IMT School for Advanced Studies, Lucca
Lucca, Italy

KU Leuven, Faculty of Economics and Business
Leuven, Belgium

Essays on Innovation Networks and Global Cities

Joint degree - "Ph.D. in Systems Science - Curriculum in Economics, Networks and Business Analytics" and "Ph.D. in Business Economics"

XXXIII Cycle

By

Samuel A. Edet

2022

The dissertation of Samuel A. Edet is approved.

PhD Program Coordinator: Prof. Dr. Rocco De Nicola, IMT School for Advanced Studies Lucca

Advisor: Prof. Dr. Massimo Riccaboni, IMT School for Advanced Studies Lucca, Italy.

Co-Advisor: Prof. Dr. Rene Belderbos, KU Leuven, Belgium.

The dissertation of Samuel A. Edet has been reviewed by:

Prof. Dr. Gaetan de Rassenfosse, EPFL, Lausanne, Switzerland.

Prof. Dr. Zhen Zhu, University of Kent, United Kingdom.

Prof. Dr. Mattia Nardotto, KU Leuven, Belgium.

Prof. Dr. Armando Rungi, IMT School for Advanced Studies Lucca, Italy.

IMT School for Advanced Studies Lucca
KU Leuven, Faculty of Economics and Business
2022

In memory of my father, Ebiefie Edet

Contents

List of Figures

List of Tables

Acknowledgements

There is an adage in my mother tongue that says "Ojú mérin ní bímo, igba ojú ní tò" which loosely interpreted means that "It takes a village to train a child." In this doctoral program, I have been privileged to be trained by a village of exceptional individuals who reflect excellence as humans and professionals.

I am deeply grateful for the scholarships from *IMT Lucca, Italy* and *KU Leuven, Belgium* that enabled me to undertake this doctoral program. I am grateful to my supervisors *Prof. Dr. Massimo Riccaboni* and *Prof. Dr. Rene Belderbos* for the guidance and advice they have provided in the last four years. *Massimo*, I am happy to have been able to work with you – your insights and feedback has been pivotal to the quality of the work, thank you for the flexibility, the career opportunities and advice you have given me. *Rene*, each time I have a work done, I eagerly look forward to your feedback because they often spur opportunities to further reflect on the work and can serve as gateway for possible research questions I never envisaged. I am very delighted to have worked under the supervision of both of you.

I am grateful for the opportunity to work collaboratively with *Prof. Dr. Pietro Panzarasa* and *Dr. Luca Verginer*. The conversation with *Luca* at the beginning of the thesis was very helpful in proactively understanding the possible questions that can be explored, and the conversation with *Pietro* and the opportunity he provided for me to visit *Queen Mary University of London* (which I could not due to COVID-19) helped refine the

writing of the thesis. Furthermore, I would like to acknowledge the efforts of all the reviewers – the internal reviewers, *Prof. Dr. Armando Rungi* and *Prof. Dr. Mattia Nardotto* who have been there all the way providing very useful feedback since the first seminar. The external reviewers *Prof. Dr. De Rasenfosse* and *Prof. Dr. Zhen Zhu* whose feedback helped enhance reflections on limitations and policy discussions in the thesis.

I would like to appreciate my colleagues and friends *Laura, Federico, Falco, Tatiana, Florence, Geon* who I have had to work with on either class projects or the thesis itself. I do not take for granted how you have all made my assimilation into the European experience an easy one either through activities such as soccer, music, apertivos, and happy hours. Also, I sincerely appreciate the exceptional work of the administrative staffs in both *IMT, Lucca* and *KU Leuven* – thank you very much *Barbara Iacobino, Maria Mateos, Sara Olson, Isabelle Theys* for guiding me through the bureaucratic process and ensuring I had a pleasant stay in both countries.

Furthermore, in the last one year writing this thesis, I have had to simultaneously work at the World Bank Group, and I can not take for granted the opportunity offered to me by *Liane Lohde* to work on different projects, some of which enabled the analytical contribution on the application of machine learning to competitiveness reflected in this thesis. I am grateful for the discussions with other bank colleagues *Samuel Rosenow*, *Miles McKenna* and *Zeinab Partow* and how they have made combining both the writing of the thesis and delivering on projects a less complicated endeavor to undertake.

Finally, my deepest gratitude goes to my wife *Funmi* who has been a great support in many ways, from the brainstorming of ideas to proofreading and providing the emotional support

to thrive at a time like this - you have always been there for me, and I love you immensely for all you do. I dedicate this doctoral work to my late father *Ebiefie* who has always inspired me to be a better person. I am also grateful for the cheerfulness of my mum *Roseline* and three wonderful sisters *Glory, Rebecca,* and *Precious* whose laughter has been succor when the road gets bumpy.

In conclusion, as I look back at the memories shared with these exceptional people in my village who have been a part of my journey in the last four years, I am hopeful that this mystic chord of memories stretching the heart of everyone I have met, will at different points in our lives resonate joy and happiness when touched, assuredly they will be, by the better angels of our nature.

Vita

December 8, 1992, Born, Lagos, Nigeria

Education

2018-2022 Ph.D. candidate in Business Economics
Faculty of Economics and Business, KU Leuven,
Leuven, Belgium

2017-2022 Ph.D. candidate in Systems Science
IMT School for Advanced Studies, Lucca
Lucca, Italy

2016-2017 M.Sc. in Mathematical Sciences
Class of degree: Distinction
University of Cape Town
Cape Town, South Africa

2010-2014 B.Sc. in Mathematics
Class of degree: First Class Honours
University of Ibadan
Ibadan, Nigeria

Experience

2021-	Economist (Consultant) World Bank Group, Washington D.C., United States
2020	Data Scientist (Data Study Group program) Alan Turing Institute United Kingdom
2017	Data Analyst Siemens Johannesburg, South Africa
2014- 2015	Research Analyst Nigeria Institute for Social and Economic Research Ibadan, Nigeria

Scholarships, Grants, Awards

1. World Bank Group Fellowship
2. Frontier Research Fellowship, IMT School for Advanced Studies Lucca, Lucca, Italy
3. Ph.D. full scholarship in Economics, IMT School for Advanced Studies Lucca, Lucca, Italy
4. Ph.D. scholarship in Business Economics, Faculty of Economics and Business, KU Leuven, Leuven, Belgium
5. ERASMUS+ mobility consortium grant (Academic Year: 2018/2019 and 2019/2020)
6. Data Study Group (DSG) sponsorship, Alan Turing Institute, United Kingdom
7. Alfred P. Sloan Young scholar conference funding, National Bureau of Economic Research (NBER).
8. University of Oxford Economics Networks summer school scholarship
9. Winning project at the University of Oxford Economics Networks summer school.
10. European School for Management Technology (AIMS-ESMT) fellowship
11. Nominee, Thomson Reuters Excellence Award for student research in data science
12. Mandela Rhodes Foundation (MRF) scholarship
13. Google Online Marketing Challenge (GOMC) funding
14. African Institute for Mathematical Sciences (AIMS) fully funded scholarship
15. University scholar award, University of Ibadan, Nigeria
16. Merit award, Nigeria Model United Nations Society

Publications

1. R. Belderbos, F. Benoit, S. Edet, G.H. Lee, and M. Riccaboni, "Global cities' cross-border innovation networks," - chapter contribution in the book *Cross-Border Innovation in a Changing World. Players, Places, and Policies,* edited by D. Castellani, A. Perri, V. Scalera, A. Zanfei. *Oxford University Press*, 2022.
2. S. Edet, P. Panzarasa, M. Riccaboni, "Global Cities in International Networks of Innovators," in *Advances in Complex Systems*, doi: 10.1142/S0219525921400026, 2021.
3. Data Study Group team. (2021, October 11). "Data Study Group Final Report: CatsAi". Zenodo. https://doi.org/10.5281/zenodo.5562660.

Presentations

1. International Conference on Rethinking Clusters, Universita degli studi Firenze, Italy, 2021
2. Second doctoral seminar, KU Leuven, Belgium, 2021
3. World Bank Group Africa BBL Series, Washington D.C., United States, 2021
4. First doctoral seminar, KU Leuven, Belgium, 2020
5. International Conference on Rethinking Clusters, Universitat Politecnica de Valencia, Spain, 2020
6. Network Science Society Conference, Rome, Italy, 2020
7. Doctoral workshop at Sant' Anna, Pisa, Italy, 2020
8. Data Science summer school, Ecole Polytechnique, Palaiseau, France, 2019
9. Economics Networks summer school, University of Oxford, United Kingdom, 2018

Abstract

In a competitive economy, technological innovation is a core element of economic growth and development, and its accumulation in a rapidly changing technological environment is key to adaptation. This thesis investigates how cities, particularly global cities can develop new technological capabilities to enter new technological fields and become competitive in new technological areas. First, a new geo-referenced patent database is developed to overcome some limitations of the existing international patent repository to make an international comparison of cities feasible. The new database provides broad coverage of cities in developed and emerging economies and allows us to tackle the following research questions: (i) how the co-patenting network of domestic and international linkages of cities has changed over time? (ii) which cities are more technologically complex? Can we predict the future evolution of cities' technological complexity? (iii) What is the effect of inter-city linkages and technological relatedness on the likelihood of cities to enter new technological areas? (iv) which cities are the most central in coordinating complex international teams of inventors on a global scale?

To geo-locate patents, an address retrieval algorithm has been applied to resolve the problem of missing addresses, exploiting the availability of address in patent family and similarity of inventors based on attributes such as working for the same applicant. A harmonized definition for functional urban areas is used to assign patents to cities on a global scale. The database provides a significant improvement from the raw PATSTAT dataset with an estimated confidence level of 84-88%.

We analyze both unweighted (extensive) and patent-weighted (intensive) linkages between cities to address the first research question. The preliminary findings show an increase of international ties both intensively and extensively and the reliance of cities in developing economies on international ties to global cities. To address the second research question, we apply the generalized economic complexity (GENEPY) algorithm to measure the economic complexity of cities based on patent production in these cities. We use machine learning models (Random Forest, XGBoost, SVM, Neural network) to forecast the future technological complexities of cities. We show that the machine learning models (especially Random Forests) have higher predictive power than the benchmark model (time-independent conditional probabilities) as they account for higher-order and non-linear interdependencies between technologies.

To address the third question, we applied a stratified semi-parametric Cox proportional hazard model to examine the likelihood of cities entering new technologies. We show that inter-city linkages and technological relatedness significantly increase the likelihood of entry into new technological areas. Inter-city linkages are more critical for non-global cities than for global cities to enter new technological areas, whereas linkages to inventors located in cities with a large pool of inventors positively moderate the effect of inter-city linkages on entry. For the last research question, we use hypergraphs structure and propose a measure based on 3-hyperedges (three cities in multiple countries) in the collaboration networks constructed from scientific publications and patents to identify the most competitive global cities in the international network of inventors. To this end, we construct a null model using the hypergeometric ensembles of random graphs and find that five US cities play a leading role in transnational networks of

researchers. San Francisco stands out as the most global city, but Shanghai is rapidly emerging as a global player.

Introduction

[1] Global cities are unique geographical entities, as they are often characterized with attractive features that make them hubs of multinational companies (MNCs) (Friedmann, 1986; Sassen, 2001; Taylor, 2001). These features include having a global reach of advanced producer service firms, and manufacturing firms, that play significant roles in global value chains. These firms serve as conduits for moving goods and services to domestic and international markets (Goerzen, Asmussen, and Nielsen, 2013; Belderbos, Du, and Goerzen, 2017). Other characteristics of these cities that make them an important location for R&D activities are the research reputation of their academic institutions, the concentration of highly-skilled workers in these cities, and their quality institutions that protect the intellectual property rights of firms (Glaeser and Gottlieb, 2009; Belderbos, Du, and Goerzen, 2017).

The competitiveness of (global) cities is driven more by their international connectedness to the globe, and in particular to other global cities, than by their locally restricted characteristics (Taylor, 2001). This notion of connectivity has been studied by economic geographers from both functional and demographic points of view. From a functional perspective, cities connectivities are spokes created by firms or collaborations between individuals with shared interests. The connectivity formed by firms is called "organizational pipelines" and they are consistent with the goals and objectives of the firms. On the other hand, connectivities created by individuals can be less focused in comparison with firm connectivities (Lorenzen and Mudambi, 2013). "Organizational pipelines" crossing country borders are often facilitated by MNCs, while individual-to-individual connectedness can be facilitated by relationships between ethnic diasporas within a country, or cross-border relationships formed between diasporas and people residing in their home country (Saxenian and Hsu, 2001; Belderbos, Du,

[1] This Introductory section is partly based on R. Belderbos, F. Benoit, S. Edet, G.H. Lee, and M. Riccaboni (2021). "Global cities' cross-border innovation networks" - chapter contribution in the book Cross-Border Innovation in a Changing World. Players, Places, and Policies, edited by D. Castellani, A. Perri, V. Scalera, A. Zanfei. Oxford University Press.

and Goerzen, 2017).

The importance of global cities in innovation networks has received minimal emphasis in the literature on global cities, as most works have focused on the interlockedness of global cities which are driven by the activities of service firms, and also the network structure created by infrastructures (such as rail, road and air transportation, and internet backbones). Analyzing the role of global cities in innovation and knowledge networks is increasingly important as the prosperity of these cities in a competitive economy is hinged on knowledge accumulation and connectivities to innovative locations that can be leveraged. It has been said that a city should leverage both external information flows and exchanges, as well as domestic knowledge (Belderbos, Du, and Goerzen, 2017; Asheim and Coenen, 2006; Bathelt, Malmberg, and Maskell, 2004). The ability to combine both intense circulations of domestic information and solid cross-border links to knowledge located in other cities and regions are characteristics of dynamic cities (Castells, 2002).

Linkages across city borders have been seen to be associated with cities' technological success (Maggioni, Nosvelli, and Uberti, 2007; Belderbos, Du, and Goerzen, 2017; Miguélez and Moreno, 2012) as they serve as conduits for diverse information, and possibly, the knowledge that may not be readily available locally (Boschma and Frenken, 2011; Malmberg and Maskell, 2002). These cross-border linkages are more important when they are connected to clusters involved in cutting-edge technological research (Bathelt, Malmberg, and Maskell, 2004). Typically, such linkages are brought about when entities in cities are involved in scientific co-authorship, R&D production through co-inventorship, and partnership frameworks that allow for the exchange of best practices between cities (Matthiessen, Schwarz, and Find, 2010; Belderbos, Du, and Goerzen, 2017). While, there has been recent analysis done on providing insights on cities global positions and growth dynamics of linkages by applying different centrality measures, clustering techniques, and kinetic models to publication and patent data, these analyses often fail to account for higher-order effects that have been observed by Neuhäuser, Mellor, and

Lambiotte, 2020 to be important given the increasing formation of teams located in more than two cities (Wagner, Park, and Leydesdorff, 2015).

The Thesis contribution to the literature of global cities is twofold. First, a new dataset is prepared in Chapter 1 to overcome some limitations of the existing international patent repository to make an international comparison of global cities feasible. The broad coverage of the dataset allows us to undertake the study of innovation activities in cities not just in developed economies, but also in emerging economies. Second, based on the new dataset we introduce in Chapter 1, we tackle a set of relevant research questions about the location of patent production in global cities (see Table 1).

Chapter	Topic	Methodology	Research Question
Introduction			
Chapter 1	Georeferencing Patent Data for Urban Studies	Database development and analysis	How to develop granular patent data and delineate cities to allow for international comparisons? How have cities' domestic and international linkages evolved?
Chapter 2	Economic complexity and Forecasting competitiveness of Global cities using Machine learning approach	Machine learning	How to measure economic capabilities of cities and forecasting competitiveness of cities in different technologies?
Chapter 3	Exploratory Innovation in Cities: Inter-city ties and Technological relatedness	Econometrics	What is the effect of collaborative ties and technological relatedness on cities' entry into new technologies?
Chapter 4	Global Cities in International Networks of Innovators	Complex System analysis	What cities are likely to emerge in higher-order international collaboration?
Conclusion			

Table 1: Dissertation outline

In Chapter 1, first, we address the issue of availability of fine-grained comprehensive patent database by building on the work of De Rassenfosse et al., 2013; De Rassenfosse, Kozak, and Seliger, 2019,Morrison, Riccaboni, and Pammolli, 2017a, and PatentView, 2018. Also, we address the funda-

mental issue of allocating patent and publication data to cities in such a way that cities boundaries are harmonized for international comparisons. The motivation for constructing a patent data with geographic coordinate information that allows patents to be allocated to cities, stems from the increasing interest by policymakers in understanding location decisions of firms (Castellani and Lavoratori, 2019), hence the construction of this database allows the identification of cities that play the role of innovation hubs and serve as attraction points for global talents. Moreover, since innovative activities are typically accounted for by pockets of cities within a country (De Rassenfosse, Kozak, and Seliger, 2019), this database further allows for a more focused analysis of the innovative activities taking place in these cities - such analysis could examine the nature of collaborations between universities and firms in patent production, the role of mobility of star inventors in innovation outcomes of these cities (Valero and Van Reenen, 2019; Verginer and Riccaboni, 2021). Finally, the research question on delineating cities such that they are internationally comparable is very important, since a cross-country analysis should account for the fact that cities and regions are defined differently across countries - sometimes this definition of cities reflect administrative or legal boundaries that do not represent the economic and functional extent of these cities, as such can bias analysis of cities across countries (Dijkstra, Poelman, and Veneri, 2019).

To address the question of "how to delineate cities such that they are internationally comparable?", we rely on the methodology introduced by the OECD to demarcate cities based on their economic reach, in a way that is uniform across countries. This approach corrects for the differences in legal and administrative boundaries adopted in different countries (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019). The resulting demarcation referred to as a "functional urban area" can be different from the actual administrative boundaries. The harmonized definition of "functional urban areas" uses both population density data and travel flow information between cities to identify "urban cores" and peripheral local units that are economically integrated with the urban core (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019). This methodology allows identifi-

cation of the right knowledge clusters (Alcácer and Zhao, 2016) and is a major advantage of research on cities in comparison with prior studies focusing on institutional and legal boundaries of regions. Based on this patent database properly allocated to cities, we draw on a list of 168 global cities as characterized in Beaverstock, Smith, and Taylor, 1999, and show changes in the formation of domestic and international linkages to global and non-global cities.

In Chapter 2, we demonstrate the use of the patent database in measuring the economic complexity of cities and forecasting the competitiveness of cities based on machine learning algorithms. Questions around the economic competitiveness of cities have been a long-standing one in the economic literature, and most analyses are geared towards unraveling explanatory variables for how countries, cities, and firms acquire and increase capabilities in different products, technologies, and scientific domains, such that their production output exceeds their fair share (Sassen, 2001; Rigby, 2015; Balland et al., 2017). However, tracking or measuring these capabilities and strategic choices made by cities can be problematic as these are generally intangible (Straccamore and Zaccaria, 2021). The usefulness of the economic complexity literature is geared at addressing the issue of uncovering these capabilities by analyzing the productive, technological, or scientific output of cities from a network perspective, with the aim to measure or forecast rather than explaining why a city like Taiwan has capability in Semiconductor (Tacchella et al., 2012; Straccamore and Zaccaria, 2021). The richness of economic complexity application is seen prominently in the economic geography and innovation literature. In this literature, economic complexity is often discussed with reference to infrastructure, product export, technological and scientific capabilities that allows a country or region to identify what products or technological areas it can readily upgrade into given its current capabilities (Hidalgo, 2021; Mewes and Broekel, 2020). Although precisely measuring complexity is challenging, there have been scientific efforts towards inferring the complexity of countries using trade, publication, and patent data. The two major methodologies for measuring economic complexity are the "Method of reflection" approach (Hidalgo et al., 2007)

and "Fitness and complexity" methodology (Tacchella et al., 2012) which consist of "non-linear coupled maps whose fixed points define new metrics for the fitness of the countries and complexity of products". We applied a recent methodology called "Generalized economic complexity index" (i.e., GENEPY) proposed by Sciarra et al., 2020 that reconciles both methodologies by applying linear-algebra tools within a bipartite network framework. Beyond measuring the economic complexity of cities and technology classes (i.e., intellectual patent classes (IPC)), we measure the fitness of cities in different technology fields. This measure captures the level of within technology field diversification and balance of more sophisticated technology competitiveness relative to other cities. Furthermore, since the GENEPY index captures the current capabilities of cities, we propose using machine learning models to forecast the future competitiveness of cities in different technologies i.e., their capability to produce patents in a given technology beyond its fair share. We show that these machine learning models provide a better forecast of future competitiveness of cities in different technologies when compared to the network approach that combines time-independent conditional probabilities with the current competitive structure of cities (Zaccaria et al., 2018). Furthermore, we try to predict extreme situations where cities leapfrog into competitive landscape in a technology (i.e., RTA exceeding 1 in 2014) from having extremely low technological advantage (i.e., RTA < 0.25) in the technology throughout the previous years (between 2000-2009). We refer to this task as the "activation task". By using the forecasted competitiveness in 2014, we reconstruct the future economic complexity index for cities in 2014.

In Chapter 3, we draw on extant literature on entry into new technology domains by examining the impact of collaboration structure and technological relatedness on the likelihood that a city enters a new technology domain. The entry process analyzed in Chapter 3 is distinctly different from the activation task discussed in Chapter 2 since you can enter a technology domain without necessarily becoming competitive. An entry process is observed if a city produced at least a patent in a technology domain, while activation is moving from an RTA less than 0.25 between

2000-2009 to an RTA exceeding 1 in 2014. The motivation for analyzing the drivers of cities' capabilities to enter new technology is inspired by existing efforts focused on unfolding the drivers of economic agents being able to acquire economically relevant technology and apply it effectively in ways that foster economic growth which over time and in different geographic areas has been lumpy and discontinuous (Rigby, 2015). The observed lumpiness of economic growth in different geographic areas has not only been attributed to reliance on localized natural resources and inability to attract capital investments but also as a result of the stickiness of tacit knowledge across geographies (Maskell and Malmberg, 1999; Rigby, 2015). While economic agents can choose to look inwardly to accumulate technology and leverage accumulated technological capabilities in a combinatorial way to develop new ideas, they may be faced with limited resources, cost and time to accumulate experience, and perhaps for some economic agents, they may have a small subset of the existing technology base that makes recombination difficult. In light of the relevance of new technology to the economic growth of a city, we examine drivers that are both internal i.e., looking into the accumulated technology within the city and their proximity to technologies not yet acquired by the city, and external i.e., looking into the collaborative ties of the city to economic agents outside the city. We posit that the likelihood that a city explores or enters a new technology domain is facilitated by the inter-city collaborative ties to external inventors who have previously invented in the new domain, and this effect is positively moderated by the size (number of inventors) of the partner city. Also, we hypothesize that the relatedness of the technology base of the city with the new domain positively affects the likelihood of entry, and this effect should be positively moderated by the focal city size (number of inventors). Based on the analysis of exploratory activities of 1220 cities (located in six continents) in 646 technological 4 digit-IPCs between the period 2005-2014, we find significant support for all hypotheses except for the moderating role of focal city size on technology relatedness. Finally, we provide evidence that entering a new domain collaboratively increases the technological performance of the city in the new domain.

Finally, in Chapter 4, motivated by Wagner, Park, and Leydesdorff, 2015's

observation of remarkable growth in international collaborative research and an increase in the formation of teams involving multiple cities, we examine positions of global cities in transnational networks of researchers (i.e., scientists and inventors) by assessing the likelihood of scientists and inventors from different cities to take part in international teams. Traditional network methodologies are not suited to analyze the rise of large international teams of inventors in scientific and technological productions. Hence, in Chapter 4 we introduce a new methodology based on recent developments in the analysis of hypergraphs. Which cities are playing a pivotal role as hubs in complex teams of inventors? To answer this question, we propose a measure based on 3-hyperedges in the collaboration networks constructed from scientific publications and patents. This measure is used to identify the most competitive global cities in the international network of researchers. To this end, we construct a null model using the hypergeometric ensembles (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018) of random graphs and find that five US cities play a leading role in transnational networks of researchers. Among them, San Francisco stands out as the most globalized city.

Chapter 1

Georeferencing Patent Data for Urban Studies

[1]Technological progress and innovation are important factors for the economic growth of urban areas and there are different ways of measuring technological innovation e.g. financial market valuation of R&D investments, scientific publications, or patents. [2]The production of patents indicates technological advancement made by individuals (referred to as inventors on the patent) working in public and private research organizations (referred to as applicants on the patent). The analysis of patent data is crucial for evaluating knowledge spillover as reflected in citation patterns (Jaffe, Trajtenberg, and Henderson, 1993; Maurseth and Verspagen, 2002), mobility of inventors (Miguelez, 2019), entry and performance in new technology domain (Leten, Belderbos, and Looy, 2016; Rigby, 2015) and knowledge production and diffusion. Patent analysis is useful for policymakers as they provide a data-driven, detailed understanding of innovation processes within and across regional and organizational boundaries.

Over the years, different patent databases have been used as a proxy for

[1]This is a joint work with R. Belderbos, F. Benoit, G.H. Lee, and M. Riccaboni

[2]There are individual inventors that are applicants as well, since not all inventors work for an organization.

studying innovation, this is because patent documents represent jurisdictional right on the ownership of intellectual property, and applicants who are willing to protect the ownership of their innovation are predisposed to going through the process of obtaining a patent, hence a patent database represents a portfolio of patent activities taking place in a jurisdiction, which include patents application from inventors home and abroad. There are different patent databases from which patent indicators can be constructed. The "United States Patent and Trademark Office" (USPTO) has been widely exploited for studies of inventive activity. However, patent indicators computed using USPTO data are subject to a strong geographic bias in favor of North-American inventors. Similar bias applies to patent indicators constructed from applications made at the European Patent Office (EPO) as several studies have shown that statistics on patent filings are biased in favor of European countries and there exists a significant difference in the transfer of national priority patent applications (i.e., initial patent application) to the EPO across European countries (De Rassenfosse et al., 2013).

To address the issue of geographic bias, the "Organisation for Economic Cooperation and Development" (OECD) proposed considering patent indicators based on triadic patent families. Triadic patent families are considered to have an international market perspective as they are filed in the three major patent offices, i.e., USPTO, EPO, and Japanese Patent Office (JPO). However, they are biased in favor of highly valued inventions that are often owned by large multinational corporations (MNCs). A more comprehensive database that provides adequate patent coverage for developed and developing economies and also patents filed by different types of applicants is the Patent Statistical (PATSTAT) database. PATSTAT is the largest database and contains patent activity information from over 90 patent offices in the world, and can be used in constructing patent indicators that are less susceptible to geographic and high-value patents bias. Nevertheless, the PATSTAT database being a compilation of other databases from different patent offices inherits data quality problems that arise in these patent offices which can limit its usage for analytical work. One drawback of using PATSTAT is the issue of ambiguous inventor

and applicant name due to a wide range of alternate spellings or misspellings on patent documents, either during multiple patent applications by inventors or during applications across multiple offices. Different disambiguation approaches have been introduced in the literature to address this problem (Morrison, Riccaboni, and Pammolli, 2017b; Kim, Khabsa, and Giles, 2016; Ventura, Nugent, and Fuchs, 2015; Pezzoni, Lissoni, and Tarasconi, 2014; Balsmeier et al., 2015). Another well-known problem is the lack of fine-grained location information of inventors and applicants on patents. Some works that have addressed the issue of providing geographic coordinate information include Morrison, Riccaboni, and Pammolli, 2017b, De Rassenfosse, Kozak, and Seliger, 2019, and PatentView, 2018.

A patent database with address information at a more detailed level can be useful in many ways: first, it can provide a high level of granularity which can foster the study of the growth or decay of knowledge spillovers at local levels such as cities and regions (Helmers, 2019). Also, a georeference database can provide insight into the distribution of complex patent activities (high impact multidisciplinary patents involving large teams and more R&D spending) and the emergence of cities in complex technological areas since innovation is increasingly concentrated in few large cities (Balland et al., 2020); a georeference information of inventors on a patent can help track productivity growth to identify star and emerging inventors (Hoisl, 2007), understand networking strategies of inventors and impact on productivity (Breschi, Lissoni, and Malerba, 2003), and mobility of inventors between private and public sectors (Crespi and Nesta, 2006). Finally, the worldwide coverage of this database makes it useful for studying innovation outcomes in developed and developing economies, as most research outcomes in previous years have largely focused on patent activities in developed economies. Other important use of this database includes the identification of innovation hubs (Castellani and Lavoratori, 2019) and improving disambiguation algorithms (De Rassenfosse, Kozak, and Seliger, 2019).

Beyond the precision required in defining the location of inventors and

applicants, there is a need to properly define the boundaries of the spatial units (i.e., cities, regions, and countries.) where the coordinates are located. The task of delineating spatial units can be challenging over time due to the dynamic tendencies of urban structures (Orellana and Fuentes, 2019). Most research work (e.g., De Rassenfosse, Kozak, and Seliger, 2019) in developing geographic information for inventors and applicants allocate geocoded data to regions and cities without taking into consideration the different administrative boundary definitions of cities and regions that are being considered by countries. Typically, boundary definitions across countries are different since they are driven by differences in administrative units. Hence, analysis done comparing the outcome of cities located in different countries will be biased (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019).

Therefore, to make sound international comparisons of cities requires delineating cities in economic terms. The aspects to consider in the delineation process is related to the concept that policymakers adopt in understanding the evolution of cities and their economic performance. The first consideration provides a uniform threshold for the density of cities in a way that is consistent with both the city's population size and land area. For cities located in more dispersed countries, the threshold for density is larger. On the other hand, the second consideration captures the economic extent of the city by allowing for the inclusion of surrounding cities that are of lower density but economically integrated with the city. These lower-density areas are typically referred to as hinterlands, peripheral cities, or commuting zones. The core city together with the integrated surrounding cities is called a "functional urban area" (FUA) (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019). By definition, an FUA "consists of a densely inhabited city and less densely populated commuting zones whose labor market is highly integrated with the city" (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019). The methodology for defining FUA was jointly developed by the "European Union" (EU) and OECD to provide a consistent demarcation of cities across OECD countries (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019). The identification of commuting zones is based on linkages between urban areas

measured based on commuting levels i.e., the number of people traveling daily between residential areas and their place of work.

In this chapter, we will develop a database of geographic coordinate information regionalized to functional urban areas for inventors and applicants on 12.3 million patent documents filed between 2000-2015. We will rely primarily on patents available in the patent statistical (PATSTAT) database and also on other sources of patent database (e.g., Morrison, Riccaboni, and Pammolli, 2017b, PatentView, 2018, De Rassenfosse, Kozak, and Seliger, 2019) which we will subsequently refer to as MRP, PV and RF databases, respectively. The RF database is introduced at a later stage of the data generating process to consolidate our final database since the RF database does not have an identifier for inventors and applicants which are important in the address and coordinate matching stage. Furthermore, the geocoded data will be assigned to the available functional urban areas in OECD, 2012 which are urban areas in OECD countries. For selected cities in non-OECD countries which we identify as important global cities based on the work done by Beaverstock, Smith, and Taylor, 1999, we will delineate these cities using the same spatial delineation process of OECD, 2012 by identifying commuting zones of cities based on (i) patent density of cities close to the focal city (ii) thresholds of the weight of linkages between cities, where we replace the definition of linkages from commuting levels between urban areas to the average traveling time between urban areas.

The rest of chapter 1 is organized as follows: Section 1.1 describes each step in the data generating process and the validation of the algorithm. Section 1.2 provides statistics on the quality of the final dataset. Section 1.3 provides background literature on domestic and international linkages and the analysis of trends in domestic linkages, international linkages to global and non-global cities. Section 1.4 provides concluding remarks on this chapter.

1.1 Data generating process

This section provides information on the data generating process which involves merging, parsing, address and coordinate matching, geocoding, algorithm validation, regionalization tasks undertaken to assign coordinate to inventors and applicants on a patent document.

Figure 1 provides an overview of the data generating process. First, we start with the collection of the primary data i.e., the Patent Statistical (PATSTAT) database version 2018, and the secondary or external patent databases which includes: (i) OECD regional patent (REGPAT) database version 2019 which contains patents that have been linked to regions using the addresses of applicants and inventors, (ii) MRP database that has patents with coordinate information based on disambiguation effort done on inventor and applicant names, (iii) PV database 2018 version, and (iv) the Institute for Intellectual Property (IIP) database that contains patent based on JPO standardized data (Akira Goto, 2007).

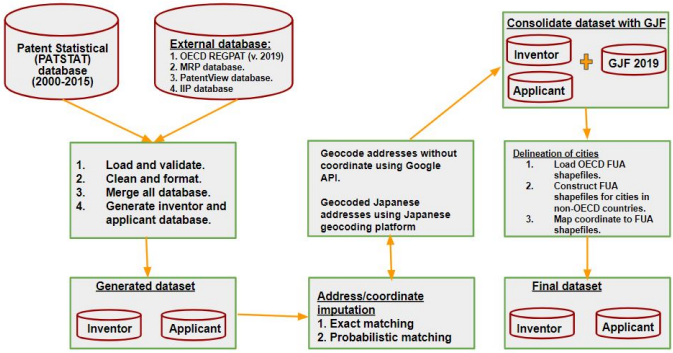


Figure 1: Flow chart of data generating process: Squares represent pre-processing and algorithms, while canisters represent databases. The final outputs of this process are regionalized locations of inventors and applicants of patents.

These data are validated, cleaned, and formatted, then all databases are merged based on primary identifiers (these are key columns in the database) identified across databases. For example, the primary identifier

in PATSTAT and OECD REGPAT is the patent application and person identifiers - implying both datasets can be merged based on these columns/identifiers. The merged database (i.e., PATSTAT, OECD REGPAT, IIP, MRP, and PV) does have rows corresponding to applicants and inventors for a given database. Therefore we split the merged database into two, one corresponding to only inventor information and the other corresponding to only applicants information. The address/coordinate imputation and batch geocoding algorithms are applied separately to both inventor and applicants databases. For some patents in which the address imputation algorithm could not allocate an address to inventors or applicants, the inventor and applicant databases are consolidated with the RF database. The coordinates in this consolidated database are further mapped to functional urban areas (FUAs) based on shapefiles from OECD, 2012, and shapefiles constructed for cities in non-OECD countries. The final datasets after this data generating process are inventor and applicant databases with 76.6% and 72.4% coordinate coverage respectively. The algorithms used in the data generating process undergo multiple validations to determine how much confidence should be attributed to the address retrieval and geocoding efforts applied to arrive at the final inventor and applicant datasets, since the address retrieval effort can be affected by name ambiguity issues, and the geocoding may provide less precise coordinates. The confidence level of the geocoding process is estimated between 84-88% (based on validation 1 and 2), while the precision of the address retrieval from patent filings in the database is estimated to be 91-93% (based on validation 3).

1.1.1 Patent statistical (PATSTAT) database

The PATSTAT database comprises patents from about 90 patent offices, but most of the patents in this database do not have address information (the country information of inventors and applicants are often available) for inventors and applicants. Specifically, not more than 30% of patents have address information for at least an inventor. Also, when addresses are available, these addresses can either have city information or not. Most of

the addresses available are in major patent offices e.g., EPO, USPTO, and WIPO. Also, major offices like the Chinese Patent Office (SIPO), Japanese Patent Office (JPO) do have a very low address coverage in PATSTAT. We sought to recover these missing addresses of inventors and applicants by first leveraging on the quality address information in the major patent office.

To do this, we consider the raw tables extracted from the PATSTAT database which are: (i) TLS201_ APPLN table, containing the key bibliographical data elements relevant to identifying patent applications. This table has 96,661,363 rows. (ii) TLS207_PERS_APPLN table, which links inventors and applicants to the patent applications in TLS201_APPLN table. (iii) TLS206_PERSON table containing the key information on inventors and applicants (e.g., the applicant's or inventor's name, address, country of residence). This table has 57,682,355 rows. The data is obtained by filtering TLS201_APPLN table based on application year between 2000-2015 (i.e., we selected only patents that were filed between 2000 and 2015). Furthermore, we selected a subset of patents with application kinds: A, F, and W which represent Utility patent, Design patent, and PCT applications respectively. These types of applications represent over 80% of the patents filed in that period. Next, we merged this sub dataset to TLS207_PERS_APPLN table using the primary key in both datasets which is the APPLN_ID (i.e. the application identifier). The resulting dataset identifies inventors and applicants involved in the sub dataset. Finally, to get the attribute of the inventors and applicants, the resulting dataset is merged with the TLS206_PERSON table using the PERSON_ID (i.e., person identifier which is the primary key in both datasets).

This final dataset extracted from PATSTAT database is split into (i) Inventor dataset: this is done by querying INVT_SEQ_NR > 0 (INVT_SEQ_NR is the inventor's sequence number assigned to the inventor name during patent application) and APPLT_SEQ_NR = 0 (APPLT_SEQ_NR is the applicant's sequence number assigned to the applicant during patent application). Also, we included inventors that are considered applicants by querying INVT_SEQ_NR > 0 and APPLT_SEQ_NR > 0. (ii)

Applicant dataset: this is done by querying INVT_SEQ_NR = 0 and APPLT_SEQ_NR > 0. Finally, we excluded patents for which they were neither INVT_SEQ_NR nor APPLT_SEQ_NR. The number of patents in the inventor and applicant dataset are 16,688,197 and 14,886,555 respectively, of which 25.8% and 29.1% of the patents have address information for at least one inventor and applicant respectively. The difference in the number of patents in the inventor and applicant databases is due to the inclusion of patents in which persons are identified as both inventor and applicants (i.e., INVT_SEQ_NR > 0 and APPLT_SEQ_NR > 0) in the inventor database.

1.1.2 External database

The dataset generated from PATSTAT is limited for the following reasons: (i) the address information of inventors and applicants is low and biased in favor of major patent offices (specifically USPTO and EPO) that tend to capture mostly patenting activities of inventors and applicants located in the country where the patent office is located and also inventors and applicants from surrounding countries. Considering the raw PATSTAT database, we have that 95.6% of patents with address information for inventors were filed in USPTO (55.5%) and EPO (40.1%) only. Similarly, 95.5% of patents with address information for applicants were filed in USPTO (63%) and EPO (32.5%) respectively. For each of the patent offices, the distribution of the patents among the top 10 countries accounts for over 90% of the address in the inventor and applicant database respectively. We have that in both USPTO and EPO, inventors from the United States, Japan, and Germany had the most patents with address information. Specifically, of the patents with address information filed in USPTO (EPO), 44.38% (25.29%) had inventors from the United States, and 45.02% (25.92%) had applicants from the United States. (ii) there are significant variations in the quality of addresses, as some addresses only have the country information and others do have complete address information (i.e., country, city, street name, and number). To improve the address quality in the dataset generated from PATSTAT, we curated secondary

or external datasets from the following external patent databases: OECD REGPAT, PV, and MRP.

OECD REGPAT database

The OECD REGPAT database contains patent data that capture patent activities in over 2000 regions in OECD countries. These data provide granular address information for firms (applicants) and persons (inventors) filing their patents in the European Patent Organization (EPO), alongside other information such as citations received by patents, and technical field the patents belong to. This information provided at the regional level can be linked to other regional databases such as labor statistics to provide insight into the connection between innovation and employment. Also, the OECD REGPAT database provides information on PCT patents regionalized for most EU 28, BRICS, and OECD countries (Maraut et al., 2008).

In our data generating process, we used the OECD REGPAT database version 2019, to improve the address information for applications made in EPO and also for "Patent Cooperation Treaty" (PCT) applications. To merge the EPO applications in our generated dataset to that available in REGPAT, we used the combination of "*application identifier*" and "*person identifier*" which are the primary key in both datasets. When there is a missing address for an inventor or applicant on a patent document in the generated dataset, but this is available on the same document in REGPAT, we input the address available in REGPAT. When both datasets have address information for an inventor or applicant on a patent document, we use the address from REGPAT as the prevailing address since REGPAT has been subjected to more cleaning effort. The string comparison and merging of the databases are done after unwanted characters are removed from the address strings. Furthermore, to merge PCT applications in the generated dataset with information from REGPAT we used the combination of "*application identifier*" and "*person name*" since there is no "*person identifier*" for PCT applications in REGPAT. This implies that we needed to make sure that the "*person name*" in both datasets are exactly matched,

hence we clean the "*person name*" in both datasets before merging. After merging we compare the string of addresses in both datasets in the same way we did for the EPO applications.

Morrison, Riccaboni and Pammolli (MRP) database

The MRP database addresses the problem of ambiguity of inventor and applicant (firms or universities) names in the patent database. This problem limits the extraction of proper information of patenting activities of an inventor or applicant. Since some patent databases may reflect different spellings of an inventor or applicant name in cases of multiple applications filed in different patent offices, to resolve this issue Morrison, Riccaboni, and Pammolli, 2017b proposed an algorithm that uses "high-resolution geolocation" information to match alternative spellings of an inventor or applicant to a unique inventor or applicant identifier. This algorithm was applied to EPO, USPTO, and PCT patents. The significant benefit of integrating this database in our data generating process is the high-quality inventor and applicants disambiguation and the high-resolution coordinate information available in the database.

MRP Patent id	*PATSTAT equivalent*
US0nnnnnnn	nnnnnnn
USREnnnnnn	REnnnnnn
USRE0nnnnn	REnnnnn
USH000nnnn	Hnnnn
USD0nnnnnn	Dnnnnnn
USPPnnnnnn	PPnnnnnn
USPP0nnnnnn	PPnnnnnn

Table 2: Modification of MRP patent identifiers.

The MRP database leverage the work done by Li et al., 2014 which covers patents granted by the USPTO between 1975-2010, as well as EPO and PCT patents filed between 1977-2011 and are in the OECD REGPAT (2014

version) database. We merge the MRP database to our generated dataset through the "*patent identifier*" and "*person name*". For PCT applications, the "*patent identifier*" in the MRP database is the "*pct number*", while for EPO and USPTO, the "*patent identifier*" is the "*publication number*". For EPO and PCT applications the "*patent identifier*" is consistent with the generated dataset from PATSTAT, while for the USPTO this is slightly different as it contains a "US" prefix and strings of "0". Therefore, the USPTO "*patent identifiers*" in the MRP database are modified to match the generated dataset from PATSTAT as seen in Table 2.

After ensuring consistency of the primary keys in both datasets for the different types of applications, we merge the MRP dataset with the generated PATSTAT dataset to extract the coordinate information from the MRP dataset. This provides an extra layer of information (i.e., coordinate information for inventor's and applicant's addresses).

PatentView (PV) database

Although the USPTO patents in PATSTAT do have the most address coverage, however, we need the corresponding geographic coordinate for these addresses. In other to get these coordinates we merge our database with the PatentView database. The data in PatentView, 2018 is based on patent data from the USPTO which includes published patent applications from 2001 till date, and granted patents from 1976 till date. The database treats the issue of inventor and applicant name ambiguities and properly geocodes addresses on patents by leveraging "location-based disambiguation" (PatentView, 2018). We use the database to improve the geographic coordinate information for USPTO patents in our generated PATSTAT dataset.

Since the patent identifier in PatentView, 2018 is consistent with what is in the generated PATSTAT dataset, we merged both datasets using a combination of "*application id*" and "*person name*". The "*person name*" is used after performing some cleaning process to account for the different ways the names are written in both datasets. Merging with PV database provides geographic coordinate information for inventors and applicants

on patents filed in USPTO. This database serves as the second source for getting geographic coordinate information (the first being the MRP database).

