GRIPS Discussion Paper 11-11

Localized knowledge spillovers and patent citations: A distance-based approach

By

Yasusada Murata
Ryo Nakajima
Ryosuke Okamoto
Ryuichi Tamura

September 2011

GRIPS
NATIONAL GRADUATE INSTITUTE FOR POLICY STUDIES

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*

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

September 27, 2011

Abstract

The existence of localized knowledge spillovers found by Jaffe, Trajtenberg and Henderson (1993) has recently been challenged by Thompson and Fox-Kean (2005). To settle this debate, we develop a new approach by incorporating their concepts of control patents into the distance-based test of localization (Duranton and Overman, 2005). Using microgeographic data, we identify localization distance for each technology class while allowing for cross-boundary spillovers, unlike the existing literature where localization is detected at the state or metropolitan statistical area level. We find solid evidence supporting localized knowledge spillovers even when finer controls are used. We further relax the commonly made assumption of perfect controls, and show that the majority of technology classes exhibit localization 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

JEL classifications: O31; R12

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

†Advanced Research Institute for the Sciences and Humanities, Nihon University (murata.yasusada@nihon-u.ac.jp).

‡Department of Economics, Yokohama National University (rn231@ynu.ac.jp).

§National Graduate Institute for Policy Studies (rokamoto@grips.ac.jp).

¶Graduate School of Humanities and Social Sciences, University of Tsukuba (tamura@dpipe.tsukuba.ac.jp).

1 Introduction

Ever since Marshall (1920), it is widely recognized that knowledge spillovers are one of the three major determinants of industry agglomeration. Of the three determinants given in his classic book, intellectual spillovers are harder to identify than trade in goods and labor pooling (Ellison, Glaeser and Kerr, 2010). Nonetheless, Jaffe, Trajtenberg and Henderson (1993) developed a matching rate method to test localized knowledge spillovers as evidenced by patent citations. By controlling for the preexisting geographic concentration of technological activities, they found evidence supporting localized knowledge spillovers at the state and metropolitan statistical area (MSA) levels. However, their finding was recently challenged by Thompson and Fox-Kean (2005*a*). The major difference between these two studies lies in the selection of control patents. In Jaffe, Trajtenberg and Henderson (1993), control and citing patents share a technology class at the three-digit level, whereas in Thompson and Fox-Kean (2005*a*), both patents share a finer technology subclass at the six-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 localized knowledge spillovers at the state and MSA levels. The existence of localized knowledge spillovers is, thus, still inconclusive (Henderson, Jaffe and Trajtenberg, 2005).[2]

Are states and MSAs relevant spatial units for testing localized knowledge spillovers? There is no *a priori* reason for the extent of knowledge spillovers to be limited by administrative boundaries. The matching rate approach, taken by Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*), however, is silent on this issue because it allocates inventors to states and MSAs while abstracting from where those aggregated spatial units are located on the map. Put differently, their approach makes the distance from Boston, MA, to New Haven, CT, equivalent to that of Boston, MA, to Los Angeles, CA.[3] To capture possible cross-boundary knowledge spillovers, we conduct distance-based tests of localization that have been recently developed by Duranton and Overman (2005). 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

[1] These case-control methods have been applied to detect localized knowledge spillovers in numerous contexts for almost two decades. See Almeida (1996) for an early application to the U.S. semiconductor industry. More recent contributions include Agrawal, Kapur and McHale (2008) and Agrawal, Cockburn and Rosell (2010), in which they explored innovation in company towns and the role of ethnicity in knowledge flows.

[2] Despite their disagreement on the selection of control patents, both Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*), as well as this paper, consider that knowledge spills over across the entire technology space.

[3] It should also be noted 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 (MAUP).

existing establishment sites, in order to assess the significance of departures from randomness.

We apply the distance-based approach to test whether knowledge spillovers, as evidenced by patent citations, are localized, and examine to what extent they are localized (if they are).[4] In doing so, we consider which technology classes are localized, and identify the localization distance that is specific to the technology class of the originating patents.[5] Our key idea is to use citation distances, computed from inventors' addresses at the census place level, instead of bilateral distances between establishments in Duranton and Overman (2005). We generate the distribution of citation distances, given that citing-cited relationships are unidirectional, unlike their establishment data. We then identify, for each citing-cited relationship, a set of control patents that could have cited the originating patent. Our novelty lies in incorporating the concept of control patents into the construction of counterfactuals in a consistent way. This can be done by randomly drawing counterfactual citations, as in Duranton and Overman (2005), while controlling for the existing geographic concentration of technological activities, as in Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*). We finally detect localized knowledge spillovers by comparing the actual and counterfactual distributions of citation distances. We thus build a new bridge between these two different strands of literature. To our knowledge, there has so far been no attempt to apply the distance-based method to citing-cited relationships.[6]

Our main results can be summarized as follows. First, distance matters. Our distance-based tests find that, even when we use six-digit controls, knowledge spillovers are localized significantly for about one-third of all 360 technology classes in question. This is in sharp contrast to Thompson and Fox-Kean (2005*a*) who use six-digit controls and found no evidence supporting localized knowledge spillovers at the state and MSA levels. In the three-digit case, more than 70% of 384 technology classes in question exhibit localization, thus confirming the result by Jaffe, Trajtenberg and Henderson (1993). 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, for example. We also find that more than 95% of all technology classes exhibiting localization are localized within 1200 km, which constitutes the widest extent of knowledge spillovers.

Second, heterogeneity across technology classes also matters. In particular, our six-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 six-digit result in

[4]Following the previous studies, we exclude self-citations and describe the detailed procedure in Section 2.1.

[5]As shown in Ellison and Glaeser (1997) and Duranton and Overman (2005), the degree of *industry* localization is known to differ across industries. Thus, we would quite naturally expect that the extent of knowledge spillovers can also differ across technology classes, and we show that this is indeed the case.

[6]Kerr and Kominers (2010) apply a similar distance-based method to patent data. However, they detect localization by using *pairwise distances among inventors* as have been done by Duranton and Overman (2005) in the context of establishment agglomeration. Their K-density tests thus abstract from the concept of control patents and explicit citing-cited relationships, both of which are at the heart of our tests.

Thompson and Fox-Kean (2005a), 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 localized knowledge spillovers, 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 three-digit case, the matching rate tests underestimate the number of localized classes for the six-digit case. Indeed, the matching rate tests fail to detect localized knowledge spillovers 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 three- and six-digit controls are perfect. However, Thompson and Fox-Kean (2005a) argue that three-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 six-digit subclass classification renders "closer" technologically matched controls. Therefore, we finally illustrate some robustness results, provided that neither the three-digit control nor the six-digit control is perfect due to technological heterogeneity within classes or subclasses. In doing so, we build on Rosenbaum's (2002) sensitivity analysis, and deal with the three- and six-digit controls simultaneously by taking into account the fact that matching on subclasses implies matching on classes (but not vice versa). This specification is general in that it encompasses the cases analyzed by Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005a) as limiting cases, while allowing for imperfect controls. We show that, even in the generalized framework, the majority of technology classes exhibit localization unless hidden biases induced by imperfect controls are extremely large.[7] 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 main results. Section 4 discusses the robustness of our results. Section 5 concludes.

2 Data and Methodology

This section describes data and methodology. Unlike the conventional matching rate tests at the state and MSA levels, we need to combine patent citations data and microgeographic

[7]In this generalized framework, where the three- and six-digit controls are placed on a common ground, the case analyzed by Thompson and Fox-Kean (2005a) constitutes a special case where hidden biases are infinitely large, as shown below.

data to conduct distance-based tests. Concerning methodology, we first identify, for each citing-cited relationship, a set of control patents that could have cited the originating patent, as in Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*), to control for the existing geographic distribution of technological activities. We then construct counterfactuals, as in Duranton and Overman (2005), while using the case-control methods. The counterfactual citations thus obtained, with which we compare the actual citations to detect localized knowledge spillovers, share common features between the matching rate and the distance-based tests. Hence, we can make a direct comparison between these two tests for localization.

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 citing-cited relationships for patents granted between 1975 and 1999. For each patent, the list of inventors, addresses of inventors, and the technological category are recorded, along with other information such as year of application, assignees, and the type of assignees. The detailed information of patent application month and *patent class* (three-digit) and *subclass* (six-digit) codes is supplemented with the United States Patent and Trademark Office (USPTO) Patent BIB database.[8]

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.

We further exclude "self-citations". We consider knowledge spillovers as knowledge flows between different inventors of different assignees. Accordingly, a citing patent is classified as self-citing (i) if it 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 not only the name of inventors recorded in the patents, but also 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

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

[9] Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*) regard only the former as self-citations. Our criterion (ii) rules out spurious knowledge spillovers associated with inventor mobility. Furthermore, in response to the comments by Henderson, Jaffe and Trajtenberg (2005), we exclude control patents that share the same inventors or the same assignees with the originating patents.

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

self-citations, we obtain 647, 983 citing patents.

2.2 Geographic Information

Our distance-based approach to localized knowledge spillovers requires microgeographic data, namely the locations at which inventions were created. In this paper, 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] They 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).[12]

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 1.70 km, which is far smaller than those for counties (22.60 km), Consolidated Metropolitan Statistical Areas (CMSAs) (59.93 km), and states (197.93 km).[13]

2.3 Control Patents and Counterfactuals

Since industries generally tend to agglomerate with one another, the mere geographic coincidence of originating and citing patents does not provide solid support for localized knowledge spillovers. For example, in the semiconductor industry, many citations are concentrated in Silicon Valley. This need not imply localized knowledge externalities. It may just reflect the fact that a disproportionate fraction of firms of the related technological area is located in that region. Hence, to test localized knowledge spillovers, we must control for the existing geographic distribution of technological activities.

[11] In the 1990 census, there are two major types of places: census designated places (CDPs); and incorporated places. These data can be obtained from 1990 U.S. Gazetteer Files.

