HARVARD | BUSINESS | SCHOOL

The Logic of Agglomeration

Gilles Duranton
William R. Kerr

Working Paper 16-037

The Logic of Agglomeration

Gilles Duranton
University of Pennsylvania Wharton School

William R. Kerr
Harvard Business School

Working Paper 16-037

The Logic of Agglomeration

Gilles Duranton and William R. Kerr

July 2015

Abstract: This review discusses frontier topics in economic geography as they relate to firms and agglomeration economies. We focus on areas where empirical research is scarce but possible. We first outline a conceptual framework for city formation that allows us to contemplate what empiricists might study when using firm-level data to compare the functioning of cities and industries with each other. We then examine a second model of the internal structure of a cluster to examine possibilities with firm-level data for better exposing the internal operations of clusters. An overwhelming theme of our review is the vast scope for enhancements of our picture of agglomeration with the new data that are emerging.

JEL: J2, J6, L1, L2, L6, O1, O3, R1, R3

Keywords: Agglomeration, clusters, cities, innovation

Author institutions and contact details: Duranton: University of Pennsylvania Wharton School, duranton@wharton.upenn.edu; Kerr: Harvard University, wkerr@hbs.edu (corresponding author).

Acknowledgments: This paper is a draft chapter for the *Oxford Handbook of Economic Geography*, edited by Gordon Clark, Maryann Feldman, Meric Gertler, and Dariusz Wojcik.

1. Introduction

A core topic in economic geography is agglomeration economies, where cities and clusters of activity boost the productivity of firms located within them. Conceptual rationales date back to Marshall (1890), and theorists have done a remarkable job of formalizing and codifying these concepts, as reviewed by Duranton and Puga (2004) and Behrens and Robert-Nicoud (2015). The empirical literature has been slower to develop, however, especially as it relates to the role of firms in agglomeration forces. Most early studies on agglomeration economies focused on measuring the wage premiums paid to urban workers, which was quite natural given the ready availability of person-level data collected through population censuses. While these data and studies could speak to the existence of agglomeration economies, they were inadequate for identifying channel(s) through which agglomeration economies percolate and how they shape behavior. Said differently, the higher wages paid in larger cities could have descended equally from several different forms of firm and worker interactions. This was not just a loss for our academic description of the world—without knowing, for example, whether agglomeration forces are arising due to knowledge flows or better matching externalities, practical advice for policy and business leaders is limited.

The most important advance for empirical research on agglomeration economies over the last two decades has been the development of firm- and establishment-level datasets of economic activity. These data have opened up the possibility of quantifying the role of firms in agglomeration economies and the productivity boost that large cities and clusters provide. Moreover, they have allowed the advent of new lenses for studying these questions: for example, continuous distance measures of geographic concentration (e.g., Duranton and Overman 2005), estimations of productivity spillover decays within cities (e.g., Rosenthal and Strange 2004, Arzaghi and Henderson 2009), firm selection mechanism and productivity (e.g., Combes et al. 2012), and dynamic perspectives that include firm entry and exit (e.g., Dumais et al. 2002, Klepper 2010, Glaeser et al. 2015). Alongside has been the development of complementary datasets for observing the actual interactions of workers and firms (e.g., employer-employee data, patents and citations). These data have also encouraged the measurement of the exchanges that we believe underlie the agglomeration economies—e.g., the many studies using patent citations to measure local knowledge spillovers following Jaffe et al. (1993)—and recent efforts have started to unite these with detailed firm location data.

This paper reviews two conceptual frameworks and proposes some interesting ways in which new micro-level data can advance our understanding of the models. Our first framework in Section 2 considers the formation of cities as the equilibrium outcome of benefits to firm agglomeration in cities against the growing cost of living that larger cities endure (e.g., land scarcity, congestion, pollution). This work pulls from Duranton (2008), and we use this lens to describe useful ways that firm-level data in rich and developing economies can advance our insights into the differences across cities in these economic forces. Our second framework in Section 3 is a conceptual model of interactions within a given cluster that pulls from Kerr and Kominers (2015). We describe through this lens the very nascent work on local interactions of firms and workers and how they give rise to agglomerative clusters. This is an area where we anticipate large advances will occur over the next two decades as we break open the black box of internal cluster dynamics and the relationships that define agglomeration economies.

There are two large boundary conditions for this piece. First, we only cover these two topics of interest out of a sea of opportunity. For example, firm-level studies of how multi-unit firms interact with local agglomeration economies versus internally sourced resources are woefully few in number (e.g., Tecu 2012, Alcacer and Delgado 2013). Likewise, a better understanding of how clusters interact with each other across countries and the role of multi-national firms is critical for today's global interconnectivity (e.g., Saxenian et al. 2002, Alfaro and Chen 2014). Other examples include the extensive margin of employer-employee relationship (e.g., spinouts and entrepreneurship), the implications of firm-level market frictions like financing constraints for agglomeration, and so on. Second, our review does not contain many references, and space constraints require that we be equally stingy towards the classics and towards recent contributions. We focus here just on ideas related to firms and these two frameworks. Previous and contemporaneous reviews contain much more extensive documentation (e.g., Feldman 2000, Duranton and Puga 2004, 2014, Rosenthal and Strange 2004, Audretsch and Feldman 2004, Feldman and Kogler 2010, Carlino and Kerr 2015, Combes and Gobillon 2015).

Finally, and most important, while we conclude in this piece that new data and fresh thinking have opened up many exciting research opportunities regarding agglomeration with firm-level data, it is important to emphasize that a mountain of data is never a substitute for a convincing research design. Agglomeration is a very complex process that involves trade-offs and equilibrium outcomes, and the best empirical progress in this field comes when researchers identify a razor-sharp way to cut through this complexity and identify causal relationships. We mention this upfront, so that we don't have to endlessly repeat the warning throughout the review! We will make the most progress when these big opportunities are met with the right methodologies.

2. Differences in Agglomeration Across Economies

We begin with a framework that depicts the formation of cities as an outcome of productivity benefits, cost factors, and labor supply/migration decisions. Duranton (2008, 2014) fully develops this framework with implications and examples that we do not repeat here, and earlier developments of this conceptual framework are contained in Combes et al. (2005). This framework visually describes the key elements of a literature dating back to Henderson (1974) and even earlier on the fundamental factors that influence city size and composition.

Figure 1 shows the three curves that are central to this framework. In both parts of Figure 1, the horizontal axis measures city size or possible labor supply as given by a population N. The upper part of Figure 1 first shows the wage/productivity curve w(N). Models of agglomeration feature an upward sloping wage curve where increases in a location's economic activity boost the productivity of firms in the city or cluster and the wages that they pay. As noted in the Introduction, this relationship has been empirically shown in many different settings and time periods. In the original formulation of Marshall (1980), agglomeration forces were expressed in terms of customer-supplier interactions, labor pooling, or knowledge exchange. Duranton and Puga (2004) provide a more theoretically amenable framework that emphasizes the sharing, matching, and learning processes of cities. Many classic studies have isolated features of these agglomeration economies, such as Helsley and Strange (1990), Porter (1990), Saxenian (1994), and Audretsch and Feldman (1996).

[Insert Figure 1 here]

Underlying these collocation benefits is the ability of firms to ship and sell their products on larger markets. An intuitive example is Hollywood. The spatial concentration of the movie production industry allows many productivity gains—e.g., better matching of actors and actresses to specific parts, the emergence of specialized law firms that support the entertainment industry, the development of schools to train employees in primary and supporting roles—with all production studios ultimately serving and competing in a global market for the best movies and largest global audiences.