Institute for Intellectual Property (IIP) database

The JPO is one of the major patent offices, however, it has low address information in PATSTAT. The IIP database is used to improve the low-quality address information for inventors and applicants located in Japan who file patents in JPO, such address information is not available in PATSTAT which implies there is no information to geocode or regionalize. The IIP database is an extraction from the "Seiri Hyojunka data" disseminated by the "National Center for Industrial Property Information and Training" (NCIPI) - an independent institute connected to JPO. JPO records arising from investigations and post-grant infringement settlements are included in the "Seiri Hyojunka data," as well as published information for individual patents. The data is structured in order of application processes (Akira Goto, 2007). This data can be accessed through the IIP platform. Since the IPC code information is only available from 1964, the IIP only provides patent applications filed after 1964 (Akira Goto, 2007).

We were granted access to download the Japanese patent records through the IIP database site: https://www.iip.or.jp/e/patentdb/data/menu.html where we extracted the inventor and applicants database having 12,381,285 and 13,581,692 patent records respectively. The addresses are in Japanese, in which case we used a Japanese geocoding server: serverhttp://newspat.csis.u-tokyo.ac.jp/geocode-cgi/geocode.cgi?action=startto which was only successful at geocoding addresses located in Japan. Merging this dataset with the generated PATSTAT dataset is done in two steps. First, we modified the "*application number*" in the generated PATSTAT dataset to be consistent with the "ida" which is the patent identifier in the IIP database. To do this, we prepend the "*application year*" and sequence of zeros to the *application number* to obtain a 10 digit identifier that is the same as the "ida" in the IIP database. Secondly, since the "*name*" in the IIP database is in Japanese, merging using a combination of "*modified application number*"

and "*person name*" will be impossible. Therefore, within the patents, we identify persons whose residential country is Japan and have the same "*seq_nr*" as in the IIP database.

Linking person names across databases

Inventor and applicant name disambiguation is a predominant issue in PATSTAT (Morrison, Riccaboni, and Pammolli, 2017a) given the fact that the same inventor or applicant can have their names written differently (e.g., abbreviation of names, and reordering of names.) when filing different patents - hence in a given database this can lead to not being able to properly identify patents that belong to an inventor. However, linking patents across databases can be less prone to this as the patents here rely on each other or come from the same source. For example, the United States patent office information in PATSTAT, as well as the PatentView database is sourced from USPTO, and the OECD REGPAT and PATSTAT databases use the same source as they are both managed by the EPO. The linking of person names on a patent in PATSTAT with the same patent in other databases such as PatentView, OECD Regpat, and MRP, is as follows (i) person names are cleaned by removing non-alphabetical characters and any leading or trailing spaces between names and then converting all strings to upper case. Also for some names that do have titles e.g., German patent offices do have some person names with academic titles such as Prof., ING, and Dr., these academic prefixes are removed. (ii) then a string similarity measure between the full names in both datasets is done using a threshold of 0.8. If the similarity score exceeds this threshold, the available information is linked to the patent in PATSTAT. (iii) There are some cases where the ordering of the full name differs (e.g., Pierre Balland Algo in PATSTAT, but Balland Algo Pierre in PatentView) - the full name is split based on space, and the similarity measure is applied to every permutation of both full names, if the threshold of 0.8 is attained for any permutation then we link the information to PATSTAT.

1.1.3 Algorithms for imputation and geocoding

Following the integration of the generated dataset in PATSTAT with the external databases, we were able to improve the address coverage on patents in the applicant database and inventor database to 35.8% (6.7 % increase) and 34.6% (8.6% increase) respectively. Not all addresses in the database do have a corresponding geographic coordinate. For example, in the applicant database, we were able to allocate coordinates to only 70.6% of addresses, hence we have to geocode about 214,100 unique addresses using the Google geocoding platform.

The next stages in the data generating pipeline are the imputation of missing address/geographic coordinate and the geocoding of addresses to geographic coordinate. These steps in the data generating pipeline are done iteratively until there is no significant improvement in how much coordinate can be allocated to missing coordinates in the database. However, before outlining the underlying algorithms for the above stages, we create an address-coordinate function mapping all addresses available to geographic coordinates available in the dataset, this is to ensure we know what addresses in the dataset are yet to be geocoded. The function first performs a text cleaning on the address string and an exact matching of a similar address is done to map geographic coordinates. There are possibilities that some group of addresses is not the same (based on string comparison) as an address that already has geographic coordinate but in principle, they are the same e.g. "Breitenseerstrasse, Vienna Austria" and "Breitenseer strasse, Vienna Austria." In such a situation, we used Libpostal library in Python which is a fast multilingual international address parser trained on Openstreet map data and currently supports normalization in 60 languages and can parse addresses in more than 100 countries. The address parser in the package is used in comparing each component of the addresses, if the city/admin1 and country components are the same and the string similarity between the street name component exceeds 90% both addresses are deemed to be similar and the differences can be attributed to systematic errors when the address was recorded in the database. This function is applied to each split of the integrated

generated PATSTAT dataset (i.e., inventor and applicant dataset). Next, we apply the coordinate-address function across the split by identifying addresses that have geographic coordinates in one split (inventor/applicant) but the same addresses or group of similar addresses in the other split (applicant/inventor) do not have geographic coordinate.

Imputation algorithm

The algorithm implemented for adding missing address and coordinate information follows the heuristics proposed by De Rassenfosse et al., 2013 and De Rassenfosse, Kozak, and Seliger, 2019 for recovering missing information for priority patents. The algorithm proposed by De Rassenfosse et al., 2013; De Rassenfosse, Kozak, and Seliger, 2019 assigns addresses to priority patents (i.e., initial or earlier application) by examining among other things the patent family (i.e., applications which cover the same or similar invention). Typically, these patent families are either classified as (i) the simple family also called the DOCDB family - in the simple family, all applications of this family lay claim to the same priority patent, and the technical contents of the applications are identical. (ii) the extended family also called INPADOC family, the applications are linked to the same priority application, but their content may differ. The De Rassenfosse, Kozak, and Seliger, 2019 algorithm explores the simple family or DOCDB in assigning missing addresses. For any given patent office in any given year, they select all priority patents, then for any of the selected priority patents with missing inventor or applicant address information, the address information is recovered from the following possible sources:

- Source 1: They used the address of the inventor and applicant from the earliest subsequent filing referring to the priority application as the sole priority.
- Source 2: If there is no address information for the inventor or applicant in the earliest subsequent filing, other subsequent filings in the order they were filed are examined.
- Source 3: If there is no address of the inventor in any of the subse-

quent filings, the address of the applicant in the priority patent is used as the address of the inventor, and vice-versa.

- Source 4: If source 3 fails, the same process in source 3 is repeated but on the earliest subsequent filings.
- Source 5: If source 4 fails, the same process in source 3 is repeated but on other subsequent filings that refer to the priority patent as the sole patent.

This heuristic for data recovery was implemented for recovering country information for priority patents, and when all the sources of recovery listed above fail, the De Rassenfosse, Kozak, and Seliger, 2019 algorithm uses the country name where the priority patent was filed as the address information on the priority patent. In our implementation, we extend the algorithm to recover addresses or coordinates for inventors and applicants on both priority and non-priority patents. We exclude the final stage of having to allocate country information of priority patents to inventors or applicants as the location of an inventor on a patent may not necessarily be the same as the location of the patent office. Our implementation goes thus, first we select all patents within a family patent (i.e., DOCDB family) and for each patent application in a family, we recover the missing address of the inventors and applicants on that patent by doing the following [3]:

- Source 1: we use the inventor's or applicant's address (coordinate) in the document if available.
- Source 2: if address (coordinate) is not available, we use inventor's or applicant's address (coordinate) information from other patents (priority or non-priority) that belong to the same family. However, we prioritize patents that were filed in the following offices USPTO, EPO, and WIPO.
- Source 3: if source 2 fails, we use the inventor's or applicant's address (coordinate) from other patents (priority or non-priority)

[3]Note: all sources are applied to the inventor dataset, while only sources 1-3 are applied to the applicant dataset.

that belong to the same family prioritizing patents that were filed in other offices.

- Source 4: if source 3 fails, to recover addresses for inventors on the patent we use the applicant's address (coordinate) from the document. We assume that the inventor should reside close to the location of the applicant. This assumption though reasonable is a bit problematic when patents are filed by MNCs headquarters, but inventors on the patents are from the subsidiaries of the MNCs which is a different location from the headquarter.

- Source 5: if source 4 fails, we use the applicant's address (coordinate) information from other patents (priority or non-priority) belonging to the same patent family. However, we prioritize patents that were made in the following offices USPTO, EPO, and WIPO.

- Source 6: if source 5 fails, we use the applicant's address (coordinate) from other patents (priority or non-priority) belonging to the same family prioritizing patents from offices other than USPTO, EPO, and WIPO.

Finally, in situations where there are variations in inventor or applicant names within the patent family, we computed a string similarity metric using a cutoff of 70% (there is no significant improvement in the address retrieval process using a higher cutoff). Also, in the address retrieval process for inventors, we considered the overall database and looked at cases where we have the same inventor (person name) in the database with some cases of missing address and coordinate information. For such cases, if the names share the following attributes: at least 70% inventor name similarity, same applicant country, and same applicant name, then these names are considered to be a unique person, and the address and coordinate imputation algorithm is applied to resolve the missing information. This probabilistic heuristic used to augment the algorithm improved the address coverage by 4.3%.

Batch geocoding algorithm

To geocode the addresses, we used the Google geocoding platform (see https://developers.google.com/maps/documentation/geocoding/overview) which geolocates addresses in different languages. We also use the Japanese geocoding platform to geocode the address from the IIP database (see: https://geocode.csis.u-tokyo.ac.jp/geocode-cgi/geocode.cgi?action=start). To some extent, these platforms can deal with common issues like spelling errors and irregular formatting. We wrote a python script that queries the database by providing addresses in batches using an API that parses the address and saves the geographic coordinate. There are different levels of accuracy of the resulting geographic coordinate. These levels can either be "Rooftop", "Range interpolated", "Geometric center", or "Approximate". If the resulting coordinate is classified as "Rooftop" this means that the coordinate provides exact location information including street information. A "Range interpolated" classification means that the address lies in an intersection of roads. A "Geometric center" classification suggests the coordinate lies in a region. While an "Approximate" level is the least classification suggesting the coordinate outcome is not exact (*Google Geocoding API*).

1.1.4 Consolidating our database with RF database

The next stage of the data generating process before delineating coordinates into functional urban areas (FUAs) is the consolidation of our database with the RF database. This is done by replacing patents in the final dataset without coordinate information with same patents with coordinate information in the RF dataset. The replacement is done due to the absence of a person identifier in the RF dataset that could have been leveraged for a proper merging. Also, the absence of a person identifier implies we can not apply the address imputation algorithm after merging with RF dataset. The inclusion of the RF database significantly improves the percentage of patents with at least an inventor with geocoded information from 55% to 76.6% and 60% to 72.4% for both inventor and applicant databases respectively, with major patent offices like the Japan, Chinese,

Germany and South Korea having the most significant improvement. This is expected as the RF database source for external datasets from China, Korea, Japan, and Germany patent offices in their data work.

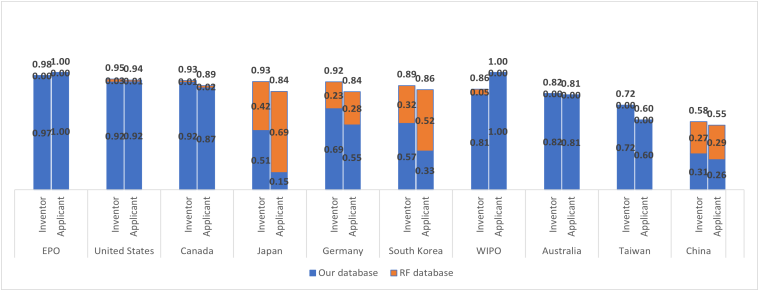


Figure 2: RF database improves the address coverage in our database (inventor and applicant) significantly for Japan, China, South Korea, and Germany patent offices.

Figure 2 shows the improvement for the top 10 patent offices with the most number of patents, we have that the RF database improved the address coverage of inventors for Japan, China, Germany, and South Korea by 41.9%, 27.3%, 22.87%, and 32.4% respectively, and the address coverage of applicants for Japan, China, Germany, and South Korea by 69.4%, 29.44%, 28.45%, and 52.4% respectively. On the other hand, the inventor database improves upon what is available in the RF inventor database by providing more geocoded patents in some selected offices. For example, in USPTO, EPO, WIPO, China, and JPO, we have that about 1442786, 797591, 641817, 540563, and 381843 geocoded information was available in the inventor database but these patents were not available in the RF inventor database.

In summary, the data generating pipeline which involve (i) linking PATSTAT to external databases (i.e., PatentView, MRP, IIP and OECD REGPAT) improves the address coverage for the inventor and applicant database to 35.8% and 34.6% respectively, representing an increase of 6.7% and 8.6% respectively from the raw PATSTAT. (ii) imputation of missing address using sources 1-6 improves the address coverage to 55% and 60% for

inventor and applicant database respectively (iii) consolidating the RF database improves the address coverage to approximately 76% and 72% for the inventor and applicant databases respectively.

1.1.5 Functional urban area demarcation

Many countries base their definitions of cities on the population size and density of a local administrative unit (Dijkstra, Poelman, and Veneri, 2019). There are two problems with this definition. The first is that some cities in large local units may have low density, while some small cities in small local units may have high density. The second problem is the polycentric nature of some cities - as they are dispersed across numerous local units (Dijkstra, Poelman, and Veneri, 2019). Therefore, in this subsection, we will use the shapefiles based on the EU-OECD methodology and modify the methodology to construct functional urban area shapefiles for cities in non-OECD countries. This notion of FUAs is better adapted to capture agglomeration economies than administrative units, as they capture the economic extent of cities (Dijkstra, Poelman, and Veneri, 2019; OECD, 2012).

The EU and the OECD collaborated to develop a framework for defining functional urban areas (FUAs) consistently across countries. FUAs are created in a series of steps. First, a population grid allows "urban centers" to be defined regardless of administrative or statistical boundaries. An urban center is a grid-based concept that consists of a cluster of contiguous cells with a high density and a population of more than 50,000 people.

EU-OECD methodology

The EU and OECD developed a framework for defining FUA. The four steps for constructing an FUA are outlined as follows. First, regardless of the administrative boundaries, an “urban center” is identified through clusters of contiguous cells of population grids (of 1500 residents per square kilometer). The cluster for an “urban center” should have a population size exceeding 50,000 people, these “urban centers” could spread over multiple local units. Secondly, a city is defined to be made up of

one or more local units with more than 50% of its residents located in the urban center (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019). This city is referred to as the core city in this "urban center". Thirdly, the peripheral cities or commuting zones are identified based on "travel-to-work" information used to determine if the commuting zone is "economically integrated" with the core city. A commuting zone is "economically integrated" with the core city if 15% of its residents commute to the core city. Finally, a functional urban area is constructed by combining the core city and the commuting zones identified (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019).

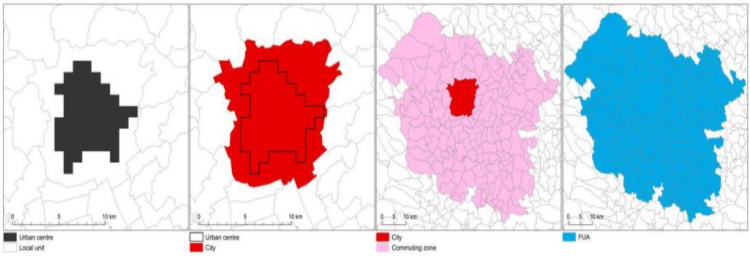


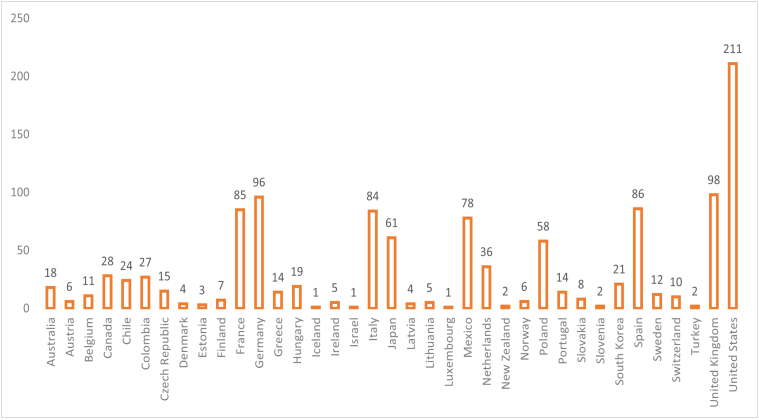
Figure 3: Urban center, city, commuting zone, and functional urban area of Graz, Austria as illustrated in Dijkstra, Poelman, and Veneri, 2019.

Defining an urban center, city, and commuting zone:

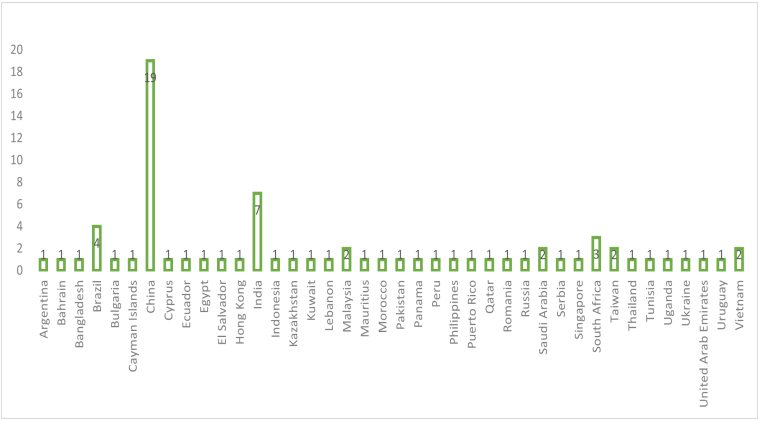
The population grids used in constructing urban centers are publicly available (e.g., GHS-POP and CIESIN GPW v4). There are different thresholds for the contiguous cells used in identifying urban centers. For cities in the EU, United States, Chile, and Canada, clusters of contiguous grid cells should have a population size exceeding 50,000, while for cities in Japan, Korea, and Mexico, these clusters should have a population size exceeding 100,000 residents (Dijkstra, Poelman, and Veneri, 2019; OECD, 2012).

Since an urban center can span multiple local units, defining core cities in the urban center depends on what local units have the most residents located in the urban center. The threshold for a local unit to be considered

a city is that it has 50% of its residents in the urban center. If the urban center consists of only a local unit, then the local unit is considered the core city. If the urban center is embedded in multiple local units meeting the threshold, then these local units are integrated if either of them is seen as a commuting zone of the other, else they are treated as separate core cities (Dijkstra and Poelman, 2012; Dijkstra, Poelman, and Veneri, 2019; OECD, 2012).



(a) 1154 FUAs in OECD



(b) 73 FUAs in non-OECD

Figure 4: Distribution of FUAs in OECD and non-OECD countries.

A commuting zone is used for either resolving the case of polycentric core cities as described above or determining if surrounding local units should be integrated into core cities. This is inferred by examining travel-to-work information between cities. A surrounding local unit is deemed a commuting zone if more than 15% of its population commute to work in the core city. In the case where the flows are to multiple cities, the surrounding local unit is integrated with the core city it has the most commuting flow with (Dijkstra and Poelman, 2012; Dijkstra, Poelman, and Veneri, 2019; OECD, 2012).

Finally, an FUA is the combination of cities identified and the commuting zones linked to these cities. Although this approach helps in addressing the modifiable unit problem by considering the economic extent of cities. However, the size of an FUA can go below or exceed the size of an urban center, and can cross state boundaries (Dijkstra, Poelman, and Veneri, 2019). This EU-OECD FUA methodology described provides shapefiles of 1154 FUAs located in OECD countries. As shown in Figure 4, the 1154 FUAs are distributed across 37 OECD countries, and the country with the most number of FUAs is the United States with 211 FUAs.

Extension of the EU-OECD methodology

Applying the EU-OECD methodology requires the following data sources: residential population data, digital boundaries of the local unit, and travel-to-work flow information which are not available in most non-OECD countries. Therefore we adapt the EU-OECD methodology for 73 cities in non-OECD countries selected from the alpha-beta list of cities based on the Globalization and world cities (GAWC) project (Beaverstock, Smith, and Taylor, 1999). These cities can be found in China, India, and Brazil. From Figure 4, the country with the most number of cities delineated to FUAs is China, having 19 cities on the list. The extension of the EU-OECD methodology is based on patent density identified in these cities, and commuting zones are identified based on average travel-to-work time from the city, which is estimated using the Open Street Map (OSM) application.

Defining commuting zones:

First, we select 78 cities in non-OECD countries from the alpha-beta list of cities in GAWC. Cities categorized as alpha are seen to link "major cities and regions to the global economy", and cities categorized as beta are seen to link "moderate economic cities and regions to the global economy". To identify commuting zones surrounding these cities, we identify urban centers in the city using the Open Street Map (OSM) application (the application provides a freely adaptable map of the world based on local knowledge, GPS data, and air/satellite photography)

Commuting zones are places around the urban center that a resident in the urban center is willing to commute to for work purposes. The average commuting time estimated does not take into account traffic jams. This time can be determined for each city using the OSM application, and the time varies between cities. Next, isochrones are constructed based on the average traveling time. These isochrones are considered FUAs as they include both the urban centers and commuting zones, which can extend beyond the city. The average traveling time used ranges between 30-60 minutes.

The OSM application only allows the calculation of standard vehicles: bicycles, cars, and trucks. We used cars in the OSM tool based on the assumption that most people use cars to go to work and these are generally faster than bicycles. However, public transport in form of trains is usually faster compared to cars. Therefore, we used either Google maps, ISO4APP or Mapnificient to check whether one could reach farther in cases where public transport is used. In the majority of cases, the isochrones were either equal or smaller compared to the isochrones of cars measured by the OSM application. In the rare case when they were not, we slightly extended the isochrones using the Quantum Geographic Information Systems (QGIS) software. In cases where we have overlapping isochrones of FUAs within a country, we reduce the shapefile of the non-capital city to eliminate the overlap. Finally, to validate the demarcation of the shapefiles, we check if the delineation matched with the hotspots of innovation. To do this, we project the geographic coordinates from the

patent database onto the shapefile and showed a heat map that captures patent intensity in hot spots. If there are spots located outside the city center that is not captured, we expand the isochrones to include such hotspots. In all cases, the isochrones never exceeded the average driving time of 60 minutes. These isochrones are stored as FUA shapefiles.

Case study (Sao Paulo):

Let us consider the construction of the Sao Paulo isochrone as a case study. Figure 5 depicts two images, the first is a shapefile of Brazil with green dots representing coordinates of inventors located in Brazil. We see a concentration of the green dots on the Southeast side of the country. By zooming in on that area we can distinguish 5 clusters (Bel Horizonte, Rio De Janeiro, Sao Paulo, Porto Alegre, and the area surrounding Florianopolis), of which 3 clusters i.e., Rio De Janeiro, Sao Paulo, and the Florianopolis region have a very high density.

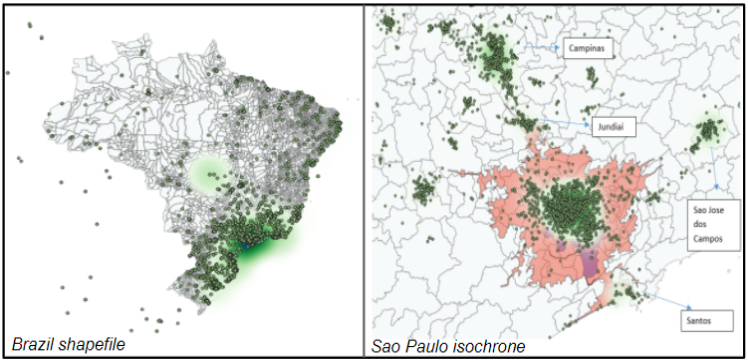


Figure 5: Sao Paulo's isochrone: The green dots are inventor locations. The purple patch capture inventors located 30 minutes from the urban center, while the orange patch capture inventors located 60 minutes away from the urban center.

Based on the country data from 2012-2014, the majority of the population (around 50%) travel only 30 minutes to work, while 30% travel between 30 minutes to an hour to work. Only 20% of the population travel longer than 60 minutes to work. We could not find any official documents on

which transportation method was dominantly used, but several reports indicate that Brazil struggles with poor infrastructure, causing people to go by bus or by car as metro or train lines are very limited. However, in more recent years, Brazil has been investing more in developing efficient public transport.

Further investigation into the public transportation reach of Sao Paulo using Mapnificient shows that the driving time approach does cover all of the public transportation routes that can be covered in the same time-span. Given this information, we established isochrones based on 30 minutes drive and a 60 minutes drive. As seen in Figure 5, the majority of the patents are captured within the 30-minute isochrones (i.e., the purple region). However, only including the 30-minute range leaves out a couple of dense areas such as Santos and Jundiai which are captured in the 60-minute isochrones. The 60-minute isochrones capture almost everything in the neighborhood of Sao Paulo. Although, some clusters can still be found in the neighborhood (Campinas and Sao Jose dos Campos), these are located too far away to be considered part of Sao Paulo.

1.1.6 Algorithm validation

The validity of the data generating pipeline depends partly on the validity of the geographic coordinate information in the external databases that were integrated. Another possible source of invalid geographic coordinate information can be intrinsic to the data generating pipeline, specifically at the coordinate or address imputation and geocoding stages. Therefore, we choose to validate the data generating pipeline by doing the following:

Validation 1:

The first validation randomly confirms if the algorithm imputing the missing address is performing the task correctly and if we recover a similar geocode should we choose to independently geocode the available address. To do this, we estimate the level of discrepancy between the geographic coordinate that was assigned based on the imputation algorithm and the geographic coordinate gotten from independently geocoding

available addresses. This process is done on 5 samples of randomly selected 1,000 addresses without replacement. The discrepancy computes the average distance (km) between both coordinates. These 1000 addresses randomly selected for observation are complete (for example, "48 Pirrama Rd, Pyrmont, NSW, Australia") and are fed into the Google geocoding platform using the batch geocoding python script (check Appendix A.0.4). Given the latitude and longitude from the imputation algorithm as (θ_1, λ_1) and the Google platform as (θ_2, λ_2). The great circle distance (in km) between the two points is computed as follows:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\theta_2 - \theta_1}{2}\right) + \cos(\theta_1)\cos(\theta_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{1.1}$$

where r is the radius of the earth (in km). The discrepancy is the percentage of the great circle distances computed that lie within a given threshold e.g. 5km, 10km. Table 3 shows the discrepancy at different threshold levels for each of the samples. For example, in Sample 5 we have that for 71.7% of the address in the sample the great circle distance between the coordinates gotten through the imputation algorithm and that gotten through the Google geocoding platform is less or equal to 5km, and for 94.6% of the address in the sample, the great circle distance is less or equal to 50km. Following the results in Table 3, we adopt a threshold of 10km as a reasonable great circle distance to validate the algorithm. Based on this threshold, the algorithm is on average 84.14% valid.

	$\leq 5km$	$\leq 10km$	$\leq 20km$	$\leq 30km$	$\leq 40km$	$\leq 50km$
Sample 1	69.5	84.1	91.8	92.8	93.4	93.6
Sample 2	71.7	86.4	94.4	94.8	95.1	95.2
Sample 3	68.3	82.7	92.2	93.1	93.9	94.5
Sample 4	69.4	82.9	91.4	92.9	94	94.2
Sample 5	71.7	84.6	92.1	93.5	94.1	94.6

Table 3: Validation of imputation algorithm: this computes the percentage of the great circle distances at different thresholds for each sample.

Although, the choice of 10km is a conservative choice for the discrepancy,

as typically, the size of FUAs exceed $10km$. For example, the FUA with the least size is required to have a density of at least 1500 per km^2 and a population exceeding 50,000, implying that the size of the FUA will typically exceed $30km^2$. Hence, using this threshold, we have a low geocoding discrepancy less than 10%.

Validation 2:

We estimate the discrepancy between the coordinates assigned based on the imputation algorithm and the coordinates in the De Rassenfosse, Kozak, and Seliger, 2019 (RF) database. Although computing the discrepancy at the level of inventors and applicants is not feasible because the RF database does not have identifiers for inventors and applicants.

	Patent office	*Number of patents*	*Patent office*	*Number of patents*
0	Switzerland	93	Luxembourg	94
1	Spain	84	Ireland	90
2	Germany	85	Poland	93
3	Mexico	85	Czech Republic	95
4	United Kingdom	90	WIPO	84
5	Norway	93	Hungary	89
6	Canada	73	Romania	92
7	China	94	South Africa	91
8	Australia	82	Finland	91
9	New Zealand	82	Denmark	92
10	Sweden	89	Chile	91
11	South Korea	91	Greece	86
12	Japan	93	India	82
13	United States	85	Latvia	94
14	EPO	94	Bulgaria	92
15	Austria	90	Turkey	68
16	France	92	Slovenia	95
17	Italy	94	Croatia	91
18	Netherlands	99	Slovakia	97
19	Russia	71	Lithuania	93
20	Portugal	87	Estonia	97
21	Brazil	68	Israel	84

Table 4: Validation of imputation algorithm: the number of patents in the final dataset whose coordinates are at most 10km away from coordinates of the same patent in RF database.

Therefore, for selected patent offices we randomly select 100 patents that

are both in the final dataset and the RF dataset. Using the established threshold of 10km in Validation 1, we check if the coordinates on the patents from the final dataset are at most 10km away from coordinates on the same patent in the RF database.

Table 4 shows the result for each patent office. For example, of the 100 patents selected from the Switzerland patent office in the final dataset, coordinates on 93 of those patents were at most 10km away from the same patents in the RF database. We have the least discrepancy for the Netherland patent office i.e. 99/100 and the most discrepancy for the Brazil and Turkey patent office i.e. 68/100 for both. On average, 88/100 patents do have coordinate information that is less than 10km away from the same patents in the RF database. Therefore the validity of the imputation algorithm based on this is 88%.

Validation 3:

Next, we validate the address retrieval process that requires obtaining address information from subsequent or priority patent filings within a patent family. This process can be impacted by ambiguity in names as the similarity of strings and co-applicants conditions used may miss out on some addresses or attribute the wrong address to an inventor or applicant. To evaluate the accuracy of this process, we consider the Brazil patent office - which is the patent office with the least accuracy based on Validation 2. We randomly select 100 patent families containing patents (priority and non-priority patent) filed in the Brazil patent office. We require that for any of the 100 patent families, patents filed in the Brazil office do have missing inventor and applicant address, but at least one patent of the same family does have address information for inventor and applicant.

The idea is to manually perform the address retrieval for both inventors and applicants using the same process of looking into available addresses in other patents within the patent family, and then use this as a benchmark to determine the accuracy of the address retrieval process implemented. The randomly selected 100 patent families do have 289 unique inventors

and 98 applicants. To determine the accuracy of the address retrieval from patents within the same patent family, we measure the precision and recall. For any of the Brazil office patents P in the selected 100 patent families, let $P(i)$ represent the set of geolocated inventors and $P^B(i)$ represent the set of geolocated inventors from the manual address allocation done. Then the true positive ($TP(P)$) is the number of inventors having the same address as the benchmark, the false positive ($FP(P)$) is the sum of all inventors having different addresses from the benchmark, and the false negative ($FN(P)$) is the sum of inventors still having missing address but the benchmark does have address information. Therefore an estimate for the precision and recall of the address retrieval process is given as:

$$\begin{aligned} Precision &= \frac{\sum_P TP(P)}{\sum_P (TP(P) + FP(P))} \\ Recall &= \frac{\sum_P TP(P)}{\sum_P (TP(P) + FN(P))} \end{aligned} \tag{1.2}$$

	Inventor dataset	Applicant dataset
1) Number of patents	100	100
2) Number of inventors or applicants	289	98
3) Precision 4) Recall	0.93 0.98	0.91 0.99
5) Address retrieval from applicant	10 patents $\approx$ 41 inventors	-

Table 5: Benchmarking of address retrieval process for patents in Brazil office

In Table 5 we find a higher recall and precision of the address retrieval process for both the inventor and applicant datasets. For the inventor dataset, the true positive, false positive, and false negative were 246, 18, and 4 inventors respectively; while for the applicant dataset, the true positive, false positive, and false negative are 87, 9, and 0 applicants respectively. The precision being lower than the recall suggests that the address retrieval process results in fewer false negatives than false positives. An

example of a false negative encountered is with the inventor name "VAN LIR Astrid Lutsiya Khelena Mariya Villemina", the same inventor on a patent belonging to the same family had the name written as "ASTRID LUCIA HELENA MARIA WILLEMINA VAN LIER." The similarity measures between strings was not robust enough to identify both inventors as the same, hence the address for "VAN LIR Astid Lutsiya Khelena Mariya Villemina" could not be filled. An example of a false positive is "Schlumberger Surenco" being geolocated to The Hague, Netherlands when this firm should have been geolocated to Bogota, Colombia, this arises due to similarity with "Schlumberger".

1.2 Results on data quality

Table 6 describes the variables in each of the final datasets (i.e., inventor and applicant).

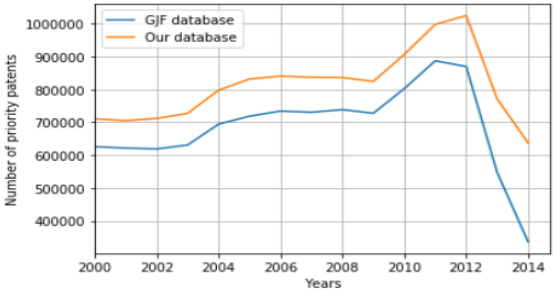
	Inventor dataset *Variable*	Applicant dataset *Variable*	*Description*
1	appln_auth	appln_auth	The competent authority (i.e., patent office) responsible for processing patent application
2	appln_filing_year	appln_filing_year	Year of the application filing date.
3	appln_id	appln_id	unique identifier for patent application.
4	coordinate	coordinate	The combination of latitude and longitude which define the geographic location of inventors and applicants.
5	docdb_family_id	docdb_family_id	Docdb family identifier means that the applications share the same priorities.
6	earliest_filing_year	-	Year of the earliest filing date.
7	fua_ctry	fua_ctry	Country of the functional urban area
8	fua_name	fua_name	This is the urban location of the inventor or applicant. FUA consists of a "densely inhabited city and a less densely populated commuting zone whose labor market is highly integrated with the city" (Dijkstra, Poelman, and Veneri, 2019).
9	latitude	lat	The geographic coordinate that specifies the North-South position of the inventor or applicant.
10	longitude	lon	The geographic coordinate that specifies the East-West position of the inventor or applicant.
11	person_address	person_address	Correspondence address of the inventor or applicant.
12	person_name	person_name	Name of the inventor or applicant.
13	person_ctry	person_ctry	Country part of the correspondence address of the inventor or applicant.
14	error	error	error is either 1 or 0. It is 1 if the person country is not the same as fua country, and 0 otherwise. In this case, always use the person country and ignore the coordinate and FUA information.

Table 6: Variables in the final datasets: Description of variables in the inventor and applicant datasets.

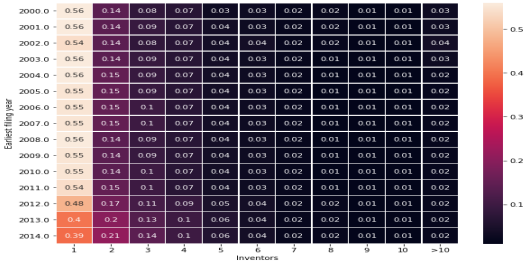
In the applicant dataset, there is the functional urban area country which is based on the data generating process, and the person country which is directly gotten from the PATSTAT database. Therefore, there can be a possible discrepancy between the person country and functional urban

area country which we estimate to be the case for 2.68% patents in the inventor database and 3.50% in the applicant database. Specifically, of these 2.68% patents with different country information in the inventor dataset, we have that 16.6% patents had inventors from Japan been given different country information (mostly United States (83.6%), Canada (8.7%), and South Korea (2.9%)). Also, 15.5%, 11%, 8.6%, 8.1%, 7.9% of these patents with discrepant country information were inventors in United States, China, Germany, United Kingdom, and Canada mostly mistaken to be located in China, United States, United States, United States, United States respectively. Most of these errors arise from sources 4-6 in the imputation algorithm where inventors are allocated address information of the applicant (the case specifically 1.6%). This discrepancy for global cities is approximately 2.01%. The patent offices accounting for the most discrepancy in the inventor dataset are the USPTO, Korea patent office, WIPO, Taiwan patent office, and EPO respectively. Similarly, for the applicant database, of the 3.5% patents with the discrepancy, most of these discrepancies are observed for applicants located in the United States (28.6%), Germany (15.9%), Japan (13.4%), France (8.7%), United Kingdom (5.5%), are often mistakenly assigned to South Korea. This mistake is mostly seen in patents offices such as WIPO, USPTO, EPO, Canada, and Germany patent offices. For cases where these errors are observed in both the inventor and applicant database, the column error is set to 1 to flag this discrepancy, hence when doing a country-level analysis we recommend sticking to the person country information in the original PATSTAT dataset, rather than the FUA country information.

This kind of discrepancy only shows that the data generating process is not 100% valid as we estimate the validity of the data generating process to be between 84-88% (considering 10km radius) based on the validations performed in the algorithmic validation section. For FUAs which are usually larger than administrative city boundaries, we can go beyond the 10km radius, in which case the validity of the algorithm is about 92.38%, 93.42%, 94.10%, and 94.42% for radius 20km, 30km, 40km, and 50km respectively.



(a)



(b)

Figure 6: (a) compares the number of geocoded patents in the final inventor dataset with the De Rasenfosse inventor dataset (b) shows the fraction of geocoded patents with specific numbers of inventors location geocoded.

Figure 6 and 7 demonstrate the quality of geocoding coverage in the final dataset that emerges from the data generating process, by showing that the final dataset does have a significant geocoding coverage when compared to the De Rassenfosse, Kozak, and Seliger, 2019 (RF) dataset, and that there is a significant improvement when compared to the raw PATSTAT dataset. Figure 6a shows the number of priority patents (i.e., first filings) with geocoded information for at least an inventor on the patent. In the period 2000 to 2014, the RF database has 10.3 million priority patents with geocoded information for at least an inventor on the patent, while in our final database we have 12.1 million priority patents with geocoded information for at least an inventor on the patent.

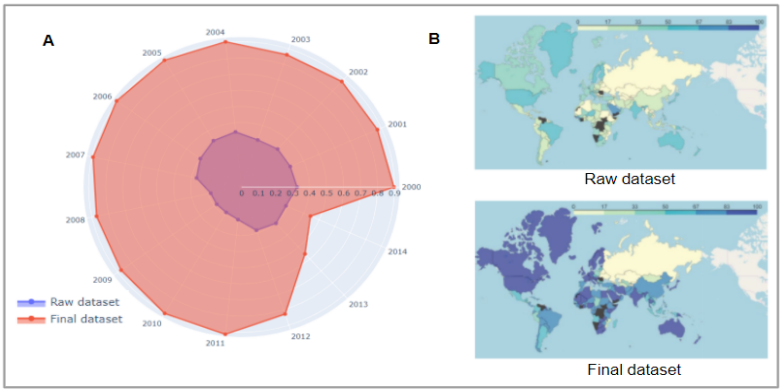


Figure 7: Significant improvement from the raw PATSTAT dataset: Panel A shows the yearly percentage of patents with at least a coordinate or address information for both the final inventor dataset and the raw inventor dataset from PATSTAT (after including addresses and coordinate from OECD database). Panel B shows the heat map for countries based on the fraction of patents with at least coordinate information. This is computed at the level of the country using all patents filed between 2000-2014.

As depicted in Figure 6a we achieved significant coverage for every year. The decrease in the last year stems from our data gathering process incorporating data truncated in 2015. In general, we have approximately 2 million more geocoded patents than what is available in the RF database for the same period. This could partly be attributed to the fact that our geocoded inventor database includes patents in which the inventors are also applicants. This set of patents may or may not be accounted for in RF database. Figure 6b shows the fraction of geocoded patents with the specific number of inventors geocoded. For example, we have that in the year 2000, 56% of the geocoded patents had coordinate information for only one inventor on the patent while 3% of the geocoded patents had coordinate information for more than ten inventors on the patent.

Panel A in Figure 7 compares the percentage of patents with at least one inventor geocoded in the inventor dataset with the percentage of patents with address information for at least one inventor in the raw inventor dataset from PATSTAT. We find a significant yearly improvement in avail-

able information for all years, and for the applicant dataset, we have an overall coverage of 72%. Panel B shows an overall improvement in the final inventor dataset across countries with strong coverage in countries like the United States, Japan, China, India, France, Germany, Belgium, Denmark all having a coverage exceeding 90%. There is significantly low coverage for countries like Moldova, Tajikstan, Russia, Ukraine, Kyrgyzstan, Georgia, and Kazakhstan all have a coverage lower than 20%. Nevertheless, this is an improvement from what was in the raw PATSTAT dataset. For example, Russia in the raw dataset had address information for less than 3% of its patent, but in the current database, we have a coverage of 12.1%.
Finally, not all geocoded patents have their coordinates assigned to the 1227 functional urban areas. For those assigned to the 1227 FUAs, we observe that these FUAs account for 59.9% of patents in the applicant database, and 58.1% of the patents in the inventor database. Cities with the most number of patents include Tokyo, Seoul, Higashiosaka, Toyota, San Francisco, and New York, while cities with the least number of patents include Poza Rica de Hidalgo, Navojoa, and Acuna having just a single patent.

1.3 Application of the database: analyzing domestic and international linkages

The role of cities as knowledge hubs has been central to research in the economic and geography literature for more than a century. This is because the knowledge generation and flows between cities contribute to the sustainable economic growth of countries, and ultimately the economic globalization of these cities. Economic globalization has been based on the extent of provision of services and flow of capital to an international market, however considering the importance of innovation in cities makes room for characterizing economic globalization as "concrete economic complexes" specific to some cities, and hence the need for interdependence of both core and peripheral cities (Sassen, 2001). In this section, we examine the trends in domestic and international interdependence

between cities based on patent collaboration network. In this network, the nodes are cities, while the links or edges between cities are formed when inventors in different cities collaborate to produce a patent. A domestic linkage is formed when the link is to a city that is located in the same country, an international global city linkage is formed when the link is to a city in another country and the city is considered a global city, and an international non-global city linkage is formed when the link is to a city in a different country and the city is not considered a global city.

We define core cities or global cities using the characterization and classification of cities in Beaverstock, Smith, and Taylor, 1999. These global cities are the 168 alpha-beta classified cities that are linked to the world economy and are responsible for linking moderate economies to the world economy. Because global cities are becoming more important as command centers and headquarters for functional activities, the diversity of these specialized activities and the attractiveness of high-skilled labor are becoming increasingly important in the valuation of today's leading sectors of capital. Therefore, global cities are seen to be significant hubs for today's most important economic sectors (Lorenzen, Mudambi, and Schotter, 2020). Also, non-global cities or peripheral cities do have a crucial role in the overall economic configuration of a country, as their specialization in low skilled knowledge or business areas sustains the local supply chain of the country (Lorenzen, Mudambi, and Schotter, 2020; Soja, 2014; Henderson and Castells, 1987).

The economic dimension of the linkages between global and non-global cities can be through trade relationships between agricultural producers and suppliers, and more recently due to the rise of the knowledge economy, knowledge collaboration such as research collaborations between institutions, firms, and inventors. These economic connections can create value chain relationships that promote flexibility and efficiency, foster spillovers that induce exploratory activities and refine the exploitative innovation process (Boschma and Iammarino, 2009; Breschi and Lenzi, 2016).