[12] We could use zip code areas. The NBER U.S. Patent Citations Data File, however, reports zip codes for only 15.4% of all U.S. patent records. As the NBER Data File reports cities for almost all cases, we could relate cities to zip codes. Yet, it is often the case that a city has several to dozens of zip codes. Also, the area for each zip code as of 1990, which is needed to compute its internal distance below, is not available. We therefore decided to link cities with census places whose relationships are uniquely determined and whose areas as of 1990 are readily available.

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

To this end, we use *control patents*, proposed by Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*), which satisfy the following two conditions. First, control patents should belong to the same *technological area* as the citing patent under consideration. Jaffe, Trajtenberg and Henderson (1993) select a control patent at the *three-digit* level, whereas Thompson and Fox-Kean (2005*a*) construct a finer control at the *six-digit* level. The latter also claims that a control should match not only with the citing patent but also with the originating patent. In what follows, emphasizing their difference in technology classification, we refer to the controls of Jaffe, Trajtenberg and Henderson (1993) as the *three-digit controls*, and call those of Thompson and Fox-Kean (2005*a*) the *six-digit controls*. Second, a control patent should be in the same cohort as the citing patent. Jaffe, Trajtenberg and Henderson (1993) choose a control patent whose application date is within a one-month window on either side of the citing patent's application date. Similarly, Thompson and Fox-Kean (2005*a*) 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 three-digit and six-digit controls, respectively.[14]

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, 60.20% of the citing patents have three-digit controls. The rate of the citing patents having six-digit controls is lower, at 18.65%. The citing patents with no controls assigned (and their originating patents) are dropped out of the samples.[15] As a result, 92.64% of the originating patents remain "in-sample" for the three-digit controls, and the corresponding number is 51.04% for the six-digit controls. In the analysis that follows, we use these in-sample patents.

Once the relevant control patents are identified, we can construct *counterfactuals* with which we compare the actual citations. For each citation, we define an admissible patent set by collecting the citing and control patents, either three or six digit, so that the admissible patent set consists of the patents that either actually cited or *could have cited* the originating patent.[16]

[14] There is one minor difference between their and our control patents. We use a fixed application date window within which control patents are searched, while Thompson and Fox-Kean (2005*a*) 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.

[15] We also drop technology classes in which originating patents are distributed across less than 10 census places. This selection of patents is required because we estimate the density of distances for each technology class in the subsequent analysis, and a sufficient number of location points are needed to obtain well-behaved estimated density functions. See Section 3.2 for more exposition.

[16] It should be noted that, in the six-digit case, we use the admissible patent set that consists only of the

We then allocate a counterfactual citation between the originating patent and a patent that is randomly drawn from the corresponding admissible patent set. In what follows, we propose tests that nonparametrically balance the actual and counterfactual citations subject to the same technological and temporal profiles, and attribute the remaining difference in geographic distributions to the localization of knowledge spillovers, which is unrelated to the preexisting concentration of technological activities.

2.4 The Matching Rate Approach

The main idea of the matching rate approach, invented by Jaffe, Trajtenberg and Henderson (1993) and refined by Thompson and Fox-Kean (2005*a*), is to compute the geographic matching rate of the actual citations, and compare it with that of counterfactual citations. Following the previous studies, 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. Thompson and Fox-Kean (2005*a*) 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 citation. By contrast, we resample patents *many times* from the admissible patent set, and consider a simulated distribution of the counterfactual matching rate. We now describe the procedure of our matching rate test in detail.

Let p^c be the population probability that a citing patent is in the same geographic unit as the originating patent, and let p^r be 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 localized knowledge spillovers) against the alternative hypothesis $H_1 : p^c > p^r$ (significant localized knowledge spillovers). 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 thus reject the null hypothesis of no localized knowledge spillovers if the p-value, $\text{Prob}(\hat{p}^c \leq \hat{p}^r)$, is less than 5%.

We first construct the observed matching rate $\hat{p}^c$ as follows. Let $\{o^i\}_{i=1}^{n^o}$ be the set of originating patents, where n^o is the number of originating patents. The set of the patents that cite o^i is defined as $\{c^{ij}\}_{j=1}^{n^{ci}}$, where n^{ci} is the number of citing patents. We compute the number of location matches, m^{ci}, between the originating patent o^i and the citing patents $\{c^{ij}\}_{j=1}^{n^{ci}}$. The total number of location matches divided by the total number of citations gives the observed matching rate $\hat{p}^c = \sum_{i=1}^{n^o} m^{ci} / \sum_{i=1}^{n^o} n^{ci}$.

citing and control patents belonging to the same technology class as the corresponding originating patent. This is a logical consequence of the additional restriction in the six-digit case that originating-citing-control triads of patents must share at least one patent subclass in common.

We then construct the distribution of the counterfactual matching rate $\hat{p}^r$ by the following Monte Carlo simulation. For each citing patent c^{ij}, we identify the admissible patent set R^{ij} that consists of the citing patent itself and the associated control patents. Suppose that we run 1000 simulations. In the k-th simulation, for each citing patent c^{ij}, we randomly select a hypothetical patent r_k^{ij} from the admissible patent set R^{ij}. We then calculate the number of location matches, m_k^{ri}, between the originating patent o^i and the randomly chosen hypothetical patents $\{r_k^{ij}\}_{j=1}^{n^{ci}}$. The total number of location matches divided by the total number of hypothetical citations gives the counterfactual matching rate $\hat{p}_k^r = \sum_{i=1}^{n^o} m_k^{ri} / \sum_{i=1}^{n^o} n^{ci}$, where the total number of hypothetical citations equals that of actual citations. The Monte Carlo process allows us to obtain the simulated distribution of the matching rate $\{\hat{p}_k^r\}_{k=1}^{1000}$. We finally compute the p-value of the matching rate test by using the standard percentile method.

Although the matching rate test is straightforward, one should be careful for multiple inventors per patent. To determine whether or not a pair of citing and cited patents falls into the same geographic unit, we use the following two matching methods. Consider, for each citing-cited relationship, all possible pairs of an inventor of the citing patent and an inventor of the cited patent. The locations of the citing and cited 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 accord with those used in previous studies. For example, Jaffe, Trajtenberg and Henderson (1993) employ a similar method as our median matching. Thompson and Fox-Kean (2005*a*) mention the minimum matching as an alternative to their random matching.

2.5 The K-density Approach

As mentioned in the Introduction, the extent of knowledge spillovers is unlikely to be limited by administrative boundaries. The matching rate approach that we have taken in the previous subsection, however, cannot address this issue because it abstracts from where CMSAs and states are located in the United States. To capture possible cross-boundary knowledge spillovers, we rely on distance-based tests of localization that were recently developed by Duranton and Overman (2005). 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.

We apply Duranton and Overman's approach to test whether knowledge spillovers, as evidenced by patent citations, are localized, and examine to what extent they are localized (if they are). As before, we allocate a counterfactual citation between the originating patent and a patent drawn randomly from the corresponding admissible patent set. Unlike the matching rate approach, however, 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 localized knowledge spillovers. 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. We thus build a new bridge between the two strands of literature, which are the matching rate test of localized knowledge spillovers and the distance-based test of industry localization.

Such an attempt, however, poses two main difficulties that we need to deal with. First, unlike establishments whose locations are usually uniquely determined, patents can have multiple addresses because their inventors are not necessarily unique. We thus compute, for each citation 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. Although census places are narrowly delineated compared with counties and states, they 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, it is readily verified that the distance between the two randomly chosen points in census place ℓ with area S_ℓ is given by $[128/(45\pi)]\sqrt{S_\ell/\pi}$ (Kendall and Moran, 1963). We use this correction for the distance between the two inventors who are in the same census place, instead of regarding the distance as to be zero.

It is also noted that, unlike the previous studies on patent citations, we analyze the localization distance that is specific to each patent class.[17] 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 these intra- and inter-class spillovers into account, we examine whether each patent class — to which originating patents belong — displays localization.[18]

We now describe the detailed procedure of our distance-based test of localized knowledge spillovers. Let $\mathscr{A}$ be the set of all technology classes, categorized at the patent class level. We denote by $\{o_A^i\}_{i=1}^{n_A^o}$ the set of originating patents for technology class $A \in \mathscr{A}$, where n_A^o is the

[17]Because the degree of localization is known to differ across industries (e.g., Ellison and Glaeser, 1997; Duranton and Overman, 2005), it seems natural to expect that the extent of localized knowledge spillovers can also differ across patent classes. As we show later, this is indeed the case.

[18]Note that this procedure is common regardless of whether we use the three- or six-digit controls. In the latter case, we could examine whether each patent *subclass* exhibits localization. Unfortunately, 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.

number of originating patents. The set of patents that cite o_A^i is denoted by $\{c_A^{ij}\}_{j=1}^{n_A^{ci}}$, where n_A^{ci} is the number of citing patents. The number of citations originating from technology class A is then given by $N_A = \sum_{i=1}^{n_A^o} n_A^{ci}$. We finally denote by d_A^{ij} the great-circle distance between patents o_A^i and c_A^{ij}, which, as mentioned above, is given by either the minimum or median distance from the inventors of the originating patent to those of the citing patent. Following Duranton and Overman (2005), the kernel density (henceforth K-density) estimator of citation distance for technology class A at any point d is

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

where f is a Gaussian kernel function and h is the bandwidth set as in Silverman (1986). Note that, expression (1) reflects the fact that, unlike Duranton and Overman (2005), we consider unidirectional relationships from the inventors of originating patents to those of citing patents.[19]

Concerning counterfactuals, we run 1000 Monte Carlo simulations.[20] The construction of counterfactuals is the same as that of the matching rate test. For each citing patent c_A^{ij}, we identify the admissible patent set R_A^{ij} that consists of the citing patent itself and the associated control patents. In the k-th simulation, we randomly draw a hypothetical patent r_{Ak}^{ij} from the admissible patent set R_A^{ij} for each citing patent c_A^{ij} to estimate the counterfactual K-density for the distribution of distances from originating patents o_A^i to hypothetical patents r_{Ak}^{ij}, using a formula similar to (1). After 1000 simulation runs, we 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.