The top panel also shows a cost curve, H(N). Costs are also rising with city size. This aspect of agglomeration can be forgotten at times by academics—to the extent that they adopt a "more is always better" view of clustering—but is very evident to those living in large cities. These costs come with higher rents and transportation costs, greater congestion, and perhaps environmental degradation (e.g., air pollution). Simply put, prices on Wall Street in New York or Sand Hill Road in Silicon Valley are extraordinary. In addition, the price of consumption goods may rise or fall with city size, depending upon the outcome of higher input prices vs. potentially greater competition and diversity in large cities. As reviewed in greater detail by Duranton (2014), the cost curve is significantly less studied than the wage curve.

Subtracting the cost curve from the wage curve, one determines a net benefit for w(N)-H(N), as shown in the lower half of Figure 1. This benefit is initially rising in city size as the early productivity benefits from larger city size dominate cost increases. After point B, these net benefits decline. To determine the stable size of a city, one must further specify how labor supply responds to the net benefits that a city provides (taken to be through migration to the city for simplicity). The figure presents a case where labor responds with some elasticity to higher net benefits, but that labor is not perfectly mobile (which absent amenities would be a horizontal line at a given net benefit level due to the spatial equilibrium). The stable equilibrium that forms will be at point C. The fact that the city is larger than what maximizes net benefits is commonly observed.

There are many possible avenues for comparative analyses across cities using firm-level data that we can illustrate with this model. (There are also many interesting questions that go beyond firm-level studies, but we focus on these opportunities exclusively given the emphasis of this review.)

As a starting point, we may investigate how the wage/productivity curves of cities are influenced by micro-interactions within clusters. The framework makes a reduced-form connection of the upward sloping wage curve due to productivity benefits, which is certainly observed in wage relationships. Yet we have collected very little evidence that quantifies the different agglomerative mechanisms in the background, especially in a comparative form across cities. Likewise, we have few studies that attempt to discern the quantitative roles of candidate explanations and assign importance. The explosion in firm-level data offers great promise here.

Take, for example, the rise of employer-employee datasets that follow both workers and firms over time. In principle, one should be able to use these datasets to describe how well workers and firms are matched to each other. That is, controlling for the cost curve, does one observe improvements in employer-employee matching that are commensurate with differences in city sizes? As a second example, many studies look at knowledge flows in local areas as an agglomerative rationale. Work in this line can turn towards tracing out how knowledge-intensive firms accrue different benefits with larger cities that make them more productive. Alternatively, the mobility of workers of workers may also act as a vehicle for local spillovers (Serafinelli, 2014). What part of the wage curve slope comes from a greater volume of ideas, more timely access to the latest ideas, the diversity of ideas circulated, and similar features?

Taken a step further, empirical work can also assess how firm production functions change with city size. Within the same industry, for example, some production models have stronger dependency on the complementary matching of team members, similar to the O-Ring model of Kremer (1993). The theory posits that the success of a group depends deeply of the strength of its weakest member. Such models may be more likely to emerge in the largest cities, given that a deeper labor pool can allow for better choices across all of a team's needs, and thereby provide added curvature to the wage/productivity relationship. Anecdotally, for example, joint recruiting of complete teams on Wall Street across investment banks seems quite common, and Silicon Valley now talks about "acqui-hiring", where start-ups are acquired mostly for their start-up teams rather than developed technologies. In a similar spirit, multi-unit firms place their headquarters frequently in larger cities, which could be in part be due to this complementarity across top management team functions.

Greater city size can also shape the industrial organization of the cluster, perhaps giving steepness to the wage curve. Fallick et al. (2006) describe how high-velocity labor markets in clusters can facilitate a modular production structure. At the start, many small firms compete to identify the best idea or design, and then in the second stage the winning firm rapidly scales by hiring employees from the losing firms. This can speed innovation and create design improvements, if the labor market conditions and modular production design features are evident. Fallick et al. (2006) provide some evidence for this industrial organization in the computer industry in California, which echoes the anecdotal accounts of Saxenian (1994). More systematic assessments of how city size changes industrial organization is most warranted and feasible with firm-level records.

Greater availability of new data, including data from previously-underrepresented regions, allows us to take a wider view and examine variations in clusters worldwide. Most empirical work on the subject considers advanced economies, but there are many firm-level datasets for developing and emerging economies that are coming into widespread use. Duranton (2014) reviews the higher elasticities for the wage curve in China and India compared to advanced economies, and there is heterogeneity in outcomes among advanced countries. Further refining these cross-country comparisons will help identify how economic development gaps emerge from an agglomeration perspective. These may be connected, for example, to the limited scaling of productive plants in India and Mexico compared to the United States that is identified by Hsieh and Klenow (2014).

Moreover, developing and emerging economies may provide interesting insight into the model overall to the extent that their settings and institutions isolate some features from others. Local infrastructure, for example, is quite deficient in many poor countries and the subject of massive investments by national governments and international organizations. Many efforts are being made to upgrade these facilities, which can allow for real-time analysis of firm behavior as investments are being made, so long as a suitable control group can be identified. Many studies are pursuing such angles in Brazil, China, India, etc. for highways and railroad access (Redding and Turner 2015). As the bulk of these investments for the United States came before the collection of firm-level datasets, the delayed start in other countries provides a unique advantage for “before and after” views of the role of infrastructure for individual firm choices (e.g., input sourcing) and for overall distributions of firm outcomes.

Similarly, developing economies face many challenges that are mostly absent in advanced countries. Three examples include dual housing markets (e.g., squatter housing without legal property rights), inefficient migration, and city favoritism, and Duranton (2008, 2014) traces out some of the conceptual implications from these three factors. To the extent that these factors isolate some features—e.g., city favoritism can be thought of as raising the wage curve of the primal city without impact on actual firm productivity—we can learn more about the function of agglomeration and city formation generally in a specialized laboratory.

At the same time, this framework has applications to related areas of economic research, including questions of interest to policy makers and business leaders. As one example, it can help to illuminate the role that migration plays in the development of clusters.

The labor supply curve plays an important role in governing the degree of equilibrium net benefits in cities. Many urban models feature a spatial equilibrium with full mobility that pins down the same net benefits across all cities, described for example in Glaeser (2008). Deviations from this framework can provide an interesting range of labor supply responses. Constraints on internal migration in some countries (e.g., the Hukou system in China) may severely limit the ability of locations to expand, or push this expansion into non-optimal forms like secondary labor markets. The extra-legal nature of these migrations may place constraints on how much firms can take advantage of agglomeration economies that would otherwise emerge. Related, progress has been made in India for how organized sector firms tap into the unorganized sector that surrounds them through sub-contracting relationships (e.g., Mukim 2015).

At the other end of the spectrum, many immigration systems allow firms to choose locations for workers who are legally tied to the employer. In the United States, for example, most skilled immigrants hired for employment come through the H-1B temporary visa program or are an intra-company transfer under the L-1 visa program. In both cases, employees are effectively assigned geographic locations while they are on the visa. This has the interesting implication of allowing firms to overcome some of the limits imposed by the labor supply curve in this conceptual framework. Indeed, Kerr and Lincoln (2010) and Ruiz et al. (2012) describe how U.S. immigration visas are used by firms in locations where they struggle to acquire labor resources, in addition to the better-known cases of high-tech clusters. Linking firm-level data with immigrant employee traits may afford opportunities to learn about the limits locations place on firms by evaluating the workarounds that firms seek.

