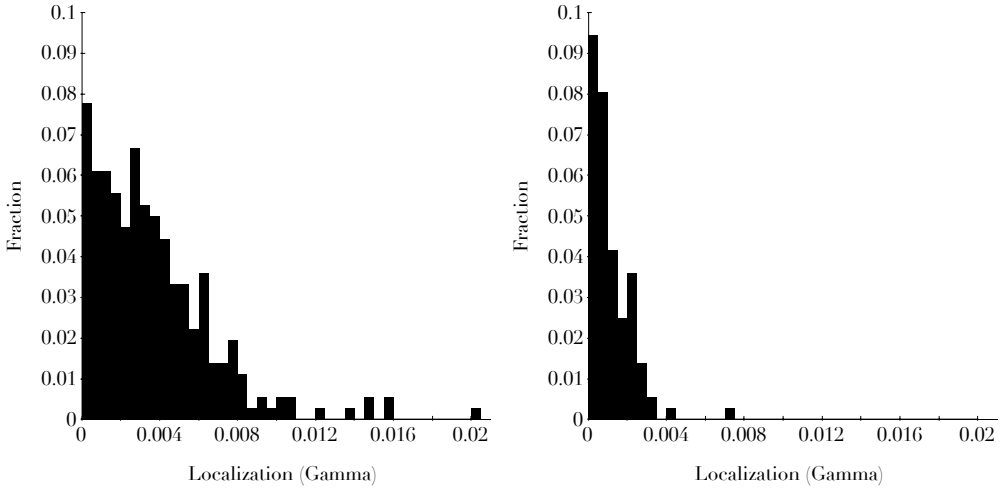
(a) 3-digit controls: $|L^1(d)|$ (b) 6-digit controls: $|L^1(d)|$

Figure 1: Distance Distribution of the Numbers of Localized Technology Classes. The solid and dotted lines represent the results for the median and minimum distance methods, respectively.



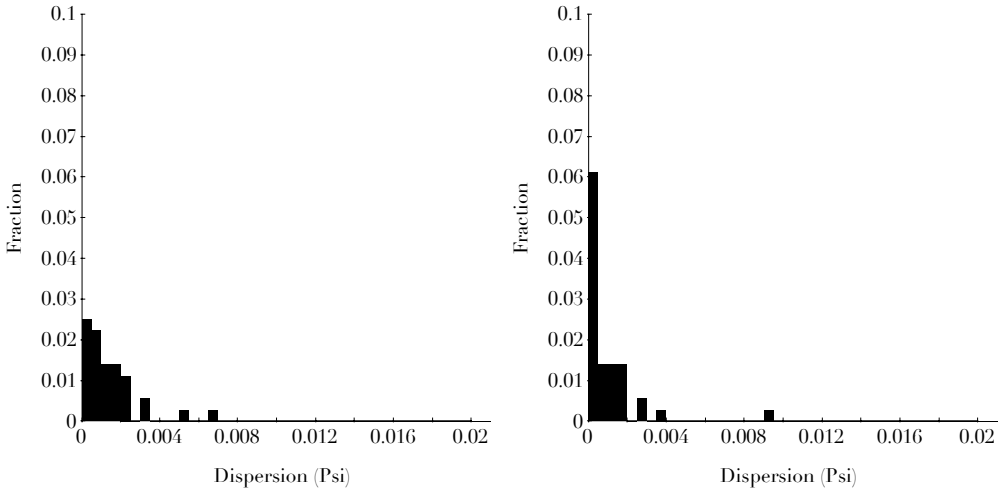
(a) 3-digit controls (b) 6-digit controls

Figure 2: Percentage of Localized Technology Classes within Each Distance. The solid and dotted lines represent the results for the median and minimum distance methods, respectively. The 50% and 95% levels of localized technology classes are depicted by the thin horizontal lines.