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. Unfortunately, this does not lead to statements about the global citation patterns of a technology class because even a technology class with randomly distributed citations will exhibit dispersion or localization with a high probability. Indeed, by construction, there is a 5% probability that a technology class displays localization for each distance, so that the probability for this to occur at least once across all distances is quite high, even though smoothing induces some autocorrelation in the K-density estimates across distances.

Therefore, we finally define the global confidence bands that we use to detect localized knowledge spillovers. Let $\bar{d}_A$ be the maximum distance for technology class A under consideration.[21] We look for the identical upper and lower local confidence intervals such that,

[19] As in Duranton and Overman (2005), we adopt the reflection method in Silverman (1986) to deal with boundary problems associated with the fact that distances cannot be negative.

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

[21] Following Duranton and Overman (2005), we define the maximum distance as the median of all distances

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 by construction. Hence, we define

$$\Gamma_A(d) \equiv \max\left\{\widehat{K}_A(d) - \overline{\overline{K}}_A(d), 0\right\}$$

as an index of global localization, and

$$\Psi_A(d) \equiv \begin{cases} \max\left\{\underline{\underline{K}}_A(d) - \widehat{K}_A(d), 0\right\} & \text{if} \quad \sum_d \Gamma_A(d) = 0 \\ 0 & \text{otherwise} \end{cases}$$

as an index of global dispersion.

Insert Figure 1

Figures 1 (a)–(b) illustrate K-densities (solid) and global confidence bands (dotted) for two patent classes, namely butchering (452) and amusement devices: toys (446), respectively. The former exhibits global localization while the latter is globally dispersed.

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 Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*), despite some differences in data and methodology. 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 six-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.[22] Following Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*), the

of all possible counterfactual citations for technology class A.

[22] As in Thompson and Fox-Kean (2005*a*), we use 16 CMSAs as defined in 1981 by excluding Puerto Rico.

matching rate tests are implemented *at the aggregate level* encompassing all technology classes. Using controls at the three- and six-digit levels, 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.

Insert Table 2

In the case of the three-digit controls, the observed matching rates are significantly higher than the counterfactual ones for all spatial scales, although the matching rates become smaller for finer geographic units. We reject the null hypothesis of no localized knowledge spillovers at a 5% significance level, and, thus, find solid evidence of localized knowledge spillovers. By contrast, the null hypothesis is not rejected for the six-digit controls, which suggests no evidence of localized knowledge spillovers. These results share the same qualitative features as those of Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005a), although our data construction and methodology are somewhat different from theirs.

3.2 The K-density Tests

We now describe the results of the K-density tests by introducing some notations. For a 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 localized knowledge spillovers 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 main results. As we consider both the minimum and median distances, as well as the three- and six-digit controls, there are four possibilities. First, concerning the three-digit case, we find localized knowledge spillovers 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 Jaffe, Trajtenberg and Henderson (1993). Turning to the six-digit controls, more than 30% of technology classes exhibit localized knowledge spillovers regardless of whether we use the median or minimum distance. Although fewer classes exhibit localization in the six-digit case, we obtain solid evidence for localized knowledge spillovers. This is surprising given that Thompson and Fox-Kean (2005a) find no evidence supporting localization at the state and CMSA levels. The matching rate tests that we have presented above also report no localization for the six-digit controls.

Insert Table 3

To investigate more closely the scope of knowledge spillovers, let $L^1(d) = \{A \in \mathscr{A} | \Gamma_A(d) > 0\}$ and $D^1(d) = \{A \in \mathscr{A} | \Psi_A(d) > 0\}$ be the sets of technology classes that exhibit localization and dispersion at distance d, respectively. Figure 2 illustrates the distributions of $|L^1(d)|$ and $|D^1(d)|$ for the three- and six-digit controls. In each case, there is no substantial difference between the median (solid) and the minimum (dotted) distance methods. We see that the number of localized technology classes is greater at smaller distances for both the three- and six-digit controls. The degree of localization decreases as the distance from the originating patents increases, thus suggesting that knowledge spillovers decay with distance. Interestingly, this is consistent with the assumption that is made in the recent theory of spatial development (Desmet and Rossi-Hansberg, 2009). By contrast, 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).

Insert Figure 2

We can delineate a boundary within which knowledge spillovers are localized. Figure 3 shows the percentages of technology classes displaying localization at least once within distance d. As shown, there are substantial differences between the three- and six-digit cases. However, no matter which control is used, more than half of the technology classes displaying localized knowledge spillovers 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 knowledge spillovers because more than 95% of all localized classes are localized by this distance, regardless of which controls are used.

Insert Figure 3

We further examine heterogeneity in the patterns of knowledge spillovers across technology classes. Figure 4 illustrates the distributions of Γ_A and Ψ_A for the median distance case because the results are fairly robust regardless of the choice between the median and the minimum distances.[23] Both distributions are skewed substantially with the localization and dispersion indices being close to zero, while there are several technology classes displaying highly localized or dispersed knowledge spillovers. Interestingly, for the three-digit controls, the fraction of localized technology classes outweighs substantially that of dispersed technology classes. By contrast, in the six-digit case, the corresponding difference between the localized and dispersed technology classes is not so large.

Insert Figure 4

Finally, Table 4 presents the top 20 technology classes with highest degrees of localization, measured by Γ_A, for the median distance case. The rankings for the three- and six-digit

[23] The results of the minimum distance method are available upon request from the authors.

controls are roughly similar, in that five out of the top 20 localized classes overlap between the three- and six-digit cases.[24] Table 4 also shows that knowledge spillovers are highly localized in "traditional" industries such as: agriculture, husbandry and food (Patent Class 452); furniture and house fixtures (256); earth working and wells (166 and 405); and apparel and textile industries (2, 36, and 112), where the categories are given by Hall, Jaffe and Trajtenberg (2001). We also find significant localization of knowledge spillovers for many mechanical industries (Patent Classes 192, 221, 239, 254, 296, 301, 303, 411, 440, 492 and 508), in particular, transportation mechanical industries (296 and 301).

Insert Table 4

3.3 Comparison

We have shown that, unlike the matching rate tests, the K-density tests provide solid evidence for localized knowledge spillovers, even for the six-digit controls. We now explore the differences, as well as the similarities, in detecting localization between these two approaches. In particular, we argue, in what follows, that the matching rate tests using the six-digit controls underestimate localization of knowledge spillovers because of 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 six-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 localized knowledge spillovers at the 5% significance level *for each technology class*. Let $L_1 = \{A \in \mathscr{A} | p^c_A > p^r_A\}$ 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 5 shows that, when the three-digit controls are used, localized knowledge spillovers are detected for 270 or 266 technology classes, depending on whether the spatial units are states or CMSAs. Interestingly, these numbers are fairly close to the 275 localized classes, obtained from the K-density tests in Table 3.[25] Hence, we conclude that the matching rate and the K-density tests detect roughly the same number of localized technology classes for the three-digit controls.

[24]In fact, the rank correlation coefficient between the three- and six-digit controls is computed as $\rho = 0.36$, and the null hypothesis of no correlation is rejected at the 1% significance level.

[25]Table 5 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.

However, for the six-digit controls, the matching rate and the K-density tests substantially differ in detecting the number of localized technology classes. Indeed, the class-specific matching rate tests identify localization for a smaller fraction of technology classes than do the K-density tests. More concretely, only 47 to 69 technology classes display localization in the former tests, depending on the spatial units, whereas more than 100 technology classes are shown to be localized in the latter tests. Yet, even in the class-specific matching rate, the percentages of technology classes with localized knowledge spillovers remain in the range between 13% and 20%. Hence, we find evidence that knowledge spillovers are localized for nonnegligible, though not overwhelming, technology classes, even in the six-digit case.

Insert Table 5

The second problem of the conventional matching rate tests is “geographic aggregation”. The matching rate tests *ex ante* allocate inventors to spatial units such as states and CMSAs. As Duranton and Overman (2005) pointed out, this aggregation treats 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 localized knowledge spillovers that cross an administrative boundary. The distance-based tests have an advantage in that they do not overlook such cross-border knowledge flows.

To investigate this possibility, we again focus on the discrepancy between the matching rate and the K-density tests for the six-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 $L^1_0 = \{A \in \mathscr{A} | p^c_A = p^r_A \text{ and } \Gamma_A > 0\}$ as the set of technology classes where the K-density tests detect significant localization, while the matching rate tests do not. Thus, it follows that $L^1_0 \subseteq L^1$. Similarly, we define $L^0_1 = \{A \in \mathscr{A} | p^c_A > p^r_A \text{ and } \Gamma_A = 0\} \subseteq L^0$.

Table 6 provides the results. First, looking at the results of $|L^1_0|$ 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 tests *underestimate* localized knowledge spillovers. 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 $|L^0_1|$ 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. This implies that the matching rate tests can also *overestimate* localized knowledge spillovers. However, the numbers of *underestimated* localized classes, $|L^1_0|$, much outweigh those of *overestimated* localized classes, $|L^0_1|$. We see that the difference ranges from 40 to 62, which explains the difference between $|L^1|$ in Table 3 and $|L_1|$ in Table 5 for the six-digit controls.

Insert Table 6

We finally investigate where we observe the downward biases of the matching rate tests using the six-digit controls in detecting localized knowledge spillovers. Figure 5 plots $|L_0^1(d)|$ for each distance d, where $L_0^1(d) = \{A \in \mathscr{A} | p_A^c = p_A^r \text{ and } \Gamma_A(d) > 0\}$. We first notice that 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 about 40 technology classes as having no localized knowledge spillovers at 200 km. This underestimation is inherent in their construction. The matching rate tests cannot discern, by their definitions, 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, localized knowledge spillovers 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 localized knowledge spillovers that can be detected by the K-density tests.