For another example, firm-level datasets have opened up new possibilities for the study of clusters and entrepreneurship, which is often viewed as a central driver for city growth. Recent examples of this work include Chinitz (1961), Michelacci and Silva (2007), Glaeser et al. (2010), Delgado et al. (2010), Behrens et al. (2014), and Glaeser et al. (2015), and yet, some big open areas remain for study.

First, there are two views of entrepreneurship supply across regions. One view is that entrepreneurs are very mobile and move across cities to opportunities. This is surely the right perspective for very high-growth entrepreneurship like that observed in Silicon Valley. A second view is that entrepreneurs are very sticky and local in nature, such as the Chinitz (1961) depiction of differences in entrepreneurial supply across U.S. cities due to their industrial legacies. This is most likely the right depiction for the lowest end of work, as the "local bias of entrepreneurship" studies observe. Behind these two models are some very natural and intuitive economic forces like effective market size, the importance of local connections for business sales or financing, winner-take-all dynamics, and similar. Silicon Valley is home to a special cluster of consumer-internet firms given the very high agglomeration benefits in this industry, while cement manufacturing sits at the opposite end of the spectrum. What we have yet to discern is the relevant ranges of industries over which these two models operate—where do we tip from a local dynamic and sticky places to a world of specialized clusters?

Second, and somewhat building on the modularity point described above, Glaeser et al. (2015) identify how industrial dynamics differ in cities and clusters with a strong entrepreneurial base. Entrepreneurship does not yield city growth through an endless replication of small firms, but instead links to local employment growth through an up-or-out scaling process that scales up the best new entrants to become the largest employers in a city. This line of work is very young, however, and needs much greater empirical attention with firm-level data. This scaling relationship and the entrepreneurial supply functions depicted above will have first-order implications for whether the wage/productivity curve takes a common shape across cities or is very heterogeneous due to differences in industrial structures.

A final important feature of this literature is the depiction of why industries move across locations and the implications for the cities that house them. Duranton and Puga (2001) describe how product and technology maturity leads industries to move out of expensive nursery cities towards cheaper locations. In other work, Duranton (2007) formalizes a model of city evolution that has industry movements across cities related to breakthrough inventions at its core, and Kerr (2010) provides some empirical evidence along the lines of this model in terms of patenting behavior. Panel datasets on firms are starting to be of sufficient time dimension for firm-level analyses to provide additional insights about these dynamics. Likewise, faster product cycles are allowing more rapid observation of these movements.

Obviously, this section can only provide a partial list of possible topics for future research, as the abundance of new kinds of data has opened up countless ways to extend this model of city formation. The view of cities as a consequence of the competing benefits and costs that arise from clustering behavior has existed for decades in one form or another, and we now have the chance to flesh it out further by illuminating some of the underlying mechanisms or extending the model to new settings.

3. Structures of Interactions within Cities

We next turn to the internal structure of agglomeration economies for cities and the implications for firms. For this exercise, we use the theoretical framework of Kerr and Kominers (2015). This model and its empirical applications have roots in the observation that most micro-level studies of how workers and firms interact (e.g., commuting patterns, patent citations) show a much shorter geographic distance of interaction than the actual footprint of cities and clusters. Said differently, commuting patterns tend to 20 miles or less, but agglomerations for labor pooling appear to stretch much farther. While perhaps obvious, a second line of work—such as Rosenthal and Strange (2001, 2003), Duranton and Overman (2005, 2008), and Ellison et al. (2010)—finds that regional-based approaches for measuring agglomeration forces yield seemingly quite reasonable and informative depictions for how agglomeration forces influence cluster size. For example, these studies tend to find that technology- or knowledge-based clusters appear smaller and more tightly knit than agglomerations building upon labor pooling or customer-supplier connections, despite examining data that are orders of magnitude larger than the underlying forces believed to cause the cluster. This is true when using political boundaries (e.g., counties vs. states) or continuous distance (e.g., 100 vs. 500 miles).

The Kerr and Kominers (2015) model conceptualizes how these two empirical pieces can be reconciled through a theory of small, overlapping regions of interaction visualized in Figure 2. There is a set of sites for businesses that is shown by the dots. Firms enter in sequence and without foresight, and the set of potential sites is fixed. Sites with black dots have already been chosen by firms, and the light dots are sites available to the next entrant. Each firm participates in the cluster by interacting with neighbors that fall within a maximal spillover radius, indicated by the dashed circles. As noted below, the network to which the firm is connected by its neighbors can have arbitrary amounts of benefit transmission for the baseline model. The maximum spillover radius has a limited and defined boundary due to fixed costs of interaction that exceed a decaying benefit to interaction at some point.

[Insert Figure 2 here]

This model pictures large-area clustering that arises due to small, contained interaction effects that overlap each other. Two important points follow. First, the introduction of fixed costs and defined effective spillover boundaries yields interesting and testable predictions regarding the internal structure of clusters that we discuss below. This framework makes clear why the short-distance interactions that we measure with commuting flows or patent citations are different from the distances that we consider with region-based data. As we trace out, this model can start to identify many useful lines of inquiry regarding firms and agglomeration going forward.

Second, the model provides a rationale for why multiple clusters form. In our simple illustration, the next entrant is indifferent among available sites, including sites X, Y, and Z, because none of the remaining sites are within reach of the populated cluster. In other words, clusters fill up, and this marginal entrant will choose randomly among the remaining available sites and start a new cluster. This foundation provides a basis for comparative statics of spillover radius size and cluster structure. Consider, for example, a second agglomerative force that has a longer maximal spillover radius than what is shown in Figure 2 because the second force has a slower rate of benefit decay (or a lower fixed cost of interaction). Such a model would increase the size of the circles in Figure 2 and transform site X into the strictly preferred next location because it can participate in the existing cluster with a longer radius. The full model formally traces out how a longer maximal radius results in fewer, larger and less-dense clusters. Thus, we can use the shapes and sizes of clusters to back-out the micro forces beneath them.

From this launching point, Kerr and Kominers (2015) test the broader model predictions by measuring how far apart patents in 36 technologies tend to be from each other when they cite each other in their patent filings, a measure of the spillover radius for each technology. Some technologies like semiconductors use very localized networking, while others exhibit much longer spatial horizons. Using estimates from the United States and the United Kingdom, they show that technologies with shorter effective interaction distances exhibit smaller and denser clusters.

Kerr and Kominers (2010) contains an extensive analysis of technology and worker flows for Silicon Valley, an iconic agglomeration that is composed of overlapping tech spaces. We select a couple of these pieces to illustrate the model in greater depth and provide some new research ideas about the internal structure of clusters. Walking through a couple of detailed maps will help us visualize the concepts that follow.

[Insert Figure 3 here]

Figure 3 describes the construction of the technology core for Silicon Valley. Looking across the entire San Francisco Bay Area, the core of Silicon Valley includes the top 10 zip codes in terms of patent filings and 18 of the top 25. Panel A shows the three most important zip codes—for each, the zip code is indicated with the star, and the other three points on the connected shape are the three zip codes that the focal zip code cites the most in its patent filings (Hall et al. 2002). On average, these three external zips contain 41% of local external citations for a zip code. Panel B does the same for the top 10 zip codes. The primary technology sourcing zones for these zip codes in the core are also fully contained in the core, even though we have made no restrictions in design and these sourcing zones could have been included anywhere in the San Francisco Bay Area. These zones are small, overlapping regions, and in the next map, we represent the core as a shaded triangle whose longest side is 25 miles in length.