(a) 3-digit controls: Γ_A

(b) 6-digit controls: Γ_A



(c) 3-digit controls: Ψ_A

(d) 6-digit controls: Ψ_A

Figure 3: Distributions of Localization and Dispersion Indices. The distributions of localization indices (Γ_A) for the 3- and 6-digit controls are shown in (a) and (b). The distributions of dispersion indices (Ψ_A) for the 3- and 6-digit controls are shown in (c) and (d). All the tests are based on the median distance.

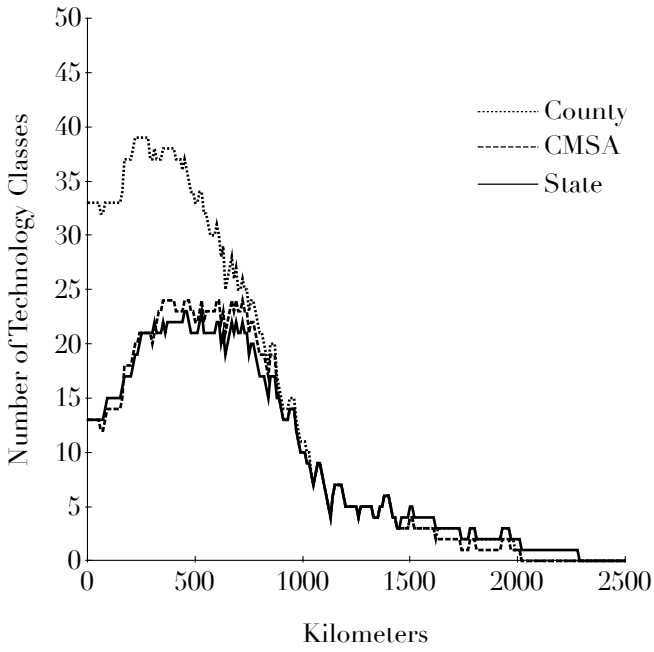
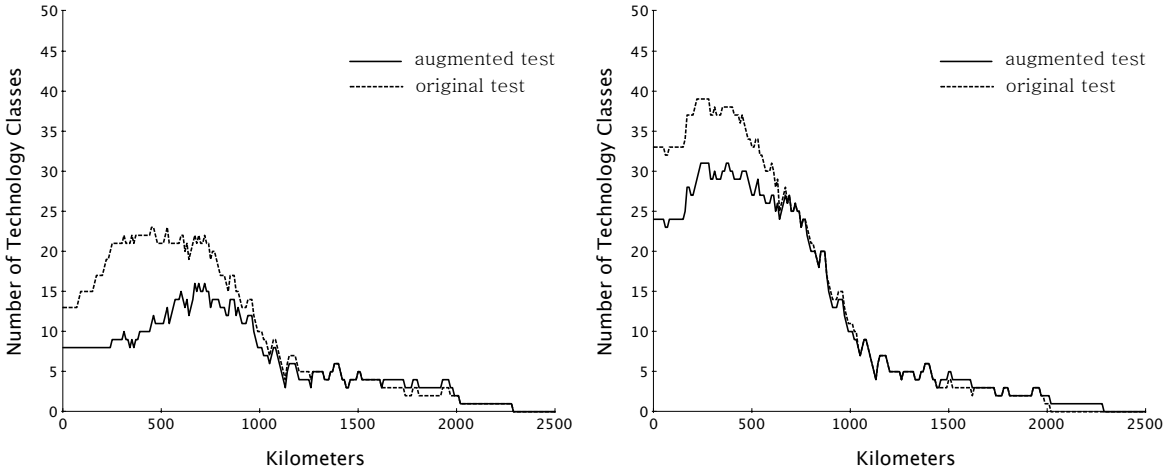
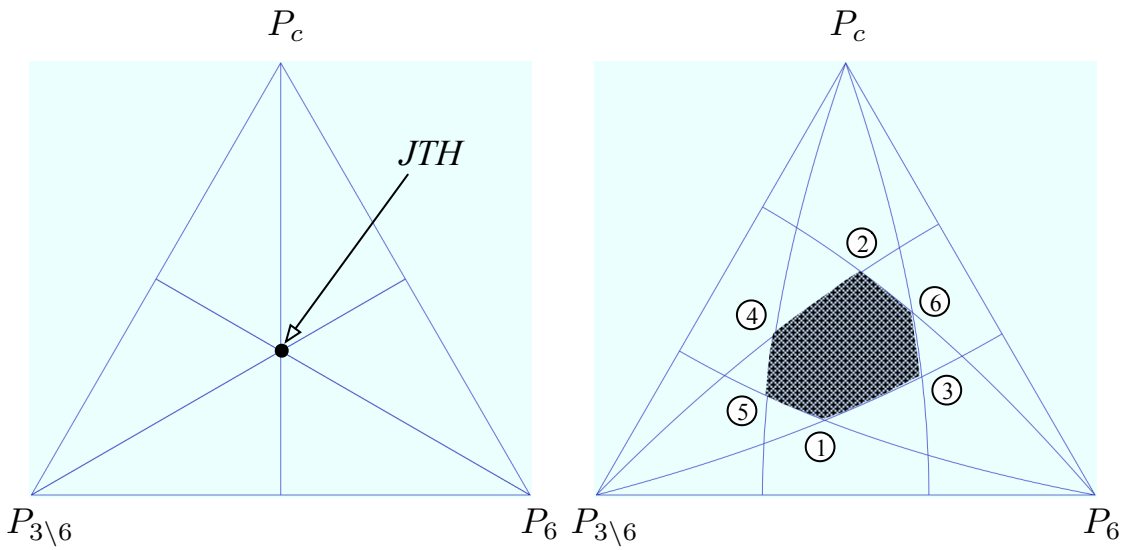