Insert Figure 5

In summary, the existing matching rate tests systematically understate localized knowledge spillovers, as evidenced by patent citations. We explain this by two aggregation problems, which are technological and geographic aggregations. If we control for heterogeneity in localization and dispersion by disaggregating technology classes, the matching rate tests provide evidence of localized knowledge spillovers 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

In the localization tests presented so far, we have constructed counterfactual citations by drawing patents randomly from the admissible set. This amounts to assuming that citing and control patents are equally likely to cite the originating patent. Indeed, we have made this assumption, relying 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. However, Thompson and Fox-Kean (2005a) argue that three-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 six-digit subclass classification renders "closer" technologically matched controls.

The aim of this section is to discuss the robustness of our localization results, provided that neither the three-digit control nor the six-digit control is perfect. More specifically, relying on Rosenbaum's (2002) sensitivity analysis, we consider how to 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 counterfactual citations, we conduct both the matching rate and distance-based tests of localization to address how sensitive our results are to various magnitudes of hidden biases induced by imperfect matching between the citing and control patents.[26] In doing so, we deal with the three- and six-digit controls simultaneously by taking into account the fact that matching on subclasses implies matching on classes (but not vice versa). This approach encompasses our previous analysis as limiting cases, and allows us to illustrate 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 hidden biases are 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 localized knowledge spillovers in terms of matching estimators.[27] Let m be a dummy variable indicating whether a pair of patents match 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 distribution of 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 patent distance d.

The basic premise of these localization tests is that the outcomes, m and d, are independent of treatment assignment t, conditional on the technology class x (this is the so-called *conditional independence assumption* in the matching estimation literature). It is also known that, if the conditional independence 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

[26]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.

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

treatment 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).

Consider an admissible set that consists of the citing and control patents that share the same three-digit patent class. Recall that R^{ij} denotes the admissible set of patent o^i and the corresponding citing patent c^{ij}. In what follows, we drop superscripts, i and j, for simplicity. In general, each citing patent has multiple control patents, but, for the moment, we assume that the control patent is unique, so that $R = \{b, c\}$, where b denotes the three-digit control patent corresponding to the citing patent c.

Let r be a patent in the admissible set R. Rosenbaum (2002) assumes that the treatment assignment probability of patent r with technology class x_r is given by

$$p_r = \text{Prob}(t_r = 1|x_r) = F(\kappa(x_r) + \lambda u_r), \tag{2}$$

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, but $u_b \neq u_c$ in general. The assignment probability p_r is nothing but the propensity score $p(x)$ because $x_r = x$ for $r \in R$.

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 set, with equal chances, in the simulation process of the localization tests. 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 set with different probabilities, reflecting the magnitudes of hidden biases.

So far, we have illustrated the effect of hidden biases on the localization tests in the case of a single control group (either the three- or six-digit control). We now turn to a general class of Rosenbaum's sensitivity analysis that encompasses multiple control groups (both the three- and six-digit controls). Let $\mathbf{b}_3$ be the set of three-digit controls that match the citing patent c at the three-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 six-digit level also match at the three-digit level by construction of three- and six-digit codes. We thus have $\mathbf{b}_3 = \mathbf{b}_6 \cup \mathbf{b}_{3\setminus 6}$, where $\mathbf{b}_6$ is the set of six-digit controls and $\mathbf{b}_{3\setminus 6}$ is the set of controls that match the citing patent at the three-digit level but not at the six-digit level.[28] Let $n_6 = |\mathbf{b}_6|$ and $n_{3\setminus 6} = |\mathbf{b}_{3\setminus 6}|$ with $n_3 = n_6 + n_{3\setminus 6}$,

[28]In this sensitivity analysis with multiple control groups, we remove the restriction, which is applicable only to the six-digit controls, that control patents must share any subclass in common with originating patents. This allows us to analyze both the three- and six-digit controls on a common ground.

$n_6 \geq 1$ and $n_{3\backslash 6} \geq 1$. Then, the admissible set at three-digit level is given by $R = \{\mathbf{b}_6, \mathbf{b}_{3\backslash 6}, c\}$.

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

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

If the three-digit control is correct, we have $p_6 = p_{3\backslash 6} = p_c$. On the other hand, if the six-digit control is correct, we have $p_6 = p_c$ but $p_{3\backslash 6} \neq p_c$. 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 (2) for $r \in R = \{\mathbf{b}_6, \mathbf{b}_{3\backslash 6}, c\}$. Let x_6 and $x_{3\backslash 6}$ be vectors of three-digit technology class dummies for $r \in \mathbf{b}_6$ and $r \in \mathbf{b}_{3\backslash 6}$, respectively. Because any patent in the admissible set $R = \{\mathbf{b}_6, \mathbf{b}_{3\backslash 6}, c\}$ shares the same three-digit technology class, the observed factors are perfectly comparable., i.e., $x_6 = x_{3\backslash 6} = x_c$. In contrast, the unobserved terms are partially comparable. As in Rosenbaum (2002), we write u_r as a weighted sum of unobserved factors, v_r and w_r,

$$u_r = (1 - \phi) v_r + \phi w_r,$$

where $\phi \in [0, 1]$, $v_r \in [0, 1]$ and $w_r \in [0, 1]$. We impose the following restriction

$$w_r = w_c \quad \text{if} \quad r \in \mathbf{b}_6, \tag{4}$$

on unobserved terms, while allowing for $w_r \neq w_c$ if $r \in \mathbf{b}_{3\backslash 6}$. In words, the six-digit controls and the citing patent share some unobserved similarities that are not shared by the three-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 (2), we compute the odds ratios as follows:

$$\begin{aligned}
\frac{q_6}{q_c} &= \exp\left[\lambda(1-\phi)(v_6 - v_c)\right] \\
\frac{q_{3\backslash 6}}{q_c} &= \exp\left[\lambda(u_{3\backslash 6} - u_c)\right] \\
\frac{q_6}{q_{3\backslash 6}} &= \exp\left[\lambda(u_6 - u_{3\backslash 6})\right].
\end{aligned}$$

Because $0 \leq u, v \leq 1$, 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 $\Lambda = \exp(\lambda)$. Because $\Lambda^{1-\phi} \le \Lambda$ for $0 \le \phi \le 1$, the bound of q_6/q_c is narrower than the others due to the restriction in (4).

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 simplex 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 case analyzed by Jaffe, Trajtenberg and Henderson (1993), where the three-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 in the simplex such that $\{(p_{3\backslash 6}, p_6, p_c) | p_6 = p_c\}$, i.e., the six-digit control and citing patent cite the originating patent with equal likelihood. Indeed, Thompson and Fox-Kean (2005a) 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 $1 \le \Lambda \le \infty$ and $0 \le \phi \le 1$ to examine how sensitive our results of localized knowledge spillovers 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 (3) can be rewritten as

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

For example, 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

$$\begin{cases} A_3 = n_6 + n_{3\backslash 6} > 0 \\ A_2 = (n_6 + n_{3\backslash 6} - 1) + n_6\Lambda^{-1} + n_{3\backslash 6}\Lambda^{\phi-1} > 0 \\ A_1 = (n_6 - 1)\Lambda^{-1} + (n_{3\backslash 6} - 1)\Lambda^{\phi-1} > 0 \\ A_0 = -\Lambda^{\phi-2} < 0 \end{cases}.$$

We can show that the equation has the unique solution for $q_c \ge 0$. 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 the formula $p_r = q_r/(1+q_r)$ for $r \in R$. The assignment probabilities for the other five vertices are analogously obtained (see the Appendix). We finally conduct the matching rate and distance-based tests of localization for each set of assignment probabilities associated with each vertex in order to examine the robustness of our localization results.

Figure 7 presents the results of the sensitivity analysis for the K-density tests.[29] 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" depicted in Figure 6. As Λ increases, the difference between the upper and lower bounds of localized knowledge spillovers gets larger, reflecting increasing uncertainty in the admissible set. If there were no hidden bias ($\Lambda = 1$), the localization would be observed for about 70% of technology classes (not graphed), which is quite comparable to the previous localization result for the three-digit controls.

Insert Figure 7

Figure 8 presents the sensitivity analysis for the matching rate tests at the state level.[30] The overall patterns are roughly similar to those for the K-density tests. However, for a given 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%. This confirms our previous finding that, with the six-digit controls, the matching rate tests understate localized knowledge spillovers in comparison to the K-density tests. Our sensitivity analysis thus shows that the underestimation result remains true, even with a more general choice of control patents allowing for unobserved factors.

Insert Figure 8

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 $\phi \in [0, 1]$.[31] As shown, the worst-case scenario bound for the matching rate tests is uniformly lower than that for the K-density tests. This again provides evidence that the matching rate tests understate localized knowledge spillovers when compared with the K-density tests. Focusing on the latter tests,

[29] In these K-density tests and in the following matching rate tests, we use the median distance and median matching, respectively.

[30] 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.

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

the worst-case scenario bound exceeds 50% even at $\Lambda = 25$. This means, even if we allow for significant unobserved factors that make the originating patents 25 times more likely cited by the actual citing patents than the control patents – an extreme departure from the assumption of no hidden factor – localized knowledge spillovers remain dominant. In light of this, the K-density tests with the six-digit controls, which show that 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 localized knowledge spillovers unless hidden biases are extremely large.

Insert Figure 9

5 Conclusion

We have proposed a distance-based approach to localized knowledge spillovers and revisited the recent debate by Thompson and Fox-Kean (2005*a*,*b*) and Henderson, Jaffe and Trajtenberg (2005) on the existence of localized knowledge spillovers. Our concern has been two aggregation problems, namely technological and geographic aggregations, both of which are ignored in that literature. Overcoming these two problems, our distance-based tests have found solid evidence supporting localized knowledge spillovers for a substantial number of technology classes, even when the finer six-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 six-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 Thompson and Fox-Kean (2005*a*), could not find any significant evidence for intranational knowledge spillovers.