[Insert Figure 4 here]

Figure 4 next shows the seven largest zip codes for patenting that are not contained in the core itself (#12, 13, 17, 19, 22, 24, 25). The Silicon Valley core depicted in Figure 3 is represented on this larger map as the shaded triangle. Similar to Figure 3, the shape of each technology sourcing zone is determined by the three zip codes that firms in the focal zip code cite most in their work. For visual ease, San Ramon and Santa Clara are indicated on the edge of the map at the location of their primary transportation route. Downtown San Francisco and Oakland, CA, are to the north and off of the map.

As observed for the core, these technology zones are characterized by small, overlapping regions. The three zip codes that are labelled with numbers are three of the four largest zip codes for patenting in the San Francisco Bay Area that are outside of the Silicon Valley core. Zone 1, which covers Menlo Park, extends deepest into the core. Zone 2, for Redwood City, CA, shifts up and encompasses Menlo Park and Palo Alto but has less of the core. Zone 3, which covers South San Francisco, further shifts out and brushes the core. None of the technology sourcing zones traverses the whole core, much less the whole cluster, and only the closest zip code (Menlo Park) even reaches far enough into the core to include the area of Silicon Valley where the greatest number of patents are issued. While technology sourcing for individual firms is localized, the resulting cluster extends over a larger expanse of land.

There are many possible avenues for further research that we can illustrate with this model.

One line of investigation addresses structural variations within a cluster. A comparison of Figures 3 and 4 suggests that the cluster's structure in the Silicon Valley core may be different from what exists at the periphery. Both areas show localized interactions, but the core exhibits great overlap among these regions in Figure 3 that resembles the density around site C in Figure 2's illustration. By contrast, the periphery in Figure 4 shows overlapping zones that resemble the A-B-C structure of sites in Figure 2. The baseline model is mostly agnostic about these features and to whether each location only derives benefits from the locations that it directly touches vs. all members of the cluster benefiting equally. Nonetheless, stronger empirical and theoretical depictions of how this networking exists, how it is priced into wages and locations, and so on is first order. Arzaghi and Henderson (2008) provide some characterization with advertising agencies in Manhattan, and Rosenthal and Strange (2003) comparatively estimate production function decays in several industries. Kemeny et al. (2015) further describe the central role of dealmakers and social capital in local areas. The availability of firm-level data with detailed geographic coding provides many opportunities here.

On a similar note, it may also be possible to determine a relationship between the types of interactions within clusters and the shapes of the zones that arise. Technology spillover zones are directional in nature. Our depictions of sourcing zones are unrestricted in the sense that the three most important zip codes could lie in any direction from the focal zip code. In Figures 3 and 4, however, the technology sourcing zone almost always lies within a 90-degree arc from the focal zip code. This pattern is very strong at peripheral locations, as there is a general flow of information or knowledge from the Silicon Valley core to Redwood City and Palo Alto, and then further to South San Francisco, and so on. Even within the tightly-knit core itself, the zones are remarkably lop-sided.

Kerr and Kominers (2015) also analyze the commuting patterns of scientists and engineers using IPUMS data. Commuting patterns tend to be more diffuse, and as a consequence the zones of interaction around labor inputs into firms appear significantly less directional. Visually, one is more likely to see workers for a firm commuting in equal measure from all sides, compared to technology sourcing that tends to be concentrated in one direction. These issues need to be traced out further, both in terms of how they reflect the nature of agglomeration forces (e.g., the pooling nature of labor inputs) and how they then affect the overall structure or operations of resulting clusters. The interactions between incumbents and entrants also deserve greater attention (e.g., Combes and Duranton 2006).

Expanding on this theme, there is a surprising gap in our knowledge about how skyscrapers affect the structure of interactions. Naturally, tall buildings allow greater density for a given land area, but do they do more by adding a third dimension to local structures? This is unlikely to impact patenting and technology firms, which tend not to locate in the core of urban areas, but it could be central to the functioning and organization of very high-end financial and professional service firms like Wall Street. Observing heavy levels of agglomeration that occur within a single large building, versus across nearby skyscrapers, may signal the relative importance of different types of agglomerative forces (e.g., knowledge sharing versus labor pooling). Many cities have imposed and later removed maximum building heights, allowing models of these forms to be tested empirically with firm-level data.

Shifting focus a bit, this model can also be extended to include the impact of physical features on clusters and their constituent firms. For example, the background to Figure 4 demonstrates the roles of geographical features (e.g., mountains, protected land) and transportation routes (e.g., highways, bridges). These forces substantially govern how the peripheral zones access the core of cluster. While not shown for visual reasons, these same features also play an important role in the technology flows evident in the core of Silicon Valley. These connections between local infrastructure and firm-level interactions are quite understudied and yet important for local policy and business choices. Agrawal et al. (2014) provide a recent contribution on roads and knowledge flows within cities that signals this importance. Moreover, new tools now exist for measuring travel time and associated costs that can translate spatial distances into real terms and account for congestion. For example, Google Maps reports that the expected lengths of time needed to drive across the three labeled technology zones in Figure 4 are comparable, which may indicate the length of the sourcing zones is determined more by interaction costs rather than true spatial distance.

Similarly, the model has straightforward extensions that allow for fixed natural advantages that firms also want to access. These could be traditional natural advantages (e.g., harbors, coal mines) or "manmade" advantages like universities, military bases, or state capitols. We have the capacity now to understand better how MIT's location affects the biotech community that surrounds it in Kendall Square and, in turn, the broader Cambridge and Boston communities.

We can also examine more closely the detailed “inner workings” of clusters and the concerns of individual firms. An important theme throughout this work is the breaking up of a city or cluster to recognize that all firms do not automatically participate equally in benefits. At one level this is obvious, but it is remarkable how little our existing work factors in where in a cluster a firm is located. Moreover, recent studies emphasize the differences in locations within clusters for women-owned firms and their networking benefits (e.g., Rosenthal and Strange 2012, Ghani et al. 2013), and an extensive literature describes similar or worse segmentation on racial lines. This segmentation means excluded groups receive unequal benefits from the cluster. Chatterji et al. (2014) describe policy issues related to this and other micro-cluster perspectives related to entrepreneurship and innovation.

New firm-level data also allow the study of entry and exit, with one major theme of recent empirical work being the high degree to which we observe entry and exit coinciding with each other in local areas. Said differently, while rapid long-term growth may favor entry over exit in a cluster, the bigger take-away from the data is that some areas are very dynamic, with lots of entry and exit, while others are less so. These features deserve greater attention. Moreover, entry and exit might allow more advanced statements about the attractiveness of places. Our simple model highlights the important information contained in the location decisions of marginal entrants, which in principle could be observed through panel data. Pricing theory emphasizes the valuable information contained at the margin, and some of these insights could be applied to the study of clusters if researchers fully learn how to harness firm dynamics in local areas.

On the management front, scholars can use new data to study how the location choices of individual firms affect their performance. This is challenging, of course, given that location choices are intimately connected with business models and strategies, as reviewed in the MBA course material of Kerr and Brownell (2011) and Kerr et al. (2011). A starting point is to note that 1) benefits differ across firm types for each location, 2) costs are mostly generic as landlords tend to rent property for the rental prices regardless of tenant’s industry, and 3) managers have limited knowledge about all of these features. Thus, one should find differences in performance depending upon the quality of site chosen for a particular business. More speculative, the presence of directional flows like the technology zones in Figure 4 open up the possibility for asymmetry over short distances—whether a firm locates in the path of flows from the cluster core towards an important periphery point may make a meaningful difference in the frequency and quality of interactions the firm makes, even holding fixed the distance from the core and the overall density/prices of the area selected.