The development of non-global cities is often hinged on domestic link-

ages to global cities, as the formation of this kind of connection requires low resources when compared to the formation of international linkages either to global or non-global cities (Lorenzen, Mudambi, and Schotter, 2020). The functional system of cities reflects a form of inter-dependence between global and non-global cities which can be useful for the development of a country due to differences in specializations of these cities. For example, a non-global city like Williston, North Dakota has helped catapult the United States to the top of the world's list of oil producers due to the shale revolution and the city being a production site of shale oil. Although global cities e.g. New York might be responsible for the business specialization of the Oil & Gas sector, Williston remains the industrial production site. Hence, the inter-dependence of these cities has improved not just the economic outcome of the United States in this sector, but also the capital flow in these cities.

From the perspective of local R&D collaborations between cities located in the same country, such collaborations could be facilitated by universities or firms. and these are important for the distribution of economic opportunities. The distribution of economic opportunities is made available as global cities evolve into more specialized technological areas that are globally competitive, which results in the need for non-global cities to fill-in knowledge areas useful to the domestic economy. The existence of domestic linkages implies that plugging into these economic opportunities will not be very difficult for non-global cities, as they can leverage domestic linkages for learning and producing domestically relevant technologies.

Although most inter-city linkages are within the border of countries, there has been significant growth in the formation of international linkages. This is driven largely by the mobility of both skilled and unskilled workers, and the reinforcing advocacy of firms for free mobility to take advantage of skills outside the location of firms and to create a geographically disaggregated value chain that allows them to undertake activities at the most efficient global location (Masey, 1984; Lorenzen, Mudambi, and Schotter, 2020). Such international linkages often time facilitated by multinational corporations lead to the creation of spatial configuration

connecting global cities in advanced economies to non-global cities in emerging economies in the most cost-efficient way. This formation of linkages is motivated by lower spatial transaction costs and the existence of technological clusters with certain specializations. Such clusters are the result of locally diverse skill bases in the city and company capabilities being reinforced in a path-dependent manner. Following a city's acquisition of the needed capabilities in a sector, successive corporate investments can reinforce these accumulated capabilities over time. (Hannigan, Cano-Kollmann, and Mudambi, 2015; Lorenzen, Mudambi, and Schotter, 2020).

1.3.1 Measurement of weighted and unweighted linkages

Formally, let $G = (V, E)$ denote network structure composed of $V(G) = \{1, ..., N\}$ representing a set of N nodes (i.e., cities), that are connected by linkages $E(G) = \{(i, j) : i, j \in V(G)\}$. These linkages exist if inventors from cities i and j are co-inventors on a patent. The network structure $G = (V, E)$ can be represented by the weighted adjacency matrix $A \in \mathbf{R}^{N \times N}$ as follows:

$$A_{ij} = \begin{cases} w_{ij} & if (i, j) \in E(G) \\ 0 & otherwise \end{cases} \tag{1.3}$$

where w_{ij} is the number of patents involving inventors from cities i and j. The unweighted linkages are simply the number of linkages a city has to other cities - it is a measure of the extensive innovative ties it has with the world, and it is given as:

$$Unweighted\ linkage\ of\ city\ i = \sum_{w_{ij} \neq 0} 1 \tag{1.4}$$

On the other hand, weighted linkages can be a measure of the usefulness of innovative ties taking place collaboratively if this is measured using the number of forward citations received. Here, we focus on the intensity

of the linkages observed between cities - it captures the average number of patents a city has with other cities, and it is given as:

$$Patent - weighted\ linkage\ of\ city\ i = \frac{1}{U_i} \sum_{j \in V(G)} w_{ij} \tag{1.5}$$

where U_i is the number of linkages of city i - unweighted linkage of city i.

1.3.2 Results

In this section, we provide results at levels of cities, countries, and continents, on the evolution of domestic linkages (weighted and unweighted), international linkages (weighted and unweighted) to global and non-global cities, and the evolution of the share of both domestic and international linkages (global and non-global). The share of a linkage (domestic or international) in a given city is the number of such linkages divided by the total linkages the city has.

Trends in domestic linkages

Figure 8 illustrates the average number of domestic linkages and average number of patent-weighted domestic linkages. First, we observe a decline in the average number of unweighted linkages, nevertheless, we see an increase in unweighted domestic linkages prior to 2007 and a steady decline in unweighted linkages after 2007. We observe different episodes of decline and rise in the average weighted domestic linkages, but between 2000 and 2014, we see a 17.3% increase in the average number of patent-weighted linkages.

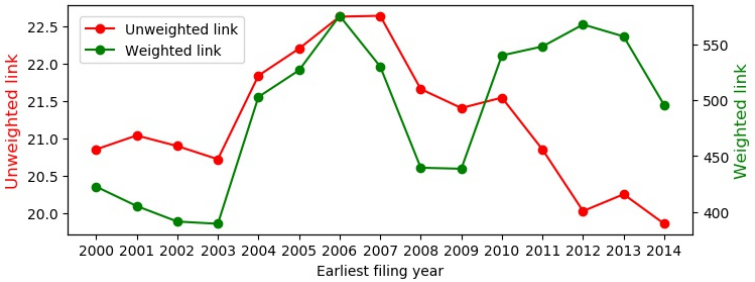


Figure 8: Domestic linkages: The green solid line represents the yearly trend of the average number of patent-weighted linkages between cities in the same country, while the red solid line is the yearly trend of the average number of links between cities in the same country. Note: intra-city connections are not counted.

Table 7 indicates the top 50 cities in terms of average weighted domestic linkages between the period 2010-2014. We observe that cities in Asia e.g., Seoul, Tokyo, and Seo are more represented at the top of the ranking. Also, we observe relatively strong changes in the ranking. For example, Beijing and Shenzen moved 185 and 163 upward in the period 2010-2014 from their ranking in the period 2000-2004 and had 966.5% and 1170.9% increase in average weighted domestic linkages, while in the same period, Mito and New Haven moved 20 and 22 downward and had a 2.06% and 16.18% decrease in average weighted domestic linkages respectively. Also, we have that San Francisco, New York, and Portland were the only top cities that maintained a stable position in their ranking. They ranked 4th, 5th, and 38th position respectively and had a 191.8%, 37.2%, and 52.2% increase in average weighted domestic linkages respectively. At the level of the country, we see that the top countries with the most average weighted domestic linkages are South Korea and Taiwan with a 175% and 242% increase in linkages between the period 2000-2004 and 2010-2014. While South Korea retained its number one position across all periods, we see a decline in the United States position by one point and a rise in the position of China by 15 points.

At the level of the continent, we see a stable ranking across all periods

with Asia consistently taking the lead followed by North America, Europe, Oceania, Africa, and South America in that order.

	Cities	Domestic linkages (2010-2014)	Ranking (2010-2014)	Domestic linkages (2005-2009)	Ranking (2005-2009)	Domestic linkages (2000-2004)	Ranking (2000-2004)	Domestic linkages (% change)	Change in ranking
0	Seoul	55896.8	1	35153.2	1	21283.2	2	162.63	1
1	Tokyo	26964.8	2	29585.2	2	28029.8	1	-3.79	-1
2	Seo	26093.0	3	16038.0	4	8941.8	6	191.80	3
3	San Francisco (Greater)	20449.2	4	16900.4	3	14902.6	4	37.21	0
4	New York (Greater)	14353.0	5	14836.0	6	13808.2	5	3.94	0
5	Higashiosaka	12924.2	6	15671.2	5	15366.2	3	-15.89	-3
6	Gimhae	12511.2	7	6840.0	12	4626.2	14	170.44	7
7	Dalseong	10378.6	8	6088.2	15	3672.6	19	182.59	11
8	Boston	9344.4	9	8889.0	8	8528.6	7	9.56	-2
9	Toyota	8914.4	10	9585.2	7	8209.6	8	8.58	-2
10	Los Angeles (Greater)	8037.0	11	7869.4	9	7666.6	9	4.83	-2
11	Nam	7727.8	12	4052.6	21	2819.2	26	174.11	14
12	Gwangsan	6955.6	13	3951.0	23	2454.2	34	183.41	21
13	Philadelphia (Greater)	6646.8	14	7258.4	11	7292.2	10	-8.85	-4
14	Chicago	6352.6	15	6309.8	13	5338.4	11	18.99	-4
15	San Diego	6204.2	16	5768.2	16	4671.4	13	32.81	-3
16	Stuttgart	5719.0	17	7333.4	10	4358.8	15	31.20	-2
17	Sebuk	5717.8	18	3433.2	31	2006.0	41	185.03	23
18	Seattle	5533.0	19	5069.2	17	3217.8	22	71.94	3
19	Taipei	5154.2	20	3230.6	35	1415.2	63	264.20	43
20	Heungdeok	5109.2	21	2759.4	43	1791.6	46	185.17	25
21	Shenzen	5106.4	22	2101.8	48	401.8	185	1170.88	163
22	Seongsan	5017.4	23	3166.0	37	1829.6	44	174.23	21
23	Munich	4688.0	24	6152.4	14	3843.0	17	21.98	-7
24	Washington (Greater)	4675.8	25	4726.4	19	4769.6	12	-1.96	-13
25	Detroit (Greater)	4474.2	26	4406.8	20	3856.8	16	16.00	-10
26	Deokjin	4048.2	27	2077.6	50	1455.6	62	178.11	35
27	Houston	4040.8	28	3383.2	32	2702.0	30	49.54	2
28	Gumi	3860.0	29	3218.2	36	2099.8	39	83.82	10
29	Frankfurt am Main	3732.4	30	4784.2	18	3812.6	18	-2.10	-12
30	Shanghai	3721.8	31	1553.2	65	441.2	169	743.56	138
31	Paris	3610.2	32	3711.4	24	3070.0	24	17.59	-8
32	Dallas	3532.6	33	3291.2	33	3047.0	25	15.93	-8
33	Suncheon	3511.4	34	2162.4	47	1198.8	74	192.90	40
34	Austin	3475.4	35	3666.4	26	2614.6	32	32.92	-3
35	Beijing	3355.4	36	850.0	125	314.6	221	966.56	185
36	Mannheim-Ludwigshafen	3332.0	37	3967.6	22	3085.0	23	8.00	-14
37	Portland	3306.4	38	2643.2	44	2171.8	38	52.24	0
38	Ruhr	3293.2	39	3556.8	27	2715.4	29	21.27	-10
39	Minneapolis	3266.6	40	3457.2	29	3238.6	21	0.86	-19
40	Dusseldorf	3226.8	41	3523.2	28	2771.2	27	16.44	-14
41	New Haven	3053.4	42	3439.6	30	3643.0	20	-16.18	-22
42	Wake	2946.0	43	3103.8	39	1973.2	42	49.30	-1
43	Atlanta	2777.8	44	2882.6	41	2322.6	36	19.59	-8
44	Cologne	2772.0	45	3143.2	38	2666.4	31	3.96	-14
45	Heidelberg	2704.0	46	3271.6	34	2555.0	33	5.83	-13
46	Taichung	2697.4	47	1703.2	61	880.2	99	206.45	52
47	Mito	2674.0	48	3669.4	25	2730.4	28	-2.06	-20
48	Nuremberg	2499.2	49	2916.2	40	1571.4	57	59.04	8
49	Washtenaw	2492.4	50	2187.8	46	1644.4	55	51.56	5

Table 7: Ranking of cities based on domestic linkages. The domestic linkage reported is the average weighted domestic linkage in the given period.

Trends in international linkages

Figure 9 illustrates the average number of international linkages and average patent-weighted international linkages to global and non-global cities. The average number of international linkages to non-global cities exceeds the average number of international linkages to global cities. However, the average weighted international linkage to global cities exceeds the average weighted international linkages to non-global cities. Specifically, the average number of weighted international linkages to global cities have increased by 73.5% in comparison to the average number of weighted international linkages to non-global cities which has only increased by

18.9%. After the decline in 2008 (corresponding to the same period as the financial crisis), the average weighted international linkages to global and non-global cities have increased by 49.3% and 30.9% respectively.

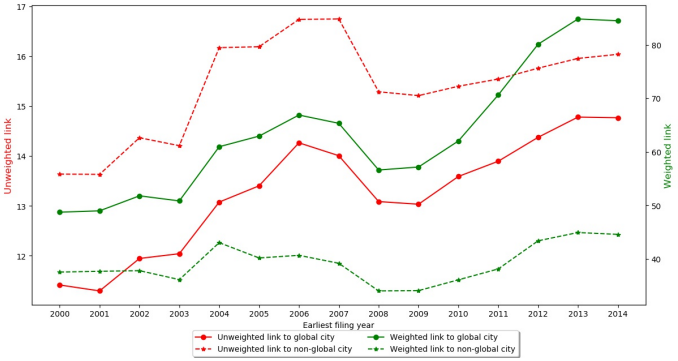


Figure 9: International linkages: The left y-axis is the average number of links, while the right y-axis is the patent-weighted linkages. The solid lines are trends of international linkages to global cities, while the dotted lines are trends of international linkages to non-global cities.

Comparing the international connectedness pre and post-2008, average weighted international linkages to global cities increased from 35.4% to 49.3% (representing +13.9 percentage point) and average weighted international linkages to non-global cities increased from 4.5% to 18.9% (representing +18.4 percentage point).

	Cities	Intl. global city linkages (2010-2014)	Ranking (2010-2014)	Intl. global city linkages (2005-2009)	Ranking (2005-2009)	Intl. global city linkages (2000-2004)	Ranking (2000-2004)	Intl. global city linkages (% change)	Change in ranking
0	San Francisco (Greater)	4268.0	1	2806.4	1	1853.2	2	130.30	1
1	Taipei	3179.0	2	1859.8	3	650.0	15	389.07	13
2	New York (Greater)	2717.0	3	2124.8	2	1819.8	3	49.30	0
3	Shanghai	2664.2	4	1355.6	6	429.4	25	520.44	21
4	Shenzen	2281.8	5	1700.8	4	360.0	28	533.83	23
5	Tokyo	2173.0	6	1699.6	5	2000.8	1	8.60	-5
6	Sedgwick	1696.0	7	857.6	14	1804.2	4	-5.99	-3
7	Bangalore	1639.6	8	761.6	18	266.4	45	515.46	37
8	Boston	1504.6	9	1008.8	11	879.2	8	71.13	-1
9	TelAviv	1404.0	10	779.0	16	518.8	18	170.62	8
10	Seoul	1315.0	11	1183.6	8	479.0	21	174.53	10
11	Worcester	1232.0	12	1274.0	7	1116.2	5	10.37	-7
12	London	1219.6	13	1087.0	9	1015.4	6	20.11	-7
13	Los Angeles (Greater)	1216.6	14	934.2	13	846.4	10	43.73	-4
14	Beijing	1172.6	15	615.0	23	272.8	44	329.83	29
15	Paris	1066.4	16	1061.6	10	801.0	12	33.13	-4
16	Philadelphia (Greater)	1065.2	17	815.4	15	730.8	13	45.75	-4
17	Toronto	1045.8	18	946.4	12	833.2	11	25.51	-7
18	San Diego	992.0	19	777.2	17	505.0	20	96.43	1
19	Chicago	906.4	20	710.2	20	466.6	22	94.25	2
20	Seattle	821.0	21	616.4	22	354.2	29	131.78	8
21	Toyota	746.2	22	409.2	34	892.6	7	-16.40	-15
22	Eindhoven	739.2	23	514.0	28	643.4	16	14.88	-7
23	Munich	728.6	24	710.2	19	557.6	17	30.66	-7
24	Brussels	728.0	25	651.4	21	434.2	23	67.66	-2
25	Houston	712.6	26	540.2	27	353.4	30	101.64	4
26	Washington (Greater)	662.8	27	557.8	26	513.0	19	29.20	-8
27	Portland	661.2	28	362.6	42	295.0	38	124.13	10
28	Singapore	611.8	29	361.8	43	301.8	37	102.71	8
29	Cambridge	598.4	30	593.0	25	686.6	14	-12.84	-16
30	Dallas	569.8	31	425.6	33	321.8	34	77.06	3
31	Montreal	567.8	32	601.0	24	425.4	26	33.47	-6
32	Zurich	564.2	33	452.4	30	294.2	39	91.77	6
33	Higashiosaka	562.6	34	469.6	29	863.8	9	-34.86	-25
34	Stockholm	538.8	35	374.2	39	221.6	57	143.14	22
35	Tulsa	507.0	36	271.2	58	376.4	27	34.69	-9
36	Ottawa	465.4	37	368.6	40	256.2	46	81.65	9
37	Detroit (Greater)	441.2	38	406.6	35	433.4	24	1.79	-14
38	Minneapolis	426.6	39	337.8	44	254.0	47	67.95	8
39	Austin	424.6	40	399.8	36	237.8	51	78.55	11
40	Stuttgart	421.2	41	451.4	31	304.8	36	38.18	-5
41	Vancouver	414.4	42	305.6	50	274.8	43	50.80	1
42	Frankfurt am Main	390.4	43	364.6	41	330.6	32	18.08	-11
43	Kitchener	390.0	44	331.0	46	176.6	66	120.83	22
44	Dusseldorf	373.4	45	299.4	51	246.8	49	51.29	4
45	Basel	368.0	46	450.0	32	286.8	41	28.31	-5
46	Mannheim-Ludwigshafen	366.0	47	379.6	37	235.0	53	55.74	6
47	Albany	338.8	48	273.4	57	139.0	84	143.74	36
48	Greater Sydney	332.6	49	314.2	48	331.0	31	0.48	-18
49	Helsinki	331.6	50	283.8	54	214.6	58	54.52	8

Table 8: Ranking of cities based on international global city linkages. The international global city linkage reported is the average international global city linkage in the given period.

Table 8 indicates the top 50 cities in terms of average weighted international global city linkages in the period 2010-2014. We observe that the top cities in the rankings are San Francisco, Taipei, New York, Shanghai, and Shenzhen. Also, we observe less fluctuation in the ranking in comparison to the weighted domestic linkage ranking. We have that Bangalore, Albany, and Beijing moved upward the most by 37, 36, and 29 points to the 8th, 48th and 15th position respectively and increased their weighted international global city linkages by 515.5%, 143.7%, and 329.8% respectively. However, cities like Higashiosaka, Sydney and Toyota experienced the most decline in ranking, with 25, 18 and 15 point change in ranking respectively. Although, Higashiosaka and Toyota had a decrease in average weighted international global city linkages by 34.8% and 16.4% respectively, we see that Sydney had a 0.48% increase in average weighted

international global city linkages.

At the level of the country, we see that the top countries with the most average weighted international global city linkages are Taiwan and Israel with a 352.3% and 170.6% increase in linkages between the period 2000-2004 and 2010-2014. While Singapore retained its third position across all periods, we see a strong decline in Japan's position by 12 points and a rise in the position of Malaysia by 12 points. At the level of the continent, we see that Asia has displaced North America from the top position by increasing its average linkages to international global cities by 119.4%.

Table 9 indicates the top 50 cities in terms of average weighted international non-global city linkages in the period 2010-2014. We have that the top cities are Tokyo, Boston, New York, London, and San Francisco. Bangalore, Beijing, and Shenzen moved up the ranking the most with a 426.8%, 278.8%, and 185.2% increase in average weighted international non-global city linkage. While Cambridge, Vancouver, and Sedgwick decline the most with a decrease of 39%, 22.5%, and 22.9% in average weighted international non-global city linkage. At the level of the country, we see that the top countries with the most average weighted international linkages to non-global cities are Israel, Singapore, and Taiwan with a stable ranking. These countries have increased their linkage by 155.5%, 81.3%, and 121.3% respectively between the period 2000-2004 and 2010-2014. We see a stronger rise in position by Luxembourg, China, and Malaysia, and a decline in position by Russia, Japan, and the United States. At the level of the continent, we have that Asia leads with an average of 67.4 weighted international linkages to non-global cities, followed by North America, Oceania, Europe, Africa, and South America with average weighted linkages of 63.3, 29.3, 28.5, 11.7, and 7.6 respectively.

	Cities	Intl. non-global city linkages (2010-2014)	Ranking (2010-2014)	Intl. non-global city linkages (2005-2009)	Ranking (2005-2009)	Intl. non-global city linkages (2000-2004)	Ranking (2000-2004)	Intl. non-global city linkages (% change)	Change in ranking
0	Tokyo	2484.0	1	1607.4	1	2585.6	1	-3.92	0
1	Boston	1530.4	2	1317.4	3	1519.0	4	0.75	2
2	New York (Greater)	1477.2	3	1119.8	4	1783.4	2	-17.16	-1
3	London	1412.6	4	1572.4	2	1585.2	3	-10.88	-1
4	San Francisco (Greater)	1353.8	5	1101.8	5	1053.2	5	28.54	0
5	Shanghai	1149.2	6	562.8	9	225.0	36	410.75	30
6	Toronto	835.0	7	500.6	13	981.6	6	-14.93	-1
7	Brussels	664.0	8	690.2	6	452.8	11	46.64	3
8	Paris	591.4	9	603.6	7	582.2	9	1.58	0
9	Bangalore	581.6	10	287.8	28	110.4	77	426.81	67
10	Los Angeles (Greater)	570.2	11	513.8	11	582.6	8	-2.12	-3
11	Philadelphia (Greater)	542.0	12	512.2	12	621.4	7	-12.77	-5
12	Seoul	536.2	13	560.4	10	509.4	10	5.26	-3
13	TelAviv	509.6	14	282.2	30	199.4	41	155.56	27
14	Chicago	495.0	15	445.2	15	341.8	18	44.82	3
15	San Diego	430.4	16	305.4	25	304.0	25	41.57	9
16	Houston	424.0	17	388.6	19	336.2	20	26.11	3
17	Munich	394.8	18	465.4	14	323.4	24	22.07	6
18	Eindhoven	385.6	19	400.2	17	342.0	17	12.74	-2
19	Washington (Greater)	363.2	20	359.0	21	450.2	12	-19.32	-8
20	Detroit (Greater)	361.0	21	404.8	16	328.2	22	9.99	1
21	Minneapolis	357.8	22	298.8	26	338.0	19	5.85	-3
22	Zurich	352.4	23	364.4	20	259.8	30	35.64	7
23	Beijing	341.0	24	170.4	54	90.0	84	278.88	60
24	Basel	338.8	25	572.0	8	392.2	13	-13.61	-12
25	Taipei	335.8	26	205.2	39	168.0	48	99.88	22
26	Worcester	335.4	27	391.0	18	293.2	27	14.39	0
27	Frankfurt am Main	329.6	28	320.0	23	255.2	33	29.15	5
28	Dallas	321.8	29	194.2	41	298.4	26	7.84	-3
29	Seattle	320.8	30	281.0	31	245.6	34	30.61	4
30	Mannheim-Ludwigshafen	294.4	31	286.2	29	210.4	39	39.92	8
31	Vancouver	287.8	32	249.2	32	371.8	14	-22.59	-18
32	Hamburg	277.2	33	222.2	35	138.4	61	100.28	28
33	Stuttgart	264.4	34	337.6	22	212.8	38	24.24	4
34	Higashiosaka	261.0	35	213.8	37	331.0	21	-21.14	-14
35	Singapore	258.2	36	171.4	52	142.4	59	81.32	23
36	Barcelona	256.2	37	173.4	48	229.0	35	11.87	-2
37	Sedgwick	251.6	38	105.8	85	326.4	23	-22.91	-15
38	Shenzen	243.0	39	172.6	50	85.2	89	185.21	50
39	Cincinnati	242.6	40	187.6	45	255.6	32	-5.08	-8
40	Tulsa	242.4	41	117.8	80	274.2	28	-11.59	-13
41	Portland	222.8	42	137.2	69	123.6	68	80.25	26
42	Aachen	220.6	43	292.0	27	195.2	42	13.01	-1
43	Copenhagen	216.0	44	192.6	43	183.0	46	18.03	2
44	Stockholm	215.0	45	223.6	34	150.6	55	42.76	10
45	Cambridge	211.8	46	223.6	33	347.4	16	-39.03	-30
46	Duesseldorf	209.8	47	173.0	49	154.8	52	35.52	5
47	Wake	206.8	48	151.0	62	99.8	79	107.21	31
48	Vienna	194.4	49	319.0	24	188.0	44	3.40	-5
49	Atlanta	192.8	50	187.0	46	128.4	66	50.15	16

Table 9: Ranking of cities based on international non-global city linkages. The international non-global city linkage reported is the average international non-global city linkage in the given period.

Changes in the share of domestic and international linkages

Finally, Table 10 examines the share of domestic linkage, international global city linkage, and international non-global city linkage. The share of domestic or international global city or international non-global city linkages in a city is the number of the city's domestic or international global city or international non-global city linkages divided by the city's total linkages respectively. We have that in period 2010-2014, North America on average had the highest share of domestic linkages with 59.3% of its total linkages being domestic, which is 4.7% short of the share it had in the period 2000-2004. While Europe and South America moved up the ranking by increasing their share of domestic linkages by 0.21% and 22.68% respectively, Asia and Oceania moved down the rank by decreasing their share of domestic linkages by 6.1% and 6.5% respectively. Africa's share of domestic linkage decreased by 28.7%.

	Continents	domestic linkages (2010-2014)	Ranking (2010-2014)	domestic linkages (2005-2009)	Ranking (2005-2009)	domestic linkages (2000-2004)	Ranking (2000-2004)	domestic linkages (% change)	Change in ranking
0	North America	0.59	1	0.61	1	0.62	1	-4.70	0
1	Europe	0.51	2	0.51	3	0.50	3	0.20	1
2	Asia	0.49	3	0.52	2	0.52	2	-6.05	-1
3	South America	0.39	4	0.32	4	0.32	5	22.67	1
4	Oceania	0.30	5	0.31	5	0.32	4	-6.49	-1
5	Africa	0.16	6	0.18	6	0.23	6	-28.65	0

(a) Share of domestic linkage

	Continents	Intl. global city linkages (2010-2014)	Ranking (2010-2014)	Intl. global city linkages (2005-2009)	Ranking (2005-2009)	Intl. global city linkages (2000-2004)	Ranking (2000-2004)	Intl. global city linkages (% change)	Change in ranking
0	Africa	0.51	1	0.45	1	0.40	2	28.65	1
1	Oceania	0.42	2	0.35	3	0.37	3	13.48	1
2	South America	0.35	3	0.38	2	0.41	1	-13.04	-2
3	Asia	0.29	4	0.27	4	0.27	4	5.77	0
4	Europe	0.26	5	0.25	5	0.25	5	4.31	0
5	North America	0.21	6	0.18	6	0.17	6	18.20	0

(b) Share of international global city linkages

	Continents	Intl. non-global city linkages (2010-2014)	Ranking (2010-2014)	Intl. non-global city linkages (2005-2009)	Ranking (2005-2009)	Intl. non-global city linkages (2000-2004)	Ranking (2000-2004)	Intl. non-global city linkages (% change)	Change in ranking
0	Africa	0.31	1	0.36	1	0.36	1	-13.31	0
1	Oceania	0.27	2	0.33	2	0.30	2	-9.74	0
2	South America	0.24	3	0.28	3	0.26	3	-7.29	0
3	Europe	0.22	4	0.23	4	0.23	4	-5.08	0
4	Asia	0.21	5	0.20	5	0.19	5	7.81	0
5	North America	0.19	6	0.19	6	0.19	6	-1.68	0

(c) Share of international non-global city linkages

Table 10: The average share of unweighted domestic and international linkages for continents.

Also, we find that Africa and Oceania increased their connectivity with international global cities by 28.7% and 13.5% respectively, while South America decreased its connectivity with international global cities by 13.04%. Asia cities on average do have a higher share of their linkages to international global cities compared to North America (29.5% versus 21.2%). Also, we see a stable ranking in terms of share of linkages to international non-global cities, with African cities at the top having an average 31.7% of its linkages with international non-global cities, while North America has the least average share of 19.5%. In general, examining the weighted and unweighted linkages, we see that developed economies e.g., cities in Africa rely more on international linkages for innovation process, specifically, they have increasingly looked at forming ties with international global cities. While Asia has pursued the formation of ties with global and non-global cities (4.3% and 7.8% increase respectively), North American cities have mostly pursued the formation of ties with international global cities (18.2% increase).

1.4 Discussion

In this chapter we describe each step of the data generating process leading to the final inventor and applicant datasets that will be used in Chapters 2-4 of the thesis. To effectively use this dataset, the data can be stored in a relational database, which can be merged with tables in PATSTAT using the application identifier column and can also be merged with other regional databases. The data work is geared at improving the quality of address information specifically in PATSTAT, this is because PATSTAT is the major database often used in analyzing patenting activities, its attractiveness stems from it being the largest patent database with patents that are filed in over 90 different patent offices in developed and developing economies. This implies that issues of data quality from the different patent offices are transferred to the PATSTAT database, some of these issues include ambiguities in inventors and applicant names (Morrison, Riccaboni, and Pammolli, 2017a), missing addresses, country and IPC information in patents (De Rassenfosse, Kozak, and Seliger, 2019; De Rassenfosse and Seliger, 2021). Previous efforts addressing this issue of missing address in PATSTAT rely on improving address coverage of priority patents by leveraging patent information available in the national patent offices and exploring the availability of addresses in subsequent filings (De Rassenfosse et al., 2013; De Rassenfosse, Kozak, and Seliger, 2019). The address retrieved in this case is simply assigned to patents with no linkage between the inventors or applicant and the retrieved address. While this is useful in studying innovation outcomes and spillovers at the level of cities, regions, and countries, it limits our ability to track emerging and star inventors at the sub-regional level. Although recent work by De Rassenfosse and Seliger, 2021 addressing the issue of missing country information allow such analysis at the level of countries, the database introduced in this Chapter provides this even at the sub-regional level. More importantly, the major contribution of the database to the literature is that the retrieved addresses are geolocated to functional urban areas, which is a harmonized definition of an urban area that captures the economic reach of urban areas. This is a major departure from previous work

in which addresses are geolocated to cities based on administrative or political boundary definitions of countries. Such non-harmonized definitions can impact analyses of innovative outcomes of cities in different countries.

The dataset provided can potentially be used in studying the productivity of inventors in different urban areas, mobility of inventors across sectors (private-public) within an urban area and how that impact productivity of these sectors, the concentration of star of emerging inventors in an urban area. This dataset is not full proof, as seen in the validation process that the validity of the data generating process is about 84%, and there are discrepancies in the country information for inventors and applicants in 2.68% and 3.5% of the geocoded patents respectively. Some of the limitations of the data generating process is the fact that sources 4-6 assigns address to inventors using address information of the applicant, relying on the assumption that inventors locate in the same geographic areas as applicants. This is not necessarily the case when the headquarter of a company file for patents on behalf of their subsidiaries where the inventors are located. The imputation of missing address based on sources 4-6 will impact the preliminary analysis on evolution of linkages done in Chapter 1 and the analysis done in Chapters 2-4 as more patents will be allocated to cities that are hubs for headquarters. In the dataset we have included an indicator called "sources_4_6" that identifies observations whose coordinates potentially come from sources 4-6 of the imputation algorithm. These observations are coded 1 if the inventor name is not the same as the applicant name (and the applicant is located in multiple FUAs) but the coordinates are the same, and 0 otherwise. We observe that at most 11% of the DOCDB patents have at least an inventor whose coordinate potentially comes from sources 4-6. Another limitation is the issue of ambiguities in inventor and applicant names which will impact the address retrieval from subsequent filings and not the linking of PATSTAT to the external patent database (since the external patents used here share similar sources as PATSTAT). We have applied similarity measure for string comparison using high-thresholds which at best can capture simple mistakes in the spelling of names, and we have also used

the harmonized names (HAN) in this process when this is available. However, future work can focus on best disambiguation methodologies relying on machine learning techniques (Li, Wei, and Wang, 2015; Cuxac and Bonvallot, 2013) that can improve the quality of the address retrieval methodology. In using this dataset, researchers should note that while some inventors do have coordinate information, we were only able to assign coordinates to 1227 functional urban areas, implying that some coordinates are yet to be assigned to a functional urban area. Further work can be done in creating FUA shapefiles for more urban areas using the steps outlined in this paper and then mapping the yet-to-be-assigned coordinates to these shapefiles.

To demonstrate the use of this database we provide a preliminary analysis of the evolution of three different kinds of linkages (i.e., domestic linkage, international linkage to global cities, and international linkages non-global cities) at the level of FUAs, country, and continent. The analysis contributes to discussions in the global city literature where the inter-dependence of cities is well documented (Lorenzen, Mudambi, and Schotter, 2020). It's been observed that non-global cities often hinge on domestic linkages to global cities to foster growth as such linkages require low resources. Crossing borders to create collaborative ties are often observed in large cities that serve as innovation hubs for multinationals responsible for the creation of cross-border ties (Lorenzen, Mudambi, and Schotter, 2020). The creation of cross-border ties be aware of skills and opportunities abroad, and allow firms to attract capacities to large cities (Verginer and Riccaboni, 2020) that can function in highly-skilled tasks, allowing for low-skilled tasks being moved to non-global cities. In the analysis, we define global cities to be cities classified as alpha-beta cities in Beaverstock, Smith, and Taylor, 1999. Our findings suggest a decline in the number of domestic linkages (weighted and unweighted), and an overall increase in the two types of international linkages (weighted and unweighted). Generally, we see a decline in the share of domestic linkages for all continents except Europe and South America. We observe that cities in developing economies (i.e., Africa, Oceania) rely on international linkages (especially linkage to global cities) to innovate. Also,

we find that cities in Asia pursue the formation of international linkages with both global and non-global cities, while cities in North America pursue the formation of international linkages mostly with global cities. The preliminary analysis done here confirms that there is an increased focus on the formation of international ties both extensively and intensively, as opposed to the formation of domestic linkages. Although we do not provide any evidence of the consequences of this trend on social or developmental outcomes in urban areas. However, increase domestic disconnectedness and focus on spawning international ties has been seen to energize populist backlash against activities of multinational corporations (Lorenzen, Mudambi, and Schotter, 2020), and such backlash can hurt the productivity of urban areas. As Lorenzen, Mudambi, and Schotter, 2020 argue in their work, global orchestration of resources can contribute to domestic disconnectedness which can lead to a populist backlash. To diffuse the resulting backlash, cities or multinational corporations can engage in creating an entrepreneurial eco-system where intra-city linkages between start-ups and multinationals can be formed (e.g., the case of Bangalore (Bala Subrahmanya, 2017)) or inter-city linkages within the boundary of countries can be formed.

Chapter 2

Economic complexity and Forecasting competitiveness of Global cities using Machine learning approach

[1] Since Porter published the "Competitive Advantage of Nations", there has been an increasing interest in what it means for cities to be competitive, given that, unlike firms and individuals, cities do not act. Nevertheless, given the increased intercity competition around the world, an extensive body of works shows that cities can be important for cross-fertilization of ideas. Concepts such as the ability of cities to learn (Asheim, 1996), presence of industrial district, and innovative milieus (Camagni, 1991; Becattini, 2017), shows the importance of externalities in maintaining competitiveness that extends the boundaries of firms but operates within the boundaries of territories (Boschma, 2004). Following the Ricardian comparative advantage, typically a city is seen to be competitive if a city is

[1] This is a joint work between S. Edet and M. Riccaboni.

producing more than its fair share in a given sector, industry, technology, or scientific domain. Unraveling the linkage between the competitiveness of cities and resource allocation and capabilities (Zhang and Liu, 2020) would be crucial to understanding why some cities are more competitive, productive, or resilient than others. Economic analysis attempting to uncover this intangible element (i.e., capabilities) pursue an economic framework that provides explanations for the heterogeneity of capabilities and competitiveness, and prescriptions for what economic pathways cities should adopt (Straccamore and Zaccaria, 2021). On the contrary, the economic complexity and machine learning literature provide useful insight on how to measure and predict the future capabilities of cities or the capabilities required to become competitive in specific sector, scientific or technology domain by leveraging the spatial pattern of production, scientific or patenting activities (Tacchella et al., 2021; Hidalgo, 2021; Balland and Rigby, 2022). The core principle of the methodology in the economic complexity literature is that complex cities accumulate know-how that leads to the diversification of patenting activities and increase the capabilities of inventors in the city to innovate in complex technology domain that is ubiquitous (Balland and Rigby, 2022). Generally, technologies differ in the knowledge and tools required to produce them, hence, since most complex technologies are highly leveraged, diverse cities endowed with economic agents with different know-hows can produce them.

Given the progress made in the development of machine learning models and endogenous growth theories, the economic complexity model was introduced (Hidalgo, 2021), and unlike traditional approaches, economic complexity combines economic input and output without aggregating output as GDP does, nor does it make an assumption on the nature of input such as capital and labor (Hidalgo, 2021; Balland and Rigby, 2022). The method relies on the fine-grained representation of economic activities and learns how they combinatorially explain the economic activities taking place in a location. This is done based on dimensionality reduction techniques applied to activities such as exporting activities, patenting activities, and scientific activities observed in different locations. This technique of dimensionality reduction relies on matrix factorization which

is a common theme in machine learning (Hidalgo, 2021; Balland and Rigby, 2022). It provides a succinct way to summarize the geography of economic activities of different locations and can be used as predictor for economic development potential of locations. The economic complexity methodology estimates the combined presence of the factors necessary for economic growth without assumptions about what these factors are or their nature (Hidalgo, 2021).

The economic complexity methodology has grown in application due to growing interest in industrial policy and the need for regional strategies for technological upgrading (Hidalgo, 2021; Rodrik, 2019). The methodology provides a quantitative perspective for designing focused industrial policies - this quantitative measure is seen to be integrated into industrial policies such as China's special economic zone (Zheng et al., 2016; Kahn et al., 2018), Europe's smart specialization strategy (Balland et al., 2018), Mexico's Smart Diversification strategy, and similar policies. The methodology has also been applied to the study of the economic structure of countries e.g., United Kingdom (Mealy and Coyle, 2019), Italy (Basile, Cicerone, and Iapadre, 2019), Australia (Reynolds et al., 2018), United States (Fritz and Manduca, 2019). Also, economic complexity has been used to study the consequences of certain outcomes such as output volatility (Güneri and Yalta, 2020), greenhouse gas emission (Neagu and Teodoru, 2019), and gender inequality (Ben Saâd and Assoumou-Ella, 2019). Other works have also examined factors such as taxation (Lapatinas, Kyriakou, and Garas, 2019), and intellectual property right (Sweet and Eterovic Maggio, 2015) impacting the growth of complexity of countries.

While the economic complexity index provides insight into the current capabilities of a location, and the required capability for economic activity, it does not forecast future capabilities. Some literature such as Zaccaria et al., 2018 and Pugliese et al., 2019 provide a framework that relies on the structure of relatedness between activities to estimate the density of a country's capability around economic activity. Recently, Tacchella et al., 2021 pointed out a shortcoming of these methodologies which rely on relatedness structure between activities. They observed that "location-

activities bipartite networks do have a strong nested structure" (Tacchella et al., 2021) which could immediately lead to communities in which case the relatedness signal is of second-order and not adequate to capture higher-order non-linear relationships. To address this shortcoming, a machine learning approach adapted for learning non-linearity can be useful.

Appropriately forecasting the competitiveness and future capabilities of cities (i.e., economic complexity index) using better predictive models can improve the specificity of strategic innovation policies of cities. The contribution of this chapter is to (i) provide insight on current capabilities of cities using state of the art economic complexity methodology introduced by Sciarra et al., 2020 (ii) provide a better predictive model for the future competitiveness of cities in all technologies, and specifically in technologies where cities have remained consistently non-competitive (iii) provide a forecast of future capabilities (i.e., generalized economic complexity index) using the predicted competitiveness structure of cities.

This chapter is organized as follows. Section 2.1 provides an overview of the literature on the concept of economic complexity and forecasting the competitiveness of cities. Section 2.2 describes the data. Section 2.3 introduces the generalized economic complexity model, the machine learning models (i.e., Random Forest, eXtreme Gradient boosting, Support Vector Machine, Neural network model), the Time-delayed co-occurrence model (which is the benchmark model for forecast), and the evaluation metrics. Section 2.4 describes the results, and Section 2.5 provides a concluding remark.

2.1 Literature Overview

This section gives an overview of existing literature on the concept and measurement of economic complexity and forecasting competitiveness of cities.

2.1.1 Concept of Economic Complexity

In the economic and geography literature, economic complexity is often discussed with reference to infrastructure, product export, technological and scientific capabilities that allow a city, region, or country to identify its current capabilities, and the capabilities required to upgrade into new technologies and sectors, to increase its growth rate (Mewes and Broekel, 2020; Hidalgo, 2021). Although precisely measuring complexity is challenging, there have been scientific efforts towards inferring the complexity of countries using trade, publication, and patent data. This methodology assumes that countries with positive comparative advantage in diverse products, or technologies that are spatially scarce do have higher capabilities (Hausmann, Hwang, and Rodrik, 2007; Hausmann et al., 2014).

The notion of economic complexity is often described using the game of Scrabble (Hausmann et al., 2014). In this analogy, players are countries, cities or regions. Words are like products, patents, or publications, and letters are like capabilities (e.g., HS6 products, IPCs classes), or modules of embedded knowledge. Each player is assumed to have several copies of letters which can be used in different ways to produce words. The economic complexity approach is aimed at uncovering the level of capabilities (i.e., how many letters) a player has by examining (i) how many words the player can form (ii) the kind of players that can form such words as well. In this sense, players with more letters can form more words, hence placing them at the forefront of how diverse their capabilities are. This measure of diversity is the first indicator of the capabilities of the player (Hausmann et al., 2014). Although, this measure in itself is a crude estimate of the complexity of the player if the composition of the words is not examined. In examining, the creation of words, longer words will be less common, since they can only be created by players with the prerequisite letters. Therefore words that require fewer letters will be more ubiquitous and vice-versa. An improved measure of complexity of the player is given by its diversity, the ubiquity of the words it makes, and the diversity of other countries that make these words (Hausmann

et al., 2014).

The analogy above forms the basis of the commonly used methodologies for measuring economic complexity, namely, "Method of reflection" (MR) and the "Fitness and complexity" (FC) algorithm. These methods are both conceptually and mathematically different. The FC and MR schemes are both nested equations connecting the complexity of words with the complexity of players. However, the key difference between both methods is that while the MR scheme argues the existence of a linear relationship between the complexity of words formed and the complexity of players, the FC scheme puts forward a non-linear relationship between both quantities. Particularly, the FC method emphasizes the fact that the complexity of words is downgraded when it is created by less competitive players, an effect that can best be captured with a non-linear relationship (Tacchella et al., 2012; Sciarra et al., 2020). Both methodologies have contributed to the interdisciplinary study of economics and complex systems. However, they have been shown to produce significantly diverging outcomes (Sciarra et al., 2020).

While the FC methodology is seen as an improvement of the MR methodology, it is still plagued by several technical and conceptual issues which have been identified through simulated networks of scale-free model of preferential attachment, as well as networks constructed from real-world datasets of trade and patent activities (Morrison et al., 2017). It has been observed that results from the FC scheme can change when there are small changes to the network, and the algorithm often overestimates the capabilities required for niche outputs having a small number of players (Morrison, Riccaboni, and Pammolli, 2017b). The instability in the results applies to weighted and unweighted networks (Pugliese, Zaccaria, and Pietronero, 2016). However, the more granular the level of estimation, the less severe the estimation of the fitness of players, as fitter players are likely to engage in producing or patenting in niche products or technologies (Morrison et al., 2017).

By recasting the MR and FC methodologies into a multidimensional framework, Sciarra et al., 2020 reconcile the MR and FC approaches, allowing

to recover and integrate the merits of both methods. The resulting measure is referred to as the Generalized Economic Complexity (GENEPY) index. The idea of GENEPY is based on allowing an interpretation of the eigenvectors of the proximity matrix of players as the multidimensional eigenvector centrality of the players in the network. In which case the eigenvectors can be combined into a unique metric where they are obtained using a least-square estimation (Sciarra et al., 2020). The information from the GENEPY index is best understood considering the meaning of its components (i.e., the first and second eigenvectors). The first eigenvector represents the eigencentrality of the players based on the proximity matrix of players due to the similarity of letters they have. While the second eigenvector cluster players according to the similarities of letters they have (Sciarra et al., 2020). Hence, GENEPY identifies the set of capabilities a player possesses and has in common with other players. In other words, complex players are found in the vicinity of other complex players, while less complex players are in the periphery. GENEPY has been applied to trade data to characterize the set of capabilities countries need to export in different products (Sciarra et al., 2020).