Figure 4: Distance Distribution of $|L_0^1(d)|$ for 6-digit Controls. $|L_0^1(d)|$ is the number of technology classes for which the K-density tests detect localization at distance d while the matching rate tests do not. All the tests are based on the median distance.



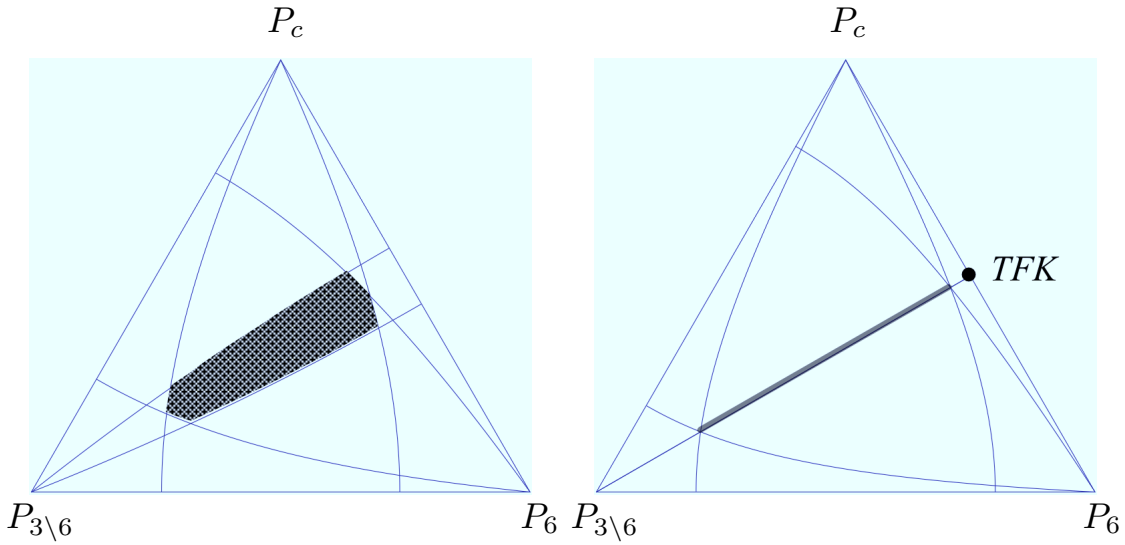
(a) states (b) counties

Figure 5: Distance Distribution of $|L_0^1(d)|$ for the Neighboring Region Augmented Tests. All the tests are based on the median distance with 6-digit controls.