Finally, while much of this discussion in this chapter focuses on clusters formed by firms within the same industry, a fruitful area for future investigation is the clustering behavior of firms in multiple industries. Recent research emphasizes the important degree to which firms in related industries interact (e.g., Ellison et al. 2010, Faggio et al. 2014) and theoretical constraints on the degree to which these joint location decisions are well aligned (e.g., Helsley and Strange 2014). There seems to be an unlimited scope for empirical advancement in this regard, and we note three pieces here. First, Jacobs (1970) and Duranton and Puga (2001) describe how local industrial diversity can give rise to new industries, and micro-data on firms provide deep scope for advancing our understanding of nursery cities and this crossfertilization process. Second, several studies describe the particular importance of supplier industries for entry choices following Chinitz (1961), including Helsley and Strange (2002) and Glaeser and Kerr (2009). With new data, one should be able to follow the material flows along these lines. Finally, model of entrepreneurship often depict founders as picking up a varied, "jack of all trades" background (e.g., Lazear 2005), which can be studied in local areas using employer-employee data and movements of future founders across firms.

4. Conclusions

The advent of micro-level data is a boon to our understanding of agglomeration economies. In the short time since data like the Census Bureau's Longitudinal Research Database first became available, researchers have fleshed out a much richer portrait of economic geography. Likewise, many European countries now have tremendous data resources for studying these issues given the depth of personal and firm-level information available. As a consequence, firms are no longer anonymous and in the background, but are instead playing a central role in our modern depictions of the internal workings of cities and the differences over locations. The future promises to be as bright with the development of new employer-employee datasets, the linking of micro-data like patents to the establishment-level records in administrative datasets, and the known and unknown tools that big data may shortly provide to researchers.

The two conceptual lenses that we developed in this review highlight the massive opportunity for empirical research that lies ahead. We won't repeat here the tactical ideas provided within this chapter, but instead circle back to two broader themes. First, agglomeration is fundamentally an equilibrium outcome where firms and workers weigh benefits against costs. Theoretical models in this regard are quite well developed, and there is a valuable tradition in this field for a heavy interplay of theory with empirics. Thus, empiricists can and should use theory to identify interesting conceptual topics that exist but have little empirical foundation. In some cases (e.g., matching externalities, infrastructure impacts), new and forthcoming data are already opening up exciting research opportunities. In these cases, our main hope is that researchers craft empirical work into frames that match theory (e.g., measuring commuting time costs vs. simple distances). We also hope that researchers make more use of powerful computing resources and estimation procedures to shed light on higher-order and non-linear effects, beyond the standard linear models. This curvature in treatment effects is central for our understanding of agglomeration, as Figure 1 shows, but poorly measured to date.

In other cases, the development of new data remains a priority and one that should be valued by the profession. To us, three issues are paramount. First, we need a better understanding of how agglomeration operates in large cities that are full non-manufacturing firms. This requires getting beyond patent data to measure knowledge spillovers, and it demands greater attention to the economics of skyscrapers. Second, we need to think harder about global interconnections and how this impacts clusters and firms in different countries. Agglomeration fundamentally connects into trade of outputs, and this does not stop at national borders for goods or services. In the theme of this chapter, the natural starting point here is richer work with multi-national firms and the operations of their many establishments. Finally, the urbanization of the developing nations is one of the biggest issues facing the world over the next decade. We are woefully behind on understanding how agglomeration is similar or different in these places, which is a first-order concern given that our existing insights are being used to define policies here and now! We are very hopeful that in 20 years, we will look back on many studies that developed new data for clusters outside of the developed world, harnessed real-time variation in the emergence and growth of these economies to garner insights that advanced economies cannot reveal given their long-standing development, and delivered a deep and beneficial policy impact.

Finally, we re-iterate that a mountain of data does not equal insight. For insight to be realized, these new data sets must be paired with strong and convincing research designs. At this point, the usefulness of the clever "natural experiment" to study these issues is well understood (e.g., Bleakley and Lin, 2012), and we hope that more of these empirical gems are unearthed. We also think the great research promise in many developing nations is to design interventions to facilitate future rigorous program evaluation. Yet, the real trade-off is not between unfounded correlations and the perfectly random shock. For important topics, we need to figure out the best ways to continually raise the empirical bar, recognizing the dual need to identify causal relationships and also acknowledge the equilibrium nature of these forces makes it very hard. We also need to learn to take better advantage of the insights that global data and research can provide. For example, we have good data for countries at many points along the spectrum of income inequality (e.g., very compressed Nordic structures to the United States and beyond). There appears to be good opportunity to learn through these global similarities and differences by embracing more meta work across places.

REFERENCES

Agrawal, A., Galasso, A., Oettl, A. (2014). "Roads and innovation." CEPR Discussion Paper 10113.

Alcacer, J., Delgado, M. (2013). "Spatial organization of firms and location choices through the value chain." Harvard Business School Working Paper 13-025.

Alfaro, L., Chen., M. (2014). "The global agglomeration of multinational firms." Journal of International Economics, 94 (2), 263-76.

Arzaghi, M., Henderson, J.V. (2008). "Networking off Madison Avenue." Review of Economic Studies, 75, 1011-38.

Audretsch, D., Feldman, M. (1996). "R&D spillovers and the geography of innovation and production." American Economic Review, 86, 630-40.

Audretsch, D., Feldman, M. (2004). "Knowledge spillovers and the geography of innovation." In: Henderson, J.V., Thisse, J-F. (Eds.), Handbook of Urban and Regional Economics, Vol. 4. North-Holland, Amsterdam, 2713-39.

Behrens, K., Duranton, G., Robert-Nicoud, F. (2014). "Productive cities: sorting, selection, and agglomeration." Journal of Political Economy, 122 (3), 507-53.

Behrens, K., Robert-Nicoud, F. (2015). "Agglomeration theory with heterogeneous agents." In Duranton, G., Henderson, J.V., Strange, W. (Eds.), Handbook of Regional and Urban Economics, Vol. 5. North-Holland, Amsterdam, 175-245.

Bleakley, H., Lin, J. (2012). "Portage: path dependence and increasing returns in U.S. history." Quarterly Journal of Economics, 127, 587-644.

Carlino, G., Kerr, W. (2015). "Agglomeration and innovation." In Duranton, G., Henderson, J.V., Strange, W. (Eds.), Handbook of Regional and Urban Economics, Vol. 5. North-Holland, Amsterdam, 349-404.

Chatterji, A., Glaeser, E., Kerr, W. (2014). "Clusters of entrepreneurship and innovation." In: Lerner, J., Stern, S. (Eds.), Innovation Policy and the Economy, Vol. 14, University of Chicago Press, 129-66.

Chinitz, B. (1961). "Contrasts in agglomeration: New York and Pittsburgh." American Economic Review, 51 (2), 279-89.

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

Combes, P., Duranton, G., Gobillon, L., Puga, D., Roux, S. (2012). "The productivity advantages of large cities: distinguishing agglomeration from firm selection." Econometrica, 80 (6), 2543-94.

Combes, P., Duranton, G., Overman, H. (2005). "Agglomeration and the adjustment of the spatial economy." Papers in Regional Science, 84 (3), 311-49.

Combes, P., Gobillon, L. (2015). "The empirics of agglomeration economies." In: Henderson, J.V., Duranton, G., Strange, W. (Eds.), Handbook of Regional and Urban Economics, Vol. 5. North Holland, Amsterdam, 247-348.