2.1.2 Forecasting competitiveness of cities

While economic complexity captures the current capabilities of cities, it is interesting to forecast the competitiveness of a city in a given technology in years to come. Typically, the competitiveness of a city in a given technology is captured in a binary matrix (i.e., city-technology matrix) based on the revealed technological advantage of the city in the technological area. This quantitative metric based on the work of Ricard's comparative advantage indicates that the overall capability of a city should be related to other cities to determine if the city is producing more or less than its "fair share." The quantity evaluates the cities number of patents in a technology relative to the total market share. The measure is stable to small fluctuations, and significant changes only occur when the relative share varies appreciably. In most literature, if the revealed technological advantage $RTA_{ij} \in [0, \infty)$ of the city i in a technology j is greater than

1, the city is considered to be competitively patenting in the technology. This is the case when the relative patenting of the city is greater than the whole market share in that technology. This means that the signal of the capabilities of the city exceeds the baseline level and it is deemed significant enough to consider the city competitive. The binary matrix representation of the revealed technology advantage matrix makes it easier to study the time evolution of the competitiveness matrix, and represent the forecasting task as a link prediction or classification problem that can be addressed using machine learning approaches.

There are different approaches in network science that tries to address the forecasting problem based on the underlying assumption of correlation of technologies. These correlations between technologies can be estimated following e.g., the product space approach (Hidalgo and Hausmann, 2009) or taxonomy network approach (Zaccaria et al., 2014). The idea is that the correlation between the ecosystem of technologies provides information on what technology serves as a gateway to patent in other technologies. Combining the correlation matrix of technologies alongside the current competitive structure of cities can be used to estimate the density of cities' competitive structure around different technological areas. The lower the density of a city's capabilities around a technology, the less likely the city will upgrade to or activate the technology. This implies that the technologies the city is currently competitive in are distant from the given technological area. We can evaluate the accuracy of the estimated densities against what we observe to be the competitive structure using different binary score measures. The criticism of this method to forecast competitiveness is that the correlation structure of technologies does not account for the high-order structure of correlations between technologies. For example, patenting an *electric car* might require capacity in *electronics* and *mechanics*, which implies the existence of a correlation between three IPCs (jointly). Evaluating this structure with pairwise correlations means that the signal would be the sum of correlations of *electric car - electronics* and *electric car - mechanics*. Hence, the correlation of either pair could lead to acknowledging the presence of a signal that in reality does not exist. Specifically, as observed in the paper "Relatedness in the era of machine

learning" by Tacchella et al., 2021 - "the binary RCA matrix does have very strong nested structure as opposed to a block-diagonal structure that would immediately lead to a definition of sector-communities, in these networks, the relatedness signal is of second order with respect to the drive towards diversification that generates the nested structure." This implies that describing the competitiveness development path of a city as the sum of binary relationships is an oversimplification, and existing methodologies can be improved upon considering higher-order interactions through more complex machine learning models that learn successfully through boosting, in which the models are learned in order, such that each new model is trained to minimize the residuals of the preceding ones, and through data augmentation by averaging the learning outcome on randomized samples (Tacchella et al., 2021). These machine learning models are algorithms of statistical inference and they are used to address tasks classified as either supervised or unsupervised. A task is considered supervised if we have a labeled output $y_i \in Y$ corresponding to an input $\mathbf{x_i} \in \mathbf{X}$, the labeled output is the ground truth against which the ability of the model to learn is evaluated. On the other hand, an unsupervised task is done when there is no labeled output - an example of such task could be clustering of observations based on similarity of features.

Specifically, for our task which involves forecasting the competitiveness of cities in technology, we are faced with a supervised task where we will be predicting binary attribute $y_i \in \{0, 1\}$ of the competitiveness matrix $\mathbf{M}_{ij}^{t+5}$ in 5 years based on the attributes $\mathbf{RTA}_{ij}^{t}$ - revealed technology advantage at time t. A 5-years forecast horizon is chosen, since most investment needed to significantly impact the competitiveness structure of cities are usually long-term. As seen in Figure 10b, $P(0 \longrightarrow 1)$ and $P(1 \longrightarrow 1)$ shows slight changes in competitiveness structure in 5-years, this eliminates the need to make a forecast for a lower number of years, and given the time coverage of the geolocated patent database in Chapter 1, a reasonable period to forecast would be 5 years, as a forecast beyond 5-years would reduce the sample size for training. The inventor dataset used is pre-processed such that the models applied will not be able to

exploit the auto-correlation of the competitive structure for prediction. The choice of inventor dataset as opposed to applicant is to focus on human capital in patent production as some firms can choose to take advantage of opportunities in emerging markets by seeking for intellectual coverage through their subsidiaries located in the emerging market, while the knowledge is created in their headquarters or abroad. The machine learning models applied include Random Forest, XGBoost, Support Vector Machine and Artificial neural network. The performance of these models is benchmarked with the performance from the time-dependent network forecasting model proposed by Zaccaria et al., 2014 that exploits auto-correlation of the competitive structure of patenting activities of cities. We provide the performance of the models in (i) predicting the full competitiveness matrix (i.e., $\mathbf{M}_{ij}^{t+5}$ matrix) (ii) predicting submatrix of $\mathbf{M}_{ij}^{t+5}$ corresponding to a city activating a technology condition on the fact that cities' revealed technological advantage in the technology has consistently fallen below 0.25. The performance of the model is evaluated using metrics that account for the imbalance in the target variable, as it is often the case that $y_i = 0$ constitutes the majority class label. Finally, since the binarized revealed technological advantage is the basis for computing the generalized economic complexity index, the forecast of the full $\mathbf{M}_{ij}^{t+5}$ matrix from the best performing machine learning model is used to compute the future generalized economic complexity index of cities.

2.2 Data

The source of data for computing the generalized economic complexity index and predicting future competitiveness and capabilities in different technological areas is the patent dataset developed in Chapter 1 which is based on the Patent Statistical (PATSTAT) database - the most extensive database that captures patent activities from over 90 patent offices located in different countries. The geocoding effort done in Chapter 1 provides a patent database with improved address information granular enough to capture patenting activities in urban areas, and an urban delineation process based on OECD, 2012's delineation methodology that addresses

the issue of differences in national boundaries, by providing a harmonized definition of cities. The analysis is done on 150 global cities selected from

Continents	Number of cities	Time period	Number of patents (thousand)	Average employment (thousand)	Average net migration (thousand)
Africa	6	2000-2004	1.62	1431.78	2.52
		2005-2009	1.76	1632.11	30.78
		2010-2014	2.82	1759.73	38.15
Asia	45	2000-2004	1320.61	3145.15	97.21
		2005-2009	1400.51	3722.14	125.21
		2010-2014	1362.48	4358.46	64.87
Europe	48	2000-2004	230.21	1210.84	10.46
		2005-2009	268.17	1284.38	11.72
		2010-2014	269.01	1310.01	8.36
North America	34	2000-2004	476.99	2069.75	1.83
		2005-2009	504.49	2136.25	2.49
		2010-2014	531.18	2171.25	4.52
Oceania	7	2000-2004	11.53	956.70	11.96
		2005-2009	11.77	1083.13	22.92
		2010-2014	12.58	1186.06	24.11
South America	10	2000-2004	2.13	3121.41	2.86
		2005-2009	3.23	3561.11	5.70
		2010-2014	5.12	4011.93	11.82

Table 11: Descriptive of dataset

Beaverstock, Smith, and Taylor, 1999's Globalization and World Cities (GaWC) research project. These cities are either alpha cities (i.e., they link major economic cities into the world economy) or beta cities (i.e., they connect moderate economic cities to the global economy). These cities are located in Africa (6), Asia (45), Europe (48), North America (34), Oceania (7), and South America (10). Also, the analysis is done at the level of 637 4-digit "international patent classification" (IPC) - a system for classifying patents produced by inventors located in these cities, according to the areas of technology the invention pertains to. We identify 6,242,439 patents filed by inventors in these cities across these 637 technological areas between 2000-2014. The analysis done in this chapter (and subsequent chapters) selects patents at the level of the DOCDB patent families (i.e., we do not distinguish between priority or non-priority patent) thereby avoiding inflation of the actual patent size of a city. For each year we construct the revealed technological advantage (RTA_{ij}) of a city i in an IPC j which gives a measure of the share of patenting activity

of a city in an IPC when compared to the global average patenting done in the given IPC. The RTA matrix is used in constructing the incidence matrix M_{ij} which is 1 if RTA_{ij} exceeds 1, else 0. The machine learning algorithm explores the RTA structure computed between 2000-2009 to predict the incidence matrix in 2014, while the generalized economic complexity index is computed based on the predicted incidence matrix M.

2.3 Methodology

This section describes the Generalized economic complexity (GENEPY) index used in measuring the level of capabilities of cities and the capabilities required in a technology. Next, we describe the machine learning models and the benchmark network science model used in forecasting the competitiveness of cities in the different technology domain. Also, we provide a description of the evaluation metrics used in determining what models perform best in the forecasting exercise.

2.3.1 Generalized economic complexity model

To describe the GENEPY algorithm, first, we construct the binarize revealed technology advantage $\mathbf{M}$ of an undirected bipartite network between 150 cities and 637 4-digit IPCs. To do this we compute the revealed technology advantage (RTA_{ij}^{t}) of *city i* with respect to *IPC j* in a given year *t* . To account for partial ownership of patents attributed to a city for an IPC, we will adopt the fractional count of patents (at the patent family level) i.e., given a patent *P* instantiated with IPCs $\{j_1, j_2, ..., j_n\}$ and done by inventors from cities $\{i_1, i_2, ..., i_m\}$. The fractional count of patents to any given city i_k is $\sum_P 1/m$, the fractional count of patents to any given IPC j_k is $\sum_P 1/n$, and the fractional count of patents produced by city i_k in a given IPC j_k is $N_{ij} = \sum_P \frac{1}{mn}$, where P is the total patents. Then, the

revealed technology advantage of city i in an IPC j is computed as:

$$^{2}RTA_{ij} = \frac{\frac{N_{i,j}}{\sum_i N_{i,j}}}{\frac{\sum_j N_{i,j}}{\sum_j \sum_i N_{i,j}}} \tag{2.1}$$

If the RTA_{ij} exceeds 1, it implies that city i positively specializes in technology *j*), hence the entry in the binary matrix $M_{ij} = 1$, otherwise $M_{ij} = 0$. From the matrix $\mathbf{M}$, the complexity of cities and technologies can be described by a system of coupled equations:

$$X_i = f(Y_1, Y_2, ..., Y_j, M_{ij}) \; ; \; Y_j = g(X_1, X_2, ..., X_i, M_{ij}) \tag{2.2}$$

where f and g are linear functions used in recasting X_i and Y_j as a solution of an eigenvalue problem given below (Sciarra et al., 2020):

$$X_i = \frac{1}{\lambda} \sum_{i^*} N_{ii^*} X_{i*} \; ; \; Y_j = \frac{1}{\lambda} \sum_{j^*} G_{jj^*} Y_{j*} \tag{2.3}$$

where $\mathbf{N} = \mathbf{W}\mathbf{W}^T$ and $\mathbf{G} = \mathbf{W}^T\mathbf{W}$ are proximity matrices for cities *i* and technologies *j* respectively. From a complex network perspective, X_i and Y_j represents the eigen-centrality in the bipartite network of cities and technologies respectively. For example, the values in X_i are derived based on the idea that connections to more central cities should matter more than the same number of connections to less central cities. The associated λ represents how much information or variance exists in the corresponding eigenvector. Recovering either MR or FC depends on how the weighted incidence matrix $\mathbf{W}$ in $\mathbf{N}$ and $\mathbf{G}$ is defined. To recover the FC method, $W_{ij} = M_{ij}/k_i k'_j$, while to recover the MR method, $W_{ij} = \frac{M_{ij}}{\sqrt{k_i k_j}}$, where k_i is the degree of the city, k_j is the degree of technology, and k'_j is the degree of technology divided by the degree of the city. In our implementation, the weighted incidence matrix is defined consistently with the FC method.

[2]We consider only patents with IPCs. Hence the computation of the revealed technology advantage assumes that the distribution of patents with missing IPCs across cities is random ("missing-at-random" condition). Check Song and Guo, 2021 on how to denoise RTA_{ij}.

The eigenvalues λ_k and related eigenvectors of the matrices N and G are solutions to the eigenproblems in Equation 2.3. In most cases, the eigenvector with the biggest eigenvalue holds the most information and is thus considered the solution to the eigenproblems. However, the GENEPY index combines the first and second eigenvalues and eigenvectors of $\mathbf{N}$ and $\mathbf{G}$ to give a measure of complexity for cities and technology respectively as:

$$
\begin{aligned}
GENEPY_i &= \Big(\sum_{k=1}^{2} \lambda_k X_{i,k}^2\Big)^2 + 2\sum_{k=1}^{2} \lambda_k^2 X_{i,k}^2 \\
GENEPY_j &= \Big(\sum_{k=1}^{2} \lambda_k Y_{j,k}^2\Big)^2 + 2\sum_{k=1}^{2} \lambda_k^2 Y_{j,k}^2
\end{aligned}
\tag{2.4}
$$

$GENEPY_i$ and $GENEPY_j$ in Equation 2.4 represent the generalized economic complexity of cities and technologies respectively. For example, $GENEPY_i$ can best be understood based on the interpretation of $X_{i,1}$ (first eigenvector) and $X_{i,2}$ (second eigenvector) being combined. The first eigenvector represents the eigencentrality measure, while the second eigenvector represents the clustering of cities based on similarity in patenting structure. The combination of the two eigenvectors to arrive at $GENEPY_i$ can be interpreted as identifying capabilities the city has and are similar to other cities. Hence cities with high GENEPY are seen in the cluster of central cities, while cities with low GENEPY are seen in the peripheral of the network. Economic complexity measure such as GENEPY is scale-free as they do not correlate with size of cities or countries. For example, the correlation of GENEPY and city size in our dataset is approximately 0.20. From an economic perspective, the economic complexity measure is interesting as it has been shown to correlate strongly with per-capita GDP in both country and regional analysis (Mealy and Coyle, 2019).

The GENEPY algorithm presented above assesses the current capabilities of cities and technologies based on a single measure - economic complexity of patenting activities. It does not consider the time evolution of

competitiveness in technological areas. For example, it might be useful to know if a city being competitive in patenting in radar would in 5 years time patent in radio broadcasting apparatus. To do this we characterize forecasting the competitiveness of cities in technological areas as a classification problem, and propose using machine learning models.

2.3.2 Models for forecasting competitiveness of cities

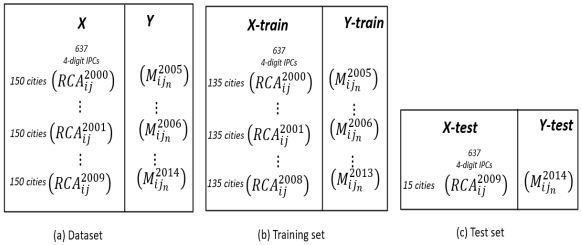
In this section, we present the *model-free* machine learning approach and the network-based benchmark model for forecasting the competitiveness of cities. First, we present the implementation strategy used in all the machine learning models.

[3]Based on the patent data, we derive the RTA_{ij}^{t} for each year $t \in [2000, 2009]$ following Equation 2.1 - which represents the revealed technological advantage of 150 cities in 637 4-digit IPCs. This implies that the machine learning model learns the 637 features to arrive at a probability score of whether a city will be competitive in 2014. The competitiveness target variable is the binarized RTA_{ij} measure in 2014 given as $M_{ij} = 1$ if $RTA_{ij} > 1$ else 0. The choice of 1 as a threshold is consistent with the literature (Tacchella et al., 2021). In this sense, a city with a unit entry in **M** is competitively patenting in the given technology, and a zero entry implies otherwise.

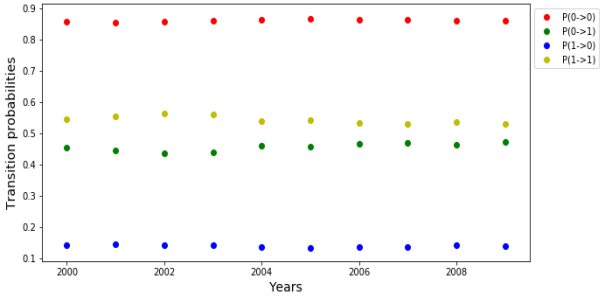
The choice of forecasting competitiveness in 5 years as opposed to shorter number of years or longer forecasting horizon can be explained partly by the observation in Figure 10b, we see that transition probabilities in 5 years are more or less flat (with exception of $P(0 \to 1)$ and $P(1 \to 1)$). The conditional probabilities suggest that only when we consider a longer forecast time horizon can we then observe changes in cities being more

[3]The complexity and competitiveness forecasts will not significantly change, since the RTA_{ij} and M_{ij} used is highly correlated with the RTA_{ij} and M_{ij} constructed from (a) same data without observations that have country discrepancies or (b) same data without observations having both country discrepancies and coordinate allocation done via sources 4-6. Specifically, the correlation of RTA_{ij} with that computed in (a) and (b) are approximately 0.94. The fraction of non-matching binaries between M_{ij} and that computed in (a) and (b) are 0.052 and 0.059 respectively.

likely to remain competitive or become competitive in technologies. This eliminates the choice of a smaller time horizon like 2 or 3 years. 5 years is the best compromise given the period of the geocoded data we have (i.e., 2000-2014), if we choose a 10-year horizon for the forecast, we will halve the sample size for training the model which will impact the capability of the ML models to arrive at meaningful inference.



(a) Implementation strategy



(b) Transition probabilities

Figure 10: (a) The structure of the training and test dataset is such that for each IPC j_n $(n = 1, ..., 637)$ and each disjoint partition $k = 10$, the training set consists of 135 cities, and the test set consists of 15 cities in partition k. The number of models trained for each classification algorithm is 6370. (b) The low stability in the transition probabilities $P(0 \longrightarrow 1)$ implies that the activation task is more difficult

The prediction exercise is in two ways: (i) the unconditional prediction of **M** in 2014 (ii) the conditional prediction of **M** which we refer to as the *activation task* requires predicting $M_{ij} = 1$ in 2014 for samples with revealed

technological $RTA_{ij} < 0.25$ consistently for period $t \in [2000, 2009]$. This task is a difficult task, as it can be seen from Figure 10b, the probability that cities transition to becoming competitive in a technology they never where is less than 0.5.

The strategy employed for implementing the machine learning models hinges on the models being able to learn not just the patenting structure of cities (since the model could exploit the high autocorrelation $P(0 \longrightarrow 0)$ and $P(1 \longrightarrow 1)$), but more importantly, the model should rather exploit the relatedness between technologies in predicting competitiveness of cities. To achieve this, the cities used in the testing phase are different from cities used in the training phase of the model. Secondly, since each IPC requires a different set of capabilities, for each of the classification models considered, we choose to train different models for each IPC.

From Figure 10a, the training set is built using patenting activities in period $t \in [2000, 2008]$ - the procedure for training the classification algorithms is the same for all models. The algorithms are to learn the structural relationship between the RTA_{ij}^{t} values (i.e., features) of city i and IPC j, with the corresponding element in the competitiveness matrix M_{ij}^{t+5} (i.e., target). Since we are training different models for each IPC, the target variable M_{ij}^{t+5} will always change, but the features will be the same. More importantly, since we do not want the model to exploit the auto-correlation present in the revealed technology matrix, but should rather use the similarities in technologies inferred, hence at the test phase, we partition the $N = 150$ cities to $k = 10$ disjoint sets of cities. For each partition k, we build the model on a training set involving $N - k$ cities and predict on a test set RTA_{ij}^{2009} consisting of cities in k partition. This implies that for each classification algorithm, we are training 6370 models - i.e., 637 for each IPCs and 10 for each partition. The result of the prediction is put together to reconstruct the complete competitiveness matrix. The performance of the model is evaluated by comparing the prediction with competitiveness matrix $\mathbf{M}$ in 2014 using the evaluation metrics described in Section 2.3.3. Each of the machine learning models is described below:

Random Forest model

The Random Forest (RF) model is an ensemble approach that uses bootstrapping and aggregation to train many decision trees simultaneously. Bootstrapping refers to the training of many independent decision trees in parallel on different subsets of the training dataset with varying subsets of accessible characteristics. Bootstrapping is done to lower the RF classifier's total variance. The RF classifier then aggregates individual tree decisions for the final decision. Thus, are usually good at generalizing to unseen data and are less prone to issues of overfitting (Misra, Li, and He, 2019).

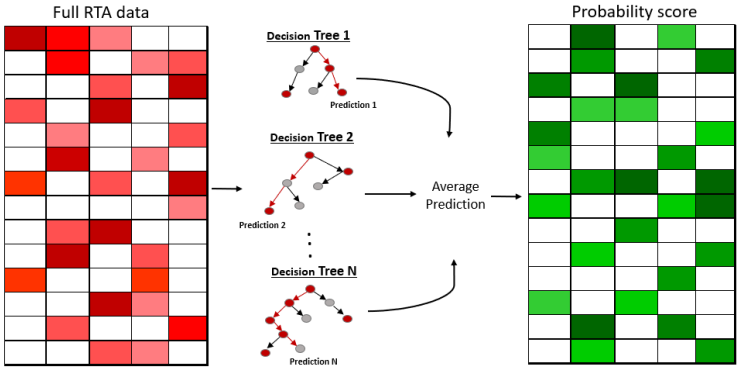


Figure 11: How the Random forest model is used to compute the probability of being competitive (i.e., $y_i = 1$) in an IPC, given the revealed technological advantage features

As depicted in Figure 11, the Random forest (RF) model consist of training multiple decision trees using features $\mathbf{x_i} \in \mathbf{RTA_{ij}^t}$ in the dataset. Each decision tree randomly selects observations $(\mathbf{x_i}, y_i) \in (\mathbf{RTA_{ij}^t}, M_{ij}^{t+5})$ and the goal is to find a rule that identifies the best subsets of features x_i relevant to each class y_i. The selection of random observations allows for training multiple uncorrelated models, enhancing the aggregate predictive power of the Random forest. The rule derived during the training phase is applied to the features of each city in the test set to predict class y_i. The prediction of the Random forest is based on the majority vote from

all the decision trees trained.; and the probability of a predicted class is the number of votes for the class divided by the number of decision trees. Since we are interested in class $y_i = 1$, if the predicted class from the RF model is $y_i = 0$ with $P(y_i = 0)$, we have that $P(y_i = 1) = 1 - P(y_i = 0)$.

For each of the decision trees, starting from the root node of the decision tree, the graph splits into binary nodes by finding the subset $\mathbf{x_j} \subset \mathbf{x_i}$ that minimize the Gini function $I(\mathbf{x_j}|\mathbf{x_i})$. The Gini function is given as: $1 - \sum_{n=1}^{N} P_n^2$ calculates the probability that a specific feature will be incorrectly associated with a class when selected randomly, where P_n denotes the probability that a feature is associated with a single class. If all the features in $\mathbf{x_j}$ are associated with a class, then the Gini coefficient is 0 and x_j is said to be pure, if they are not then the Gini coefficient is 1. This minimization process to achieve the best split is continuously done at every node until no further split can be done or if the stop condition is reached (e.g., if the decision tree reaches its specified depth set apriori). Upon reaching the stop condition, the leftover features from the input feature are associated with the predicted attribute.

Each of the decision trees that makes up the ensemble in the Random Forest model is trained following this process, and the prediction is based on the average class probabilities estimated by each decision tree. The parameters to be tuned to achieve an optimal performance include the number of trees, the maximum depth of each tree, criteria for evaluating the quality of each split in each node. In our implementation, we used the default parameters in the Random Forest classifier algorithm in the sklearn library available in python.

eXtreme Gradient Boosting model

The eXtreme Gradient Boosting (XGBoost) model is an algorithm introduced by Chen and Guestrin, 2016 to increase speed and address the issue of over-fitting by introducing a regularization parameter. Unlike the Random forest model which trains several decision trees in parallel, the XGBoost is based on a sequential learning process by sequentially combining regression trees which are considered weak learners. These

regression trees unlike decision trees assign continuous scores w_i to each $i-th$ leaf (i.e., the last node). For each input in the training set, a decision rule in the tree is applied to classify the input into specific leaves. When this is done for all input, the sum of the scores in all the leaves is taken, and this represents the final prediction of the tree. Formally, this tree ensemble model is trained in an additive manner as follows, suppose $\hat{y}_i^t$ be the prediction of the $i-th$ instance at the $t-th$ iteration, we will need to greedily add a regression tree f_t to minimize the regularized objective function:

$$\mathcal{L}^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \tag{2.5}$$

The optimal values for each leaf and the overall score of the tree are then computed using gradient descent. The score is also known as the "impurity of a tree's predictions". The loss function in Equation 2.5 contains a regularization term $\Omega \geq 0$ whose objective is to reduce the complexity of the regression tree function f_t. Besides the regularized objective function described above, there are two other methods applied to prevent overfitting. The first technique is the feature column sub-sampling similar to what is used in the Random Forest model, and the other is the introduction of shrinkage (Friedman, 2002) - which involve scaling the newly added weight at each step of the boosting process (Chen and Guestrin, 2016). The advantage of XGBoost is its speed compared to other boosting algorithms like Adaboost, and other characteristics like shrinkage, regularization, and feature subsampling allow it to generalize better. However, in comparison to the Random forest model it has more hyper-parameters (e.g., regularization rate, shrinkage, maximum depth of trees, and number of estimators) that need to be tuned to achieve better performance. In our implementation, we used the default parameters of the *XGBoost* package in python.

Support Vector Machines model

The Support Vector Machines (SVM) algorithmic paradigm introduced in Boser, Guyon, and Vapnik, 1992 and Cortes and Vapnik, 1995 can be used to learn predictors in high dimensional feature spaces similar

to our application setting. The algorithm does this by searching for a hyperplane with the largest margin separating the training set $(x_i, y_i) \in (RTA_{ij}^t, M_{ij}^{t+5})$. Let $S = (x_1, y_1), ..., (x_n, y_n)$ be training set such that $x_i \in \mathbf{R}^d$ and $y_i \in \{0, 1\}$, we say that the training set is linearly separable if there exists a hyperplane $(\mathbf{w}, b)$, such that $y_i(< \mathbf{w}, x_i > +b) > 0$ for all $i \in \mathbf{N}$. For any given training set there are many hyperplanes $(\mathbf{w}, b)$ that satisfies the condition i.e., many hyperplanes are able to correctly separate the training samples into positive and negative predictions. However, the choice of an optimal hyperplane is hinged on the notion of margin - the margin of a hyperplane is the minimum distance between points in the training set and the hyperplane. If the margin of the hyperplane is large then such hyperplane will be robust to perturbations in the training set - this implies that such a hyperplane will be less prone to error. There are two SVM learning rules - the Hard SVM and the Soft SVM used to infer the optimal hyperplane. Intuitively, the Hard SVM examines the vector space $\mathbf{w}$ separating the training set, and identifies the vector $\mathbf{w}$ with the least norm and also satisfying the condition $| < \mathbf{w}, x_i > +b| \geq 1$ for all i. Formally, it is simply solving the optimization problem:

$$(w_0, b_0) = argmin_{(w,b)} ||\mathbf{w}||^2 \ \ni \ y_i(< w, x_i > +b) \geq 1 \tag{2.6}$$

The complexity in identifying $\mathbf{w}$ is directly proportional to the number of features in the training set. While the Hard SVM approach assumes linear separability of the training set, the Soft SVM relaxes this strong assumption. The optimization problem is then adjusted by introducing a non-negative parameter ψ_i that measures how much the constraint $y_i(< w, x_i > +b) \geq 1$ is being violated. The optimization problem for the soft SVM is formalized as follow:

$$min_{(w,b,\psi)} \left(\lambda ||\mathbf{w}||^2 + \frac{1}{n} \sum_{i=1}^{n} \psi_i \right) \tag{2.7}$$

$$\text{such that } y_i(< w, x_i > +b) \geq 1 - \psi_i \ \forall i \ \textbf{and} \ \psi_i \geq 0. \tag{2.8}$$

The Soft SVM jointly minimizes $\mathbf{w}, \psi$. The trade-off between the two parameters is captured in λ. Most problems in principle are non-linearly separable i.e., they fall under the Soft SVM paradigm. This is the default

implementation available in most SVM packages. In our implementation, we used the default SVM implementation in *sklearn*, however, we specify the probability to be True, in order to obtain both the prediction of class labels and probability estimates of predicted classes.

Neural network model

In our implementation, we used a simple feed-forward neural network model in which there are no loops (i.e., no self connections) in the network. A feed-forward neural network is described by a directed network, $G = (V, E)$, where V is a set of nodes also known as neurons, and E is a set of edges between neurons across layers, the weight of the edges are defined as $w : E \longrightarrow \mathbf{R}$. The input into a neuron is the weighted sum of all other neurons connected to it, and the activation of the neuron which is a signal of information in the neuron is characterized by an activation function σ define as $\sigma : \mathbf{R} \longrightarrow \mathbf{R}$, which could assume different functional forms depending on the task.

In our implementation we built a neural network with 3 hidden layers (each layer has 32 neurons), and the activation function used in the hidden layers is *ReLu*, while we chose a *sigmoid* activation function in the output layer in other to have as an output the prediction probabilities - these probabilities are the probability of the neural network model classifying an observation as positive (i.e., $\mathbf{M}_{ij}^{t+5} = 1$). Once the neural network model is built, the goal is to learn the optimal embedding $f_{V,E,\theta} : \mathbf{R}^{input} \longrightarrow \mathbf{R}^{output}$ from a class of possible embeddings $\mathcal{F} = \{f_{V,E,\theta}\}$ by adjusting hyper-parameters θ - example of such parameters include learning rate, number of epochs (which specifies the number of times the learning algorithm will go through the entire training set), and batch size (the number of samples of the training set to go through before updating the model parameters). A typical learning algorithm that is used for training the feed-forward neural network is the "stochastic gradient descent" (SGD) which is an iterative algorithm for minimizing the loss of a predictive model. To update the weight w at each iteration what is being used is the back-propagation algorithm. The hyper-parameters

and settings of the neural network model used in our implementation include: batch size= 5, epoch size = 10, dropout = 0.1, l1_reg = 0.001, optimizer=rmsprop, loss=binary_cross_entropy.

Benchmark model: Time-delayed co-occurrence model

To benchmark the performance of the machine learning models, we use a network model that combines the time-delayed correlation between technological areas, and the current competitiveness of cities to estimate a score on what the competitiveness of the city would be in a given technological area in the future. To measure this time-delayed correlation, we will use the idea proposed in Pugliese et al., 2019 in constructing the enabling or progression network between technologies by computing the conditional probability between technologies. To do this, we first compute the time-delayed co-occurrence between technology j and j' given as

$$C_{j,j'}(t, t + \Delta t) = \sum_i M_{i,j'}(t + \Delta t) M_{i,j}(t) \tag{2.9}$$

$C_{j,j'}(t, t + \Delta t)$ counts the number of cities that patented in j at time t and also patented in j' at time $t + \Delta t$. The higher the value the stronger the signal that j is a prerequisite to j'. However, if a city is very diversified the information coming from the city is less, similarly if a technology is ubiquitous. Therefore the co-occurrence matrix is normalized with respect to the diversity of cities and ubiquity of technology (Pugliese et al., 2019; Zaccaria et al., 2018) given as :

$$d_i = \sum_j M_{i,j}(t) \; ; \; u_j = \sum_i M_{i,j}(t) \tag{2.10}$$

The expression for the normalized co-occurrence matrix which is interpreted as the conditional probability of being competitive in technology j' given the city is competitive in j Δt (we choose $\Delta t = 5$) years ago is given as:

$$P_{j,j'}(t, t + \Delta t) = \frac{1}{u_j} \sum_i \frac{M_{i,j'}(t + \Delta t) M_{i,j}(t)}{d_i(t + \Delta t)} \tag{2.11}$$

To arrive at a time-independent conditional probability, we can remove the noise in the technological network by averaging over the years so that any spurious connections $P_{j,j'}(t, t + \Delta t)$ will not affect the estimation of the density.

$$P_{jj'}(\Delta t) = \frac{1}{N} \sum_{t=2000}^{2009} P_{j,j'}(t, t + \Delta t) \tag{2.12}$$

To discriminate the real inter-dependencies between technologies from random occurrences that can be driven by the ubiquity of technologies or diversity of cities, we estimate an ensemble of 100 $\hat{P}_{j,j'}$ computed from random realizations of the binary matrix such that the diversity and ubiquity is kept fixed. The estimation is done using the hypergeometric ensemble model (Casiraghi et al., 2017). The entries in $P_{j,j'}$ are kept if they are significantly (at an alpha level of 0.05) above the distribution from the ensembles. Combining the discriminated time-independent conditional probabilities with the competitive structure M_{ij}^{2009}, we compute the density of the cities capabilities around technological areas j' in 2014 by doing the following:

$$d_{ij'}(\Delta t) = \sum_{j} M_{ij}^{2009} P_{jj'}(\Delta t) \tag{2.13}$$

The density can be characterized as a score that captures the strength of connections between the cities' current competitive technologies with technologies in the future. The density is between 0 and 1, a stronger score will imply that there are many current technologies connected to future technologies. If the density of a city in a given technology is 0 it implies the city is not currently competitive in any technology that can allow the city to move into the given technology in the future.

2.3.3 Evaluation metrics

Choosing the right evaluation metrics for a machine learning model is as important as building the model itself, as a wrong metric could either make a good model seem to perform poorly or a bad model perform

excellently. The choice of an evaluation metric depends on the specific problem considered. We will first motivate the choice of metrics and then describe the practical meaning of each of the performance indicators chosen to compare the ML algorithms with the benchmark. For the benchmark model, the output (i.e., the density of cities

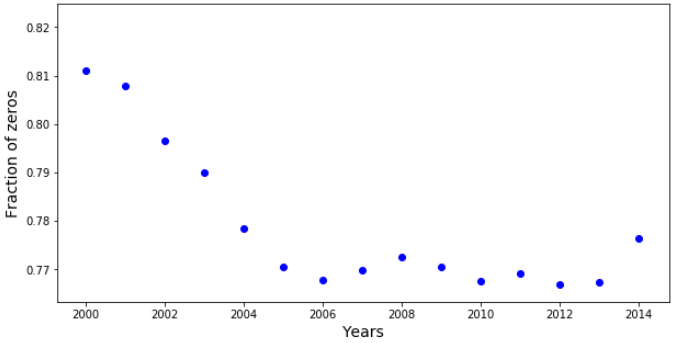


Figure 12: Imbalance class labels: The fraction of zero elements in the competitiveness matrix show that the class labels are imbalanced.

capabilities in a technology) is a continuous variable we need to define a threshold above which the density is associated with a prediction of 1, below which is associated with a prediction of 0. To do this, we choose the threshold that maximizes the F1-score. Similarly, for the ML algorithms, we output the probability of a city being classified as competitive in a technology, which is then binarized such that the F1-score is maximized.

Figure 12 shows the density of the competitiveness matrix M as a function of time, we have that on average the occurrence of zeros is about 77.8%, and 77.6% specifically for the year 2014, implying that we have an imbalance in the class labels. This observation does have consequences in the choice of performance metrics. For example, we can not use metrics like the ROC-AUC to evaluate the classifiers as this metric does have a strong bias when there is a class imbalance in the labels. With this understanding, we will be using the following metrics suitable for addressing class imbalances - F1-score, PR-AUC score, and Mathews correlation coefficient.

These metrics are described below:

F1-score

F1 score penalizes extreme values by combining both recall and precision using harmonic mean rather than arithmetic mean (Sasaki, 2007). The precision of the model simply captures what fraction of the actual positives (i.e., entries with value 1 - where the cities competitively patent in IPCs) after Δt years were correctly predicted. If the precision score is high it implies the model had a low number of false positives. On the other hand, recall is the ratio between true positives and the sum of true positives and false negatives. It captures how well the positive class was predicted and is synonymous with the sensitivity of the prediction. Combining both precision and recall provides a single score that captures how well the positive class is predicted and the sensitivity of the prediction. The F1-score takes values between 0 and 1, and the higher the F1 score the better the model.

Precision Recall curve (PR-AUC)

The PR-AUC curve is an alternative to the ROC-AUC curve when dealing with imbalanced data. The ROC-AUC curve is simply the area under the curve computed from a plot of the false positive rate (x-axis) against the true positive rate (y-axis). The problem using this is that for imbalance data, the area under the curve might not reflect the true performance of the classifier, since the false positive rate will always be seen to have small values (which is good) not because this is the case, but because we have a large number of negative values - hence making the ROC-AUC not appropriate. By replacing the false positive rate with precision, and computing the area under the curve, we obtain PR-AUC score. A larger PR-AUC curve signals a better model performance.

Matthews correlation coefficient (MCC)

The MCC is a good measure for evaluating the quality of a binary classification model especially when there is a class imbalance, as it accounts for the model's precision and ability to recall. It is essentially a correlation coefficient that takes values between -1 and 1. A coefficient of 1 implies

a perfect prediction, 0 implies that the model is no better than a random prediction, and -1 implies a prediction completely different from the observation. The MCC is computed from the confusion matrix as follows (Boughorbel and El-Anbari, 2017):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.14}$$

Other scores such as accuracy score, ROC-AUC score, precision and recall are also reported.

2.4 Results

First, we will present the result of the economic complexity of cities, and then the result for the machine learning models used in forecasting the competitiveness of cities in 5 years. In Figure 13 we show the changes in the complexity ranking of cities between the period 2000-2004 and 2010-2014. In computing the economic complexity index of cities, we aggregate the patenting activities using a window of 5 years to eliminate fluctuations in patenting activities that might emerge as a result of our data generating process. In the period 2000-2004, the top 10 most complex cities include Los Angeles (Greater), Tokyo, Paris, Stuttgart, Boston, New York (Greater), Chicago, Beijing, Munich, and Washington (Greater), while in period 2010-2014, the top 10 most complex cities include: Taipei, Shenzhen, Tokyo, Shanghai, San Diego, Paris, Seoul, Munich, Beijing, and Minneapolis. Between these periods, we see a drastic change in complexity ranking, with cities in Asia moving up the rank while cities in North America and Europe moving down the rank. The top movers include cities like Bangalore, Istanbul, and New Delhi, and cities that declined the most include Lyon, Geneva, Montreal, and Los Angeles (Greater). Specifically, we have that 33, 19, 8 cities moved up the ranking from Asia, Europe, and North America respectively.

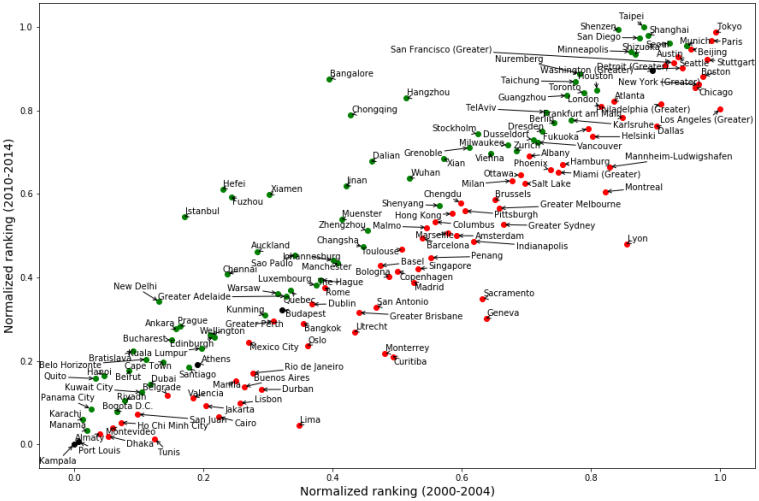


Figure 13: Changes in the complexity ranking of cities: The green dot represents cities that moved up the ranking between 2000-2004 and 2010-2014, the red dot represents cities that moved down the ranking between 2000-2004 and 2010-2014, and the black dot represents cities whose complexity ranking remains unchanged.

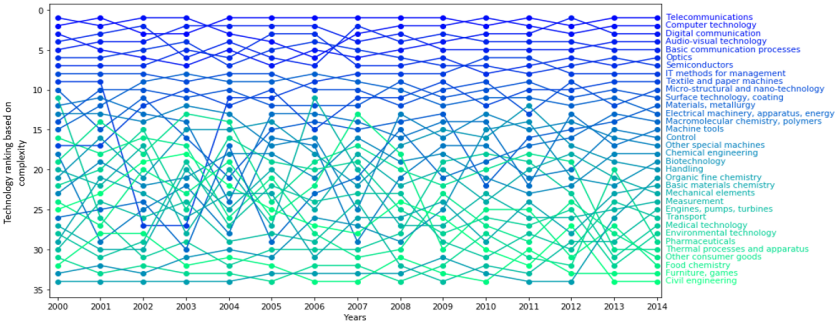


Figure 14: Technology complexity: The evolution of technology complexity ranking between the 2000-2014 shows some level of stability over time.

Figure 14 shows the time evolution of the complexity of technology fields. On the right, we have the technology fields to be ranked based on their

complexity in 2014. We observe that technology fields such as Telecommunications, Computer technology, Digital communications do have high complexity values, while technology fields such as Furniture, games, Food chemistry, and Other consumer goods do have low complexity values.

Complexity rankings in each year are connected by lines and the color of the lines is based on the ranking in the year 2014. What we find is that the time evolution of the ranking is stable which reflects the validity of theoretical assumptions around stability in capabilities needed to upgrade into technological field. While some fluctuations are expected, however a complete overturning of the rankings would have meant that capabilities needed to be competitive in a technology field vary over time – which would be at odds with the general idea of capabilities (Zaccaria et al., 2018).

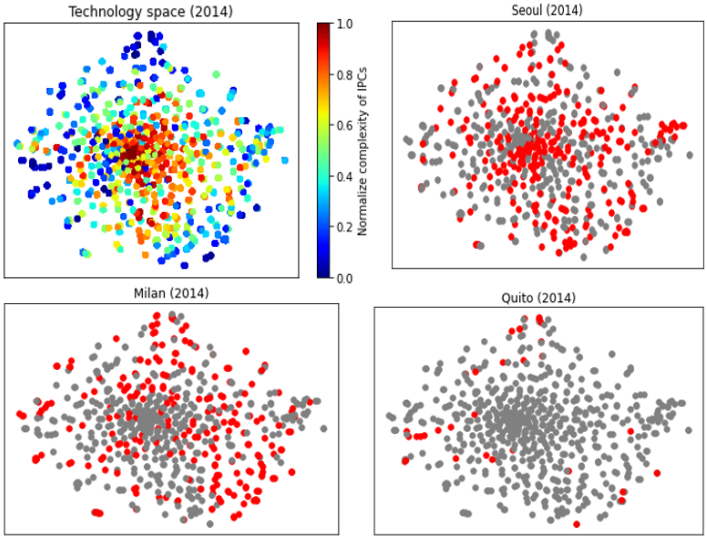


Figure 15: Technology space in 2014: The figure on the top left shows the embedding for the 634 IPCs. While the top right and bottom figures identify the IPCs in this technology space where Seoul, Milan, and Quito are competitively producing patents. More complex cities produce patents in central IPCs, and central IPCs are more complex.