To compare our distance-based tests with the conventional matching rate tests by Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*), 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 three-digit control nor the six-digit control is perfect due to technological heterogeneity within classes or subclasses. Therefore, we have developed a new approach 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 localized knowledge spillovers between Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005*a*,*b*) who rely on the 1975-1999 data for which that information is not available.[32]

[32] To avoid imperfect controls, Thompson (2006) develops an alternative way that does not involve case

Our findings from class-specific distance-based tests have implications for cluster policies. First, policy makers need to select the "right" technology classes. Second, for each "right" technology class, the "right" scope must also be taken into account. Since the majority of technology classes that display localization are localized within 200 km, knowledge cluster policies can generally be made within this distance in order to enhance knowledge externalities. As administrative boundaries need not limit knowledge spillovers, such policies would require coordination among adjacent administrative units. Although we have mainly focused on cross-boundary knowledge spillovers to illustrate the biases generated by the matching rate tests, our K-density tests can also be applied to localized knowledge clusters in smaller scales.

Finally, our distance-based measure of localized knowledge spillovers can be used to explore the determinants of industry agglomeration. Some studies (e.g., Rosenthal and Strange, 2001; Ellison, Glaeser and Kerr, 2010) already attempted to include proxies for the importance of knowledge spillovers, which are constructed more directly from patent data, into their regression analysis. However, we are not aware of any study that incorporates a measure of *localized* knowledge spillovers for explaining industry agglomeration. Using such a localization measure would lead to a better understanding of the relationship between industry agglomeration and knowledge spillovers.

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, Ajay, Devesh Kapur, and John McHale. 2008. "How do spatial and social proximity influence knowledge flows? Evidence from patent data." *Journal of Urban Economics*, 64: 258–268.

Agrawal, Ajay, Iain Cockburn, and Carlos Rosell. 2010. "Not Invented Here? Innovation in company towns." *Journal of Urban Economics*, 67: 78–89.

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

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools." *Journal of Political Economy*, 113: 151–184.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

Combes, Pierre-Philippe, and Gilles Duranton. 2006. "Labour pooling, labour poaching, and spatial clustering." *Regional Science and Urban Economics*, 36: 1–28.

Desmet, Klaus, and Esteban Rossi-Hansberg. 2009. "Spatial development." Working Paper 2009-18, Instituto Madrileño de Estudios Avanzados.

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

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

Ellison, Glenn D., Edward L. Glaeser, and William R. Kerr. 2010. "What causes industry agglomeration? Evidence from coagglomeration patterns." *American Economic Review*, 100: 1195–1213.

Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg. 2001. "The NBER patent citation data file: Lessons, insights and methodological tools." NBER Working Paper #8498.

Henderson, Rebecca, Adam Jaffe, and Manuel Trajtenberg. 2005. "Patent citations and the geography of knowledge spillovers: A reassessment: comment." *American Economic Review*, 95: 461–464.

Ichino, Andrea, Fabrizia Mealli, and Tommaso Nannicini. 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, Guido W. 2003. "Sensitivity to exogeneity assumptions in program evaluation." *American Economic Review*, 93: 126–132.

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

Jaffe, Adam, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic localization of knowledge spillovers as evidenced by patent citations." *Quarterly Journal of Economics*, 108: 577–598.

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

Kerr, William R., and Scott Duke Kominers. 2010. "Agglomerative forces and cluster shapes." NBER Working Paper #16639.

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

Marshall, Alfred. 1920. *Principles of Economics.* London: MacMillan.

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

Rosenthal, Stuart S., and William C. Strange. 2001. "The determinants of agglomeration." *Journal of Urban Economics*, 50: 191–229.

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

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

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

Trajtenberg, Manuel, Gil Shiff, and Ran Melamed. 2006. "The NAMES GAME: Harnessing inventors' patent data for economic research." NBER Working Paper #12479.

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

Appendix

In order to obtain the assignment probabilities, p_r, for the vertexes of the "hexagon", we need to solve cubic equations of the odds ratios, q_r. Each equation corresponds to a pair of bounds given by (5)–(7). We need to consider the following six cases:

Case ①: (5) max, (6) max

Case ②: (5) min, (6) min

Case ③: (5) max, (7) min

Case ④: (5) min, (7) max

Case ⑤: (6) max, (7) max

Case ⑥: (6) min, (7) min.

As we have already considered the first case, we will consider the other cases.

Case ②: (5) min, (6) min The bounds are

$$\begin{aligned}\frac{q_6}{q_c} &= \Lambda^{\phi-1}\\ \frac{q_{3\backslash 6}}{q_c} &= \Lambda^{-1}.\end{aligned}$$

The cubic equation of q_c has the following coefficients:

$$\begin{cases} A_3 = n_6 + n_{3\backslash 6} > 0 \\ A_2 = (n_6 + n_{3\backslash 6} - 1) + n_6\Lambda + n_{3\backslash 6}\Lambda^{1-\phi} > 0 \\ A_1 = (n_6 - 1)\Lambda + (n_{3\backslash 6} - 1)\Lambda^{1-\phi} > 0 \\ A_0 = -\Lambda^{2-\phi} < 0 \end{cases}.$$

Case ③: (5) max, (7) min The bounds are

$$\begin{aligned}\frac{q_6}{q_c} &= \Lambda^{1-\phi}\\ \frac{q_{3\backslash 6}}{q_6} &= \Lambda^{-1},\end{aligned}$$

which implies

$$\begin{aligned}\frac{q_6}{q_c} &= \Lambda^{1-\phi}\\ \frac{q_{3\backslash 6}}{q_c} &= \Lambda^{-\phi}.\end{aligned}$$

The cubic equation of q_c has the following coefficients:

$$\begin{cases} A_3 = n_6 + n_{3\backslash 6} > 0 \\ A_2 = (n_6 + n_{3\backslash 6} - 1) + n_6\Lambda^{\phi} + n_{3\backslash 6}\Lambda^{\phi-1} > 0 \\ A_1 = (n_6 - 1)\Lambda^{\phi} + (n_{3\backslash 6} - 1)\Lambda^{\phi-1} > 0 \\ A_0 = -\Lambda^{2\phi-1} < 0 \end{cases}.$$

Case ④: (5) min, (7) max The bounds are

$$\begin{aligned} \frac{q_6}{q_c} &= \Lambda^{\phi-1} \\ \frac{q_{3\backslash 6}}{q_6} &= \Lambda, \end{aligned}$$

which implies

$$\begin{aligned} \frac{q_6}{q_c} &= \Lambda^{\phi-1} \\ \frac{q_{3\backslash 6}}{q_c} &= \Lambda^{\phi}. \end{aligned}$$

The cubic equation of q_c has the following coefficients:

$$\begin{cases} A_3 = n_6 + n_{3\backslash 6} > 0 \\ A_2 = (n_6 + n_{3\backslash 6} - 1) + n_6\Lambda^{-\phi} + n_{3\backslash 6}\Lambda^{1-\phi} > 0 \\ A_1 = (n_6 - 1)\Lambda^{-\phi} + (n_{3\backslash 6} - 1)\Lambda^{1-\phi} > 0 \\ A_0 = -\Lambda^{1-2\phi} < 0 \end{cases}.$$

Case ⑤: (6) max, (7) max The bounds are

$$\begin{aligned} \frac{q_{3\backslash 6}}{q_c} &= \Lambda \\ \frac{q_{3\backslash 6}}{q_6} &= \Lambda, \end{aligned}$$

which implies

$$\begin{aligned} \frac{q_6}{q_c} &= 1 \\ \frac{q_{3\backslash 6}}{q_c} &= \Lambda. \end{aligned}$$

The cubic equation of q_c has the following coefficients:

$$\begin{cases} A_3 = n_6 + n_{3\backslash 6} > 0 \\ A_2 = (n_6 + n_{3\backslash 6} - 1) + n_6\Lambda^{-1} + n_{3\backslash 6} > 0 \\ A_1 = (n_6 - 1)\Lambda^{-1} + (n_{3\backslash 6} - 1) > 0 \\ A_0 = -\Lambda^{-1} < 0 \end{cases}.$$

Case ⑥: (6) min, (7) min The bounds are

$$\begin{aligned}\frac{q_{3\backslash 6}}{q_c} &= \Lambda^{-1} \\ \frac{q_{3\backslash 6}}{q_6} &= \Lambda^{-1},\end{aligned}$$

which implies

$$\begin{aligned}\frac{q_6}{q_c} &= 1 \\ \frac{q_{3\backslash 6}}{q_c} &= \Lambda^{-1}.\end{aligned}$$

The cubic equation of q_c has the following coefficients:

$$\begin{cases} A_3 = n_6 + n_{3\backslash 6} > 0 \\ A_2 = (n_6 + n_{3\backslash 6} - 1) + n_6\Lambda + n_{3\backslash 6} > 0 \\ A_1 = (n_6 - 1)\Lambda + (n_{3\backslash 6} - 1) > 0 \\ A_0 = -\Lambda < 0 \end{cases}.$$

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

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
Non-localized Classes $\lvert L^0\rvert$	109	111	251	251
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%)