Delgado, M., Porter, M., Stern, S. (2010). "Clusters and entrepreneurship," Journal of Economic Geography, 10 (4), 495-518.

Dumais, G., Ellison, G., Glaeser, E. (2002). "Geographic concentration as a dynamic process." Review of Economics and Statistics, 84 (2), 193-204.

Duranton, G. (2007). "Urban evolutions: the fast, the slow, and the still." American Economic Review, 97, 197-221.

Duranton, G. (2008). "Viewpoint: from cities to productivity and growth in developing countries." Canadian Journal of Economics, 41 (3), 689-736.

Duranton, G. (2014). "Growing through cities in developing countries." World Bank Research Observer, 30 (1), 39-73.

Duranton, G., Overman, H. (2005). "Testing for localization using micro-geographic data." Review of Economic Studies, 72, 1077-106.

Duranton, G., Overman, H. (2008). "Exploring the detailed location patterns of U.K. manufacturing industries using micro-geographic data." Journal of Regional Science, 48, 213-243.

Duranton, G., Puga, D. (2001). "Nursery cities: urban diversity, process innovation, and the life cycle of products." American Economic Review, 91, 1454-77.

Duranton, G., Puga, D. (2004). "Micro-foundations of urban agglomeration economies." In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Urban and Regional Economics, Vol. 4, North-Holland, Amsterdam, 2063-117.

Duranton, G., Puga, D. (2014). "The growth of cities." In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth, Vol. 2, North-Holland, Amsterdam, 751-843.

Ellison, G., Glaeser, E., Kerr, W. (2010). "What causes industry agglomeration? Evidence from coagglomeration patterns." American Economic Review, 100 (3), 1195-1213.

Faggio, G., Silva, O, Strange, W. (2014). "Heterogeneous agglomeration." SERC Working Paper 0152.

Fallick, B., Fleischman, C., Rebitzer, J. (2006). "Job-hopping in Silicon Valley: some evidence concerning the microfoundations of a high-technology cluster." Review of Economics and Statistics, 88 (3), 472-81.

Feldman, M. (2000). "Location and innovation: the new economic geography of innovation, spillovers, and agglomeration." In: Clark, G., Feldman, M., Gertler, M. (Eds.) The Oxford Handbook of Economic Geography. Oxford University Press, Oxford, 373-394.

Feldman, M., Kogler, D. (2010). "Stylized facts in the geography of innovation." In: Hall, B., Rosenberg, N. (Eds.), Handbook of the Economics of Innovation, Vol. 1. Elsevier, Oxford, 381-410.

Ghani, E., Kerr, W., O'Connell, S. (2013). "Local industrial structures and female entrepreneurship in India." Journal of Economic Geography, 13 (6), 929-64.

Glaeser, E. (2008). Cities, Agglomeration and Spatial Equilibrium, Oxford University Press, Oxford.

Glaeser, E., Kerr, S., Kerr, W. (2015). "Entrepreneurship and urban growth: an empirical assessment with historical mines." Review of Economics and Statistics, 97 (2), 498-520.

Glaeser, E., Kerr, W. (2009). "Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain?" Journal of Economics and Management Strategy, 18 (3), 623-63.

Glaeser, E., Kerr, W., Ponzetto, G. (2010). "Clusters of entrepreneurship." Journal of Urban Economics, 67 (1), 150-68.

Hall, B., Jaffe, A., Trajtenberg, M. (2002). "The NBER patent citation data file: lessons, insights and methodological tools." In: Jaffe, A., Trajtenberg, M. (Eds.) Patents, Citations, and Innovations: A Window on the Knowledge Economy, MIT Press, Cambridge, MA, 403-60.

Helsley, R., Strange W. (1990). "Matching and agglomeration economies in a system of cities." Regional Science and Urban Economics, 20 (2), 189-212.

Helsley, R., Strange, W. (2002). "Innovation and input sharing." Journal of Urban Economics, 51, 25-45.

Helsley, R., Strange, W. (2014). "Coagglomeration, clusters, and the scale and composition of cities." Journal of Political Economy, 122 (5), 1064-93.

Henderson, J.V. (1974). "The size and types of cities." American Economic Review, 61, 640-56.

Hsieh, C.T., Klenow, P. (2014). "The lifecycle of plants in India and Mexico." Quarterly Journal of Economics, 129, 1035-84.

Jaffe, A., Trajtenberg, M., Henderson, R. (1993). "Geographic localization of knowledge spillovers as evidenced by patent citations." Quarterly Journal of Economics, 108, 577-98.

Kemeny, T., Feldman, M., Ethridge, F., Zoller, T. (2015). "The economic value of local social networks." Working paper.

Kerr, W. (2010). "Breakthrough inventions and migrating clusters of innovation." Journal of Urban Economics, 67 (1), 46-60.

Kerr, W., Brownell, A. (2011). "Location choice for new ventures: choices within cities." Harvard Business School Background Note 812-036. Harvard Business School, Boston, MA.

Kerr, W., Kominers, S. (2010). "Agglomerative forces and cluster shapes." NBER Working Paper 16639.

Kerr, W., Kominers, S. (2015). "Agglomerative forces and cluster shapes." Review of Economics and Statistics. Forthcoming.

Kerr, W., Lincoln, W. (2010). "The supply side of innovation: H-1B visa reforms and U.S. ethnic invention." Journal of Labor Economics, 28 (3), 473-508.

Kerr, W., Nanda, R., Brownell, A. (2011). "Location choice for new ventures: cities." Harvard Business School Background Note 811-106, Harvard Business School, Boston, MA.

Klepper, S. (2010). "The origin and growth of industry clusters: the making of Silicon Valley and Detroit." Journal of Urban Economics, 67, 15-32.

Kremer, M. (1993). "The O-Ring theory of economic development." Quarterly Journal of Economics, 108 (3), 551-75.

Lazear, E. (2005). "Entrepreneurship." Journal of Labor Economics, 23, 649-80.

Michelacci, C., Silva, O. (2007). "Why so many local entrepreneurs?" Review of Economics and Statistics, 89 (4), 615-33.

Moretti, E. (2012). The New Geography of Jobs. Houghton Mifflin Harcourt, New York, NY.

Mukim, M. (2015). "Coagglomeration of formal and informal industry : evidence from India." Journal of Economic Geography, 15 (2), 329-51.

Porter, M. (1990). The Competitive Advantage of Nations. The Free Press, New York, NY.

Redding, S., Turner, M. (2014). “Transportation costs and the spatial organization of economic activity.” In: Henderson, J.V., Duranton, G., Strange, W. (Eds.), Handbook of Regional and Urban Economics, Vol. 5. North Holland, Amsterdam, 1339-98.

Rosenthal, S., Strange, W. (2001). “The determinants of agglomeration.” Journal of Urban Economics, 50, 191-229.

Rosenthal, S., Strange, W. (2003). “Geography, industrial organization, and agglomeration.” Review of Economics and Statistics, 85 (2), 377-93.

Rosenthal, S., Strange, W. (2004). “Evidence on the nature and sources of agglomeration economies.” In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, Vol. 4. North-Holland, Amsterdam, 2119-71.

Rosenthal, S., Strange, W. (2012). “Female entrepreneurship, agglomeration, and a new spatial mismatch.” Review of Economics and Statistics, 94 (3), 764-88.

Ruiz, N., Wilson, J., Choudhury, S. (2012). “Geography of H-1B workers: demand for high-skilled foreign labor in U.S. metropolitan areas.” Brookings Institute Report.