Figure 15 provides a visualization of the technology space in 2014. This embedding is constructed based on the M_{cp} matrix for 2014 using the t-distributed stochastic neighbor embedding (t-SNE) - a non-linear dimensionality reduction technique used for embedding high dimensional data in low dimensional space. IPCs that have similar cities competitive in them are assigned high probabilities and are more clustered together in the technology space, while dissimilar IPCs are assigned lower probabilities.

From the top left image in Figure 15, we see that central IPCs are more complex, while IPCs in the peripheral are less complex. Comparing Seoul which is more complex than Milan, and Quito, we observe that complex cities like Seoul tend to patent competitively in IPCs that are central and complex in the technology space. On the other hand, Milan patent competitively in less central IPCs, and Quito patent in peripheral IPCs.

2.4.1 Results of forecasting competitiveness of cities

Table 12 present the evaluation metrics of each machine learning model and the benchmark model for the prediction task on the full matrix. First, we observe that all machine learning models were able to learn the interdependence better than the benchmark model for this task irrespective of what evaluation metrics being used. Considering the target variable in the dataset is imbalanced, as there is a larger proportion of zeros than ones, we focus on interpreting the performance of the model based on F1 score. The machine learning model with the least performance i.e., Neural network has an F1 score that is 11.9% better than the benchmark model, while the machine learning model with the best performance i.e., the Random forest model has the best F1 score (0.56) indicating that it does a better job than all other models in minimizing both false positives and false negatives. Specifically, it has an F1 score that is 33.3% better than the benchmark model. The Random forest model did not only show to be better at learning but also has a lower computational time in comparison with other machine learning algorithms. The performance of the Random forest model can be attributed to its capability to combine relatedness

and anti-relatedness signals to provide better predictions (Tacchella et al., 2021), as this can be useful given the possibility that the presence of a technology in a city's technology base can negatively impact the likelihood of patenting in certain target technologies. Also, in Table 12 the systematic gap between the recall and precision scores for all the models suggests that the models do a better job at minimizing false negatives. Specifically, the benchmark model does a better job at minimizing false negatives in comparison with other machine learning models, but is not as good as the machine learning models in minimizing false positives. The expectation is that precision of the model should be better with less complex cities since these cities are not diversified and predicting an upgrade will be less likely, while the recall capability of the model should be better the more complex the city becomes partly due to diversification of the city. Considering this expectation alongside the cohort of global cities which are on average diversified in 141 IPCs, the recall capability will be greater than the precision.

	Benchmark model	Random Forest model	XGBoost model	SVM model	Neural network model
Best F1 score	0.42	**0.56**	0.54	0.51	0.47
PR-AUC score	0.34	**0.58**	0.56	0.53	0.38
Mathew's coefficient	0.20	**0.40**	0.39	0.33	0.26
Accuracy score	0.57	**0.81**	0.79	0.69	0.65
ROC-AUC score	0.62	**0.73**	0.71	0.69	0.65
Precision score	0.30	**0.47**	**0.47**	0.40	0.36
Recall score	**0.72**	0.69	0.63	0.70	0.67

Table 12: Predictive performance of algorithms on the full M_{ij} matrix

Comparing the evaluation metrics in Table 12 and Figure 17, we observe that the evaluation metrics for the full matrix task are higher than that of the activation task suggests that predicting the unconditional presence of new technology in a city is a relatively simple task, as this is driven by the stability in the transition probabilities seen in Figure 10b.

Figure 16 shows how well we predict the presence of new technologies in cities viz-a-viz their economic complexity in the given year - we present the relationship between the economic complexity of cities as at the year the prediction was done i.e., 2009, and the F1-score of the Random Forest

model for each city. We observe a positive relationship, suggesting that the more economically complex the city is the better the Random Forest model was able to make better predictions for the full matrix task. The Pearson correlation between the GENEPY index of cities and the F1-score is 0.67.

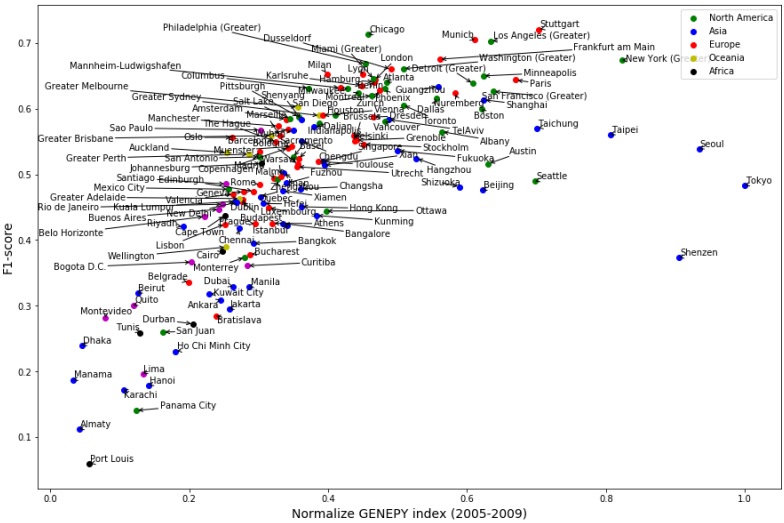


Figure 16: Generalize economic complexity vs Prediction performance: On the x-axis is the Generalized economic complexity index of cities computed between period 2005-2009, and on the y-axis is the F1-score that captures the performance of the Random Forest model in forecasting the competitiveness (full matrix) in 2014 for each city. The correlation between the GENEPY index and the F1-score of cities is 0.67.

Next, we examine the prediction on becoming competitive in new technologies condition on the fact that the RTA level of the city in the technology has consistently been below 0.25 (i.e., *activation task*). Based on our observation in Figure 10b the probability that a city will remain competitive in a technology is approximately 0.53, while the probability that a city will remain non-competitive in a technology is approximately 0.86. The strong auto-correlation implies that we can safely predict the presence of 0 and 1 by exploiting the competitive structure in previous years. However,

being competitive in a new technology is a rare event, specifically, the probability of this event to occur after 5 years is about 0.45. This implies that predicting the competitiveness of a new technology that the city was previously not competitive in is a hard task. Such a task is more economically useful and has practical policy implications as it depends on the presence of suitable but unrevealed capabilities in the city that can be determined by looking at the other technologies the city is patenting competitively.

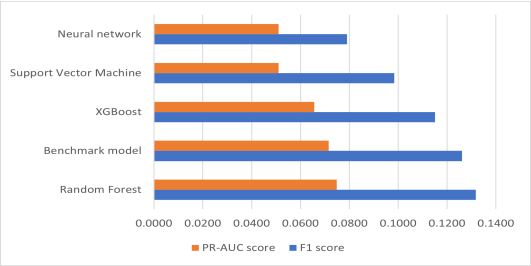


Figure 17: Only the Random Forest model performs better than the benchmark model on the activation task. We evaluate how well the algorithms can predict there will be an activation of new technology (i.e. $M_{ij}^{2014} = 1$) condition on the fact that $RTA_{ij} < 0.25$ between 2000-2008.

From Figure 17, we observe that the performance of the models on the activation task is lower compared to the full matrix task. Specifically, Random forest, XGBoost, SVM, Neural network and benchmark model had a a reduction in F1 score by 76.5%, 78.7%, 80.7%, 83.2%, 70% respectively. Also, we observe that only the Random Forest model is better than the benchmark model considering both the F1 and PR-AUC scores. The performance of the Random forest model is 4.5% better than the benchmark model. To put the performance metrics in context, the model is as good or better than what has been observed in the literature applying similar methodologies to firms and countries. For example, Straccamore and Zaccaria, 2021 used patent data to forecast what the next technology of a firm would be using a similar implementation strategy, they found that the Random Forest model achieved better predictive power with the best F1 score less than 0.12. This score is slightly lower than our Random Forest

model with a best F1 score of 0.135. Also in comparison with Tacchella et al., 2021 work on forecasting countries competitively exporting in new product class observe that the XGBoost model outperforms other models with a best F1 score of about 0.139, slightly outperforming the Random Forest model presented.

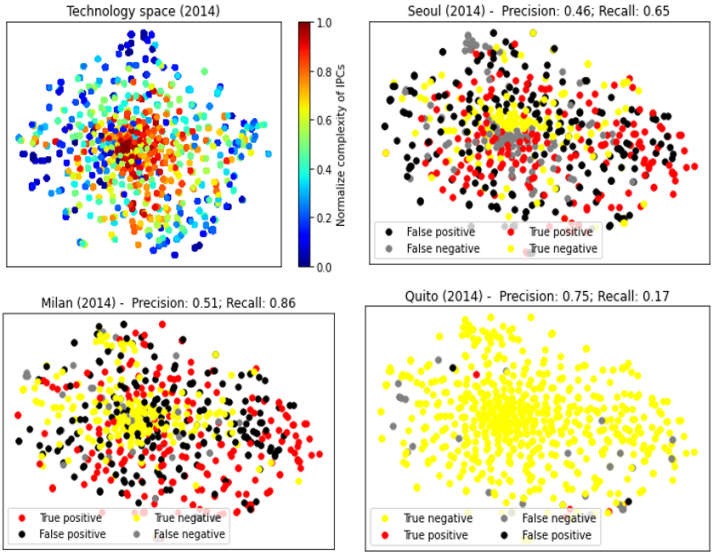
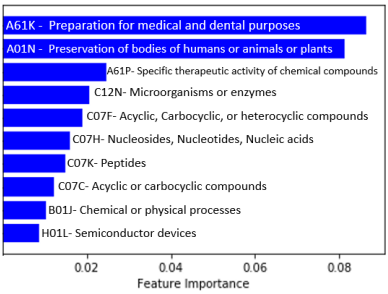


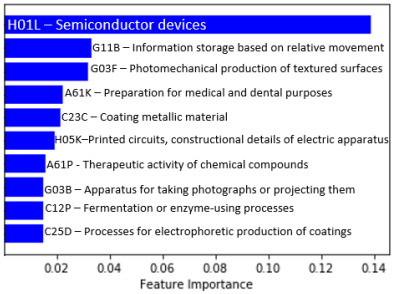
Figure 18: Random Forest prediction quality for Seoul, Milan and Quito

Figure 18 shows the prediction quality for the case of Seoul, Milan and Quito. We observe better precision for low complex cities, these less complex cities do not diversify a lot, hence the ML model is less likely to predict upgrade into new technologies, since typically the pre-requisite technologies are missing in the city. This implies that there will be fewer false positives for less complex cities. In the case of Quito, the ML model had false positives for IPCs CO7H and C22B representing Sugar derivatives and Production or refining of metals respectively. The recall measure which captures how well the actual positives are predicted is seen to have a better result for Milan than for Seoul this is because the more complex you become the more diversified you are, hence the ML model will be

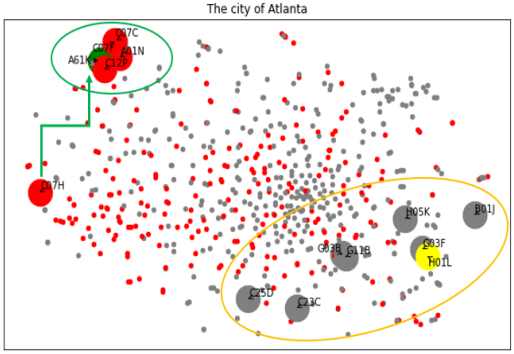
more prone to recording false negatives. Milan had actual positives in 239 IPCs and the model recorded false negatives in 33 IPCs, while Seoul had actual positives in 276 IPCs, and the model recorded false negatives in 96 IPCs (i.e. approximately 300% of false negatives in the case of Milan).



(a) A61K-Preparation for medical and dental purposes



(b) H01L-Semiconductor devices



(c) Atlanta upgrade to A61K and HO1L

Figure 19: (a-b) Feature importance for A61K and H01L based on Random Forest model. These are the top 10 IPCs useful to predict upgrade into A61K and H01L IPCs. (c) Competitive upgrade of the city of Atlanta to A61K and HO1L in 2014. Starting from the big red IPCs in 2009, Atlanta in 2014 patented competitively in A61K IPC but did not in H01L.

Furthermore, we used SHAP (SHapley Additive exPlanations) a game-

theoretic approach to identify what capabilities are useful across cities in becoming competitive in an IPC. Figure 19 shows an example of the top capabilities useful to transition competitively into "A61K - Preparation for medical and dental purposes" and "H01L - Semiconductor devices." The top capability needed are "A01N - Preservation of bodies of humans or animals or plants" and "G11B- Information storage based on relative movement" respectively. We also see that capabilities in A61K can be more useful for transitioning to H01L competitively in comparison to the importance of capabilities in H01L for transitioning into A61K.

Figure 19 shows the case study for the city of Atlanta which as of 2009, was neither competitive in "A61K" nor "H01L". However, the RF model predicted that Atlanta would competitively patent in "A61K" (big green dot) but not in "H01L" (big yellow dot) in 2014 which is consistent with what we observed. The red and grey dots represent the IPCs in which Atlanta was competitively ($RTA_{ij} > 1$) and not competitively ($RTA_{ij} < 1$) patenting in as at year 2009. The enlarged red and grey dots are those specifically needed for an upgrade into "A61K" and "H01L". From the figure we see that Atlanta was competitive in most of the features important for predicting competitiveness in "A61K" but was not competitive in the features important for predicting competitiveness in "H01L."

2.4.2 Reconstructing future generalized economic complexity of cities

Finally, based on the predicted competitiveness structure $\widehat{M}_{ij}^{2014}$ from the Random Forest model (best performing classification algorithm), we construct the generalized economic complexity index for 2014 i.e., $\widehat{GENEPY}_{ij}^{2014}$, and compare this with the actual generalized economic complexity index $GENEPY_{ij}^{2014}$. The Kendall rank correlation coefficient which measures the ordinal association between the two quantities shows a high correlation ($\tau = 0.67$) between the actual GENEPY index and the GENEPY index computed based on the predicted competitive matrix. Also, the $p - value < 0.05$ indicates a significant correlation between both quantities. This result is indicative of the predictive capabilities of

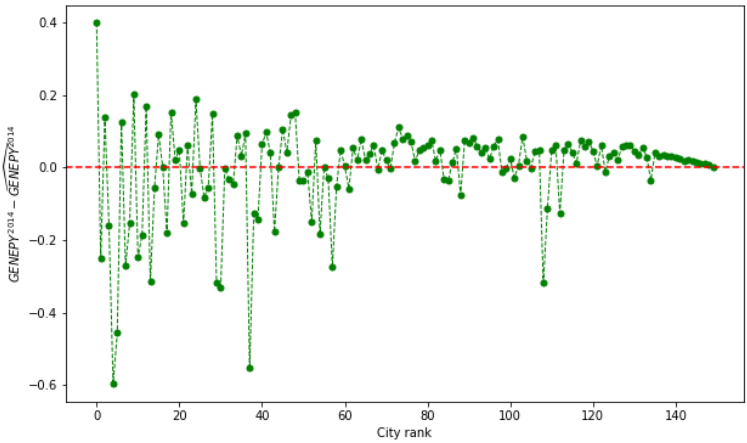


Figure 20: Differences between actual GENEPY and GENEPY based on predicted M_{ij} in 2014: Cities on the x-axis are sorted from the most complex to least complex based on the actual GENEPY index in 2014

the Random Forest model useful in both forecasting the competitiveness structure of cities in new technologies and also the future capabilities of cities as measured by the generalized economic complexity index. Figure 20 shows the difference between the actual GENEPY index and GENEPY index based on the predicted M_{ij}. On the X-axis is the city rank - cities are ranked from the most complex to least complex based on the actual GENEPY index. We see that the more complex a city is the higher the difference between the actual and predicted GENEPY index. Also, we see that for most of the least complex cities, the actual GENEPY index exceeds the predicted GENEPY index.

2.5 Discussion

We demonstrate the use of the geocoded patent database in examining the economic complexity of cities and technologies, and to forecast the competitiveness of cities in new technologies using machine learning models and network models. Our findings suggest that over time Asia cities have

become more economically complex with top movers being cities such as Bangalore and Istanbul. We also identify that there is stability in the complexity of technologies, which suggests that capabilities needed to be competitive in technologies do not vary over time. Also, in examining the complexity of the two most prominent cities in North America and Asia - San Francisco and Shanghai, across different technologies, we observe that Shanghai appears to be a global leader in more technological areas when compared to San Francisco (see Figure 32 in Appendix B.0.1).

Finally, we forecast the competitiveness of cities in new IPCs by considering different machine learning models - Random Forest, XGBosst, SVM, and Neural net, alongside a discriminated time-independent model (the benchmark model). We find that for the full matrix prediction task, all machine learning models were better than the benchmark model, while for the more difficult task - activation task, only the Random forest model was better than the benchmark model. Specifically, the Random forest model had an F1 score 4.5% higher than the benchmark. Given that there was no hyper-parameter tuning done to select optimal parameters for the machine learning models, and that the machine learning model is built such that it is not allowed to exploit auto-correlation like the benchmark model, and yet had better performance for both the full matrix task (all machine learning models), and activation matrix task (only Random forest model), it goes to show that machine learning approaches can be better at extracting information relevant to determining technologies that will be activated in a given city in the future. The major contribution of this work involves the use of machine learning models to predict competitiveness. These ML models provide a way to uncover relationships between technologies that are beyond linear and pairwise. This is an improvement in the way notions such as technological relatedness have been measured in previous works (e.g., (Pugliese et al., 2019; Engelsman and Raan, 1992)). The technological relatedness uncovered by the machine learning models can be combined with the understanding of the current competitive landscape of a city to forecast future competitiveness. Similar machine learning approach has been applied to forecast trade competitiveness of countries (Tacchella et al., 2021) and firm entry into new technology IPCs

(Straccamore and Zaccaria, 2021). Considering that predicting future competitveness condition on having low RTA can be a more difficult task than predicting future entry, our model still slightly outperforms Straccamore and Zaccaria, 2021. Generally, we find that in our work and other literature (Tacchella et al., 2021; Straccamore and Zaccaria, 2021) the predictive performance is low especially when prediction is done condition on very low RTAs (in our case $RTA < 0.25$) - this is largely because such jump in competitiveness are rare events. Future work can improve upon the quality of the models by consolidating patent or trade features with other policy, economic and industrial indicators.

Being able to predict what technological areas a city or region will most likely become competitive in the future can better help the region align itself with other regions that can facilitate the realization of this goal. For example, such regions can put in place policy instruments that incentivize firms and other economic agents to engage in projects consistent with being competitive in the identified technological areas. Also, they can identify the technological pieces missing in their technological portfolio which are pre-requisite for competitively upgrading into another technology (Balland and Rigby, 2022). Finally, the predicted competitiveness structure in the full matrix task is used to reconstruct the future generalized economic complexity index. The result shows that we can reasonably reconstruct the future generalized economic complexity index of cities, but this is more challenging for economically complex cities. The challenge is that this measure depends on being able to correctly predict the competitiveness structure (i.e., M_{ij}), which can be difficult for economically complex cities that are dynamic and do have changing competitive landscape.

Chapter 3

Exploratory Innovation in Cities: Inter-city ties and Technological relatedness

In a competitive economy, technological innovation is a core element of economic growth and development, and the accumulation of it in a rapidly changing technological environment is key to adaptation. Such accumulation of technological capabilities leads to distinctions in the specialization and diversification patterns observed in countries, regions, and cities, hence the emergence of different capabilities. This disparity in competencies has an impact on the roles that different regions play in the global value chain (Balland et al., 2019). To capture a greater share of the value-added in the global value chain, regions develop strategies that enhance their ability to introduce novel ideas or recombine already existing ideas to foster new technological trajectories. This search for new technological trajectories constitutes what is known as "exploration" which according to March, 1991 is defined as an innovation process characterized by "search, variation, risk-taking, experimentation, flexibility, and discovery". Exploratory innovation specifically captures the pursuit of new technology, as opposed to exploitative innovation, which is characterized by the development of existing technology stock. The pursuit

of new knowledge has been seen to lead to the accumulation of diverse knowledge, hence increasing the diversification of a region, and also the scope of utilizing a technology (Levinthal and March, 1993; Lavie, Stettner, and Tushman, 2010).

For economic geographers, there has been increasing interest as to understanding the diversification and evolution of the technology base of regions over time and the environmental characteristics facilitating the upgrade to more fruitful technological areas. This diversification process results in regions accumulating specialized technology(ies) that may be located in the "technology space". The technology space is seen to provide insights into the interaction of technology pieces and the association of regions with specific technology, which reflects communities of similar innovation processes (Maskell and Malmberg, 1999; Lawson and Lorenz, 1999; Gertler, 2003). The main feature of this technology space for different regions is that they are irreversible and are path-dependent (Rigby, 2015). While some regions "lock-in" to technologies with declining value, others are better able to explore new technological areas thereby diversifying their technology base and increasing their capacity to innovate through more related technological structures or connections to external knowledge (Bathelt, Malmberg, and Maskell, 2004; Agrawal, Kapur, and McHale, 2008; Rigby, 2015; Breschi and Lenzi, 2016). Although the argument of related technological architectures and social linkages to external knowledge are reasonably well-accepted there is little empirical evidence at the level of cities that jointly examine the impact of these mechanisms (i.e., social linkages and technological relatedness) on the likelihood of exploring new technologies.

Recent empirical work that examines the relationship between technological relatedness and exploration of new technology illustrates a positive correlation between them and shows a long-run relationship between the size of countries and the diversification of knowledge (Cantwell and Vertova, 2004; Hausmann, Hwang, and Rodrik, 2007). Detailed analysis of this by Hidalgo et al., 2007 and Hidalgo and Hausmann, 2009 using trade data, measured relatedness based on patterns of co-exporting of products

depicted in the "product space." Their findings suggest that specialization in one product hinders diversification into related items and that the likelihood of a country starting to export a product rises in proportion to the number of related products the country already exports. Furthermore, in studying the exploratory activities of Korea and Chile, Hidalgo et al., 2007 simulated different thresholds of relatedness needed to explore new products, in which case Korea was shown to be more effective in exporting new products at higher thresholds. This simulated analysis confirms the theoretical notion of the effect of degree of relatedness on the likelihood of exploration (Lavie, Stettner, and Tushman, 2010). At the regional level, the notion of relatedness has been extended to explain industrial branching and technological cohesion (Essletzbichler, 2015), path dependence in the evolution of Sweden's and China's industrial landscape (Neffke, Henning, and Boschma, 2011; Wu et al., 2019), and the diversification of US cities to technology class where they have related technological expertise (Boschma and Frenken, 2011; Boschma, Minondo, and Navarro, 2013). In examining how industrial regions evolve in the long-term, these studies highlight that the likelihood that a new industry will enter a region has a direct relationship with its relatedness with other industries in that region. However, only a few of these studies examined have accounted for characteristics of cities such as their specialization, or density, that may impact branching (Rigby, 2015). Beyond the study of industrial branching into regions, several studies have examined the significant role of knowledge variety and relatedness on productivity growth across European regions (Balland et al., 2019; Balland and Boschma, 2021; Colombelli, Krafft, and Quatraro, 2014). In analyzing entry made by European regions, Balland et al., 2019 show that in increasing the relatedness density of a region by 10%, the likelihood of entry increases by 23-26%, and an increase in relatedness by 10 points results in technological growth from 2-4.64%. This shows that technical expansion is constrained by relatedness, necessitating the importance for regions to enhance their capabilities in new technological areas by leveraging already existing capabilities.

Although, entry into new technology domains is not only conditioned by relatedness, as there is also an effect of inter-city ties since cities and

regions are not necessarily independent spatial units. The exploration of new technology by agents in cities are influenced by exploratory activities and technology structures in other cities, as there is an increasingly strong social linkage between economic actors (inventors, firms, and universities) across cities through co-inventions (Breschi and Lenzi, 2016; Agrawal, Kapur, and McHale, 2008; Singh, 2005). When there are fewer intermediates between economic participants in a collaboration network, there is a rapid flow of less distorted information between participants, as opposed to when participants are connected by longer chains of relationships in the network (Breschi and Lenzi, 2016). Therefore, new technology generated outside the city can rapidly flow to economic actors within the city through existing social ties with economic actors outside the city who have pre-requisite understanding of the new technology. The existence of such ties can effectively improve the inventive productivity of cities (Breschi and Lenzi, 2016).

Despite the extensive empirical work, we still have a limited understanding of how cities explore or enter new technological domains. Quantitative studies examining patterns of technological exploration at the level of cities are few (except Boschma, Balland, and Kogler, 2015) and are oftentimes restricted to the study of European or U.S. cities. Therefore, in this chapter we focus on technological entry done by 1220 cities (these cities are defined as functional urban areas i.e. they encompass the full extent of the city's labor market) in 646 4-digits International Patent Classifications (IPCs). These cities are located in Europe (56.4%), North America (25.8%), Asia (10.9%), South America (4.6%), Oceania (1.6%), and Africa (0.6%). The analysis of these numbers of cities provides robust and comprehensive evidence on the general pattern of entry that is not necessarily specific to developed economies. Specifically, we ask if inter-city ties and relatedness condition technological entry. This chapter contributes to the geography of innovation literature by providing insights into the enablers of technological entry done at the level of cities. The contributions include (i) analysis of technological entry considering a more heterogeneous collection of cities. This is possible due to the effort in developing a novel geo-referenced PATSTAT database in Chapter 1

that provides more granular information of inventor activities in cities (ii) the joint analysis of technological relatedness and inter-city ties as a determinant of the likelihood that a city will enter a new technology domain (iii) and the moderating role of partner city size and focal city size on the effect of inter-city ties and technological relatedness respectively. Furthermore, we provide a supplementary analysis showing that entering a new technology collaboratively increases the performance of the city in the new technology.

3.1 Theory and Hypotheses

In this section, we give an overview of the literature on relatedness and the literature on inter-city collaboration. We proceed by suggesting two primary propositions: (i) there is a positive relationship between the ties of city to external inventors (who have engaged in new technology) and the likelihood that the city will enter the new technology (ii) there is a positive relationship between the relatedness of a city's technological base to a new technology domain and the likelihood that the city will enter the new technological domain. Furthermore, we propose a positive moderation of the size (i.e. number of inventors located in the city) of the partner city on the effect of inter-city ties on entry and a positive moderation of the size of the city (which we call focal city size - the number of inventors in the city) on the effect of technological relatedness on entry into new technology domain.

The focus of this work in analyzing the effect of inter-city ties and technological relatedness on exploratory activities of economic agents in urban areas as opposed to exploitative activities is based on the fact that exploration require some level of flexibility and choices geared towards long term innovation of technological piece that can later be exploited. These different modes of technological entry (i.e., exploitative and exploratory) are usually not at odds with each other as they are often combined such that one reinforces the other (Lavie, Stettner, and Tushman, 2010). Although the organizational literature has shown how exploratory activities

of firms lead to long term performance and exploitative activities lead to short term performance, the balance of both activities can enhance the performance outcome in the short-long term depending on the mode of balancing (Lavie, Stettner, and Tushman, 2010). The preference for activities that enhance long-term performance can be useful to hedge lock-in-effect economic agents may experience by choosing to refine existing technologies.

3.1.1 Inter-city ties in co-invention network.

The ease with which cities explore new technology in different fields do have implications for innovation and growth (Grossman and Helpman, 1991). Many existing theories of diversification into new technology emphasize external effects and the resulting agglomeration economies (Krugman, 1991; Saxenian, 1994; Porter, 1998). The dense web of social relationships in cities, as Glaeser and Gottlieb, 2009 points out, creates agglomeration economies, which can lead to significant variance in innovative events throughout time and space. For example, co-locating creative individuals in the same area is attributed with encouraging interactions in which "tacit knowledge" crucial to imaginative creativity is conveyed and exchanged (Breschi and Lenzi, 2016). This web of connections fosters the flow of localized knowledge and ideas between inventors and firms, thereby boosting the innovative output of all local actors (Breschi and Lenzi, 2016; Jaffe, Trajtenberg, and Henderson, 1993). Recent work identifies the different structure of local network that enhances access to new technology to include spatial proximity, and clique density (Breschi and Lenzi, 2016). The argument for social proximity is that inventors are more ready to share information with persons living in close proximity based on trust, and because they believe there is a higher chance of reciprocation (Agrawal, Kapur, and McHale, 2008; Breschi and Lenzi, 2016). On the other hand, when inventors are part of a tight-knit group, in which there is an extensive collaboration among inventors, this facilitates the faster spread of information in the network, allows for multiple pathways for the verification of the usefulness of new knowledge (Schilling

and Phelps, 2007), and the monitoring of opportunistic behaviors among actors (Breschi and Lenzi, 2016).

However, the reliance of cities on the local structure of their network to develop new ideas may lead to loss of position in global urban ranking (Burt, 2004; Neal, 2011). Also, since economic actors in the same cities tend to converge towards a homogeneous pool of the same knowledge, this will decrease the search space of economic actors in cities to develop new varieties and increase the tendency of cities to get locked into specific path of technological process. External relationships to inventors in other locations can therefore bring significant information that may not be available within the focal city, hence broadening the scope of the focal city's search space (Breschi and Lenzi, 2016). These external relationships can provide exposure to specialized skills and human capital, new information on market opportunities, and access to a larger stock of technological solutions (Breschi and Lenzi, 2016; Kerr, 2010; Owen-Smith and Powell, 2004). Furthermore, these direct external ties of cities imply there is an existing collaboration between economic actors in the city and the direct external inventors, as such there is an established professional relationship between them that ensures mutual trust, familiarity with working relationship, the existence of learning structure, and guaranteed reciprocity. More importantly, these external inventors have been engaged in inventive processes in the new technology domain, hence are repositories of knowledge in the domain. They understand the scope and use of the new technology and can inform the inventors in the focal city on how best to appropriate this new technology in a way that reduces search and experimentation cost, and can eventually lead to economic actors in the focal city successfully innovating in the new technology.

Hypothesis 1a: The ties of a city to external inventors engaged in a new technology domain is positively associated with the likelihood that a city will explore the new technology domain.

Nevertheless, these external inventors do not exist independent of the innovative capabilities and other inventors in the city it is located in, as

recent literature has shown that knowledge spillovers between inventors and firms in a city can be facilitated by the cross-fertilization of ideas being made possible by the city. The city plays a social and economic role in attracting best talent and creating social space for cross-fertilization of ideas that leads to knowledge spillovers between inventors located in the city, subsequently, the increased capability of the inventor to innovate leads to more innovative production and economic growth of the city. This self-reinforcing relationship between the inventor and city can be useful especially when the city has a large pool of inventors located in the city (i.e., large "inventive size"). The external inventor of a focal city being embedded in a city with large pool of inventors do have a second-order effect on the focal city's ability to explore new technology in two ways: (i) it can refine the quality of technological information the city receives from its ties by virtue of a more active social interaction or "local buzz" taking place in the partner city. This local buzz can allow for quick validation of technological flows (ii) increased opportunity to access a global pipeline of heterogeneous inventors around the world, in which the formation of links could provide information of new technologies.

Hypothesis 1b: The association between inter-city ties and the likelihood that a city enters a new technology domain is positively moderated by the size of the partner cities.

3.1.2 Principle of relatedness

Existing views pose that technology growth is path-dependent, and as such, there is a need for inventors, firms, cities, or countries to examine the cost and benefit of entering into new technologies taking into account their existing capabilities (Boschma, Balland, and Kogler, 2015; Essletzbichler, 2015; Petralia, Balland, and Morrison, 2017). In the view of Frenken, Van Oort, and Verburg, 2007, the extent to which technology found in different regions are related is more important than the overall stock of technology, suggesting the need for a path-dependent innovation process for economic growth and development. The relatedness concept relies on the premise that technology has a structure based on similarities

and variations in how different forms of technology might be applied (Petralia, Balland, and Morrison, 2017). Based on this concept, proximate technology in what is called the "technology space" do share similar sets of cognitive capabilities. These proximate technologies are close substitutes, and the shared capabilities imply that a region competent in one should be competent in the other (Breschi, Lissoni, and Malerba, 2003; Balland et al., 2018). The exploration of proximate technologies by cities is viewed as a "higher-order reflection" of the dynamics at the micro-level, where inventors and firms located in these cities expand the scope of their activities around competencies endowed in the cities (Balland et al., 2017; Balland et al., 2018). Therefore entry into a technology is not random but rather consistent with cities' technology profiles. These economic actors in regions compete through exploration of new technology to increase their technology stock for further recombination. To do this requires a search cost that increases around the cognitive distance between the technology to be explored and current expertise (Atkinson and Stiglitz, 1969). Petralia, Balland, and Morrison, 2017 identified two main channels through which the proximity of existing expertise to new technology affects economic actors' possibilities for technology exploration: "economies of scope" and "absorptive capacity".

With respect to the "economies of scope", related knowledge do share common scientific principles, and as such having one of the technology in the technological stock increases the possibility of exploring the other. This is simply because the cost of capability acquisition is eliminated or reduced given the scientific experience in innovating in the other technology (Breschi, Lissoni, and Malerba, 2003; Penrose and Penrose, 2009; Petralia, Balland, and Morrison, 2017). "Absorptive capacity", on the other hand, relates to how past technology helps economic actors to perceive the importance of new knowledge, the challenges that can be envisaged while undertaking innovation in the new technology, how best to assimilate it and combine it with existing technology piece, and how to utilize it for commercial purposes (Cohen and Levinthal, 1990; Petralia, Balland, and Morrison, 2017). It is the case, that a higher "absorptive capacity" increases the ability to arrive at an informed decision on entry into new

technology (Petralia, Balland, and Morrison, 2017).

Hypothesis 2a: The relatedness between a new technology domain and city's existing technology stock is positively associated with the likelihood that the city enters into the new technology domain.

The endowment of technological pieces in a city is not only the outcome of economic actor's learning process and accumulation of capabilities, it is also determined by larger systems like norms and institutions in cities that support them (Niosi et al., 1993; Edquist and Lundvall, 1993; Patel and Pavitt, 1994; Metcalfe, 1995). Not only do the potential and costs of investigating new technologies depend on the inventive environment, but so do how economic players perceive opportunities and estimate costs (Petralia, Balland, and Morrison, 2017). Many studies have shown differences in the quality of capabilities endowment in countries and regions do influence diversification activities of economic actors (Furman, Porter, and Stern, 2002). The scope and quality of capabilities in a geographic domain affect the economic actor's abilities to quantify the benefit and cost of new technological pieces and appropriate economics of scale. Also, the lack of indigenous capabilities will put upward pressure on the cost of exploration, since economic actors will be forced to invest in technological accumulations which might delay the acquisition of new technology. Therefore access to a pool of technological actors strengthens the economic actor's capacity to produce and provide more technological space that can be explored thereby avoiding "lock-in traps" (Levitt and March, 1988; Levinthal and March, 1993). Furthermore, the ability of economic actors in a city to adequately interpret technological flow coming into the city is hinged on the inventive size of the city, this is because the co-location of creative individuals within the city will allow for formal and informal interaction on the scope of the new technology. This network of relationships allows for localized technological flows between inventors within the city and ensures rapid diffusion of ideas within the city (Jaffe, Trajtenberg, and Henderson, 1993).

Hypothesis 2b: The association between a city's technological relatedness and the

likelihood that the city enters into a new technology domain is positively moderated by the size of the city.

3.2 Data and Sample

Based on the dataset developed in Chapter 1, we examine the technological activities of 1220 cities located in Europe, North America, Asia, Oceania, Africa, and South America between the period 2005-2014. This list of cities includes both global and non-global cities in developed and developing countries, and they account for over 50% of the patent activities observed between 2005-2014. The list of cities is partly drawn from the OECD, 2012 list of functional urban areas (FUA) which define a FUA to "consist of densely inhabited city and of a surrounding area (commuting zone) whose labor market is highly integrated with the city". Also, of the 1220 cities, 78 cities in non-OECD countries were drawn from the GaWC research project (Beaverstock, Smith, and Taylor, 1999). These cities are either alpha cities or beta cities (see Chapters 1 and 2 for definition). These 78 cities are regionalized similarly as the functional urban areas selected from OECD, 2012. On average each country does have approximately 16 cities. The United States has the highest number of cities on the list with 211 cities. At the level of continents, 56.4% of the cities are in Europe, 25.8% in North America, 10.9% in Asia, 4.6% in South America, 1.6% in Oceania, and 0.6% in Africa. Table 13 shows the number of 4-digit IPC in each of the 34 technology fields. These technology fields classification is the same as that used by EPO and WIPO for their statistics. Each of the 646 4-digit IPCs considered in this study belongs to one technology field. From Table 13 we have that Textile and paper machines do have fifty 4-digit IPCs, representing the technology field with the most number of 4-digit IPCs. While Semi-conductors do have only one 4-digit IPC (i.e., HO1L).

We used the georeferenced patent data developed in Chapter 1 to construct all the variables in this analysis. The georeferenced patent data draw from multiple patent sources which are available publicly and doc-

	Technology fields	*Number of 4-digit IPCs*	*Technology fields*	*Number of 4-digit IPCs*
0	Audio-visual technology	6	Macromolecular chemistry, polymers	7
1	Basic communication processes	10	Materials, metallurgy	16
2	Basic materials chemistry	33	Measurement	25
3	Biotechnology	8	Mechanical elements	21
4	Chemical engineering	23	Medical technology	12
5	Civil engineering	31	Micro-structural and nano-technology	4
6	Computer technology	12	Optics	10
7	Control	13	Organic fine chemistry	8
8	Digital communication	2	Other consumer goods	49
9	Electrical machinery, apparatus, energy	30	Other special machines	39
10	Engines, pumps, turbines	35	Pharmaceuticals	2
11	Environmental technology	9	Semiconductors	1
12	Food chemistry	19	Surface technology, coating	12
13	Furniture, games	17	Telecommunications	10
14	Handling	13	Textile and paper machines	50
15	IT methods for management	1	Thermal processes and apparatus	30
16	Machine tools	43	Transport	45

Table 13: 4-digit IPC aggregation into technology fields.

ument information on the content of inventions done by persons or firms. Some shortcomings have been identified in using patent databases of this kind for analysis: (a) patented propensities vary across cities and technology fields, (b) patent databases often do not capture all inventions happening in a location, as not all inventions are patented, (c) there are differences in the economic and technical values of inventions (Griliches, 1990). The first concern implies there is a need to control for city and technology field specific effects, and the other concern on technology and economic value only concerns the supplementary analysis done on post-entry performance. Another disadvantage specific to our georeferenced database is the differences in the quality of georeference work done across countries where cities are located, and this data quality needs to be controlled for. Despite these limitations, the use of patent indicators are an approximation for technological indicators in cities, and has been used in different empirical work on regional innovation (Rigby, 2015; Breschi and Lenzi, 2016; Vlčková, Kaspříková, and Vlčková, 2018; Balland et al., 2020; Balland and Boschma, 2021) and some of these patent indicators correlate with important economic growth indicators of cities like employment rate, and GDP (Mewes and Broekel, 2020). In our case, patent applications provide a good measure of technological search done by inventors in cities, and these can also be seen as an indication that inventors in the city

are pursuing capacity development in a specific technological area. This effort can either lead to a successfully granted application or otherwise.

3.3 Measures and Method

The dependent variable "entry" into technology class is measured using the geo-referenced patent statistical database developed in Chapter 1. Patents in the database are originally classified in at least one of the 64,000 eight-digit technology classes based on the IPC system. These 64,000 IPCs represent a particular technical function (Leten, Belderbos, and Looy, 2016), and can be aggregated into 646 4-digit IPC classes. The aggregated IPCs into 4-digit is what is being used in this study. Also, we considered 1220 cities given the available number of cities in the georeferenced patent statistical database developed in Chapter 1. At the beginning of the study i.e., 2005, cities do have a heterogeneous number of IPCs available to explore (see Figure 21a). Table 14 provides an overview of the top 100 cities with the highest percentage of exploration done given their exploratory space.

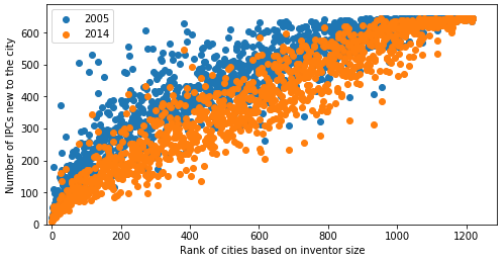
From Table 14 we have that cities in Asia predominantly explored more IPCs in the period considered. For example, Shenzen explored 71.27% of the IPCs available in its exploratory space. In Europe, we have that Berlin explored the most number of IPCs, exploring 59.06% of the IPCs available in its exploratory space as at the beginning of the period examined (i.e., 2005). The observation is that there is a significant variation in how many IPCs are available to be explored by cities. As of 2005, a city like San Francisco only had 54 IPCs to be explored, while Beijing had 77 IPCs to be explored. We also observe more variation in exploratory space between cities in the United States (149.2) compared to cities in China (107.1).