Table 4: Top 20 Localized Technology Classes

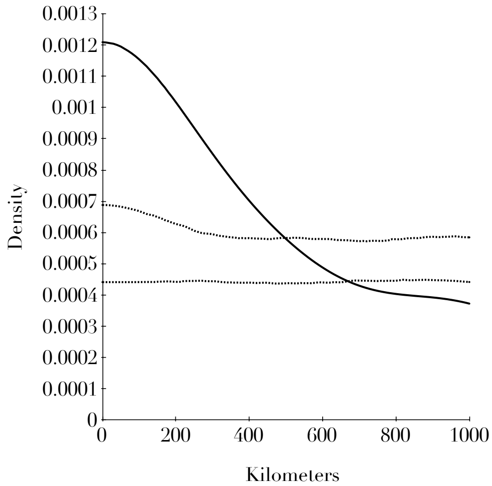
Class ID	Patent Class Name	Γ_A	Overlapped
3-digit controls			
405	Hydraulic and Earth Engineering	0.0201	■
452	Butchering	0.0155	
36	Boots, Shoes, and Leggings	0.0153	
223	Apparel Apparatus	0.0145	
606	Surgery	0.0143	
367	Communications, Electrical: Acoustic Wave Systems and Devices	0.0135	
296	Land Vehicles: Bodies and Tops	0.0122	■
285	Pipe Joints or Couplings	0.0106	■
492	Roll or Roller	0.0103	
181	Acoustics	0.0100	
30	Cutlery	0.0098	
501	Compositions: Ceramic	0.0095	
411	Expanded, Threaded, Driven, Headed, Tool-Deformed, or Locked-Threaded Fastener	0.0089	■
254	Implements or Apparatus for Applying Pushing or Pulling Force	0.0088	
256	Fences	0.0087	■
239	Fluid Sprinkling, Spraying, and Diffusing	0.0082	
290	Prime-Mover Dynamo Plants	0.0081	
303	Fluid-Pressure and Analogous Brake Systems	0.0078	
192	Clutches and Power-Stop Control	0.0078	
112	Sewing	0.0077	
6-digit controls			
256	Fences	0.0070	■
221	Article Dispensing	0.0038	
248	Supports	0.0030	
433	Dentistry	0.0029	
222	Dispensing	0.0025	
137	Fluid Handling	0.0024	
141	Fluent Material Handling, with Receiver or Receiver Coacting Means	0.0023	
296	Land Vehicles: Bodies and Tops	0.0023	■
301	Land Vehicles: Wheels and Axles	0.0023	
405	Hydraulic and Earth Engineering	0.0022	■
440	Marine Propulsion	0.0022	
411	Expanded, Threaded, Driven, Headed, Tool-Deformed, or Locked-Threaded Fastener	0.0022	■
166	Wells	0.0022	
285	Pipe Joints or Couplings	0.0022	■
508	Solid Anti-Friction Devices, Materials Therefor, Lubricant or Separate Compositions for Moving Solid Surfaces, and Miscellaneous Mineral Oil Compositions	0.0021	
2	Apparel	0.0020	
261	Gas and Liquid Contact Apparatus	0.0019	
198	Conveyors: Power-Driven	0.0019	
218	High-Voltage Switches with Arc Preventing or Extinguishing Devices	0.0019	
118	Coating Apparatus	0.0018	

Table 5: Matching Rate Test Results for Disaggregated Technology Classes

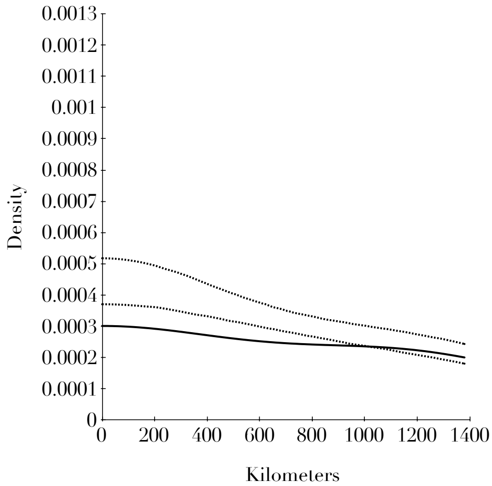
	3-digit Control			6-digit Control		
	State	CMSA	County	State	CMSA	County
All Classes $\lvert\mathscr{A}\rvert$	384	384	384	360	360	360
Localized Classes $\lvert L_1\rvert$	270	266	247	68	69	47
Non-localized Classes $\lvert L_0\rvert$	114	118	137	292	291	313
$\lvert L_1\rvert/\lvert\mathscr{A}\rvert \times 100$ (percent)	(70.31%)	(69.27%)	(64.32%)	(18.89%)	(19.17%)	(13.06%)

Table 6: Matching Rate Tests Conditional on K-density Tests for Six-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%)

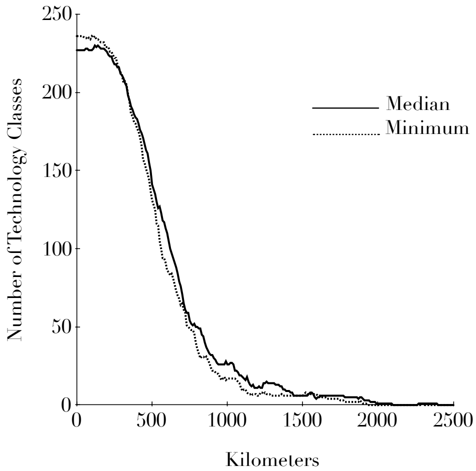


(a) Butchering (452)

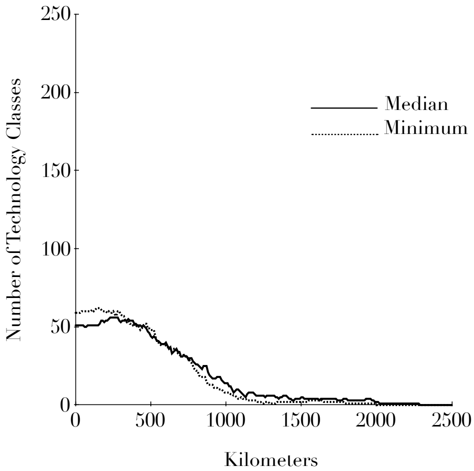


(b) Amusement Devices: Toys (446)

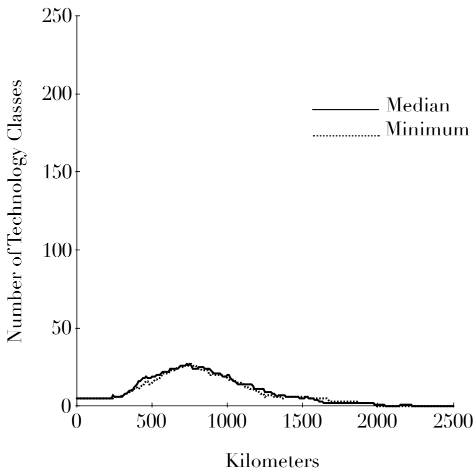
Figure 1: K-density and Global Confidence Bands for Two Illustrative Patent Classes