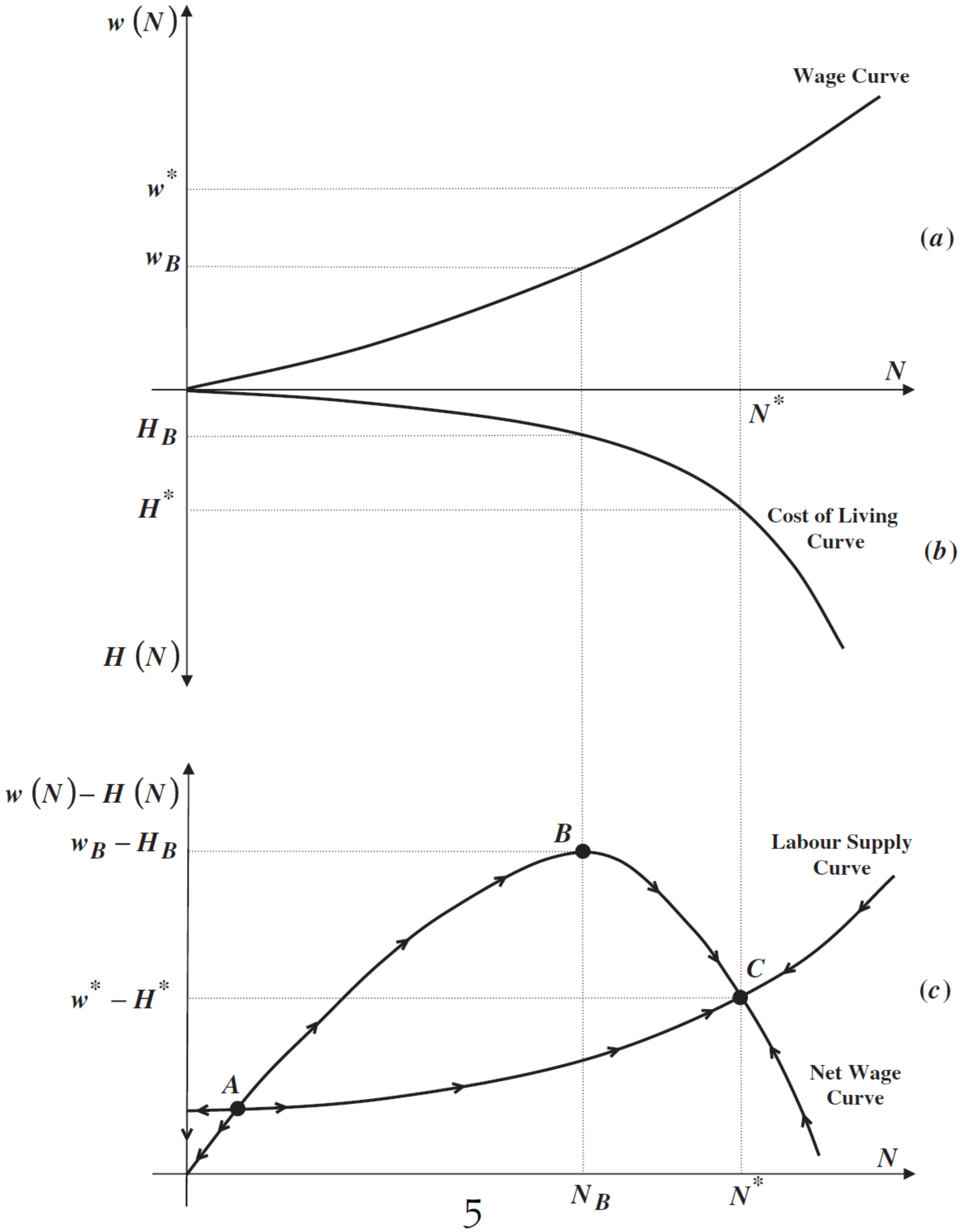
Saxenian, A. (1994). Regional Advantage: Culture and Competition in Silicon Valley and Route 128, Harvard University Press, Cambridge, MA.

Saxenian, A., Motoyama, Y., Quan, X. (2002). Local and Global Networks of Immigrant Professionals in Silicon Valley, Public Policy Institute of California, San Francisco, CA.

Serafinelli, M. (2014). “'Good' firms, worker flows and local productivity.” Working paper, University of Toronto, Ontario, Canada.

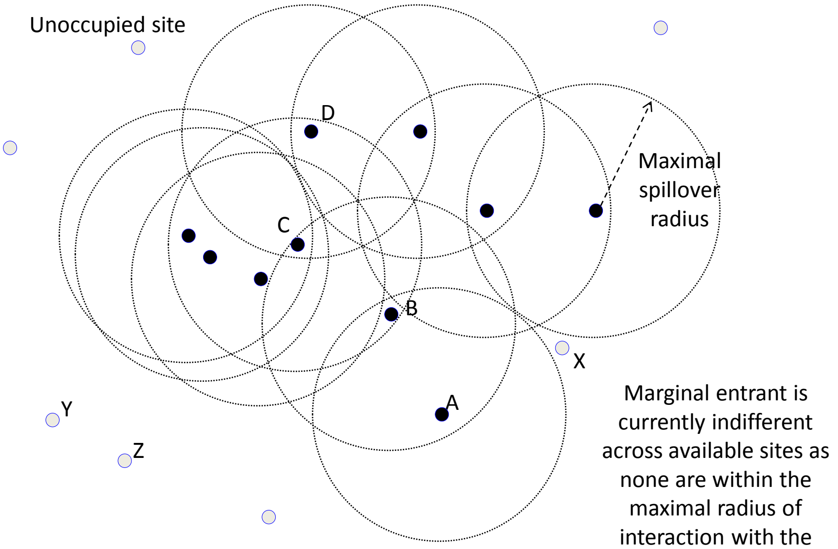
Tecu, I. (2012). “The location of industrial innovation: does manufacturing matter?” PhD. Thesis, Brown University, Providence, RI.

Figure 1: Agglomeration and City Formation



Notes: Figure represents the conceptual model of Duranton (2008) for the formation of cities. The horizontal axis measures city size as given by a population N. The upper part shows a wage/productivity curve w(N) that is increasing a location's economic activity due to agglomerative forces. The top panel also shows a cost curve H(N) that is similarly growing more acute in city size. Subtracting the cost curve from the wage curve, one determines a net benefit for w(N)-H(N), as shown in the lower half of the figure. This benefit is initially rising in city size as the early productivity benefits from larger city size dominate the cost increases. After point B, these net benefits decline. Through migration, labor supply increases to a city with rising net benefits. The stable equilibrium that forms will be at point C.