(a) $\Lambda = 1$

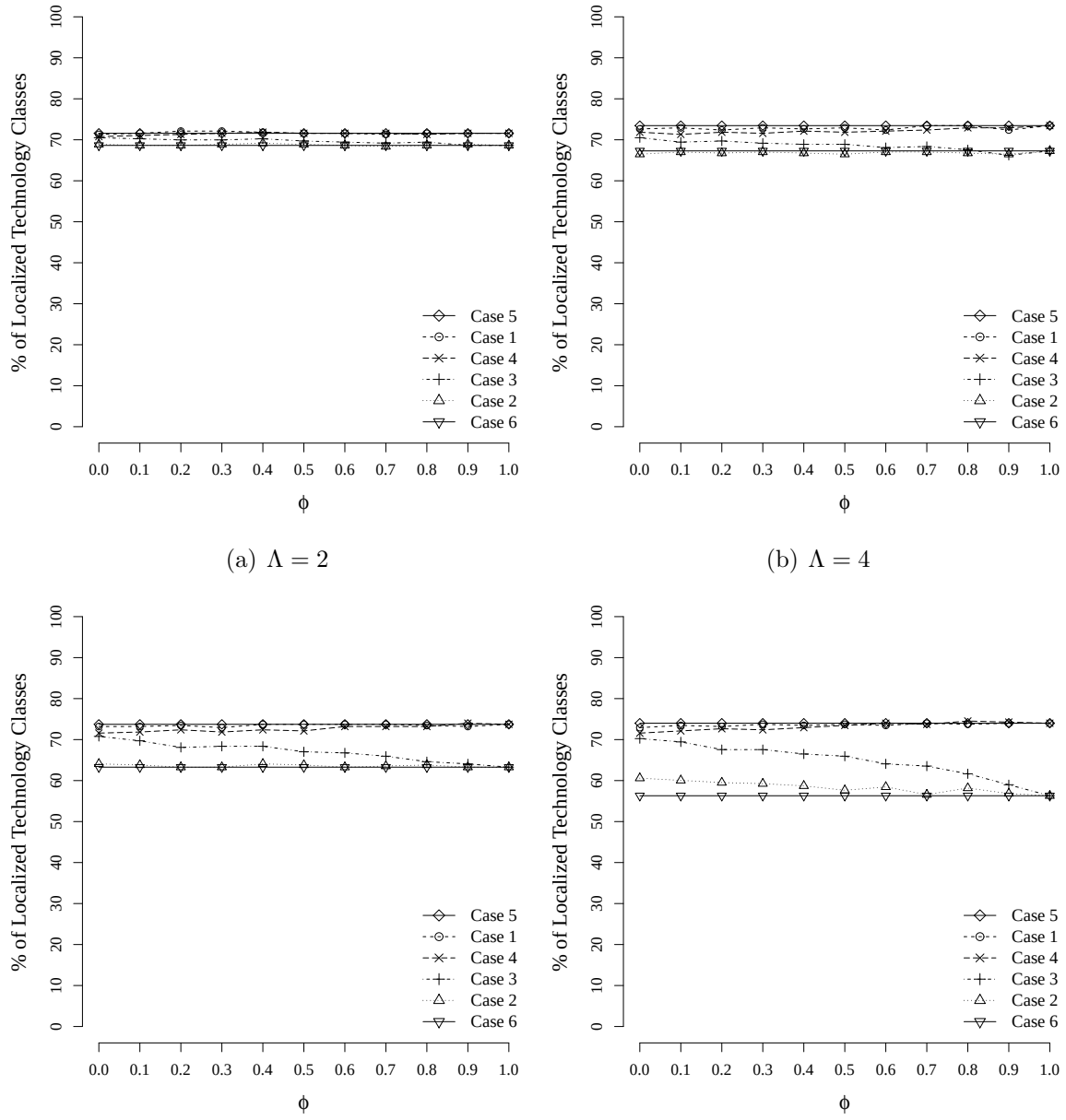
(b) $\Lambda = 4$ and $\phi = 0.25$



(c) $\Lambda = 8$ and $\phi = 0.75$

(d) $\Lambda = 16$ and $\phi = 1$

Figure 6: Feasible Probability Sets for Various Sensitivity Parameters. The probability simplices of $(p_{3\backslash 6}, p_6, p_c)$ are depicted, where the vertices are given by $P_{3\backslash 6} = (1, 0, 0)$, $P_6 = (0, 1, 0)$, and $P_c = (0, 0, 1)$. The sets of feasible probability distributions are given by the shaded hexagons, and each of the six vertices is characterized by a pair of bounds given by Eq. (5)-(7). The point denoted by JTH in (a) is the centroid $(1/3, 1/3, 1/3)$ of the probability simplex, which is the case analyzed by JTH. The point denoted by TFK in (d) corresponds to $(0, 1/2, 1/2)$, which is the case analyzed by TFK.