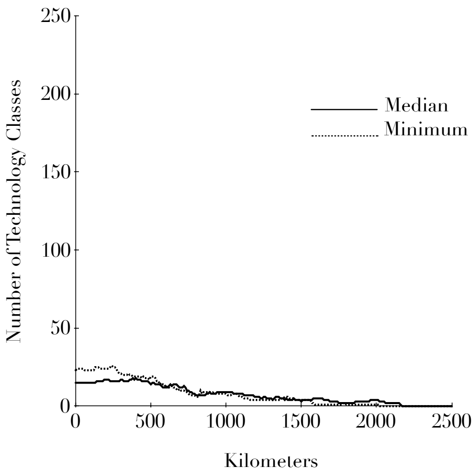
(a) three-digit controls: $|L^1(d)|$



(b) six-digit controls: $|L^1(d)|$

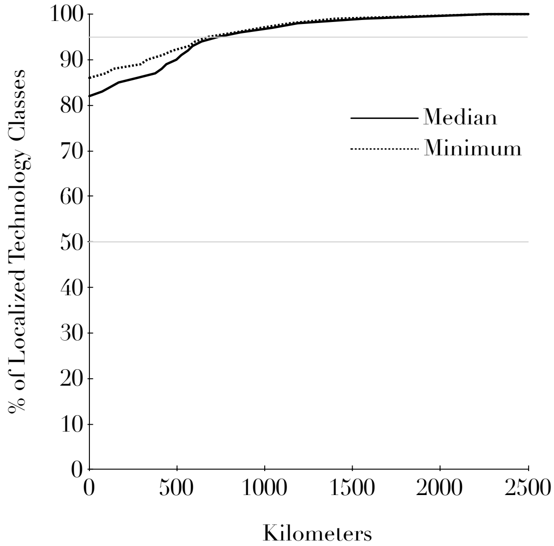


(c) three-digit controls: $|D^1(d)|$

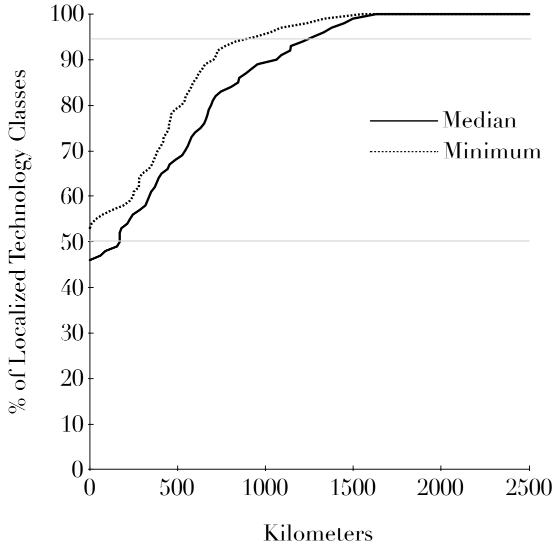


(d) six-digit controls: $|D^1(d)|$

Figure 2: Distance Distribution of the Numbers of Technology Classes

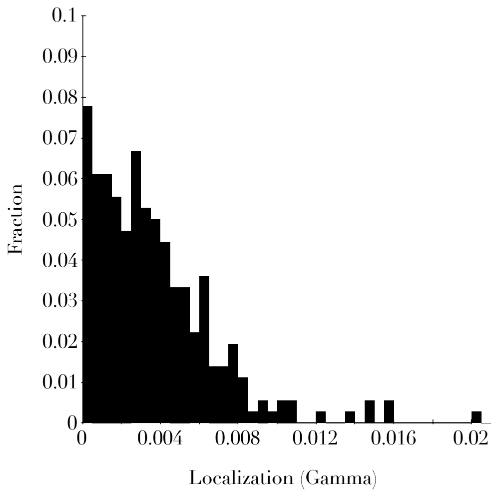


(a) three-digit controls

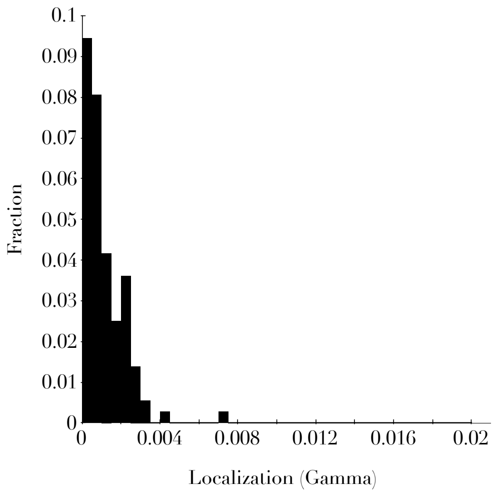


(b) six-digit controls

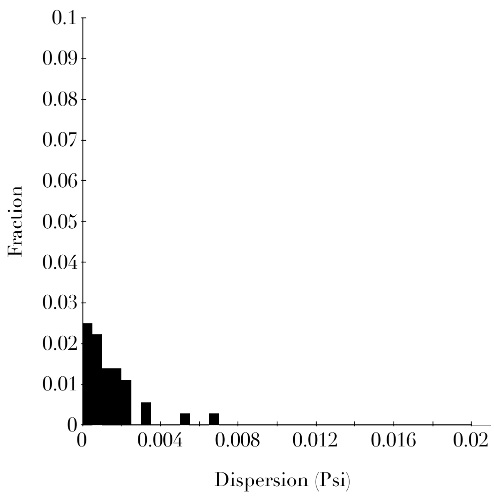
Figure 3: Percentage of Localized Technology Classes within Each Distance



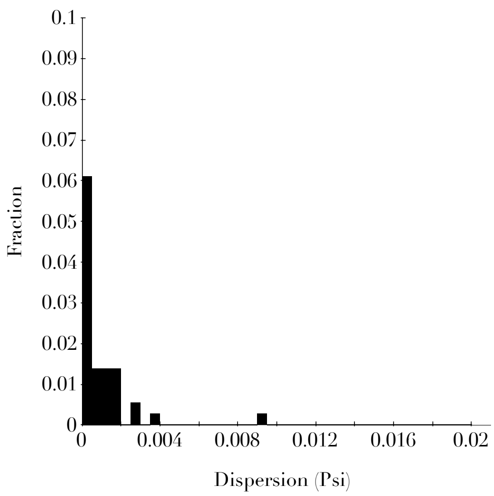
(a) three-digit controls: Γ_A



(b) six-digit controls: Γ_A



(c) three-digit controls: Ψ_A



(d) six-digit controls: Ψ_A

Figure 4: Distributions of Localization and Dispersion Indices

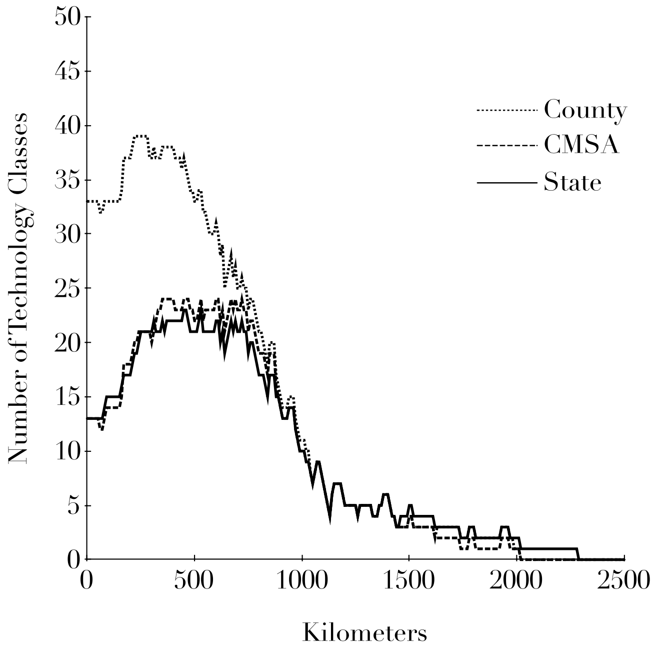


Figure 5: Distance Distribution of $|L_0^1(d)|$ for Six-digit Controls

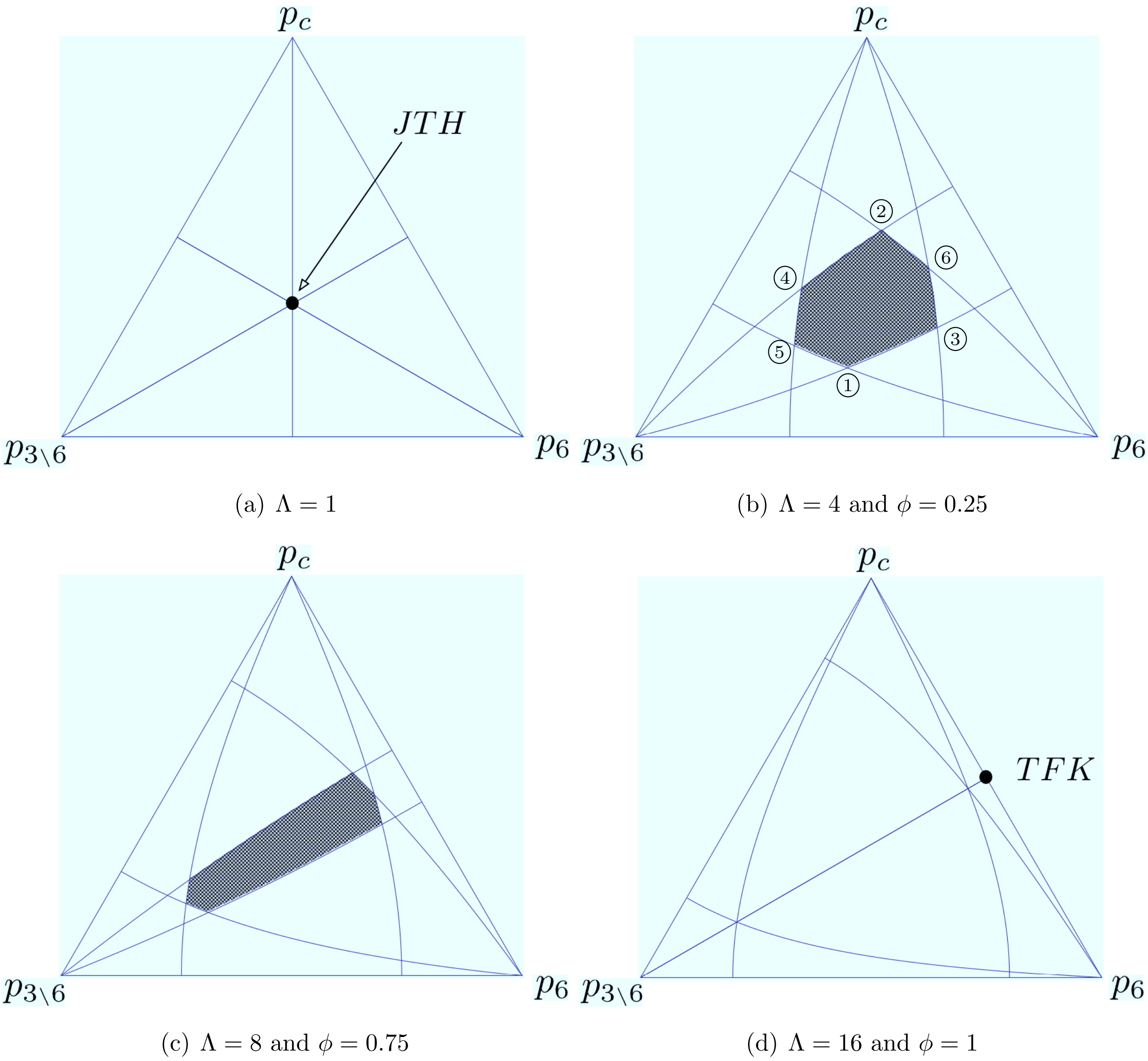
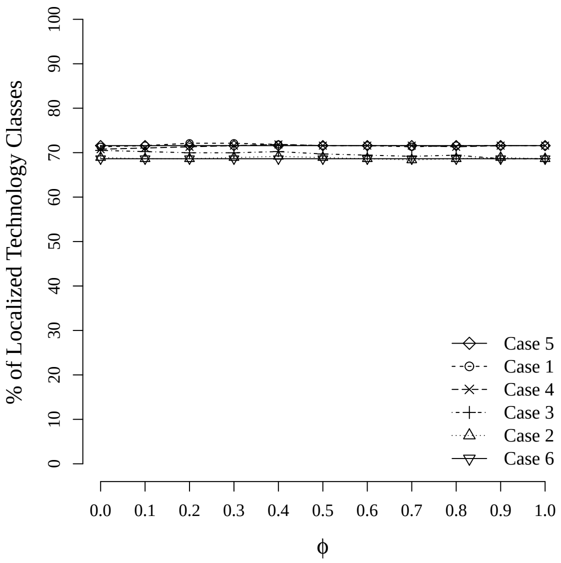
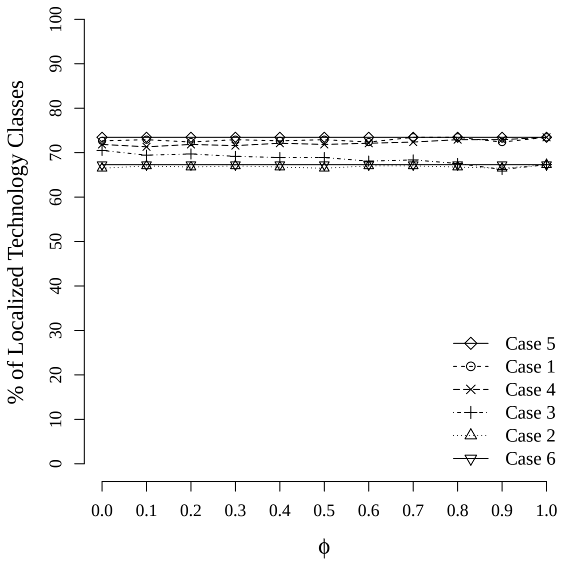