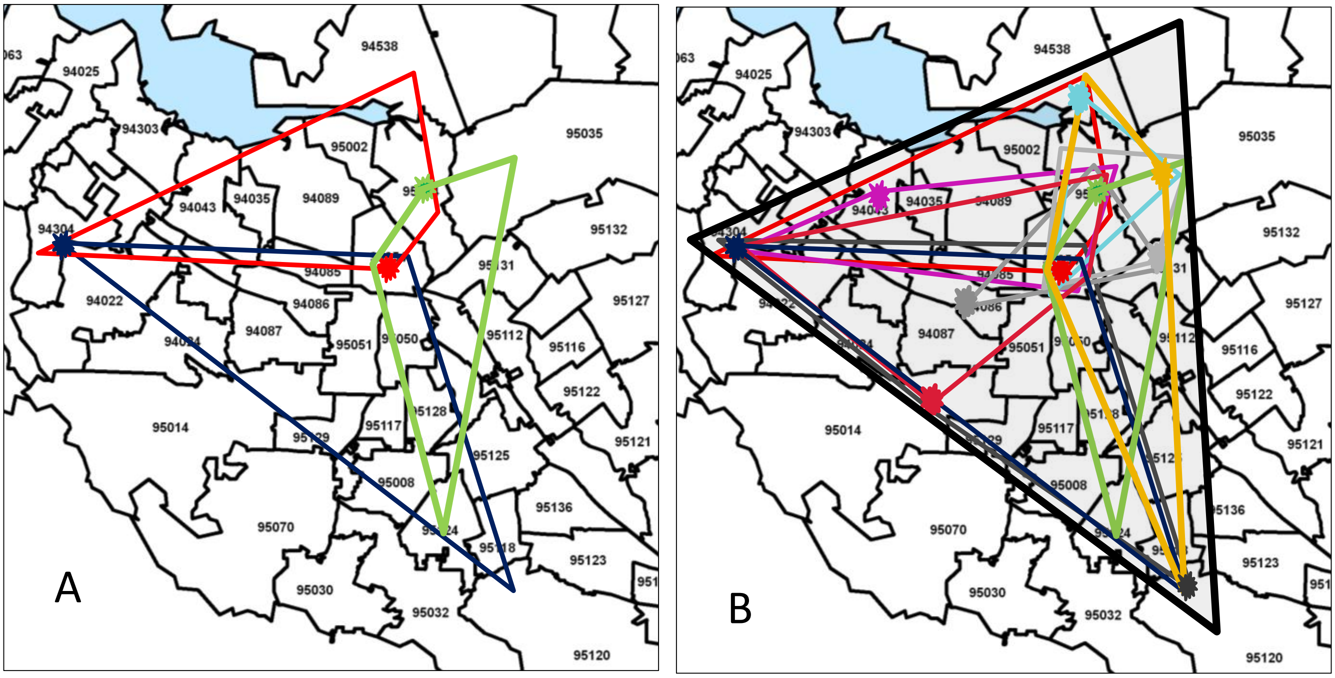
Fig. 2: Internal Structure of Clusters



Marginal entrant is currently indifferent across available sites as none are within the maximal radius of interaction with the existing cluster

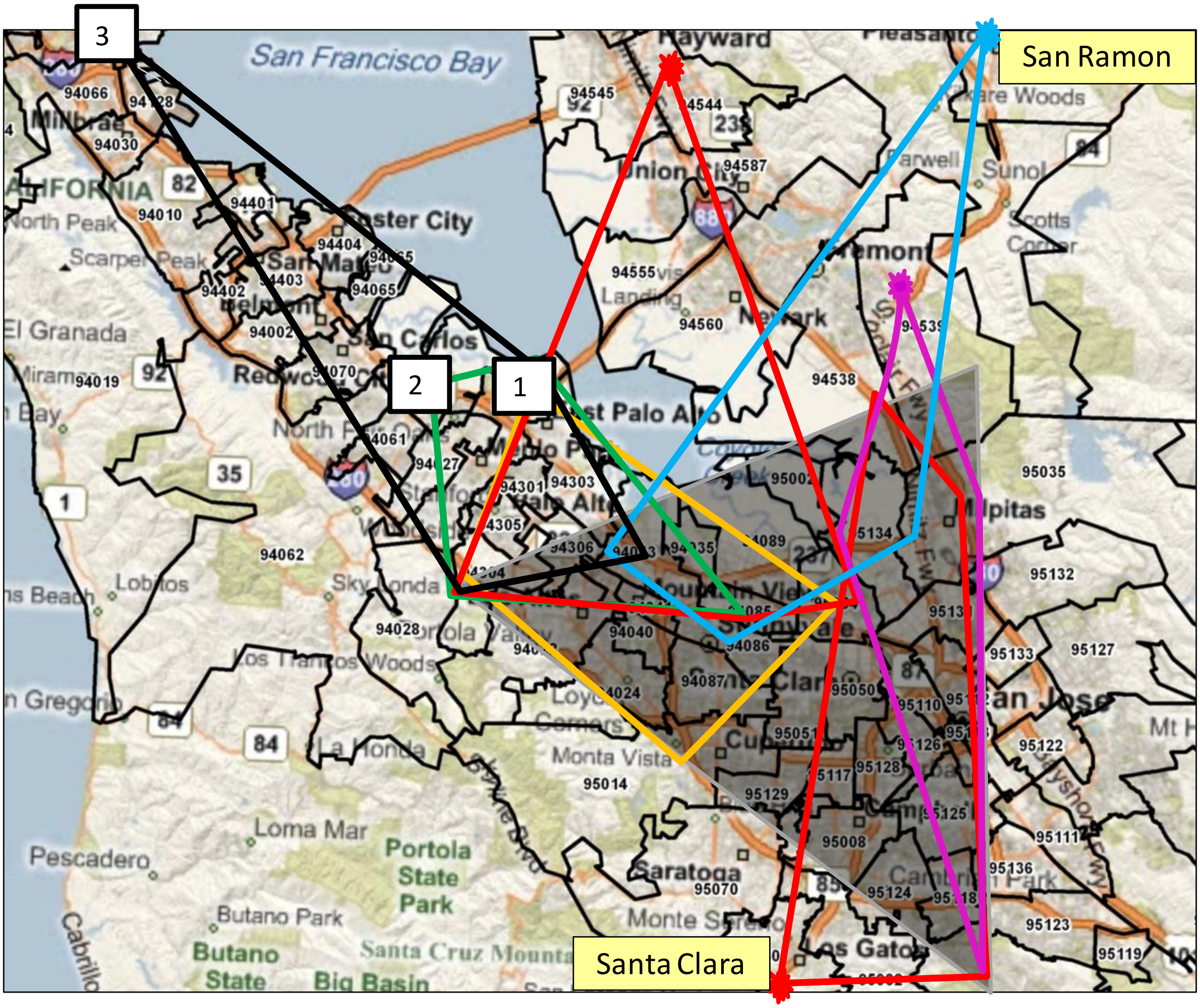
Notes: Image illustrates the Kerr and Kominers (2015) theoretical framework for the formation of an agglomeration cluster due to small, contained interaction effects among firms that overlap each other.

Figure 3: Tech Sourcing in Silicon Valley Core



Notes: Figure describes the construction of the technology core for Silicon Valley. The core includes the top 10 zip codes in the San Francisco Bay Area for patenting and 18 of the top 25. For each zip code, we present a technology sourcing zone that depicts the three zip codes that the firm in the focal zip code cite the most in their patent filings. Panel A shows the three largest patenting zip codes and their sourcing. Panel B shows the top 10 zip codes. While unrestricted in design, the primary technology sourcing zones are all contained in the core. These zones are small, overlapping regions that often exhibit directional transmission.

Figure 4: Tech Sourcing around Silicon Valley



Notes: Figure continues to characterize technology flows for the San Francisco area. The Silicon Valley core depicted in Figure 3 is represented on this larger map as the shaded triangle. The Silicon Valley core contains 18 of the top 25 zip codes for patenting in the San Francisco area. This figure includes the seven largest zip codes for patenting that are not contained in the core itself. Similar to Figure 3, the shape of each technology sourcing zone is determined by the three zip codes that firms in the focal zip code cite most in their work. For visual ease, San Ramon and Santa Clara are indicated on the edge of the map at the location of their primary transportation route. As observed for the core, these technology zones are characterized by small, overlapping regions that exhibit directional transmission.