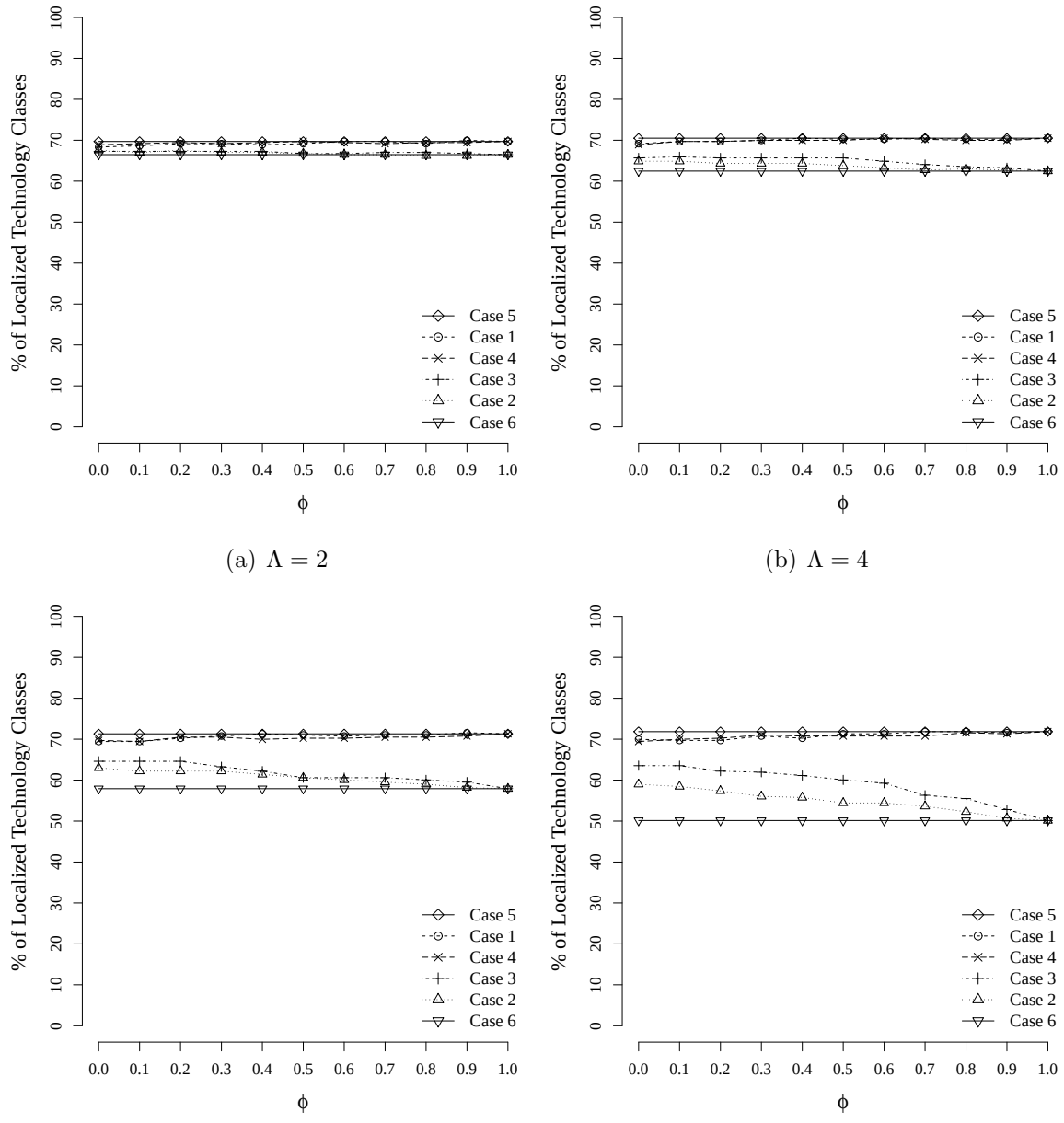
(a) $\Lambda = 2$

(b) $\Lambda = 4$

(c) $\Lambda = 8$

(d) $\Lambda = 16$

Figure 7: Sensitivity Analysis: K-density Tests. The upper and lower bounds for the percentages of localized technology classes are plotted for various sensitivity parameters. Cases 1-6 refer to the corresponding vertices of the hexagon in Figure 6.



(a) $\Lambda = 2$

(b) $\Lambda = 4$

(c) $\Lambda = 8$

(d) $\Lambda = 16$