	Cities	Country	New IPCs	Explored IPCs	Relatedness	External ties
0	Shenzen	China	181	0.7127	0.0108	0.0969
1	Hangzhou	China	223	0.6951	0.0105	0
2	Shanghai	China	85	0.6824	0.0131	0.1034
3	Chongqing	China	313	0.6805	0.0091	0
4	Dalian	China	287	0.6655	0.0089	0
5	Jinan	China	343	0.6647	0.0085	0
6	Xian	China	260	0.6615	0.0085	0
7	Guangzhou	China	120	0.6583	0.0132	0
8	Seo	South Korea	100	0.65	0.0191	0.0769
9	Dalseong	South Korea	135	0.6444	0.0125	0.0345
10	Deokjin	South Korea	236	0.6441	0.0104	0
11	Sebuk	South Korea	260	0.6385	0.0097	0.0361
12	Heungdeok	South Korea	257	0.6381	0.0092	0.0244
13	Seoul	South Korea	30	0.6333	0.0212	0
14	Chengdu	China	232	0.6207	0.0091	0
15	Fuzhou	China	407	0.6192	0.0078	0
16	Gimhae	South Korea	105	0.6095	0.0148	0
17	Gwangsan	South Korea	184	0.6087	0.0115	0.0268
18	Seongsan	South Korea	230	0.6087	0.0116	0.0071
19	Suncheon	South Korea	313	0.6006	0.009	0.016
20	Gumi	South Korea	275	0.6	0.0099	0.0121
21	Wuhan	China	251	0.5976	0.0092	0
22	Taipei	Taiwan	109	0.5963	0.0146	0.0615
23	Nam	South Korea	172	0.593	0.0128	0.0196
24	Berlin	Germany	127	0.5906	0.0138	0
25	Dresden	Germany	211	0.5877	0.0107	0.0081
26	Bangalore	India	372	0.586	0.008	0.0367
27	Beijing	China	77	0.5714	0.0123	0.0455
28	Tokyo	Japan	21	0.5714	0.0182	0
29	Shenyang	China	226	0.5708	0.0096	0
30	Gyeongsan	South Korea	417	0.5659	0.0078	0.0042
31	Vancouver	Canada	172	0.564	0.0114	0
32	Zurich	Switzerland	158	0.557	0.0125	0.0114
33	Hefei	China	419	0.5561	0.0073	0
34	Vienna	Austria	207	0.5556	0.0112	0
35	Hanover	Germany	236	0.5551	0.0098	0
36	Munich	Germany	74	0.5541	0.0171	0.0488
37	Stuttgart	Germany	87	0.5517	0.014	0.125
38	Jinju	South Korea	322	0.5466	0.0091	0.0227
39	Leipzig	Germany	341	0.5455	0.0079	0.0161
40	Pune	India	507	0.5365	0.0068	0.0037
41	Cuyahoga	United States	140	0.5357	0.0113	0.0133
42	Wake	United States	183	0.5355	0.0099	0.0102
43	Dusseldorf	Germany	146	0.5342	0.0123	0
44	Charlotte	United States	190	0.5316	0.0111	0
45	Minneapolis	United States	79	0.5316	0.0152	0.0238
46	Changsha	China	322	0.5311	0.0082	0.0117
47	Zhengzhou	China	373	0.5308	0.0079	0

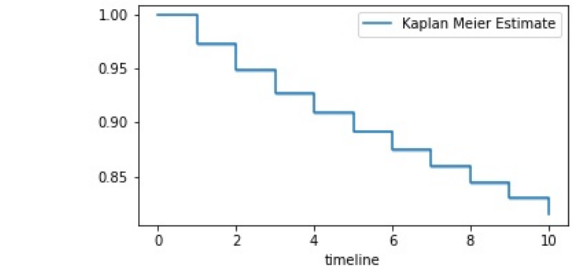
	Cities	Country	New IPCs	Explored IPCs	Relatedness	External ties
48	Istanbul	Turkey	463	0.5292	0.0073	0
49	Xiamen	China	386	0.5259	0.008	0
50	Detroit (Greater)	United States	97	0.5258	0.0138	0
51	Frankfurt am Main	Germany	111	0.5225	0.0145	0
52	Los Angeles (Greater)	United States	48	0.5208	0.0186	0.08
53	Taichung	Taiwan	169	0.5207	0.0104	0
54	Warsaw	Poland	412	0.517	0.008	0
55	Karlsruhe	Germany	189	0.5132	0.0117	0
56	Gangneung	South Korea	435	0.5126	0.0072	0.0045
57	Miami (Greater)	United States	125	0.512	0.0121	0
58	Cologne	Germany	155	0.5097	0.0116	0.0127
59	Ingolstadt	Germany	294	0.5068	0.009	0
60	Hamburg	Germany	119	0.5042	0.0128	0
61	Qingdao	China	365	0.5041	0.0083	0.0054
62	New Delhi	India	478	0.5021	0.0067	0.0042
63	Atlanta	United States	102	0.5	0.0169	0
64	Dallas	United States	90	0.5	0.016	0.2222
65	Ruhr	Germany	112	0.5	0.0123	0
66	Iksan	South Korea	349	0.4986	0.0085	0.0057
67	Grenoble	France	227	0.4978	0.0098	0.0177
68	Helsinki	Finland	165	0.497	0.0105	0.0244
69	Chuncheon	South Korea	379	0.496	0.0079	0
70	Cincinnati	United States	123	0.4959	0.0128	0.0492
71	Constance	Germany	303	0.495	0.008	0
72	New Haven	United States	120	0.4917	0.0149	0.0593
73	Columbus	United States	171	0.4912	0.011	0
74	Wonju	South Korea	387	0.491	0.0082	0.0105
75	Houston	United States	108	0.4907	0.0126	0.0189
76	TelAviv	Israel	147	0.4898	0.0111	0.0139
77	Seattle	United States	96	0.4896	0.0159	0.0851
78	Chennai	India	510	0.4882	0.0073	0
79	Nuremberg	Germany	138	0.4855	0.0126	0
80	Edmonton	Canada	327	0.4832	0.0097	0
81	Madrid	Spain	292	0.4829	0.0082	0
82	Heidelberg	Germany	164	0.4817	0.0106	0.0253
83	Jeju	South Korea	432	0.4815	0.0078	0
84	Tulsa	United States	133	0.4812	0.0119	0.0156
85	Washington (Greater)	United States	79	0.481	0.0206	0.0526
86	Greater Sydney	Australia	192	0.4792	0.0102	0.0217
87	Sao Paulo	Brazil	350	0.4771	0.0082	0.006
88	Greenville	United States	250	0.476	0.0086	0.0336
89	Aachen	Germany	225	0.4756	0.0092	0.0093
90	Portland	United States	143	0.4755	0.0121	0.0147
91	Hartford	United States	162	0.4753	0.0118	0
92	Mannheim-Ludwigshafen	Germany	148	0.473	0.0109	0.0429
93	Calgary	Canada	282	0.4716	0.0086	0
94	Basel	Switzerland	293	0.471	0.0083	0.0072
95	San Diego	United States	102	0.4706	0.012	0.0833
96	Toronto	Canada	85	0.4706	0.0149	0.1
97	Phoenix	United States	137	0.4672	0.0116	0.0625
98	Boulder	United States	213	0.4648	0.009	0
99	San Francisco (Greater)	United States	54	0.463	0.0185	0.4

Table 14: Top 100 cities: These are the top cities that explored most of the available IPCs new to them. The number of explored IPCs are presented as a percentage of the IPCs available for exploration at the beginning of the study period i.e. 2015.

3.3.1 Exploring new technology domain



(a)



Kaplan Meier Estimate						
At risk	571916	540161	514960	487344	454369	0
Censored	0	2558	5110	13453	29447	468867
Events	0	29197	51846	71119	88100	103049

(b)

Figure 21: (a) **Exploratory space vs Cities rank:** This shows the number of IPCs new to the city in 2005 and 2014. The cities are ranked (from left to right) based on the number of inventors located in the city in 2005. We see that highly ranked cities do have lower exploratory space, and exploratory space for cities in 2014 is smaller than what was available in 2005. (b) **Kaplan Meier plot:** The y-axis is the probability of cities not entering a technology domain after time t.

For each of the 1220 cities, we examine entries made to IPCs that are new to the city between the period 2005-2014. An IPC is said to be new to the city in year t, if no economic actor in the city has produced a patent in that IPC during the last 5 years. The choice of 5 years window as a yardstick

for "newness" of IPC to the city is consistent with the literature (Leten, Belderbos, and Looy, 2016; Ahuja and Morris Lampert, 2001; Belderbos et al., 2010). The assumption is that economic actors not being active in an IPC leads to a depreciating knowledge base of the city in that IPC. Figure 21a shows the distribution of the number of IPCs new to the city as of 2005 and 2014 as a function of the rank of cities based on inventor sizes. Generally, we observe that cities with a large number of inventors do have few IPCs to explore.

The dataset used in the analysis of the likelihood of entering a technology consists of all city-IPC combinations that are new to the 1220 cities. Before the study period, the cities had patents in 177 IPCs on average based on the patents produced between 2000-2004. This implies that cities had 469 IPCs on average available for exploration, resulting in a dataset with 571,916 observations between the period 2005-2014 representing new to the city IPCs which will be potentially explored by the city. From Figure 21b, we observe that exploration was done in 18.02% of the observations i.e. 103,049. This exploration was done by 1217 cities in all the 646 IPCs. On average each of these 1217 cities explored 85 IPCs. Also, we observe that a city is less likely to enter a new technology domain the closer we get to the end of the period considered in the analysis. The cohort of cities entering a new technology is steady across time. For any given year, entry is observed for approximately 2% of the dataset. The cumulative density of entry observed in the data suggests that most entry was done in the second year.

The dependent variable "entry" takes the value 1 if a city starts to explore a new technology domain by filing for a patent instantiated by the IPC corresponding to the technology domain, and 0 if the city is yet to explore or patent in the new technology domain. When exploration is done by a city in a given IPC, such IPC is no longer considered for exploration in the city - in essence, we do not consider cases of re-entry in subsequent years. The identification of time of entry might be impacted by censoring. For example, it might be the case that entry for some city-IPC combinations were not observed due to the limited study period, and the possibility

of entry after the study period is unknown. The latter which is referred to as right censoring is taken into account in the empirical model. If we consider the year 2004 as the first year at risk, we have that on average successful exploration into IPCs took cities approximately 4 years and 10 months. For cities in Asia, North America, Europe, Oceania, South America, and Africa the average time spent by cities in exploring an IPC are 4 years and 6 months, 4 years and 8 months, 4 years and 10 months, 4 years and 10 months, 5 years and 7 months, 5 years and 4 months respectively.

3.3.2 Measure of technological relatedness

Scherer, 1982's method for constructing inter-industry technology relatedness has been the most influential way of capturing technological relatedness. In this methodology, a link exists between two industries, if a significant portion of the R&D carried out in one industry is employed in the other industry (Breschi, Lissoni, and Malerba, 2003; Scherer, 1982). Another approach in the relatedness literature is that proposed by Jaffe, 1989 that captures the relatedness in the technological structure of firms using the cosine index to measure the correlation of firm's technological structure in different industries (Breschi, Lissoni, and Malerba, 2003).

In this study, we are interested in measuring the relatedness of the technology base of cities with new technology available to the city for exploration. Particularly, we identify that not all technology pieces in a city's technology base are important as some of them will be distant from the new technology in the technology space. Therefore we look at the relatedness measure of technology pieces in the cities technology base that are directly connected to the new technology in the technology space. We start by first computing a relatedness measure for each pair of IPCs in a given year. We adopted the Engelsman and Raan, 1991 approach for measuring relatedness as the number of co-occurrence of pair of IPCs on all patents in that given year. The premise for this is that two IPCs co-occurring on a patent document is a signal of similarity in scientific principles and there is a possibility of spillovers between the IPCs (Breschi, Lissoni, and

Malerba, 2003).

Formally, let M be the universe of patents at a given year t. Let $F_{im} = 1$ if patent document $m \in M$ contains the 4-digit IPC code $i = 1, 2, ..., 646$, and $F_{im} = 0$ otherwise. The number of patents with IPC i in year t is given by $N_{it} = \sum_m F_{im}$, while the number of patents classified in both IPCs i and j in year t is $C_{ijt} = \sum_m F_{im}F_{jm}$. Therefore, we have a co-occurrence square matrix $\mathbf{C(t)}$ for year t with size 646×646. This matrix is used to derive a relatedness measure for cities technology base (cities' IPCs stock) to new technology (new IPCs), in such a way that the measure does not depend on the size of patents in the IPCs, otherwise we will be overestimating the linkages for technology piece or IPCs with larger sizes. The relatedness measure for a city's technology base $\mathcal{I}(t)$ to a technology k new to the city at any given year t is given as:

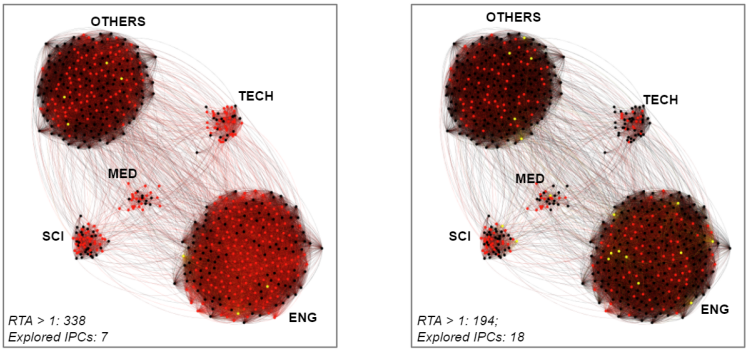
$$R_{ikt} = \frac{1}{|\mathcal{I}(t)|} \sum_{i \in \mathcal{I}(t)} \frac{C_{ikt}}{\sqrt{N_{it}N_{kt}}} \tag{3.1}$$

Equation 3.1 is known as the cosine similarity measure. The equation shows the relatedness of IPCs after adjusting for the number of patents produced in an IPC in a given year (i.e., size effect). The measure intends capturing the degree of similarity between a city's existing technology stock with the new technology. This measure is lagged one-year when introduced into the empirical model.

Other propositions for capturing technological relatedness include the work done by Leten, Belderbos, and Looy, 2016. They computed relatedness between technology pieces by considering citation patterns between technologies. A pair of technology domains is deemed related if the citation pattern between patents in the two domains is more than random. These citations across domains are indicative of common heuristics (Leten, Belderbos, and Looy, 2016). Also, recent efforts in constructing relatedness of products in the trade literature propose measures of relatedness relying on the use of machine learning models to build relatedness matrix that is beyond an indicator of shared knowledge but more importantly a matrix that is predictive of what product is feasible to a country considering an

out-of-sample prediction task (Tacchella et al., 2021). As demonstrated in Chapter 2, such machine learning based embedding techniques obtained by considering higher-order interactions through complex but less interpretable models are shown to provide better results in terms of predicting feasible product space of countries.

Figure 22 shows the technological communities of IPCs identified based on the technological relatedness matrix in 2014. The nodes are IPCs and the edges are a measure of relatedness between IPCs, reflecting the common heuristics and scientific capabilities between IPCs. The technology relatedness matrix is almost fully complete, hence we prune the matrix for better visualization by discarding edges with weight less than the 65th percentile of the distribution of the weight of edges.



(a) Vancouver (b) Rio de Janeiro

Figure 22: Relatedness of IPCs in 2014. First, we identify communities of IPCs based on the relatedness matrix. The black nodes represent IPCs the city has no comparative advantage in, while the red nodes represent IPCs where the city does have a comparative advantage, and the yellow nodes are the IPCs the city explored.

The Louvain community detection algorithm (implemented in Gephi) is applied to the matrix to identify communities of IPCs. We observe that there are five communities optimally identified by the community detection algorithm. These communities are: (i) scientific fields (SCI) which include predominantly IPCs in "Basic materials chemistry", "Food

chemistry", "Macromolecular, chemistry, polymers", and "Organic fine chemistry" (ii) engineering fields (ENG) which predominantly include IPCs in "Chemical engineering", "Civil engineering", "Control", "Electrical machinery,apparatus, energy", "Engines, pumps, turbines", "Machine tools", "Materials, metallurgy", "Mechanical elements", "Semiconductors", "Surface technology, coating", "Thermal processes and apparatus", "Measurement", and "Other special machines" (iii) technological fields (TECH) which predominantly include IPCs in "Audio-visual technology", "Basic communication processes", "Computer technology", "Digital communication", "Environmental technology", "IT methods for management", "Micro-structural and nano-technology", and "Telecommunications" (iv) medical fields (MED) which predominantly include IPCs in "Medical technology", "Biotechnology", "Optics", and "Pharmaceuticals" (v) other fields (OTHERS) which predominantly include IPCs in "Furniture, games", "Handling", "Other consumer goods", "Textile and paper machines", and "Transport". The inter-community linkages between communities shown in the figure are only limited to linkages that maximally spans the graph (i.e., only strong links). The results show that engineering fields do have the most number of IPCs in its communities, while the medical field do have the least number of IPCs. Figure 22 compares the technological structure of two cities with different number of inventors - Vancouver and Rio de Janeiro. In 2014, the number of inventors located in Vancouver and Rio de Janeiro is 1997 and 385 respectively. In examining the technological structure we observe that Vancouver does have a comparative advantage in more IPCs than Rio de Janeiro and has entered into 578 IPCs compared to 305 IPCs done by Rio de Janeiro. Due to the smaller exploratory space left for Vancouver, we see that in 2014, only 7 IPCs were explored compared to 18 IPCs explored by Rio de Janeiro. Most of Rio de Janeiro's new IPCs explored in 2014 were from the engineering field (ENG).

3.3.3 Measure of inter-city ties

Generally, cities are inter-connected, and economic actors rely on other actors to create value and access information. The information flow in this collaborative network could entail technical possibilities yet to be explored in certain locations (Rigby, 2015). The greater the interaction of economic actors of the city with economic actors in other cities, the more likely it is that the new technology located in other cities will be added to the city's technology base. The implication of this is that not all external economic actors are useful for entering the new technology, as some external economic actors might only have experience inventing in technology piece that already exists in the city's technology base, hence such ties to the city are redundant.

The measure of external ties in this study is specific to external inventors that have previously produced a patent in the technology new to the city. Formally, let v_{ijt} be the number of external inventors (that have produced a patent in IPC $j = 1, ..., n$ new to *city i*) located in other cities but having ties to inventors in *city i* in a given year t. For each of the ties to the external inventors, the weight of each tie w_{ijt} is the number of patents inventors in the city produced with the external inventor up to $t-1$, and $\bar{w}_{ijt}$ is the average weight across all ties. Therefore at any given year t, a *city i* do have a profile of external ties for technology piece new to the city, which is given as vector $M_{it} = (\bar{w}_{i1t}, ..., \bar{w}_{int})$. Hence, external tie of *city i* relevant for IPC j new to the city at any given year t is $\bar{w}_{ijt}$. This measure is lagged one-year when introduced into the empirical model. According to Hypothesis 1a, we expect that inter-city ties will positively affect the likelihood that the city enters a new technology domain.

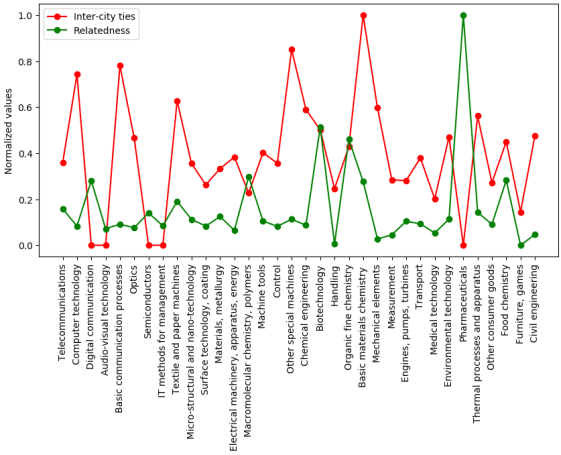


Figure 23: Distribution of relatedness and inter-city ties: The Y-axis is the normalized average value of relatedness and inter-city ties for observations where entry was made specifically in IPCs belonging to the technology field. We use a min-max normalization.

Figure 23 shows predominantly the importance of inter-city ties for exploration, we see that this is importantly the case for Basic Materials chemistry and less so for Pharmaceuticals, and Semiconductors. For these technology fields, for a city to enter into IPCs belonging to them they need to have developed technological stock that is related to them, this is moreso the case for Pharmaceuticals.

There are other characterizations of technological flows from neighboring cities. For example, Rigby, 2015 in his work assumed that the likelihood that a city enters a new technology is affected by its proximity to other cities, and if these other cities are competitively producing the new technology. To quantify this, they compute a co-invention matrix between cities and multiplied this by a bi-adjacency matrix that captures the technological advantage of cities in a given technological area. Entries in the bi-adjacency matrix are 1 if the revealed technological advantage of the city in the technology domain is at least 1, otherwise, they are 0. Hence the product matrix only captures flows of information from partner cities with competence in the technology. Another measure of inter-city ties

examines external social proximity between inventors living within and outside the focal city (Breschi and Lenzi, 2016). This is computed as the sum of the reciprocal of the geodesic distance between every inventor in the focal city and inventors elsewhere normalized by the number of inventors in the focal city. The measure being 0 implies that the focal city has no connection to other cities.

Measure of focal and partner city sizes.

An appropriate measure of city sizes would be the population size of the city. However, we could not obtain this information for more than half of the cities across the period in the study. Therefore, we focus on the productive population size of the city, which in our case is the number of inventors in the city in a given year. This measure has been shown to behave similarly to the population size. For example, the study by Bettencourt, Lobo, and Strumsky, 2007, showed similar findings when examining the effect of the number of inventors in the city and the population size of the city on patenting activities in the city. Both measures showed a superlinear effect on patenting activities in cities. Therefore a focal city size is simply the number of inventors located in the city, while a partner city size is the average number of inventors in the partner cities of the focal city. We expect that the focal city size and partner city size should positively moderate the effect of relatedness and inter-city ties on the likelihood of entry respectively.

3.3.4 Control variables

In the empirical model, we included some control variables to capture the importance of agglomeration activities on exploratory activities of cities. Firstly, the *inventor density* in each of the cities accounts for the city scale effect (Bettencourt, Lobo, and Strumsky, 2007; Lobo and Strumsky, 2008). This is computed as the ratio between the total number of inventors in a given city at a given year and the city land area in square kilometers. This variable is lagged one year and the log of the variable is inserted in the empirical model. To control for the effect of multinational corporations

(MNCs) in enabling the formation of inter-city ties, we compute the *share of foreign patents* in a city. Although, an appropriate measure of accounting for MNCs effect would be to identify the share of relevant ties that resulted from the same firms located in different cities, to do this would require a well-disambiguated applicant name in the dataset, which is a current limitation of the patent database we are using.

In addition, we compute the cognitive distance between the cities core technological area from the new technology (this is called the *distance from core IPC*). Since cities that are cognitively close to a new technology area will explore the new technology. To do this, for any given year, we identify the IPC a city has the most number of patents in (which we refer to as the core IPC of the city), then we compute the shortest network distance between the core IPC and the technology domain new to the city in the technology space (i.e., co-occurrence network of IPCs). The shortest distance is simply the least number of IPCs between them in the technology space. Also, we include a control for the depreciating number of IPCs left for the city to enter by counting the number of *existing IPCs* the city has in its technology base at every given year. We also include the *IPC growth rate* to control for the attraction of some IPCs at some point in time. The *IPC growth rate* is computed as the annual growth rate of patents produced in the given IPC. We included a measure of economic complexity of cities, which is shown to be associated with diversification behaviors of firms, regions, and countries. We use the "method of reflection" proposed by Hidalgo and Hausmann, 2009 to measure the economic complexity of cities.

Furthermore, we include a measure of *knowledge diversity* to capture if the city is a technologically diverse city, by computing the Blau index of diversity. First, we compute the share of patent of each IPCs in the city based on the city's patent stock in the last five years. Next, we took the sum of the squares of the reciprocal of each IPC patent share resulting in the Herfindahl Index, then we took the reciprocal of the Herfindahl index to obtain the Blau index. We also included controls for *city size*, which we define as the number of inventors that produced a patent in the city

in that year. Also, we control for the quality of the geo-referenced patent statistical database i.e., the *data quality* of the country the city is located in. The *data quality* in a country in a given year is the ratio between the number of patents in the country that have geo-referenced information for at least one inventor and the total number of patents in the country.

Finally, the empirical specifications include city, technology, and year fixed effects. In the empirical specification, we stratified the Cox PH estimation at the level of cities. This is done to control for any city-specific effect and it is equivalent to running a Cox proportional hazard with shared frailty model. Although, the latter explicitly test the hypothesis that there is a city-specific effect.

3.3.5 Empirical model.

The empirical model used is the Semi-parametric Cox proportional hazard (PH) model (stratified at the level of cities) - this is because the dependent variable is time-dependent entry process. The Cox PH model is generally used in survival analysis studies for modeling outcome variables that are time-dependent. This time is termed failure/event/survival time, or in our case entry or exploratory time. The analysis of time to event appears in many applied fields, such as medicine and public health (time to death), finance (time to loan default), and engineering (time to failure of an electronic component). In survival analysis, there are three requirements needed for the proper definition of the outcome variable: (i) a well specified time of origin (ii) scale for measuring time (iii) a clear definition of an event. Generally, we take the first year of risk to be the year 2004, and we assume that the exploration done by cities in an IPC is a one time event as re-entry into the same IPC is not allowed, and as such the statistical problem can not be characterized as a recurrent event (i.e., ordered multivariate exploratory time problem).

An important feature of survival analysis models is that it is characterized by a process called censoring. This is when we have information about the survival time of the city but the exact time is unknown. This sometimes arises from the city not exploring an IPC before the end of the study or the

city or IPC not being observed after some period, either due to the absence of patenting activities. There are two major types of censoring: (i) *right censoring:* this arises when cities enter the study at different times, the real entry time could be greater than the observed time of entry. Therefore the city's entry time becomes incomplete, as the city might enter the IPC at a time we do not know which is beyond the end of the observed period (i.e., 2014), resulting in a right censored data. This type of censoring is accounted for by the empirical model (ii) *left censoring:* this arises when the true time of entry for the city is before the period of observation. Although, this is difficult to deal with given that we have a georeferenced database that starts from the year 2000, in which case certain cities may have explored certain IPCs before the year 2000. For this reason, the analysis was done for the period 2005-2014, in other to give an interval of 5 years i.e., 2000-2004, to observe entries made by cities into IPCs. IPCs that were not explored by the city within the period 2000-2004 constitute a pool of IPCs to be explored between 2005-2014. The literature suggests this is a reasonable number of years to determine exploratory activities (Leten, Belderbos, and Looy, 2016).

Semi-parametric Cox proportional hazard model.

[1] We used the Semi-parametric Cox proportional hazard (PH) model Cox, 1972 as opposed to other survival models because when the PH model is used to estimate the hazard function $h(t)$, it does not require apriori an assumption of how $h(t)$ is distributed. Also, the model fits the baseline hazard h_0 from the data. An important attribute of the model that we expect to find is that the baseline hazard should decline over time. This attribute reflects the stability in cities' choice of entry over time (Leten, Belderbos, and Looy, 2016).

Formally, to define the hazard of entry, let $\mathbf{X_{ij}}$ be a vector of explanatory variables (i.e., relatedness and inter-city ties), and $\mathbf{Y_{ij}}$ be a vector of control variables that may affect the distribution of the time a city enters an IPC.

[1] Although an alternative continuous model is the cloglog, since the cloglog model with period-specific intercepts is equivalent to grouped-duration of Cox PH model (Hess and Persson, 2010). However, given the scale of the dataset, we had initial value problem issues.

The hazard of entry $h(t)$ in the Cox PH regression model (Cox, 1972) is given as:

$$h(t|\mathbf{X}, \mathbf{Y}) = h_0(t) \exp(\beta \mathbf{X} + \gamma \mathbf{Y}) \tag{3.2}$$

An important observation in the model relevant for the proportional hazard assumption is that the baseline hazard h_0 is a function of time and does not involve either explanatory or control variables, while the exponentiated expression involves $\mathbf{X}$ and $\mathbf{Y}$ but not time. $\mathbf{X}$ and $\mathbf{Y}$ have multiplicative effects. The model is assuming that the hazard for any *city i* entering an *IPC j* is a fixed proportion of the hazard for any other *city* $\hat{i}$ entering the *IPC j* i.e.,

$$\frac{h_{ij}(t|\mathbf{X_{ij}}, \mathbf{Y_{ij}})}{h_{\hat{i}j}(t|\mathbf{X_{\hat{i}j}}, \mathbf{Y_{\hat{i}j}})} = \exp\left[(\beta \mathbf{X_{ij}} - \beta \mathbf{X_{\hat{i}j}}) + (\gamma \mathbf{Y_{ij}} - \gamma \mathbf{Y_{\hat{i}j}})\right] \tag{3.3}$$

The interpretation of the coefficients β_k of variable X_k is the log hazard ratio. Therefore, the $\exp(\beta_k)$ is the hazard ratio associated with a change in variable $X_k \in \mathbf{X}$. Furthermore since $P(t \leq T < t + \delta t | T \geq t, X_k) \approx h(t|X_k)\Delta t$, we have that:

$$\exp(\beta_k) \approx \frac{P(t \leq T < t + \delta t | T \geq t, X_k + 1)}{P(t \leq T < t + \delta t | T \geq t, X_k)}, \ for\ all\ t \geq 0.$$

This implies that $\exp(\beta_k)$ can be loosely interpreted as the conditional probabilities of city exploring an IPC in the near future given the city survives at time t. Finally, $\exp(\beta_k) - 1$ is interpreted as the percentage change in hazard, when X_k increases by one unit, after adjusting for other variables.

Although, the Cox PH model has been used extensively in the analysis of time-dependence dynamics in trade and innovation (Leten, Belderbos, and Looy, 2016; Besedeš, 2008; Nitsch, 2009). However, not accounting for unobserved heterogeneity in the econometric specification can result in "spurious negative duration dependence of the estimated hazard function", and a bias parameter (Hess and Persson, 2010). Although

augmenting the Cox PH model by specifying a frailty term seems to take care of this issue, but this is often computationally time consuming when applied to big dataset like ours. In light of this, our analysis accounts for the unobserved heterogeneity of cities by stratifying the Semi-parametric Cox PH model at the level of cities. This is a similar methodological approach employed by Besedeš, 2008 and has the advantage of controlling for unobserved heterogeneity by "allowing for group-specific variation in the baseline hazard" (Hess and Persson, 2010).

3.4 Results

Table 15 shows the descriptive statistics and correlations for the dependent and explanatory variables in the empirical model. Inter-city ties and technological relatedness show a positive relationship with entry. The table shows a higher positive correlation between technological relatedness and entry when compared to the correlation between inter-city ties and entry. Also, other control variables do have a significant positive relationship with entry, except for the variables distance from core IPC and share of foreign patent. The negative correlation between distance from core IPC and entry suggests that the closer the city's core technology area is to the new technology domain the more likely the city will enter the new domain. The negative correlation between share of foreign patent and entry suggests that cities are less likely to enter a new IPC if a large share of their patent is foreign. The statistics in the table provide some prima facie evidence that the likelihood of a city to enter new technology domain increases with an increase in ties to external inventors who have previously invented in the new domain, and also with an increase in the city's technology relatedness with the new technology domain. The highest correlation in the table is 0.368 - which is found between focal city size and inventor density and shows a strong significant positive relationship, implying that cities with large pool of inventors have more geographically dispersed inventors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Entry	1												
(2) Inter-city ties	0.010*	1											
(3) Technological relatedness	0.049*	0.027*	1										
(4) Distance from core IPC	-0.159*	0.008*	0.169*	1									
(5) Knowledge diversity	-0.039*	-0.003	0.056*	0.034*	1								
(6) Existing IPCs	0.076*	0.060*	0.254*	0.053*	0.235*	1							
(7) IPC growth rate	0.034*	0.005*	0.082*	0.068*	0.006*	0.056*	1						
(8) City complexity	0.006*	-0.007*	0.034*	0.045*	0.304*	0.208*	0.006*	1					
(9) City's inventor density	0.055*	0.064*	0.079*	0.025*	-0.053*	0.270*	0.022*	-0.018*	1				
(10) Share of foreign patent	-0.048*	-0.003	-0.029*	-0.027*	0.006*	-0.147*	-0.010*	-0.160*	-0.005*	1			
(11) Focal city size	0.074*	0.094*	0.126*	-0.004*	-0.017*	0.352*	0.038*	-0.046*	0.368*	-0.044*	1		
(12) Partner city size	0.261*	0.016*	-0.003	-0.046*	-0.0003	0.084*	0.012*	0.009*	0.049*	-0.047*	0.0366*	1	
(13) Data quality	-0.061*	0.015*	0.072*	0.025*	0.226*	0.328*	0.010*	0.093*	0.082*	-0.030*	0.056*	0.028*	1
Mean	0.180	0.003	0.007	1.701	16.958	184.345	-0.035	-0.037	0.164	0.385	231.845	545.625	0.807
Std. Dev.	0.384	0.081	0.009	0.574	13.441	154.150	0.624	0.999	0.496	0.268	1157.469	4333.390	0.191
Min.	0	0	0	0	0	2	-0.954	-8.838	0	0	5	0	0.07
Max.	1	12.333	0.764	4	205.920	636	10	8.718	18.594	1	117279	261920	1

Notes: * indicate significance at 5% level.

Table 15: Descriptives and correlations of variables in the analysis of entry.

Tables 16 and 17 present the results of the Cox proportional hazard model without stratification and with stratification respectively. The model without stratification is used as a benchmark, and what we observe in

comparison with the stratified model is that (i) the stratified model has a better explanatory power when we examine the pseudo-likelihood of both models (ii) the moderating role of focal city size on the effect of relatedness on entry is significant in the non-stratified model, while in the stratified model this is not the case.

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.006	1.009*	1.006*	1.009**
		(0.00305)	(0.00343)	(0.00304)	(0.00342)
Technological relatedness		1.053***	1.053***	1.053***	1.053***
		(0.0106)	(0.0106)	(0.0106)	(0.0106)
Inter-city ties × Partner city size			1.009*		1.009*
			(0.00374)		(0.00377)
Technological relatedness × Focal city size				1.005*	1.005*
				(0.00226)	(0.00227)
Distance from core IPC	1.078***	1.077***	1.077***	1.077***	1.077***
	(0.0154)	(0.0154)	(0.0154)	(0.0153)	(0.0153)
Knowledge diversity	1.062***	1.063***	1.063***	1.062***	1.062***
	(0.0110)	(0.0110)	(0.0110)	(0.0110)	(0.0110)
Existing IPCs	1.659***	1.650***	1.650***	1.657***	1.657***
	(0.0312)	(0.0309)	(0.0309)	(0.0317)	(0.0317)
IPC growth rate	1.134***	1.129***	1.129***	1.129***	1.129***
	(0.00515)	(0.00549)	(0.00549)	(0.00551)	(0.00551)
City complexity	1.042*	1.045*	1.044*	1.045*	1.045*
	(0.0206)	(0.0205)	(0.0205)	(0.0205)	(0.0205)
Inventor density	1.003	1.002	1.002	1.003	1.003
	(0.00362)	(0.00350)	(0.00350)	(0.00347)	(0.00347)
Share of foreign patent	1.003	1.004	1.004	1.004	1.004
	(0.0287)	(0.0285)	(0.0285)	(0.0285)	(0.0285)
Focal city size	1.028***	1.026***	1.026***	1.017*	1.018**
	(0.00451)	(0.00461)	(0.00455)	(0.00679)	(0.00675)
Partner city size	1.022	1.022	1.022	1.022	1.022
	(0.0255)	(0.0252)	(0.0252)	(0.0252)	(0.0252)
Data quality	1.117**	1.118**	1.118**	1.118**	1.118**
Data quality	(0.0420)	(0.0418)	(0.0418)	(0.0418)	(0.0418)
City FE	No	No	No	No	No
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	521 339	521 339	521 339	521 339	521 339
Log pseudolikelihood	-1 094 239.2	-1 094 014.7	-1 094 013.1	-1 094 006.1	-1 094 004.7
Wald chi2	1.66e+09***	1.78e+09***	1.78e+09***	1.78e+09***	1.78e+09***

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 16: Results of semi-parametric Cox PH model without stratification.

The coefficients displayed in both Tables 16 and 17 do not have any causal interpretation. These coefficients are exponentiated which means we can interpret them as hazard ratios, and the regressors in the model are mean-centered, which means the reported coefficient is due to a standard deviation change observed in the independent variable. If the coefficient is larger (smaller) than 1, it implies an increase by one standard deviation

of the value of the independent variable results in an increase (decrease) in the likelihood that a city will explore a new technology domain. Model 1 includes all control variables. Model 2 in addition to the control variables includes inter-city ties and technology relatedness. Model 3 in addition to the variables in Model 2 includes the interaction term between inter-city ties and the size of the focal city's partners. In addition to the variables in Model 2, Model 4 includes an interaction term between relatedness and focal city size. Finally, Model 5 includes all control variables, inter-city ties, relatedness, and both interaction terms. The models reported in Table 17 are highly significant.

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.004*	1.008***	1.004*	1.008***
		(0.00208)	(0.00217)	(0.00208)	(0.00217)
Technological relatedness		1.074**	1.074**	1.074**	1.074**
		(0.0257)	(0.0258)	(0.0269)	(0.0269)
Inter city ties × Partner city size			1.012***		1.012***
			(0.00317)		(0.00317)
Technological relatedness × Focal city size				0.998	0.998
				(0.00348)	(0.00347)
Distance from core IPC	1.139***	1.137***	1.137***	1.137***	1.137***
	(0.00572)	(0.00567)	(0.00568)	(0.00567)	(0.00567)
Knowledge diversity	1.064***	1.068***	1.069***	1.068***	1.069***
	(0.0184)	(0.0186)	(0.0186)	(0.0186)	(0.0186)
Existing IPCs	1.788***	1.747***	1.747***	1.745***	1.745***
	(0.103)	(0.102)	(0.102)	(0.102)	(0.102)
IPC growth rate	1.137***	1.130***	1.130***	1.130***	1.130***
	(0.00497)	(0.00594)	(0.00594)	(0.00588)	(0.00588)
City complexity	0.993	0.993	0.993	0.993	0.993
	(0.00729)	(0.00730)	(0.00730)	(0.00731)	(0.00731)
Inventor density	1.019	1.017	1.017	1.017	1.016
	(0.0178)	(0.0174)	(0.0174)	(0.0173)	(0.0172)
Share of foreign patent	0.983	0.982	0.982	0.982	0.982
	(0.0135)	(0.0137)	(0.0137)	(0.0137)	(0.0137)
Focal city size	1.036	1.037	1.037	1.040	1.040
	(0.0245)	(0.0247)	(0.0246)	(0.0237)	(0.0237)
Partner city size	0.945***	0.947***	0.947***	0.947***	0.947***
	(0.0148)	(0.0149)	(0.0149)	(0.0149)	(0.0149)
Data quality	1.000	1.001	1.001	1.001	1.001
	(0.0184)	(0.0185)	(0.0185)	(0.0185)	(0.0185)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	521 339	521 339	521 339	521 339	521 339
Log pseudolikelihood	-396 209.92	-396 035.8	-396 033.38	-396 035.27	-396 032.83
Wald chi2	60 277.77***	61 120.64***	61 107.52***	61 126.95***	61 112.55***

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 17: Results of stratified semi-parametric Cox PH model.

From Model 2-5 in Table 17, the ties a city have to external inventors who have previously invented in a technology domain is positive and

significantly associated with the likelihood that the city will enter the new technology domain. The estimated hazard ratios indicate that a standard deviation increase in inter-city ties would increase the likelihood of entry by 0.4%-0.8%. Also, the relatedness of a city's technology stock with a new technology domain is positive and significantly associated with the likelihood of entering a new technology domain. Specifically, a standard deviation increase in relatedness would increase the likelihood that a city will enter a new technology domain by 7.4%. Both results confirm the prediction in hypotheses 1a and 2a respectively. Also, we observe that the association between inter-city ties and the likelihood of entry is approximately ten times smaller than the association between technological relatedness and the likelihood of technological entry. This can be due to the direct and indirect linkage between both variables and entry process. For example, the link between technological relatedness and entry is more direct - hence we can say that the shared heuristics between the new technology and the cities' capabilities imply that economic agents in the city will require little resources to patent in the new technology. On the other hand, the link between inter-city ties and entry is not direct as not all linkages (even to large partner city) are important for entry if the knowledge flow from these linkages are not complementary to the capabilities in the city. Hence, the correlation of inter-city linkages with the likelihood of entry may be downsized by the existence of linkages with knowledge flow not complementary to the capabilities in the focal city.

From Model 3 and 5 in Table 17, the hazard ratio of the interaction term between inter-city ties and the size of partner cities is significant and greater than 1 (i.e., 1.012), which implies that the effect of inter-city ties on the likelihood of entering a new technology domain is positively moderated by the size of partner cities. Specifically, the average marginal effect of inter-city ties for a city with average size of partner cities (i.e., approximately 545 inventors) is 0.0024, which is positive and significant. Although from the profile plot in Figure 24, we observe that the positive moderation is significantly large for cities with very large partner size i.e., focal cities' partners should have at least 55,880 inventors (equivalently a

standardized value of 10.56). For cities linked to partners with small size (less than 425 inventors), the effect on relative hazard is negative but not significant. This result generally confirms the prediction in hypothesis 1b. Furthermore, from Model 4-5 in Table 17, the interaction term of focal city size and technological relatedness shows a non-significant hazard ratio less than one (i.e., 0.998). This implies that the number of inventors in the city negatively moderates the effect of technological relatedness on the likelihood of entry. Hence we do not find evidence in support of hypothesis 2b.

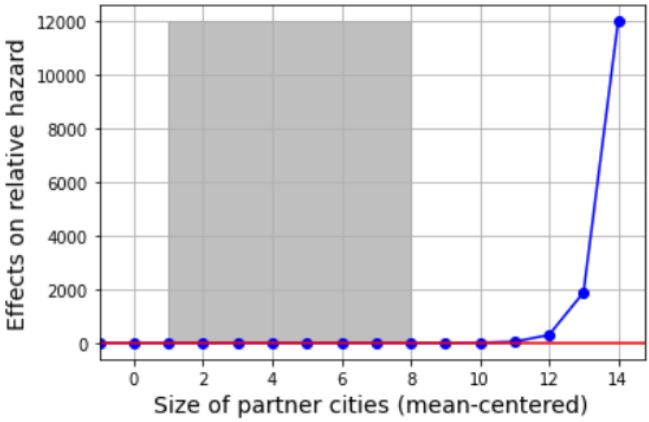


Figure 24: The average marginal effect of inter-city ties: The partner size on the X-axis is mean-centered. The horizontal red line is at 0 which determines if the moderator (Partner size) is positive or otherwise. The grey area indicates the significant region. The positive moderation is more pronounced when partner size exceeds 10.56 but it is not significant. However, for partner size between 0 and 8, the positive moderation is small and significant (check Table 49 in Appendix C for the actual relative hazard).

From Model 1-5 in Table 17, hazard ratios for the control variables show that cities whose core technological domain is proximal to new technology domains are more likely to explore the new domain. Also, cities with diverse knowledge from previous patent activities and with large technology stock are likely to enter a new technology domain. Furthermore, the

measure of attraction of IPCs based on patent growth rate in IPCs shows that cities are likely to enter an attractive new technology domain. Finally, the quality of the geocoding effort in building the dataset for this analysis does not significantly impact the results.

Supplementary analysis

To ascertain the robustness of our findings, first, we checked if the result obtained is sensitive to a flexible definition of "newness of technology to the city", by reducing the window during which the technology should not exist in the city to three years instead of five years. Thus, the number of city-IPC combinations increases to 707,323, and the number of observed entries is 186,647 (i.e., 26.3%). The empirical result did not alter materially as seen in Table 37 in Appendix C.0.3. Next, we changed the measure of inter-city ties from being the average weight of ties to inventors in other cities who have previously invented in the new technology to a measure of inter-city ties proposed by Rigby, 2015 - which is the average weight of ties to cities having a positive revealed technological advantage in the new technological area. To do this, we construct an inter-city matrix C_{ijt} of size 1220×1220 with entries being the number of co-invented patents between *city i* and *city j* in year t. Next, we construct a city-IPC capability matrix Q_{jkt} of size 1220×646 with entries being 1 if the revealed advantage of *city j* in *IPC k* in year t is greater than 1, otherwise 0. Therefore, the resulting product matrix $M_{ikt} = C_{ijt}Q_{jkt}$ of size 1220×646 captures technology flow from cities in proximity to the focal city. The empirical results shown in Table 38 in the Appendix is consistent with results discussed so far, we have that *Hypothesis 1a, 1b* and *2a* are accepted, and we do not find significant evidence in support of *Hypothesis 2b*. The robustness check shows a stronger effect for the new measure of inter-city ties used, which suggests that having inter-city ties with cities having a positive advantage in a technological area is stronger as compared to ties with external inventors who have previously invented in the new technological field.

Furthermore, considering that cities are heterogeneous, we include a supplementary analysis examining the effect of inter-city ties and tech-

nological relatedness separately for cities in Europe, Asia, and North America, global and non-global cities. For European cities, we have that both inter-city ties and technological relatedness is positive and significantly associated with city entering a new technology domain. Specifically, inter-city ties and technological relatedness increase the likelihood of entry by 0.16% and 5.5% respectively. This is also the case for cities in Asia. In Asia, inter-city ties and technological relatedness increase the likelihood of entry by 2.8% and 21.9% respectively. For cities in North America, only technological relatedness do have a positive and significant effect on entry - technological relatedness increases the likelihood of entry by 15.2%. A similar result is seen in global cities - technological relatedness increases the likelihood of entry by 16%. While for non-global cities, inter-city ties increase the likelihood of entry by 8%, and technological relatedness increases the likelihood of entry by 6.5%.

Finally, we examine the likelihood that a city whose core technological area is ICT will enter technological areas relevant to Pharmaceuticals, and vice-versa. ICT technologies comprise of IPCs in Digital Communications, telecommunications, Computer technology, and IT methods of management. Given that the size of the dataset for this analysis is reasonably small it can accommodate using Semi Parametric Cox PH model with shared frailty. The shared frailty term in the model control for city-specific effect. We find that, generally, inter-city ties and technological relatedness increase and decrease the likelihood of entering a new technology in Pharmaceuticals respectively, and the effect of technological relatedness is positively moderated if a city's core technological area is ICT. For entry into ICT, generally, we find that both inter-city ties and technological relatedness increase the likelihood of entry, and both effects are positively moderated if the city's core technological area is Pharmaceuticals.