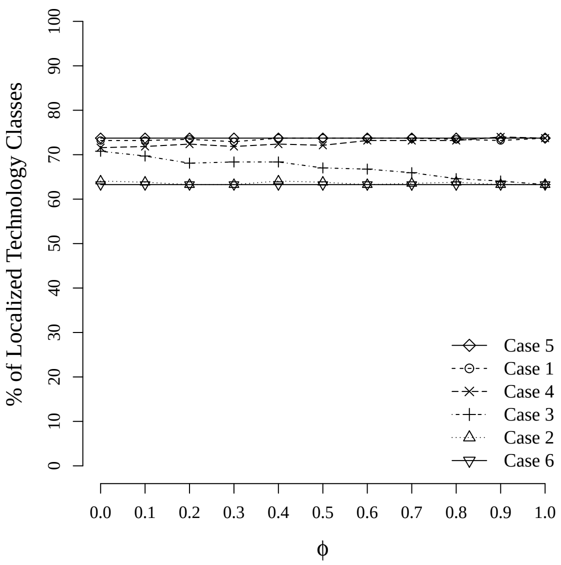
Figure 6: Feasible Probability Sets



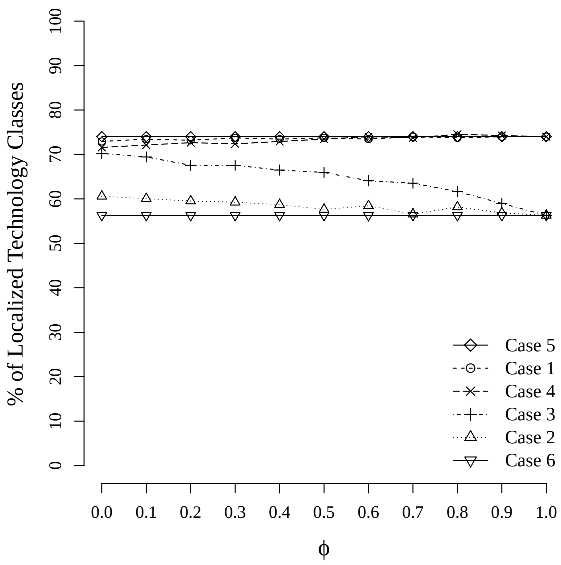
(a) $\Lambda = 2$



(b) $\Lambda = 4$

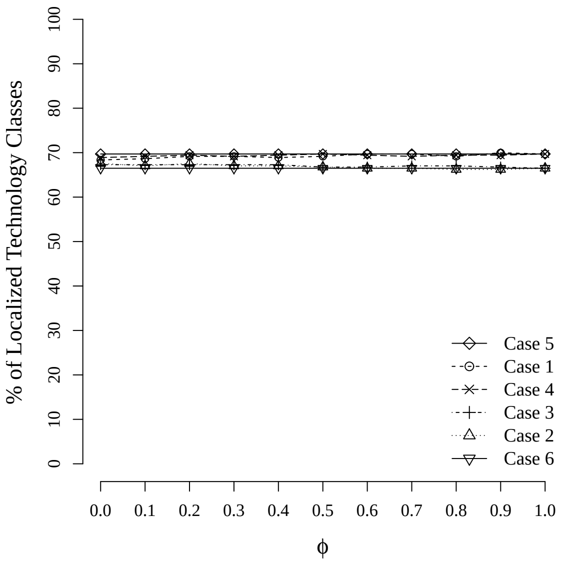


(c) $\Lambda = 8$

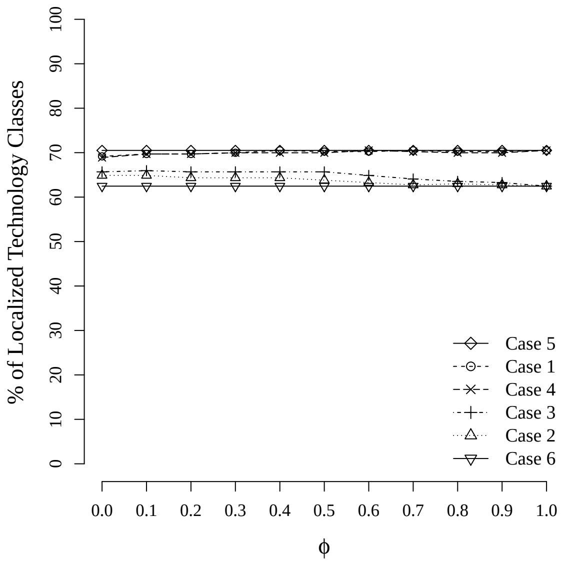


(d) $\Lambda = 16$

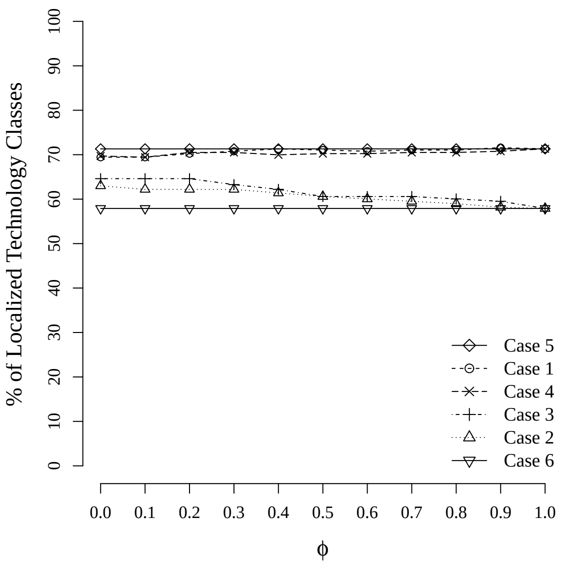
Figure 7: Sensitivity Analysis: K-density Tests



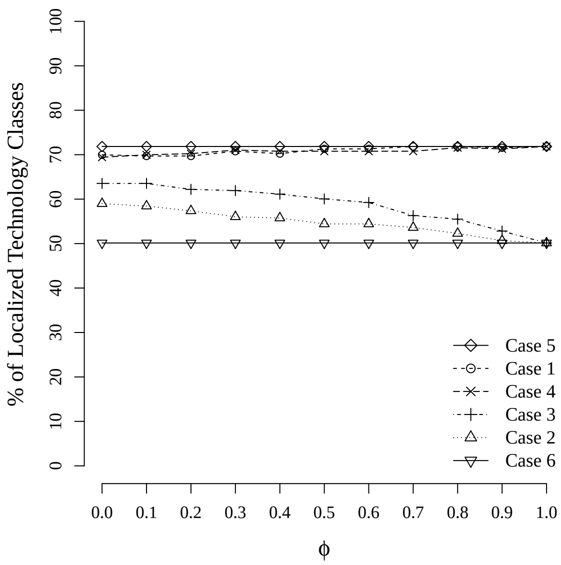
(a) $\Lambda = 2$



(b) $\Lambda = 4$



(c) $\Lambda = 8$



(d) $\Lambda = 16$

Figure 8: Sensitivity Analysis: Matching Rate Tests

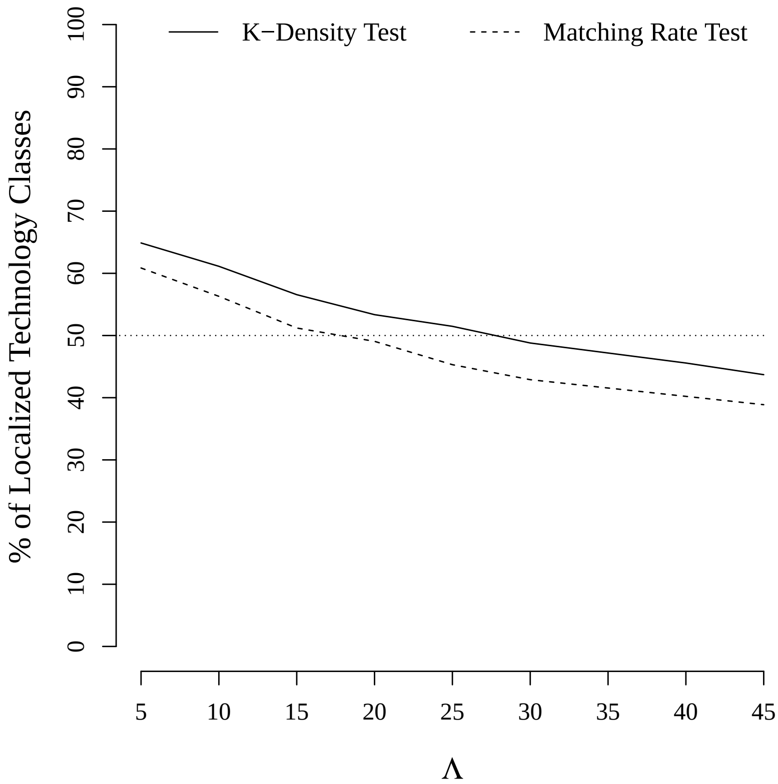


Figure 9: Worst-case Scenario Bounds