Figure 8: Sensitivity Analysis: Matching Rate Tests. The upper and lower bounds for the percentages of localized technology classes are plotted for various sensitivity parameters. The geographic units are chosen at the state level. Cases 1-6 refer to the corresponding vertices of the hexagon in Figure 6.

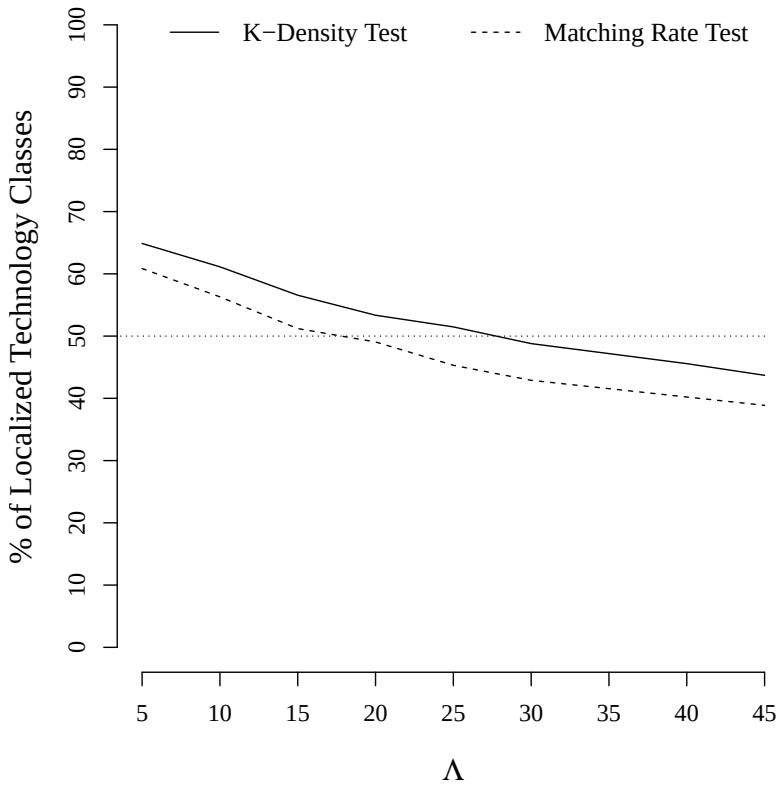


Figure 9: Worst-case Scenario Bounds. The lowest percentages of localized technology classes within the range of $\phi \in [0, 1]$ are plotted for various values of Λ. The geographic units for the matching rate tests are chosen at the state level.