3.5 Discussion

Developing competence in new technological areas is important for long-term viability of cities and also for the integration of cities in the global

inter-city network. The faster development of these competencies requires the formulation of policies that allow not just the cross-fertilization of ideas between economic actors within the city but also cross-boundary collaboration efforts with economic actors who have previously invented in technologies new to the city. Also, there is a need to pursue policies geared towards a path-dependent innovation process, that allows economic actors to build on existing local expertise within the city in such a way that it becomes easy for them to access proximal innovative technologies. This policy process involves considerable resources, hence there is need for a proper evaluation of the cost-benefit analysis of proposed mechanisms that will be adopted for increased likelihood of diversification into new technological areas by the city.

The literature on regional innovation and economic development suggest the usefulness of social proximity within and without the city in increasing the probability of knowledge flow into cities (Balland and Boschma, 2021; Breschi and Lenzi, 2016; Rigby, 2015; Breschi and Lissoni, 2001; Agrawal, Kapur, and McHale, 2008; Singh, 2005). There has been literature that show that regions build on existing capabilities to develop new ideas (Balland and Boschma, 2021; Boschma, Balland, and Kogler, 2015) and inter-regional linkage can be a source of information that can be useful for cities to explore new ideas, hence preventing them from a lock-in effect (Miguélez and Moreno, 2012; Boschma and Iammarino, 2009). Recent works have further looked into the usefulness of developing inter-regional linkages with complementary capabilities (Balland and Boschma, 2021). What we have done differently is to provide a joint analysis of both existing capabilities and inter-city linkages at a geographical level lower than regions, and specifically looking at entry process rather than specialization. The contributions to the literature are in three folds: (i) we present a joint analysis of the effect of inter-city ties and technological relatedness on the likelihood of entry. While the former reflects the outward search of the city for capacity to enter new domains, the latter reflects building on inward capacities to enter nearby technology domain. (ii) extending the analysis to wider coverage of cities across different continents (i.e., North America, Europe, Asia, South America, Oceania and Africa), which

is an improvement on previous research work restricted to developed countries, EU and OECD countries (iii) examining the role of the inventive size of partner cities in moderating the network effect of inter-city ties on the likelihood of entry, and the moderating role of focal city size on the effect of technological relatedness on the likelihood of entry. This analysis of cities in emerging and developed economies is important for development studies, as observations highlight the usefulness of linkages to cities in Asia and non-global cities in general, which means for cities to move up the developmental ladder they need to learn the know-how from other places through collaborations.

Our findings suggest that both inter-city ties and technological relatedness increase the likelihood that a city will enter a new technology domain, and the effect of inter-city networks is enhanced when a city's ties are located in partner cities with a larger pool of inventors (i.e., large partner city size). On the other hand, we do not find any evidence suggesting that the effect of technological relatedness is positively moderated by focal city size. The main result in this paper re-visits the argument of localized technological flows identified in Jaffe, Trajtenberg, and Henderson, 1993 - which predominantly put forward the stickiness of knowledge. While it might be true that technological flows between economic actors are better facilitated when they are located within the same regions, however, the prevalence of technological development has enhanced collaboration across borders, and this is even more reflected in the rise of cross-border collaborative effort in major scientific and technological experiments. More importantly, inter-city ties are not just useful for getting to know and learn about new technologies, but they can also serve as a gateway to global pipeline especially when such ties exist with cities with more inventors. The existence of a large pool of inventors enhances the quality of direct ties the focal city has and the focal city can experience a second-order effect on the likelihood of access to information. Access to global pipelines and know-how to refine information can be very useful especially for small cities that do not have the internal capacity to verify the authenticity of information or how well to integrate new information with their existing know-how. This can be seen when we compare the

result of entry analysis for global cities in Table 46 with that of non-global cities in Table 47 and also in Table 48 when we interact inter-city ties with global city status. The observation of a positive moderation of partner city size on the effect of inter-city network being more evident in non-global cities is not surprising as non-global cities are not attraction points for global scientific and technological talent in comparison to global cities (Verginer and Riccaboni, 2021), and as such might not have the internal capacity to a assess the cost and benefit of new technology, hence they need ties that are embedded in locations with rich network of inventors.

Our findings on the positive effect of technological relatedness on likelihood of entry using a broad range of cities further confirms previous research on technological relatedness and diversification done for OECD regions, EU regions and cities in the United States (Vlčková, Kaspříková, and Vlčková, 2018; Rigby, 2015). It shows that irrespective of the continents, cities' decision to upgrade into new technologies is consistent with their technological profile (check Table 43-45). This is also the case whether or not a city is global (check Table 46-47). While cities in some regions leverage relatedness to transition rapidly into new domains (e.g., 21.8% increase in the likelihood of entry for cities in Asia), transition process for other cities is much slower (e.g., 5.5% increase in likelihood for cities in Europe). We see that cities follow a well defined technological pathway for development, by leveraging on proximate technologies that share similar scientific heuristics. This means that entry into new technology depends on current practices in the city. This choice for proximal technology as highlighted by Breschi, Lissoni, and Malerba, 2003 could be a result of unintended spillovers or intentional process of local learning and preference for certain technologies due to its scope or complementarities with the city's core expertise. We observed that even the most global cities characterized by diversified patenting activities are more "coherent" in terms of leveraging technological relatedness for entry. This might be because as inventors in global cities expand their patenting activities, they gradually increase their technological coherence by systematically patenting in technology class consistent with what they began with, thus filling most technological gaps between classes.

There are certain limitations identified in this empirical work. The first is the definition of city size (i.e., focal city size and partner city size). Although, we have pointed out that the notion of city size is meant to capture the cities' inventive capabilities, in which case it might be appropriate to focus on the inventive population size and not the population size of the city. Further research can be done by specifically looking at the moderating role using population size as a measure of city size. We are agnostic as to what measure is adopted for city size as both population size and inventor size of cities have been shown to have a similar effect on patenting activities in cities (Bettencourt, Lobo, and Strumsky, 2007). However, the measure of city sizes using the number of inventors will potentially be impacted by ambiguities in inventor names, as inventor names tend to be misspelled, or ordered differently. Also, it remains unclear what kind of phenomenon is been captured by the inter-city ties. It is worth investigating if the significant positive association of inter-city ties with cities entering new technology domain reflects knowledge-spillover or knowledge acquisition. In the case of knowledge spillover, the city enters a new IPC but the applicant is not located in the city, such technological entry are usually facilitated by multinational firms. On the other hand, for technology acquisition, the city enters the new IPC and the applicant is located in the city. To distinguish these types of inter-city ties in future research work, the inventor and applicant database can be combined. Also, the analysis of the effect of inter-city linkages on likelihood of entry is biased against cities that have a larger share of external ties to inventors in non-FUA. To address this is to extend the analysis to both FUAs and non-FUAs.

Also, it is still unclear what notion of relatedness best captures the structure of technological proximity, more research effort can be geared towards properly capturing technological relatedness in such a way that we can adequately measure the density of cities capabilities around new technologies, that are more predictive on out-of-sample data, as the process of entry is a forward-looking event. Finally, we introduced the notion of technological core of cities as a control variable in our analysis (which we identify as an IPC the city has the most patent in) when indeed the city's

technological core could be a collection of IPCs. Hence more work can be done using the k-core analysis in identifying city's technological core and periphery, and also in identifying if the positive effect of relatedness is driven by technological core or peripheral IPCs in the city, and what possible combination of IPCs are most productive for a city to efficiently enter new technology.

The need for a systematic methodology for identifying linkages that are relevant to city implies that further work can be done on examining how exactly inter-city linkages facilitate entry. We have shown that for non-global cities to invent in new technological fields they need ties more than it is the case for global cities. Does it matter if these ties are complementary to the existing capabilities of non-global cities? if it is the case that complementarity matters, is the effect different across the world. Another dimension that can be looked into is how similar cultural norms alongside complementarities refine linkages (Balland and Boschma, 2021). Finally, further study can examine how weak institutions negatively moderate the effect of knowledge flows. The dataset constructed in Chapter 1 used in this chapter can be very useful for such analysis, especially in developing economies. Since even if a city has the required capabilities and inter-city linkages needed to upgrade to a new technology domain, weak institutions may still hinder exploratory innovation activities. Another potential area of research is to use the machine learning methodology developed in Chapter 2 to build a measure of the proximity of technological capabilities of cities in new technologies. Since this model unravel relationships between technologies non-linearly unlike existing measures (e.g., Engelsman and Raan, 1991) in the regional economics literature. Although Chapter 2 predicts specialization, the method can also be adapted for entry - hence the class probability for a city entering a technology domain can be used as a measure of technological relatedness of the city's existing capabilities with the new technology.

In conclusion, we show in this study that the likelihood to explore new technology by cities is governed by structural characteristics of the cities. These structural characteristics: inter-city ties and technological related-

ness combine to explain the heterogeneous patterns of entry into new technology. The study confirms that both inter-city ties and technological relatedness are strong determinants for the likelihood of entering new technology domains. While these structural properties are important, the partner size - measured as the average number of inventors located in the focal cities' partner cities plays a positive role in moderating the effect of inter-city ties on the likelihood of exploration. These results suggest that policies such as the EU smart specialization, Canada's super cluster initiative, and similar initiatives. that foster collaborations and leverage place-based capacities to identify strategies that increase technological diversification and growth should be encouraged.

Chapter 4

Global Cities in International Networks of Innovators

[1] Since 1990, international collaborative research has witnessed a remarkable growth (Adams et al., 2005). Wagner, Park, and Leydesdorff, 2015 showed that 25% of papers published in the Web of Science in 2011 were co-authored, which is an increase from the 10% observed in 1990. Several countries and cities now participate in global collaboration, many more than two decades ago (Bornmann, Wagner, and Leydesdorff, 2015; Adams, 2012; Verginer and Riccaboni, 2021; Wagner, Whetsell, and Leydesdorff, 2017). The growth in international collaboration partly results from the rise in team research and "big science". Nevertheless, "big science" does not completely explain such growth (Wagner, Whetsell, and Leydesdorff, 2017), as there exist several "small science" projects executed by international teams whose members have aligned interests and can collaborate despite geographic distance. (Jones, Wuchty, and Uzzi, 2008; Wagner, Whetsell, and Leydesdorff, 2017).

[1] This Chapter is partially based on S. Edet, P. Panzarasa, M. Riccaboni, Global Cities in International Networks of Innovators, *Advances in Complex Systems*, doi: 10.1142/S0219525921400026

Although long-distance collaboration can be costly for authors and inventors in terms of coordination, scientists increasingly co-author papers, and multi-authored papers attract more citations than single-authored ones (Wagner, Whetsell, and Leydesdorff, 2017; Jones, 2021). As observed by Barjak and Robinson, 2008, international partnership brings added benefits that exceed transaction and coordination costs, as reflected in the higher-than-expected citations received by international publications in most scientific fields (Persson, Glänzel, and Danell, 2008; He, 2008). Today the negative effect of distance on scientific collaboration has become weaker since digitalization and globalization facilitate long-distance collaboration compared to the past (Sonnenwald, 2007). Specifically, recent contributions have shown an increase in long-distance collaboration and a decrease in the proximity of collaborative ties over time (Frenken et al., 2009; Choi, 2012; Morescalchi et al., 2015). Hence, geographical constraints on collaboration have become less important (Hoekman, Frenken, and Tijssen, 2010; Chessa et al., 2013; Cerina et al., 2014). Moreover, the rise of global cities as hubs in the network of researchers facilitates the circulation of ideas and promotes international collaborations (Verginer and Riccaboni, 2020; Verginer and Riccaboni, 2021).

An important driver of international collaboration is the increased mobility of researchers (Jonkers and Cruz-Castro, 2013; Verginer and Riccaboni, 2021), as academic visits help them create new connections that can be activated in future innovation activities. The increasing complexity of scientific and technical problems also implies larger teams and more co-authors per publication (Wagner, Whetsell, and Leydesdorff, 2017; Jones, 2021). The mobility of scientists alongside other mechanisms (e.g., high-quality research institutions, the presence of multinational corporations, the quality of the entrepreneurial ecosystem) also contribute to the scientific and technological success of a city (Catini et al., 2015).

International networks of researchers are characterized by pipelines for global reach, as global cities serve as catalysts of prolific researchers with international connections, which in turn can boost their competitiveness in the innovation network (Verginer and Riccaboni, 2021; Scholl, Garas,

and Schweitzer, 2018). For this reason, it is not sufficient to have disjoint international collaboration structures for cities to qualify as global players. Global cities should be integrated into international complex teams of inventors with scientists located in different countries. This is important to understand better how to absorb knowledge from distant places and to take part in international knowledge production. An important reason why the networks of researchers should not be studied as a set of dyadic collaborations is that a dyadic approach oversimplifies the nature of collaborations that take place in teams when the production of patents and publications emerges from multiple interactions between researchers located in different places. Recent literature on the analysis of networks has begun to analyze networks based on the notion of multi-body interactions using hypergraphs or simplicial complexes (Neuhäuser, Mellor, and Lambiotte, 2020). In this chapter, we analyze the international network of researchers by focusing on the hypergraph of publications and patents team production. In particular, we focus on international collaborations involving at least three cities to identify global players in the multiplex innovation network (i.e., patent and publication networks).

Recent innovation studies from a network-based perspective have analyzed multiplex networks constructed from patents and publications. For example, Angelou et al., 2020a have examined the evolution of temporal multiplex innovation networks consisting of collaboration layers - EU framework program (FP) and patent. The dynamical processes behind the growth in collaboration for both layers were described using kinetic models. The analysis was carried out using the notion of multilinks that captures the differences and similarities between the links connecting the same nodes in both layers. The analysis shows that patents drive the creation of common multilinks at an early stage, but this process is reversed at a later stage. Further work by the same authors has concentrated on patterns of formation of triangles in these multiplex networks (i.e., with European FP and patent layers). These patterns have been uncovered by comparing the formation of triangles in the observed multiplex network with what emerges from the randomized network. Findings suggest that triangular FP collaborations are more frequent than random ones, while

the reverse is true for patents (Angelou et al., 2020b). Other studies from a multiplex network perspective have examined the evolution of urban innovation networks in China based on publications and patents (Li, Wei, and Wang, 2015). Their findings suggest that both networks are characterized by a preferential attachment growth mechanism, disassortative traits, and a hierarchical diffusion process.

In this chapter, we contribute to the literature on global cities and the analysis of international networks by specifically analyzing hypergraphs involving international collaboration of researchers located in three cities (i.e., hyperedges of size 3). We propose a new measure that relies on the hypergeometric ensemble model (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018) to identify cities that are global players in both the publication and patent hypergraphs. To do this, we used PubMed and the geo-referenced patent statistical data set we developed in Chapter 1, which allows for the analysis of the network of researchers between cities located in different countries. Cities are defined as functional urban areas consisting of densely inhabited areas and surrounding peripheral areas whose labor markets are highly integrated with the city. The definition of a city has been harmonized based on OECD, 2012 approach to defining functional urban areas, in order to allow for international comparisons. The analysis involves 153 global cities drawn from the Global and World Cities (GaWC) research project (Beaverstock, Smith, and Taylor, 1999), and these cities are in Asia (47), America (44), Europe (48), and other regions (14).

The rest of the chapter is organized as follows: Section 4.1 provides a literature review on global cities and international collaboration networks; Section 4.2 describes the methodology used to analyze the network of researchers as a hypergraph. We introduce a measure of the global reach of cities (i.e., the probability that a city will be a third player in an international collaboration), and we generate a null model using the hypergeometric ensemble (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018) which allows us to determine if the measure of global reach is statistically significant. Section 4.3 describes the geo-referenced patent data

and scientific publication data used in constructing a hypergraph of size 3, consisting of international research teams. Section 4.4 illustrates the results of the analysis, and Section 4.5 provides a final discussion of the findings.

4.1 Global cities and International collaboration networks

In this section, we provide an overview of the literature on global cities and the literature on international collaboration network.

4.1.1 Global cities network

The publication of Sassen, 1991's "*The Global City*" brought to attention the notion of global flows facilitated by critical services that take place across subsidiaries of multinational corporations located in different cities. These global flows in which global cities are a conduit made the cities more alluring to advanced producer service firms specializing in offering professional and financial services (Sassen, 1991). The creation of offices by advanced producer firms in different global cities led to the formation of interconnected service complexes between cities, which according to Sassen, 1991 is the basis for the creation of the world city network (Derudder and Witlox, 2008). The differences in technological specialization across regions, technological advancement, and ease of mobility of highly skilled labor, have increased the dispersion of production networks of firms especially multinational enterprises across the globe. However, this economic globalization has seen only a few cities being endowed with a disproportional share of economic activities, foreign investment, scientific and technological production (Sassen, 2001; Storper and Scott, 2009). For this reason, there has been an increased interest in examining economic globalization specifically at sub-national levels such as regions, clusters, and cities - particularly large cities, as they are seen as hubs for innovation and an attraction point for skilled workers (Friedmann, 1986; Sassen, 1991; Verginer and Riccaboni, 2021).

The literature on cities considers either a demographic or functional approach to measuring cities (Chakravarty et al., 2021). The consequences of enormous population concentrations in mega-cities are examined in the demographic approach. The functional approach, however, concentrates on the distinctive attributes of cities commonly referred to as being "global" (Sassen, 1991; Friedmann, 1986; Cohen, 2018; Chakravarty et al., 2021). There are two commonly used methods in the functional approach to measure the "globalness" of cities. These methodologies include the "corporate organization methodology" which focuses on the interlocking activities of firms across global cities, and the "infrastructure methodology" which relies on how global cities are connected based on infrastructures such as telecommunications, air transportation, and railway transportation (Derudder and Taylor, 2005; Otiso et al., 2011; Cheng and LeGates, 2018).

The "corporate organization" methodology

The "corporate organization" methodology is hinged on the observation that firms that are involved in international strategic activities are responsible for the formation of the world city network. This methodology was first developed in the GAWC project and hinges on the premise that service firms with global presence are responsible for the linkages between cities through intra-firm activities taking place. This approach constructs inter-city matrices based on the location strategies of firms. The location strategy of a firm is formalized by a "service value" v_{ij}, which captures the relevance of a city i in the international service network of a firm j (Taylor, 2001). Therefore, for any two cities a and b, the inter-city flow between them due to the activities of a firm j is given as $w_{ab,j} = v_{aj}v_{bj}$. In this sense, if either of the cities is very important in the international service network of the firm, it is expected that there will be more knowledge flows from the city to any other city (Chakravarty et al., 2021). Hence, the total flow to a city a is the aggregate flow coming from firm activities in other cities, and this is given as: $w_a = \sum_j \sum_b w_{ab,j}$. Based on this specification and correlation analysis, Beaverstock, Smith, and Taylor, 1999

measured the global connectivity of 315 cities based on location strategies of 100 firms, and concluded that cities such as London and New York are "exceptions" and not necessarily "exemplars" amongst world cities. This specification has been extended in other research work such as Derudder and Taylor, 2005 and Taylor et al., 2007.

Beyond service firms, other works such as Alderson and Beckfield, 2004 and Rozenblat and Pumain, 2005 rely on multinational companies to be responsible for the formation of inter-city linkages. Their approach deviates from the GAWC approach in two ways (i) it does not focus only on producer service firms (ii) it derives inter-city flows based on ownership structure in MNCs. The inter-city linkages are constructed by "a directed interaction between the city where the headquarters are located and the city where the subsidiary is owned" (Derudder and Witlox, 2008). The position of cities is inferred from this ownership network aggregated to the level of the city. The problem with the corporate organization techniques for assessing global city relevance is that they do not include data on actual inter-city flows (Derudder and Witlox, 2008).

The infrastructure methodology

The infrastructure methodology for measuring global cities is motivated by the observation that infrastructure networks determine the economic potential of urban areas (Derudder and Witlox, 2008). This implies that most global cities are seen to have important infrastructures such as airports, railway connections, and a strong fiber backbone network. These telecommunication and transportation infrastructures fundamentally enable pipeline of connectivity of major cities, such connectivity between cities in the world can be spatially represented, and different spatial analysis can be done to determine the importance of cities. Similarly, like the corporate organizational approach, there are streams of literature that specifically analyze telecommunication (i.e., the internet backbone) distinctly from physical transportation infrastructures such as airlines and railways (Derudder and Witlox, 2008).

Early studies on the infrastructure approach examine the "network of

networks" (i.e., the telecommunication network between locations) to analyze the world city network (Malecki, 2002; Rutherford, Gillespie, and Richardson, 2004). In the work of Malecki, 2002, he identified that internet backbones are located in major cities, and as such the formation of inter-city network based on fiber connectivity between cities can uncover the hierarchy of urban areas in the world. However, Malecki, 2002 also acknowledges that the analysis of the formation of transnational world city network from fiber connectivity can be misleading since the position of certain urban areas can be over-estimated due to their strategic role as points of interconnection between regions (Derudder and Witlox, 2008). Beyond telecommunication network, more studies have increasingly built inter-city network based on airline, rail, and sea flow, with the airline being used mostly (Derudder and Witlox, 2008). The advantage of airline data is that, unlike studies using the "corporate organizational" methodology, the measurement of flows between cities in this analysis are quantitatively observed and not inferred. While this is a key advantage of the airline data, however, the use of this kind of data is majorly plagued by the lack of source or destination information (Derudder and Witlox, 2008). Another drawback is that we are less likely to observe direct flight connections between cities at the top and below the hierarchy of urbanization, hence restricting more detailed geographical analysis of all cities in the world (Derudder and Witlox, 2008). An example of empirical research that applies airline data is the work by Derudder, Witlox, and Taylor, 2013 which contrasted the "connectivity profiles" of U.S. cities, to determine the relevance of a city's connections within and outside the U.S.

The corporate organization and infrastructure approaches of characterizing global cities have been applied extensively in classifying cities based on how they are economically integrated with the global economy. For example, the GaWC project published in Beaverstock, Smith, and Taylor, 1999 provides an empirically validated hierarchy of global cities in the world, which is widely adopted by researchers working on global cities. However, there have been some criticisms of the inadequacies of this approach. For example, Boschken, 2008 identifies the need to integrate more inclusive dimensions of what it means to be global, and Bassens,

Derudder, and Witlox, 2011 pointed out the western and capitalist bias in the global city list. In response, recent research (e.g., Belderbos, Du, and Goerzen, 2017 and Parnreiter, 2019) now explore other definitions of global cities integrating dimensions such as livability, innovation, and the ease of doing business. Some of the lists resulting from these city definitions include Global Power City Index, and MasterCard's Global Power List.

The scientific and technological collaboration networks of cities

Scientific and technological collaboration networks have been extensively studied predominantly in the network science domain. Typically, they represent interactions between entities (e.g., persons, cities, regions, countries), where such interactions are formed through co-authorship (scientific collaboration) or co-inventorship (technological or R&D collaboration). The analysis of such networks provides insight into the relevance of entities in the network, and patterns in the formation. There are different reasons for collaboration in either academic research or patenting. Collaboration can leverage the different skill-sets of authors and inventors, and can drive inter-disciplinary research output. Typically, R&D are characterized by high investment (expensive laboratory equipment), high risk and uncertainty (changing markets), and long payback period (Laperche and (Ed.), 2013; Li, Wei, and Wang, 2015), collaboration can be useful in minimizing these risks. Also, collaboration is useful for access to new knowledge and prevents lock-in-effect (Li, Wei, and Wang, 2015).

Early studies on scientific collaboration network (e.g., Newman et al., 2004) examined the formation of co-authorship network. In this network, two scientists are connected if they jointly write a paper. Newman et al., 2004 constructed networks of such connections by drawing from the "Los Alamos e-print Archive (Physics)", "MEDLINE (biomedical research)", and "NCSTRL (Computer science)" databases. They observe a "small world" effect in the collaboration network, which implies that any scientist can reach any other scientists in the network by a few intermediate collaborators (Newman et al., 2004). This work has been extended to examine

the structure of scientific collaboration by studying non-local statistics such as network distance between scientists, and centrality measures in these networks.

Beyond single-layer network, there have been studies focusing on analyzing multi-layer network to uncover how dynamics in one network affect the other. For example, Li, Wei, and Wang, 2015 examined the topological and spatial features of urban innovation networks in China and showed evidence of hierarchical diffusion and contagious diffusion in both the scientific and technological networks. Also, Angelou et al., 2020a described the growth in multi-layer network (i.e., European Framework program and Patent) using a kinetic-model of three differential equations and six parameters describing the system dynamics. They found that all multi-links exhibit similar growth patterns over-time and patents were seen as driving forces for the creation of common multi-links early on in time (Li, Wei, and Wang, 2015).

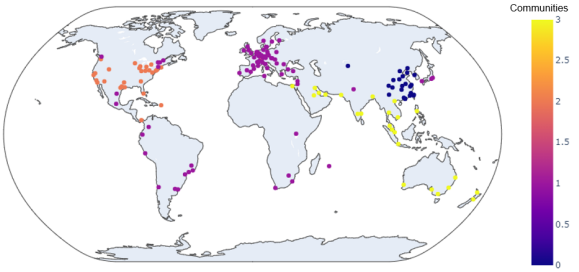
4.1.2 Multi-body collaboration in traditional network and hypergraph

In addition to multi-layer network analysis, there has been an effort at analyzing multi-body interactions, with more focus on triadic collaborations involving the formation of closed or open triangles in the network. Triangles are simply composed of three nodes in the network that are connected. They have been studied both in multi-layer networks and single layer network settings. For example, Lambiotte et al., 2008 identified inherent geographic communities and studied their network cohesion, by examining triangles constructed from a mobile phone communication network. Angelou et al., 2020a studied the formation patterns of triangles (closed triangles) in research and innovation collaboration networks, taking place between urban areas in Europe, by comparing the z-scores and clustering coefficient of the empirical and shuffled network data.

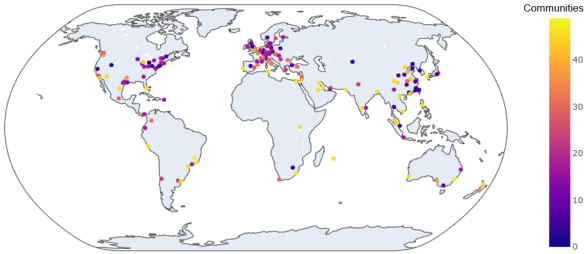
The core criticism of analyzing multi-body interactions from traditional networks built based on dyadic interactions is the fact that multi-body representations are then seen as combinations of dyadic interactions,

which can represent a false signal of such multi-body interaction taking place (Neuhäuser, Mellor, and Lambiotte, 2020). For example, consider triangles in a traditional network describing three cities (three nodes) connected (pairwise) with each other in the network. These connections are not necessarily formed because the three nodes were seen in a scientific paper or patent. It could be the case that the triangle identified in the traditional network resulted from the combination of three pairwise interactions (i.e., from three patents or publications) of the nodes. Such triangle is not a genuine multi-body interaction. Indeed, recent literature has identified this and proposed the use of simplicial complexes (Carletti, Fanelli, and Nicoletti, 2020) and hypergraphs (Neuhäuser, Mellor, and Lambiotte, 2020; Arruda, Petri, and Moreno, 2020; Arruda, Tizzani, and Moreno, 2021) for modeling multi-body interactions. An hypergraph simply consists of nodes (e.g., cities, countries, individuals, or firms), and hyperedges (a paper or patent) connecting nodes. An hyperedge is of size N, if it involves only N nodes (e.g., a hyperedge of size 3 will refer to only papers and patents having three nodes).

The differences between traditional network and hypergraphs for analyzing multi-body interactions can be seen in Figure 25a and 25b where we show the different outcomes when a community detection analysis on patent network is done using the different methodologies. In Figure 25a the network is built in the traditional sense i.e., cities are nodes, and there is a dyadic connection between cities if inventors in both cities are co-inventing, and the weight of the connection is the number of co-invented patents. The clustering analysis is done using the Louvain method for community detection (Blondel et al., 2008).



(a) Patent traditional network



(b) Patent hypergraph

Figure 25: Differences in community detection analysis of traditional patent network and patent hypergraph. More communities are identified in hypergraph compared to a traditional network, and community structure in the traditional network is national or regional, while community structure in hypergraph is internationally dispersed.

The key finding from the figure is that community structure is regional and national. This result is consistent with what has been shown in the literature (Chessa et al., 2013). While in Figure 25b the network is built using hypergraph representation, the cities are nodes, the connections are hyperedges (the analysis is done on hyperedges of all sizes), and the clustering is done using the algorithm proposed in Carletti, Fanelli, and Nicoletti, 2020. The community structure in this case is more internationally dispersed. This approach of properly capturing multi-body interactions shows a community structure different from the traditional approach.

4.2 Measures and Method

So far, in the analysis of social networks, the focus has been primarily on dyadic relationships even when the original interaction structure is more complex, such as in team production. Typically, most network representations lack information about complex interaction structures involving more than two nodes. This is problematic considering the rise of big science and the increase in the size of teams in scientific production and patenting. The use of pairwise interaction to analyze such complex networks of collaborations becomes inadequate since the information about team assembly and the structural mechanism that gives rise to innovation is lost. For instance, most publications and patents are outputs of groups and teams rather than of pairwise interactions between actors. The notion of multi-body interactions abounds across different scientific fields (e.g., social networks, functional brain networks, protein interaction networks, ecological networks) (Neuhäuser, Mellor, and Lambiotte, 2020; Lord et al., 2016; Estrada and Ross, 2018; Grilli et al., 2017). The analysis of complex interaction structures is better modeled using higher-order systems, which are generalized versions of networks.

In this chapter, we analyze the network of researchers as a hypergraph, which is one of the most common approaches for analyzing higher-order systems of interactions[2]. Hypergraphs encode interactions as hyperedges, which can be described as a set of nodes. For instance, a team of four scientists can be represented as a set of four nodes. This representation is flexible and powerful as it allows a specific encoding of all possible n-hyperedges. When n=2 (i.e., hyperedges are of size 2), the hypergraph reduces to a standard network or a 2-body system model. There are several advantages to using hypergraphs: (i) they can handle very large hyperedges efficiently, and (ii) they capture the heterogeneous distribution of the size of hyperedges (Carletti et al., 2020; Neuhäuser, Mellor, and Lambiotte, 2020). In addition, we can use a matrix representation to store information embedded in higher-order structures, thus avoiding a tensor

[2]Another common approach is the use of simplicial complexes to characterize the shape of the data in terms of the presence of holes between points (Carletti et al., 2020)

representation (Carletti et al., 2020). Recent applications of hypergraphs in the analysis of higher-order systems include studying the dynamics of social contagion (Arruda, Petri, and Moreno, 2020), modeling of random walks (Carletti et al., 2020), phase transition and stability of dynamical processes (Arruda, Tizzani, and Moreno, 2021), and nonlinear consensus dynamics (Neuhäuser, Mellor, and Lambiotte, 2020). For simplicity, our focus here is on a three-body dynamical system. In other words, we are interested in a system of interactions on a hypergraph consisting of 3-hyperedges. Unlike edges in the traditional network, the object hyperedge in a hypergraph can be seen as a patent or publication, and the size of the hyperedge is the same as the number of cities on the publication or patent, respectively. For example, a patent (publication) with three cities is represented as a 3-hyperedge, and the weight of a 3-hyperedge in the hypergraph is the number of patents (publications) jointly produced by researchers in those three cities. The object 3-hyperedge is distinctly different from triangles in a traditional network. Triangles in a traditional network can be described as three cities (three nodes) connected (pairwise) with each other in the network. Moreover, these connections are not necessarily brought about by researchers working together on a patent or publication in those three cities. It could be the case that the triangle identified in the traditional network results from the combination of three pairwise interactions (i.e., from three patents or publications) of cities. In this sense, triangles made by disjoint pairwise collaborations cannot be characterized as a genuine multi-body interaction involving three researchers. Indeed an ideal object for describing a multi-body interaction of size three is the 3-hyperedge which explicitly identifies patents or publications involving researchers from three cities. A motivation for focusing on 3-hyperedges in the analysis of international collaborations in patents and publications is because hyperedges of size 3 already represent a large number of multi-body interactions satisfying the condition of being international. For example, in the patent dataset they represent 84.83% and in the publication dataset, they represent 65.25% of international multi-body interactions. Considering the computational time of building null model used in identifying significant global cities in the empirical

hypergraphs, the choice of restricting the analysis to 3-hyperedges is reasonable. Moreover, this allows us to compare the hypergraph analysis with the usual triadic closure analysis.

Formally, for two-body interactions, the traditional edge-based networks denoted by $\mathcal{G} = (V, E)$, is composed of $V(\mathcal{G}) = \{1, ..., N\}$ representing a set of N nodes (i.e., cities), that are connected by a set of edges $E(\mathcal{G}) = \{(i, j) : i, j \in V(\mathcal{G})\}$. These edges exist if researchers from cities i and j are co-authors or co-inventors on a publication or patent, respectively. The network structure can be represented by the adjacency matrix $A \in \mathcal{R}^{N\times N}$ as follows:

$$A_{ij} = \begin{cases} w_{ij} & \text{if } (i, j) \in E(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

where w_{ij} is the number of patents or publications involving authors from cities i and j. For an undirected network, A is a symmetric matrix.

For three-body interactions, we encode the interaction using a hypergraph $\mathcal{H}$ which consists of a set $V(\mathcal{G}) = \{1, ..., N\}$ of N nodes (i.e., cities) connected by a set of 3-hyperedges (i.e. triplet of cities) $T(\mathcal{G}) = \{(i, j, k) : i, j, k \in V(\mathcal{G})\}$. Hence, we represent the structure of the hypergraph by the adjacency tensor $A \in \mathcal{R}^{N\times N\times N}$:

$$A_{ijk} = \begin{cases} w_{ijk} & \text{if } (i, j, k) \in T(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where w_{ijk} is the number of observed patents or publications involving inventors or authors from the three cities i, j and k. Figure 26 shows the types of hypergraph we analyze. The adjacency tensor is required to be symmetric and the graph should not include self-loops. This prevents the misclassification of edge combinations involving self-loops as triangles (Neuhäuser, Mellor, and Lambiotte, 2020).

Furthermore, each city $i \in \mathcal{G}$ is endowed with a dynamical variable $p_{ijk} = p(i|jk) \in [0, 1]$ describing the probability that city i is involved in a patent or publication in which cities j and k are also involved. This can

be represented as a probability tensor $P \in \mathcal{R}^{N\times N\times N}$ with entries

$$P_{ijk} = P(i|jk) = \begin{cases} \frac{w_{ijk}}{w_{jk}} & \text{if } (i,j,k) \in T(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

where w_{jk} is the number of patents or publications involving cities j and k and $\frac{w_{ijk}}{w_{jk}}$ is the fraction of patents or publications with cities j and k to which city i belong.

Given the hypergraph as described above, we are interested in knowing which of the observed fractions in the tensor P_{ijk} is statistically significant (at the significance level of 0.01) compared to what would be expected by chance. To this end, we compared the observed fractions with those computed from the ensemble networks (100 realizations) generated based on the hypergeometric ensemble model.

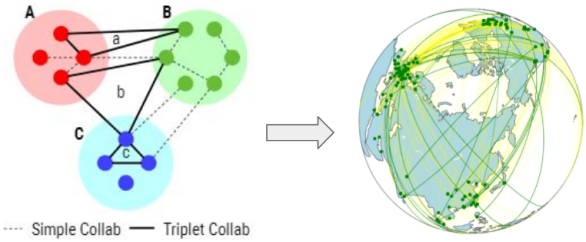


Figure 26: The global network of researchers as a hypergraph: the big circles are countries, and the small circles are cities that are nodes in the hypergraph. The hypergraph considered is composed of triplet collaborations, i.e., triplets *a,b,c*. Considering any red city in *triplet a* as the focal city, we have that the focal city is a third party in international collaboration. This is not the case if the green city is the focal city. In *triplet b*, any node in the triplet can be seen as a third party in international collaboration. *Triplet c* is discarded in the analysis as it refers to purely domestic collaborations. The objective of the analysis is to determine the global reach of cities. For example, the orthographic map shows the significant probabilities (yellow and green lines are probabilities less than and greater than 0.5, respectively) of San Francisco being a third party in different international teams (triplets of cities) of patent inventors.

4.2.1 Hypergeometric ensemble model

We use the hypergeometric ensemble model proposed by Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018 as a benchmark (null) model. The model is constructed based on the Molloy-Reed configuration model (Molloy et al., 2011), which requires randomly shuffling the topology of the network $\mathcal{G}$ while preserving the expected degrees of nodes in $\mathcal{G}$ (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018). In the standard configuration framework, each multi-edge is generated sequentially by sampling uniformly at random a node with an available "out-link" and a node with an available "in-link" until all links are exhausted (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018). The sampling or rewiring of these links is done uniformly to obtain a random realization of the null model. The drawback of the Molloy-Reed configuration model is that it is computationally expensive, not analytically tractable (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018), and does not allow the bias of the sampling process. By contrast, in the hypergeometric framework, the sampling of nodes can be biased based on the built-in propensity of nodes connecting with each other. Also, the hypergeometric framework follows a different sampling process. Rather than the sampling of nodes, the framework samples m multi-edges from a maximum combination of M multi-edges while preserving the sequence of expected nodes' degrees (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018). The hypergeometric ensembles can be formally defined as follows.

For each pair of nodes i and j, the degrees of i and j are $\bar{k}(i)$ and $\bar{k}(j)$ respectively, hence the maximum combinations of multi-edges between nodes i and j are $\psi_{ij} = \bar{k}(i)\bar{k}(j)$. In matrix form, the maximum possible combinations of multi-edges between pairs of nodes is given by $\Psi := (\psi_{ij})_{i,j \in V(\mathcal{G})}$. Specifically, for an undirected graph $\mathcal{G}$, we have that the total number of link combinations between vertices i and j is defined as follows:

$$\begin{cases} 2\psi_{ij} & \text{if } i \neq j \\ \psi_{ii} & \text{if } i = j \end{cases} \tag{4.4}$$

Let $\mathcal{G}$ be a network with n nodes and m multi-edges. Let $\mathbf{k} \in \mathbb{N}^n$ represent the degree sequence of $\mathcal{G}$ and $\hat{\mathbb{A}}$ the random model induced by $\mathcal{G}$. If the probability distribution underlying $\hat{\mathbb{A}}$ depends on $\mathbf{k}$ and m, and multi-edges have the same probability to be present, then $\hat{\mathbb{A}}$ has a multivariate hypergeometric distribution (MHD) defined as:

$$Pr(\hat{\mathbb{A}}) = \binom{M}{m}^{-1} \prod_{i<j\in V} \binom{2\psi_{ij}}{\hat{A}_{ij}} \prod_{k\in V} \binom{\psi_{kk}}{\frac{\hat{A}_{kk}}{2}} \tag{4.5}$$

For each pair of nodes i and j, the probability that $\hat{\mathbb{A}}$ there are exactly $\hat{A}_{ij}$ edges connecting i and j can be computed as the marginal distributions of the MHD in Equation 4.6:

$$Pr(\hat{A}_{ij}) = \begin{cases} \binom{M}{m}^{-1} \binom{2\psi_{ij}}{\hat{A}_{ij}} \binom{M-2\psi_{ij}}{m-\hat{A}_{ij}} & \text{if } i \neq j \\ \binom{M}{m}^{-1} \binom{\psi_{ij}}{\frac{\hat{A}_{ij}}{2}} \binom{M-\psi_{ij}}{m-\frac{\hat{A}_{ij}}{2}} & \text{if } i = j \end{cases} \tag{4.6}$$

A statistical ensemble $\Gamma := \{\hat{\mathbb{A}}\}$ is then defined as simply a set of sampled random models. Each of the random models $\hat{\mathbb{A}} = \left(\hat{A}_{ij}\right)_{i,j\in V}$ with entries such that $m = \sum_{ij} \hat{A}_{ij}$ results from sampling m edges from a possible total of $M = \sum_{ij} \psi_{ij}$ edges (Casiraghi et al., 2017; Casiraghi and Nanumyan, 2018).

In our analysis, the ensemble model described is applied to each matrix in the tensor, and the sampling of edges is purely combinatorial. A possible extension is to bias the sampling process by incorporating the spatial distance between the triplet cities. However, this is not important for our application as we are considering global cities that are deemed to have extensive economic reach, and we are focused on international collaboration. While distance can be a significant constraint on collaboration at the national level, however, beyond the national level, the increasing coordination cost of collaboration seems to become less important for global cities when collaborators can benefit from access to complementary knowledge, resources, or facilities (Scholl, Garas, and Schweitzer, 2018; Hennemann, Rybski, and Liefner, 2012).

4.3 Data

The primary sources of data for the analysis of publication and patent hypergraphs are *PubMed* and *PATSTAT*, respectively. *PubMed* is an open-access repository maintained by the U.S. National Library of Medicine. Pubmed covers more than 30 million publications, and disambiguated information on the location of authors is available (Li et al., 2014; Catini et al., 2015; Verginer and Riccaboni, 2021). The Patent Statistical (*PATSTAT*) database is the most extensive database that captures patent activities from over 90 patent offices located in different countries. The indicators constructed from this database have been used as proxies for technological activities in cities, regions, countries, and firms. However, less than 30% of the patents in the database do have address information for at least an inventor reported on a patent, and when addresses are available, sometimes they are not granular enough to capture technological activities in cities (Morrison, Riccaboni, and Pammolli, 2017a). Therefore, for the analysis of the patent network, we rely on a patent database that we developed in Chapter 1 specifically to address the problem related to the quality of address information (Belderbos et al., 2021). The hypergraph analysis is based on co-authorship and co-inventorship activities between 2005 and 2014 divided into two periods of 5 years: 2005-2009 and 2010-2014. The hypergraphs are constructed from patents and publications involving at least three cities (see Fig. 26). We have 21 293 (4 148) publications involving authors located in at least three cities in the period 2010-2014 (2005-2009), whereas we have 40 253 (32 971) patents involving inventors located in at least three cities in the period 2010-2014 (2005-2009). We opted to concentrate on 3-hyperedges because increasing multi-body interaction beyond size three would filter out most patents or publications from the analysis since there must be at least four global cities involved in producing a single paper or patent. In particular, in our data set and for the period 2010-2014, setting the size of hyperedges larger than three would exclude 84.83% and 65.25% of patents and publications, respectively. The distribution of hyperedges of different size constructed from patents in period 2010-2014 (2005-2009) are as follow: hyperedges of sizes

3, 4, 5 and above comprise of 84.8% (73.8%), 11.8% (20.3%), 2.6% (4.5%) and 0.8% (1.4%) of the total number of n-hyperedges ($n \geq 3$). While for hyperedges constructed from publications, we have 65.25% (62.5%), 22.6% (19.8%), 8.7% (8.1%), 3.5% (9.6%) respectively.

The nodes in the hypergraphs are cities, and the hyperedges are collaborations between cities. The weight of the hyperedges is the number of publications or patents involving the cities. We analyze the triangles (3-hyperedges) that emerge from the construction of this network by measuring the probability that a focal city will be the third party in a dyadic collaboration. In doing this, we selected 153 global cities drawn from the GaWC research project (Beaverstock, Smith, and Taylor, 1999). These cities are either *alpha* cities or *beta* cities (Beaverstock, Smith, and Taylor, 1999). As described in Chapter 1, in order to avoid the impact of differences in administrative boundary methodologies across countries on our analysis, we regionalized these cities into functional urban areas (FUAs) based on OECD, 2012 delineating methodology. According to the OECD definition: "An FUA consists of a densely inhabited city and of the surrounding area (commuting zone) whose labor market is highly integrated with the city" (OECD, 2012; Dijkstra, Poelman, and Veneri, 2019). The OECD approach for defining FUA uses population density to define urban cores and travel-to-work flow information to define periphery cities that are economically integrated with the urban core. For FUAs not in OECD countries, we approximate the travel-to-work flow used in identifying periphery cities with the patent density of the areas surrounding the focal city. In this study, there are 153 FUAs, located in Asia (47), America (44), Europe (48), and other regions (14).

Table 18 shows that the share of publications and patents with more than two inventors has increased. Furthermore, a growing share of collaborations involves inventors located in more than two cities.

R&D output	Time period	Share of output with more than two inventors (%)	Share of output with more than two inventors in multiple cities (%)
Patents	2005-2009	40.67	31.68
	2010-2014	48.79	40.46
Publications	2005-2009	11.06	5.08
	2010-2014	25.67	13.65

Table 18: The rising frequency of teams: the increasing share of scientific publications and patents with more than two authors.

4.4 Results

In this section, we analyze the position of global cities in the international networks of researchers. First, we analyze international collaborations involving at least three cities in different countries (international network), then we restrict the analysis to collaborations with researchers located in at least three cities in different continents (intercontinental network). Second, since in 3-hyperedges a city can be involved in an international collaboration with another city in the same country, we specifically examine international collaborations and intercontinental collaborations in which a given city joins other cities from different countries or continents, respectively. That is, we focus on hyperedges with cities (i, j, k), in which city i does not belong to the same country or continent as the ones of cities j and k (note that cities j and k could belong to the same or different countries and/or continents). Third, we examine each of the three main continental blocs (Asia, Europe, America) to which both cities j and k belong and identify which city i from outside the continental blocs has the highest probability of joining the collaboration. Furthermore, we examine the temporal changes between the periods 2005-2009 and 2010-2014 in the ranking of cities joining intercontinental collaborations (i.e., Asia-Europe, Europe-America, and America-Asia). Finally, we show the matrix representation of the collaborations of San Francisco, London, and Shenyang, and compare the global reach of San Francisco and London specifically for the patent hypergraph.

	International		Intercontinental	
	Top cities with high probabilities	*Top cities with significantly high probabilities*	*Top cities with high probabilities*	*Top cities with significantly high probabilities*
1	**San Francisco (Greater)** (0.0073)	**San Francisco (Greater)** (0.0038)	**San Francisco (Greater)** (0.0076)	**San Francisco (Greater)** (0.0037)
2	**New York (Greater)** (0.0047)	**New York (Greater)** (0.0025)	**New York (Greater)** (0.0047)	**New York (Greater)** (0.0024)
3	**Houston** (0.004)	**Paris** (0.0017)	**Houston** (0.0036)	**Shanghai** (0.0017)
4	**Paris** (0.0034)	**Houston** (0.0016)	**Paris** (0.0025)	**Houston** (0.0014)
5	**Shanghai** (0.0025)	**Shanghai** (0.0016)	**Los Angeles (Greater)** (0.0025)	**Paris** (0.0013)
6	**Los Angeles (Greater)** (0.0024)	**Tokyo** (0.001)	**Shanghai** (0.0024)	**Tokyo** (0.0011)
7	Chicago (0.0021)	Munich (0.0009)	Boston (0.0023)	Chicago (0.0011)
8	Boston (0.0021)	Taipei (0.0009)	Chicago (0.0021)	Boston (0.001)
9	**Tokyo** (0.002)	Chicago (0.0009)	**Tokyo** (0.0018)	Taipei (0.0008)
10	Taipei (0.0017)	Frankfurt am Main (0.0008)	Philadelphia (Greater) (0.0017)	**London** (0.0007)
11	**Washington (Greater)** (0.0016)	**Washington (Greater)** (0.0008)	San Diego (0.0014)	**Washington (Greater)** (0.0007)
12	San Diego (0.0016)	**Los Angeles (Greater)** (0.0007)	Dallas (0.0013)	Lyon (0.0007)
13	**London** (0.0015)	Boston (0.0007)	Taipei (0.0013)	**Los Angeles (Greater)** (0.0007)
14	Philadelphia (Greater) (0.0015)	**London** (0.0007)	**Washington (Greater)** (0.0013)	Philadelphia (Greater) (0.0007)
15	Munich (0.0015)	Seattle (0.0006)	**London** (0.0013)	Vancouver (0.0007)
16	Shenzen (0.0014)	Berlin (0.0006)	Montreal (0.0012)	Bangalore (0.0007)
17	Toronto (0.0014)	Toronto (0.0006)	Toronto (0.0012)	Munich (0.0006)
18	Dallas (0.0013)	Philadelphia (Greater) (0.0006)	Lyon (0.0012)	Frankfurt am Main (0.0006)
19	Beijing (0.0012)	Vancouver (0.0006)	Beijing (0.0012)	Seattle (0.0006)
20	Montreal (0.0012)	Brussels (0.0006)	Austin (0.0011)	Shenzen (0.0006)

(a) Patent hypergraph

	International		Intercontinental	
	Top cities with high probabilities	*Top cities with significantly high probabilities*	*Top cities with high probabilities*	*Top cities with significantly high probabilities*
1	**Paris (0.0017)**	Beijing (0.0015)	**Paris (0.0017)**	**Washington (Greater)** (0.0014)
2	Beijing (0.0017)	**Paris** (0.0012)	Beijing (0.0016)	Beijing (0.0013)
3	**London** (0.0015)	**Washington (Greater)** (0.0012)	**Washington (Greater)** (0.0014)	**Paris** (0.0012)
4	**Washington (Greater)** (0.0014)	**London** (0.0011)	**London** (0.0014)	**London** (0.0012)
5	**New York (Greater)** (0.0013)	**New York (Greater)** (0.001)	**Los Angeles (Greater)** (0.001)	**Los Angeles (Greater)** (0.0007)
6	**San Francisco (Greater)** (0.001)	**Tokyo** (0.0008)	**New York (Greater)** (0.001)	**Houston** (0.0007)
7	**Houston** (0.0009)	**Houston** (0.0007)	**San Francisco (Greater)** (0.001)	**Tokyo** (0.0007)
8	San Diego (0.0009)	**Los Angeles (Greater)** (0.0006)	**Houston** (0.0008)	**New York (Greater)** (0.0007)
9	**Tokyo** (0.0008)	Guangzhou (0.0006)	Copenhagen (0.0008)	**Shanghai** (0.0006)
10	Berlin (0.0008)	**Shanghai** (0.0006)	**Tokyo** (0.0007)	**San Francisco (Greater)** (0.0006)
11	**Los Angeles (Greater)** (0.0007)	San Diego (0.0006)	Berlin (0.0007)	Riyadh (0.0005)
12	Copenhagen (0.0007)	**San Francisco (Greater)** (0.0006)	**Shanghai** (0.0007)	Guangzhou (0.0005)
13	Rio de Janeiro (0.0006)	Berlin (0.0005)	Riyadh (0.0006)	Copenhagen (0.0005)
14	**Shanghai** (0.0006)	Copenhagen (0.0005)	Greater Adelaide (0.0006)	San Diego (0.0005)
15	Guangzhou (0.0006)	Hanoi (0.0004)	San Diego (0.0006)	Greater Adelaide (0.0004)
16	Greater Adelaide (0.0005)	Rio de Janeiro (0.0004)	Rio de Janeiro (0.0005)	Philadelphia (Greater) (0.0004)
17	Zurich (0.0005)	Zurich (0.0004)	Zurich (0.0005)	Berlin (0.0004)
18	Sacramento (0.0005)	Mannheim-Ludwigshafen (0.0004)	Sacramento (0.0005)	Zurich (0.0004)
19	Riyadh (0.0005)	Riyadh (0.0004)	Guangzhou (0.0005)	Mannheim-Ludwigshafen (0.0004)
20	Mannheim-Ludwigshafen (0.0004)	Hangzhou (0.0003)	Hangzhou (0.0004)	Rio de Janeiro (0.0004)

(b) Publication hypergraph

Table 19: Top twenty cities with high probabilities and significantly high probabilities to be third parties in international and intercontinental collaborations. In this analysis, one of the partnering cities can be in the same country as the focal city (case *a* in Fig. 26). The numbers in brackets represent the fractions of international or intercontinental collaborations in which the cities have high probabilities or significantly high probabilities of being third parties. Cities that are present in all the rankings are in bold.

Table 19 shows the top twenty cities with the highest probability of being present in 3-hyperedges (as defined in Equation 4.3). Table 19(a) shows the results for the patent network, whereas Table 19(b) reports similar results for publications. We compare the probabilities to be present in

international and intercontinental collaborations with those based on the hypergeometric null model. The threshold for high probability is a value greater than or equal to 0.75 i.e., $P(i|jk) \geq 0.75$, and the probability is said to be significant if it is estimated higher than the probabilities from the null model at a significance level of 0.01. The number in bracket represents the fraction of 3-hyperedges in which the city has a high probability or significantly high probability. Table 19(a) shows that San Francisco ranks at the top of the list both in terms of high probabilities and significantly high probabilities of being a third party in an international collaboration, in 0.73% and 0.38% of cases, respectively. San Francisco also tops the list for intercontinental collaborations considering both high probabilities and significantly high probabilities. Table 19(b) shows that Paris tops the list of cities with a high probability to be a third party in international collaboration. However, Beijing tops the list when we consider significantly high probabilities to become a third party in an international collaboration.

Similarly, Paris tops the list of cities with a high probability of being in an intercontinental collaboration, whereas Washington tops the list of cities with a significantly high probability of joining intercontinental collaborations. Only nine cities are present in all the rankings: five US cities (San Francisco, New York, Washington, Los Angeles, and Houston), two Asian cities (Tokyo and Shanghai), and two European cities (London and Paris). Beijing scores very high in the publication network, but it is not a significant top player in the intercontinental patenting network.

In the analysis summarized in Table 19, we considered all international and intercontinental collaborations (both types a and b of Fig. 26 are included). Now we limit the analysis to the cases where both partnering cities (i.e., cities j and k) are from outside the focal city's (i.e., city i's) country or continent (case b in Fig. 26 only). Table 20 shows the top-ranking global cities by the probability of taking part in international teams when co-inventors are located abroad or on different continents. Table 20(a) shows that, for the patent network, San Francisco is still the city with the highest probability to take part in international and intercontinental collaborations. Table 20(b) shows that, for the publication network,

London tops the list of cities with the highest probability to take part in international collaborations.

	International (outside focal city's country)		Intercontinental (outside focal city's continent)	
	Top cities with high probabilities	*Top cities with significantly high probabilities*	*Top cities with high probabilities*	*Top cities with significantly high probabilities*
1	**San Francisco (Greater)** (0.0068)	**San Francisco (Greater)** (0.0038)	**San Francisco (Greater)** (0.0065)	**San Francisco (Greater)** (0.0037)
2	**New York (Greater)** (0.0042)	**New York (Greater)** (0.0021)	**New York (Greater)** (0.004)	**New York (Greater)** (0.0023)
3	**Houston (0.0033)**	Chicago (0.0011)	Chicago (0.0026)	Chicago (0.0017)
4	**Los Angeles (Greater)** (0.0021)	Paris (0.001)	**Los Angeles (Greater)** (0.0023)	**Los Angeles (Greater)** (0.0011)
5	Paris (0.002)	Munich (0.0009)	**Houston (0.0023)**	Tokyo (0.0009)
6	Boston (0.0018)	**Houston** (0.0009)	Boston (0.0017)	Philadelphia (Greater) (0.0009)
7	Chicago (0.0018)	**Washington (Greater)** (0.0008)	Minneapolis (0.0014)	**Washington (Greater)** (0.0009)
8	Toronto (0.0014)	**Los Angeles (Greater)** (0.0008)	Dallas (0.0014)	Seattle (0.0006)
9	London (0.0013)	London (0.0007)	Philadelphia (Greater) (0.0014)	Paris (0.0006)
10	Munich (0.0012)	Bangalore (0.0007)	San Diego (0.0011)	**Houston** (0.0006)
11	Dallas (0.0012)	Tokyo (0.0006)	Shanghai (0.0009)	Boston (0.0006)
12	Shanghai (0.0011)	Malmo (0.0006)	**Washington (Greater)** (0.0009)	Shanghai (0.0006)
13	**Washington (Greater)** (0.0011)	Brussels (0.0006)	Paris (0.0009)	Bangalore (0.0006)
14	Brussels (0.0011)	Shanghai (0.0006)	Austin (0.0009)	Minneapolis (0.0006)
15	Philadelphia (Greater) (0.0011)	Boston (0.0005)	Bangalore (0.0009)	Dallas (0.0006)
16	Tokyo (0.0011)	Taipei (0.0005)	Tokyo (0.0009)	Greater Melbourne (0.0003)
17	San Diego (0.0011)	Seattle (0.0005)	Seattle (0.0006)	Pittsburgh (0.0003)
18	Bangalore (0.0009)	Philadelphia (Greater) (0.0005)	London (0.0006)	Albany (0.0003)
19	Minneapolis (0.0008)	Toronto (0.0005)	Detroit (Greater) (0.0006)	Taipei (0.0003)
20	Copenhagen (0.0008)	Vancouver (0.0005)	Phoenix (0.0006)	Bucharest (0.0003)

(a) Patent hypergraph

	International (outside focal city's country bloc)		Intercontinental (outside focal city's continent bloc)	
	Top cities with high probabilities	*Top cities with significantly high probabilities*	*Top cities with high probabilities*	*Top cities with significantly high probabilities*
1	London (0.0015)	London (0.0012)	London (0.0012)	**Los Angeles (Greater)** (0.0009)
2	Paris (0.0011)	**San Francisco (Greater)** (0.0008)	**San Francisco (Greater)** (0.0009)	**Houston** (0.0009)
3	**San Francisco (Greater)** (0.0009)	**New York (Greater)** (0.0008)	Copenhagen (0.0009)	London (0.0009)
4	**New York (Greater)** (0.0009)	Paris (0.0007)	**Houston** (0.0009)	**New York (Greater)** (0.0006)
5	Copenhagen (0.0007)	Copenhagen (0.0005)	**Los Angeles (Greater)** (0.0009)	**Washington (Greater)** (0.0006)
6	**Washington (Greater)** (0.0007)	Tokyo (0.0005)	**Washington (Greater)** (0.0006)	**San Francisco (Greater)** (0.0006)
7	Riyadh (0.0005)	**Washington (Greater)** (0.0005)	Prague (0.0006)	Prague (0.0003)
8	San Diego (0.0005)	Riyadh (0.0004)	Paris (0.0006)	New Delhi (0.0003)
9	Tokyo (0.0005)	**Los Angeles (Greater)** (0.0004)	**New York (Greater)** (0.0006)	Copenhagen (0.0003)
10	**Houston** (0.0005)	**Houston** (0.0004)	San Diego (0.0003)	Riyadh (0.0003)
11	Greater Sydney (0.0004)	Greater Sydney (0.0003)	Oslo (0.0003)	Sacramento (0.0003)
12	**Los Angeles (Greater)** (0.0004)	Hanoi (0.0003)	Riyadh (0.0003)	Karlsruhe (0.0003)
13	Rio de Janeiro (0.0004)	Sacramento (0.0003)	New Delhi (0.0003)	San Antonio (0.0003)
14	Berlin (0.0004)	San Antonio (0.0003)	Chongqing (0.0003)	Chongqing (0.0003)
15	Sacramento (0.0003)	San Diego (0.0003)	Karlsruhe (0.0003)	Philadelphia (Greater) (0.0003)
16	Athens (0.0003)	Philadelphia (Greater) (0.0003)	Sacramento (0.0003)	San Diego (0.0003)
17	Bologna (0.0003)	Zurich (0.0003)	Berlin (0.0003)	Albany (0.0003)
18	Pittsburgh (0.0003)	Rio de Janeiro (0.0003)	San Antonio (0.0003)	Greater Adelaide (0.0001)
19	Hanoi (0.0003)	Taipei (0.0002)	Albany (0.0003)	Greater Sydney (0.0001)
20	San Antonio (0.0003)	TelAviv (0.0002)	Philadelphia (Greater) (0.0003)	Munich (0.0)

(b) Publication hypergraph

Table 20: The ranking of top twenty cities with the highest probabilities to be third parties in international and intercontinental collaborations with other cities *from different countries and continents*, respectively. The numbers in brackets represent the fractions of such collaborations in which the cities have high probabilities or significantly high probabilities of being third parties. Cities in bold are present in all rankings.

However, when we consider the significantly high probabilities of intercontinental collaborations, Los Angeles tops the list. More interestingly,

only five cities are present in all rankings, and they are all US cities: San Francisco, New York, Washington, Los Angeles, and Houston. This finding further confirms the central role of the US innovation system.

Next, we analyze the probability that a focal city (i.e., city i) will take part in 3-hyperedges where cities j and k belong to the same continental bloc, but different from the bloc of city i. Indeed a city playing a significant role outside its continental bloc signals global reach and the ability to absorb distant knowledge effectively.

Asia	Europe	America
[1] **San Francisco (Greater)** (0.0044)	[1] Paris (0.0067)	[1] Houston (0.0066)
[2] Shanghai (0.0038)	[2] Mannheim (0.0033)	[2] New York (Greater) (0.0041)
[3] Taipei (0.0038)	[3] Stockholm (0.0024)	[3] Washington (Greater) (0.0025)
[4] Tokyo (0.0033)	[6] San Francisco (Greater) (0.0024)	[7] Amsterdam (0.0017)
[5] Houston (0.0022)	[12] Los Angeles (Greater) (0.0014)	[8] Beijing (0.0017)
[6] New York (Greater) (0.0022)	[13] New York (Greater) (0.0014)	[11] Brussels (0.0017)

(a) Patent hypergraph, three main continental blocs

Asia	Europe	America
[1] Beijing (0.0027)	[1] **New York (Greater)** (0.0014)	[1] New York (Greater) (0.0025)
[2] Guangzhou (0.0016)	[2] Berlin (0.0014)	[2] Houston (0.0025)
[3] Tokyo (0.0016)	[3] **San Diego** (0.0010)	[3] Washington (Greater) (0.0017)
[4] Paris (0.0011)	[4] Greater Sydney (0.0011)	[9] Zurich (0.0008)
[7] Washington (Greater) (0.0005)	[5] London (0.0010)	[12] Bologna (0.0008)
[8] San Francisco (Greater) (0.0005)	[6] Paris (0.0010)	[13] London (0.0008)

(b) Publication hypergraph, three main continental blocs

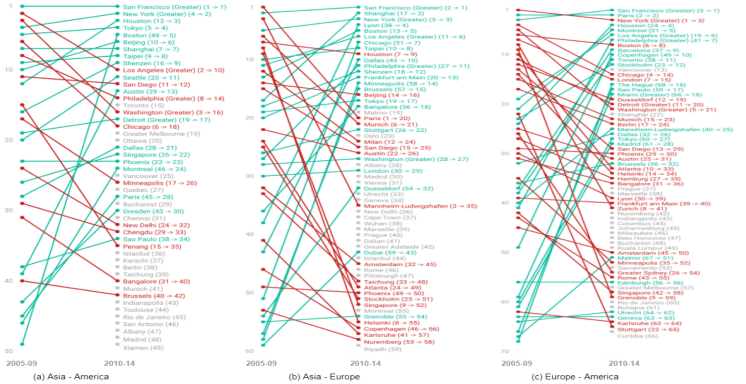
Table 21: Top cities with significantly high probabilities of being third parties in the main continental blocs (Asia, Europe, and America). Domestic collaborations are not considered. The American bloc includes North and South America. Rankings are in square brackets, while the fraction of dyadic international collaborations in the bloc the city enters with a high probability is reported in parentheses. Cities in the top three positions from different continents are highlighted in bold.

For each continental bloc (i.e., Asia, Europe, and America), Table 21 lists the top three cities in the ranking located in the continental bloc and the top three cities from outside the continental bloc. Table 21(a) shows that, for the patent hypergraph, San Francisco dominates all cities in the Asia bloc, including Asian cities. San Francisco is the top non-European city in the European bloc and ranks higher than many other European cities (except for Paris, Mannheim, and Stockholm). Finally, in the American

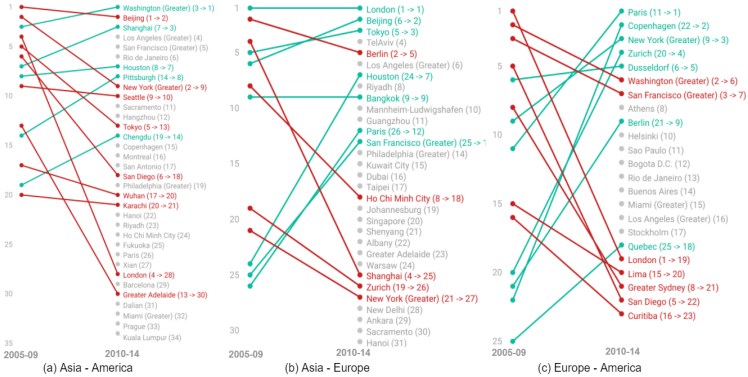
bloc, Amsterdam is the topmost non-American city. Table 21(b) shows that for the publication hypergraph, local cities dominate the Asian and American blocs. However, in the European bloc, New York City plays a significant role in being a third party in collaborations taking place within the European bloc. In particular, New York City shares this prominent role with Berlin (i.e., in 0.14% of 3-hyperedges of this kind), but fairs better than London. In summary, we find that American cities (specifically San Francisco and New York) play a pivotal role in Asian teams of patent co-inventorships and European scientific collaborations, respectively.

4.4.1 Temporal changes in the ranking of cities emerging in intercontinental collaborations

Figure 27a and 27b show the changes in the ranking of cities between the periods 2005-2009 and 2010-2014 for the patent and publication hypergraphs, respectively. The ranking of cities is based on the fraction of 3-hyperedges where cities have a high probability (i.e., $P(i|jk) \geq 0.75$) of being third parties for the period 2010-2014. Findings show three sub-panels (Asia-America, Asia-Europe, and Europe-America) of more disaggregated intercontinental collaborations for both patent and publication hypergraphs. For example, the ranking of city i in an Asia-Europe collaboration is based on the fraction of 3-hyperedges (specifically cases where city j and city k are in Asia or Europe and $P(i|jk) \geq 0.75$) in which city i appears. Cities in green moved up or maintained their ranking, while cities colored in red experienced a decline in ranking between the periods 2005-2009 and 2010-2014, and cities colored in grey only satisfied the condition of high probabilities for the period 2010-2014.



(a) Patent hypergraph



(b) Publication hypergraph

Figure 27: Changes between the periods 2005-2009 and 2010-2014 in the ranking of top cities with high probabilities (i.e., $P(i|jk) \geq 0.75$) of being third parties in intercontinental collaborations (Asia-America, Asia-Europe, and Europe-America). The green cities and red cities moved up and down the ranking respectively, while the grey cities were not in the ranking for 2005-2009 (i.e., their probability was less than 0.75 for all 3-hyperedges they appeared in). For each sub-panel, the cities are ordered based on their ranking for the sub-panel in 2010-2014. Note: America comprises of cities in North and South America.

For the patent hypergraph, Figure 27a suggests that only 49, 59, and 66 cities had high probabilities of being in Asia-America, Asia-Europe,

Europe-America collaborations, respectively. Figure 27a shows that San Francisco has remained the city with high probabilities in most hyperedges consisting of Asia-America collaborations. For Asia-Europe collaborations, San Francisco and Shanghai moved from 2nd and 17th position to 1st and 2nd, respectively. Paris retained its second place for Europe-America collaborations, while San Francisco moved from 3rd place to 1st. The observed significant leap of Shanghai in both Asia-America and Asia-Europe hyperedges, is an indication of the city being a contender for San Francisco's top position in the future. The cities in grey are new entrants whose probabilities to emerge as third parties in the relevant collaboration (e.g., Asia-America, Asia-Europe, and Europe-America) are less than 0.75 in the period 2005-2009, but larger than 0.75 in the period 2010-2014. For Asia-America, Asia-Europe, and Europe-America collaborations, we have more new entrants from America, Europe, and America, respectively.

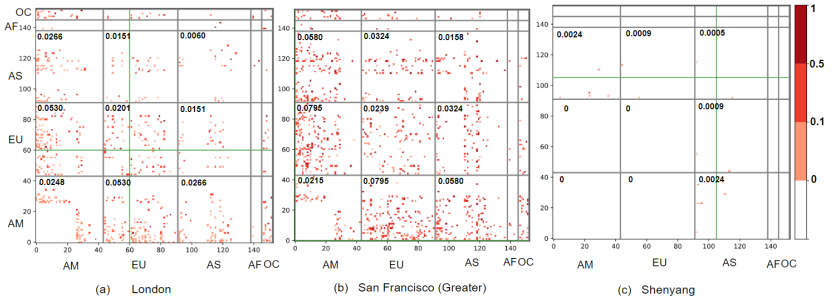
Generally, top cities in the US are well positioned to be third parties in technological partnerships across any dimension of intercontinental collaborations. In the ranking of cities for Asia-America collaborations, US cities moved up the rank on average by approximately 4 points, with Boston making the most significant leap from 49th position in 2005-2009 to 5th position in 2010-2014. Although top Chinese cities such as Beijing and Shenzen moved up the rank, Chinese cities declined by 2 points on average. Also, we observe that European cities are not third parties with high probabilities of being in Asia-America collaborations. In ranking cities with high probabilities to be third parties in Asia-Europe collaborations, we find more European cities than Asian cities. However, on average, European cities declined by 3 points, while Chinese cities moved up on average by 1 point, and cities in the US moved up on average by 9 points. Similarly, for Europe-America collaborations, cities in Asia are less likely to act as third players. We have more European cities than American cities in the list, and European cities moved up the rank on average by 1.34 points, while US cities moved up the rank on average by 0.9 points.

For the publication hypergraph, Figure 27b shows that only 34, 31, and

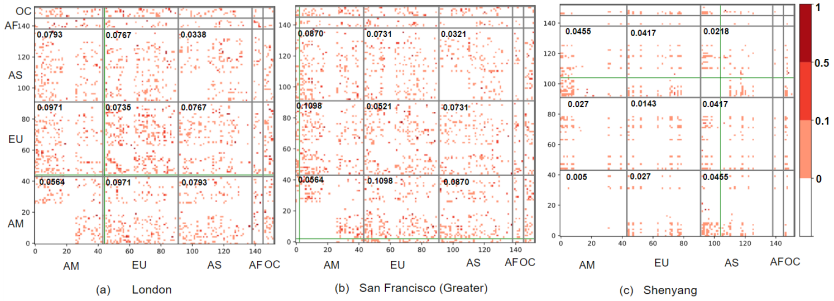
23 cities had high probabilities to be in Asia-America, Asia-Europe, and Europe-America collaborations, respectively. Figure 27b shows the changes in the ranking of cities between the periods 2005-2009 and 2010-2014. This is based on the fraction of 3-hyperedges where cities have a high probability (i.e., $P(i|jk) \geq 0.75$) of being third parties in either Asia-America, Asia-Europe, and Europe-America collaborations. With a high probability of joining the highest fraction of Asia-America scientific collaborations as a third party, Washington (Greater) moved from the third position in 2005-2009 to the top position in 2010-2014. The leading cities in Asia-Europe and Europe-America scientific publications are London and Paris, respectively. Generally, on average, cities in Asia moved up the ranking for Asia-America collaborations, while more new cities from America appeared in the ranking. Similarly, for Asia-Europe collaborations, cities in America moved up the ranking while more new cities from Asia entered the ranking in 2010-2014; and for Europe-America collaborations, cities in Europe moved up the ranking while more new cities from America entered the ranking in 2010-2014.

4.4.2 The case of San Francisco, London, and Shenyang

Figure 28 shows the projection of the adjacency tensor of patent and publication hypergraphs for selected cities: London, San Francisco, and Shenyang. Having fixed the focal city of reference (e.g., London), the matrix shows the significant probabilities (i.e., cases where $P(i|jk)$ in the observed network exceeds that computed from the network ensembles generated using the hypergeometric model) of observing the fixed city as the third party in the different 3-hyperedges. Cities are grouped in continental blocs for the sake of readability, e.g., the AM bloc consists of cities in North and South America, the EU consists of cities in Europe, and so on. The figure is organized into blocks (e.g., the bottom left block AM-AM refers to the case of observing the fixed city in America-America international collaborations), and in each block, it shows the fraction of 3-hyperedges where the fixed city has significant probabilities.



(a) Patent hypergraph



(b) Publication hypergraph

Figure 28: The adjacency tensors: Each matrix represents the projection for three cities (London, San Francisco, and Shenyang) in the patent hypergraph (subplot a) and the publication hypergraph (subplot b). Cities are grouped in continental blocs, while the matrix is divided into collaboration blocks. The green line depicts the position of the focal city in the matrix; each dot refers to the actual probability of collaboration of the focal city with a given pair of global cities (in rows and columns) when this probability is higher than the hypergeometric benchmark, and the number in each block is the fraction of 3-hyperedges in that block where the focal city's actual probabilities exceed the hypergeometric benchmark.

As suggested by the fraction in each block of the patent and publication hypergraphs, London is a top city in the publication hypergraph, San Francisco is a top city in the patent hypergraph, and Shenyang is not a top city in either hypergraph. When comparing the probability to be a third party in global city collaborations (see Figure 28) San Francisco stands out as the city with the highest probabilities and the largest fraction

of 3-hyperedge collaborations. It consistently dominates London across every block considered. In the publication hypergraph in Figure 28, we notice that London is better at being significantly present in collaborations where one city is from Europe (i.e., Europe-Asia and Europe-Europe blocks), whereas San Francisco is better at being significantly present when a city in America is involved (i.e., America-Asia, America-America, and America-Europe blocks). Shenyang is significantly present in Asian collaborations, but London and San Francisco always dominate it. For both the publication and patent hypergraphs, we see sparser matrices for less prominent global cities, such as Shenyang, compared to San Francisco and London. Furthermore, we observe sparser matrices in the patent hypergraph than in the publication hypergraph for all three cities (London, San Francisco, and Shenyang). Indeed, this pattern does not only apply to the three cities but is also observed in 78% of the global cities in our data set.

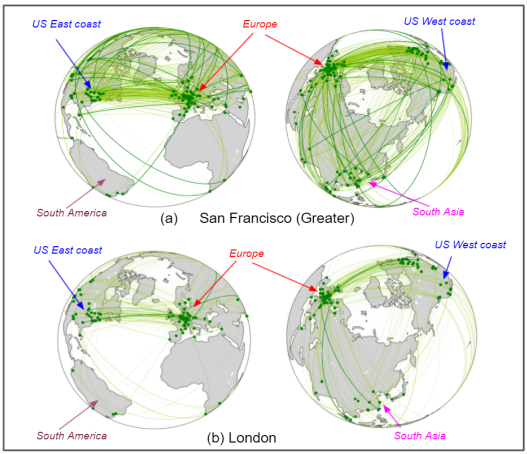


Figure 29: The global reach of San Francisco and London in the patent network: The yellow and green lines represent significant probabilities less and greater than 0.5, respectively.

To sum up, we find that US cities play a pivotal role in international networks of inventors. Among them, San Francisco is the leading city

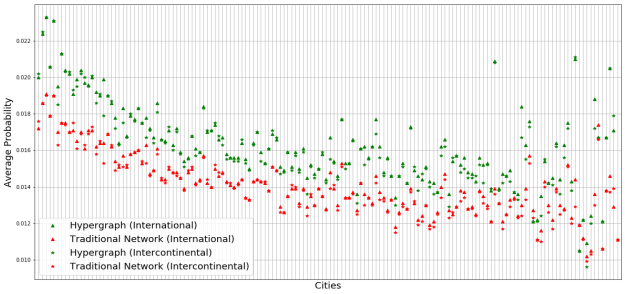
in international collaborations. Figure 29 represents the orthographic projection of the significant probabilities of San Francisco and London to take part in 3-hyperedges for the patent hypergraph. We find that San Francisco has a greater global reach than London. Specifically, San Francisco is significantly present in 536 international collaborations, 44 of which have a probability exceeding 0.5. Most of the significant probabilities exceeding 0.5 refer to cities in America and Europe, while the lowest probabilities are of cities located in Asia. On the other hand, London has a significant global reach in 357 international collaborations, of which only 4 have probability exceeding 0.5: Helsinki-Seoul, Atlanta- Istanbul, Amsterdam-Luxembourg, and Columbus-Madrid.

4.4.3 Comparison of probabilities based on traditional network and hypergraph approach

Following the traditional network approach, we built a network where cities are nodes, and linkages are formed when there is collaboration between researchers located in these cities. Using the traditional network approach implies we relaxed the condition that there should be at least three cities on a patent or publication to at least two cities. By doing this we increased the number of patents and publications in the period 2010-2014 from 40 253 and 21 293 to 182 946 and 231 479 respectively. From these patents and publications, we identified 99 745 and 169 478 triangles respectively. In a similar approach as in the hypergraph, we compute $P(i|jk)$ which is the probability that a city i is involved in a patent or publication in which cities j and k are also involved. Similarly, we used the hypergeometric model described in Section 4.2.1 to determine if these probabilities are significant (i.e., overestimated than what is obtained in the ensemble of randomized networks).

Based on the traditional network approach, first, we find that for both patent and publication networks, none of the cities had a probability $P(i|jk)$ exceeding 0.75 (i.e., high probability to be in a triangle). This can be attributed to the fact that the traditional network approach overestimates the number of genuine triangles (or genuine multi-body interac-

tions) making the fractional share very small. Since we cannot compare the result from the traditional network with what is reported in the hypergraph (in Table 2 and 3), we choose to analyze both international and intercontinental collaborations by comparing the average probabilities for each city that are computed based on both hypergraph and traditional network approaches.



(a) Patent

(b) Publication

Figure 30: Comparison of average probabilities computed in the traditional network and hypergraph. The cities on the horizontal axis are ordered (decreasingly) based on the fractional share of 3-hyperedges they emerge in.

Figure 30a and 30b show the average probability (in cases where the

probability from the null model is underestimated) of a city to emerge in international or intercontinental collaborations using both the traditional network and hypergraph approaches. The cities on the horizontal axis are ordered (decreasingly from left to right) based on the fractional share of 3-hyperedges they emerge in. For both approaches applied to patent and publication we observe that for either international or intercontinental collaborations, cities with more innovative reach do this with higher probabilities, and this is more evident for publication than it is for patent. Generally, we observe that for both publication and patent, the average probability in most cities is higher in hypergraph compared to the traditional network. Specifically, for analysis of international collaborations, 134 (for publication) and 148 (patent) cities do have an average probability (significant) being underestimated in the traditional network approach, while for inter-continental collaborations, it is the case for 132 (for publication) and 148 cities (for patent). This implies that the traditional network approach used in identifying multi-body interactions underestimates these kinds of collaborations.

	Patent	**Publication**
Mean Probability (International)	0.4934 (3.66e-17)	0.1447 (0.0828)
Std. Probability (International)	0.2829 (9.19e-6)	0.1316 (0.1440)
Mean Probability (Intercontinental)	0.4802 (3.05e-16)	0.1645 (0.0325)
Std. Probability (Intercontinental)	0.2632 (4.91e-5)	0.1184 (0.2374)

Table 22: Kolmogorov-Smirnov test: Comparing the distribution of mean probability and standard deviation computed in both traditional network and hypergraph. The number in the bracket is the p-value. Generally, the distribution is not identical for the case of patents, and for the mean probabilities of intercontinental collaborations in publication.

Table 22 shows the Kolmogorov-Smirnov (KS) test comparing the distribution of the probability mean and standard deviation computed using both approaches. The KS test shows that at an alpha level of 0.05, for patent, the distribution of the average probability in the traditional network is not identical as in the hypergraph, this result is the same for both international and intercontinental analysis. Similarly, for publication, the distribution of the average probabilities is not identical in both hypergraph and the

traditional network for intercontinental analysis, but identical for the analysis of international collaboration.

4.5 Discussion

Global cities are important hubs of innovation and cultural and economic activities: they are composed of most of the world's population, and serve as an engine of creativity (Verginer and Riccaboni, 2021). For this reason, their ability to plug into international networks of researchers secures them access to distant knowledge that can be absorbed and diffused to other cities within the country.

In light of this, this work has analyzed the international network of researchers based on publications and patents. Traditionally, the analysis of the international network of researchers has been conducted at the level of specific countries; when the focus is on regions and cities, the analysis is restricted to regions and cities within a given country or continent (Sachini, Sioumalas-Christodoulou, and Chrysomallidis, 2021; Angelou et al., 2020b). Also, recent work has concentrated primarily on dyadic interactions (Sachini, Sioumalas-Christodoulou, and Chrysomallidis, 2021; Wagner, Whetsell, and Leydesdorff, 2017), with few exceptions such as Li, Wei, and Wang, 2015 and Angelou et al., 2020b. Our study contributes to this literature by examining the international network of researchers for both patents and publications from a hypergraph perspective. To this end, we constructed a measure from 3-hyperedges in the hypergraph and used a configuration model to build null models in order to identify the cities that act as global players. Recent work that has focused on patent and publication networks from a similar perspective has carried out the analysis of triangles on networks in the traditional sense. For example, Angelou et al., 2020b examined the multiplex network of patents and European "Framework Programmes" (FPs) with the aim of uncovering temporal variations in the formation patterns of triangles. This traditional approach does not properly capture multi-body interactions, as a triangle in a network could emerge from a combination of pairwise inter-

actions, and therefore lead to an over-estimation of the actual three-body interactions observed in the network system.

In this study, we observe that the share of patents with more than two inventors and the share of patents with more than two inventors in multiple cities have increased by approximately 8% and 9%, respectively, between the periods 2005-2009 and 2010-2014. Similarly, for publications, the increase has been 14% and 8%, respectively. Second, we found a sparser but strong significant matrix of collaborations in the patent network than in the publication network, which reflects the fact that complex collaborations in patents require more effort than publications and do not take place as often as scientific publications because they typically involve greater coordination costs. Third, we found that San Francisco is a significant third-party global player in complex collaborations in the patent network, irrespective of whether such collaboration occurs within or outside the US. This observation is not surprising as San Francisco is home to the Silicon Valley where there are headquarters of major patenting companies with global reach. We also see that San Francisco maintains the top position even when we restrict the analysis to inter-continental collaborations. An interesting observation is that San Francisco plays a top role within the Asian bloc and also serves as a pivotal non-European player within the European bloc.

Generally, we have a proliferation of cities in the US (compared to Europe and Asia) as top third parties in intercontinental collaborations. This is likely due to the abundance of multinational companies and top research centers in the US. Also, the rise of China, often observed in the patent network, cannot be detected when we consider a complex collaboration system, as only a handful of top cities such as Shenzen and Beijing moved up the ranking over time. Similarly, in the publication network of researchers, we found London to be the most global third player. However, the role of London becomes weaker when we consider intercontinental collaborations and collaborations within the Asian bloc.

A limitation of the analysis is that we focus on global cities, and as such we exclude 3-hyperedges from publications and patents formed between

global and non-global cities, hence the bias in our result will be greater for cities with more distributed innovation capabilities in non-global cities. Also, notice that the results of our analysis of the patent and publication hypergraphs should be interpreted with caution as both hypergraphs represent multi-body interactions in different fields. While patents in PATSTAT are filed across different technological fields, publications in PubMed consist of papers published mainly in the life sciences. The implication of the difference in technology fields covered in patents and publications multi-body analysis is that the cities that emerge to be important in the patent multi-body analysis can be said to be embedded in higher-order structures with variety of technological flows allowing for such cities to have more piece in the technology space that can be used for combinatorial innovative purpose. On the other hand, we cannot say the same for cities emerging in the publication multi-body analysis, since the scientific space is limited, the piece for combinatorial purpose may be limited as well. Another limitation of the analysis pertains to the null model used in determining if the likelihoods estimated in the empirical network are underestimated or not. The null model is built based on randomization of the network adjusting for city size, other propensities for a network formation such as cultural norms, language similarity, and distance between innovators can be used to bias the randomization of edge formation in the network. Adequately representing the effect of distance in the randomization process can be tricky especially for cross-border analysis, as geographical borders have been shown to significantly hinder geographic span of networks. For example, linkages in the US has been shown to have faster exponential decay, while linkages in the US decay as power of distance (Cerina et al., 2014), this observation may be as a result of a more homogeneous innovation system in the US. Future research work can explore the best functional representation for the distribution of distance (power law or exponential) that can be used to bias the randomization done in the null model. This is important as this can shed more light on whether the different national innovation systems in Europe becoming more integrated have impacted its ability to source ideas globally (i.e., beyond the EU). As we currently observe in the analysis for

both patent and publication, European cities were less likely to emerge as core cities in hyperedges involving other continents - the challenge might be that the EU is still far away from having an homogeneous innovation system that allow it to strongly source knowledge from the world. Policies such as the European Research Area (ERA) around the creation of borderless market can be strengthened to achieve this.

In conclusion, by examining a hypergraph structure consisting of hyperedges of size three, our findings contribute to a better understanding of how to properly identify the cities that can be characterized as global players in various types of international and intercontinental collaborations with other cities. Our results on identifying cities with significant global reach in the system of international multi-body interactions for patents and publications have important policy implications as they can help governments and policy-makers to better design effective policies that can spur innovation in cities. Future direction of this research can examine identification of cities that are "rising stars", the clustering of cities in networks of researchers of n-hyperedges ($n \geq 3$), and can also investigate the role of exogenous shocks (e.g., economic crisis) on the clustering of cities and emergence of cities as global hubs in complex collaborations.

Conclusion

Historically, cities have contributed significantly to cross-border economic transactions. In the past, cities such as Venice and Genoa, and those located in the Baltic and Arab regions facilitated global trade through their ports, and this remains the case, as huge containers still transport goods across cities in different continents (WEF, 2014). These activities of cities in facilitating trade revolutionized Europe and Asia, as many European cities prior to the industrial revolution were seen as major commercial centers of the world (WEF, 2014). In this new century, the difference in economic development has been attributed primarily to differences in technological development between cities and regions.

Technological progress and innovation have increasingly become important for economic growth of cities, and similar to the intensity of commercialization that happened in cities centuries ago, cities are at the forefront of scientific and technological innovation as they institute more productive and targeted policies. Policies put together by policymakers at the city and regional levels have become more flexible and effective, as the proximity of policymakers to residents in the city, and the sizeable number of residents located in the city has allowed for better policy experimentation and adaptation (WEF, 2014). Therefore, to build on the successes of innovation policies in cities, being able to quantitatively and qualitatively evaluate the effect of policy instruments on innovation outcomes might be very useful. A granular geolocated patent data can lend itself useful for quantitative evaluations.

Previous research (such as (Balland and Boschma, 2021; Boschma, Balland, and Kogler, 2015)) have relied on the use of patent data, however, some of the available patent database either lack granularity or do have limited coverage and are prone to national bias, or allocate patent to cities based on national boundary definitions which imply that differences in national boundaries across countries will impair sound international comparisons of technological output of cities. Recent effort such as (De Rassenfosse, Kozak, and Seliger, 2019; De Rassenfosse and Seliger, 2021) in address-

ing the issue of imputing missing addresses do a good job at allocating addresses to patents but not necessarily identifying which inventor or applicant on the patent this address is associated with. Moreso, these works regionalize the given address to geographical areas with boundary conditions that are administratively defined, and as such the regions do not necessarily represent the economic extent of the geographical areas. In Chapter 1 of the thesis, we contribute to existing work that addresses allocating missing address to inventors and applicants on patent, as well as employing an harmonized way of delineating cities to functional urban areas that represent the economic reach of cities. The importance of this is uniformity in the definition of geographical boundaries enhances international comparison of innovation outcomes of geographical areas. When compared to the raw PATSTSTA dataset, the final dataset for both inventors and applicants shows a significant improvement in the availability of georeference addresses. This new dataset can be applied to research that examines networking strategies of inventors or applicants and the impact on productivity, distinguishing between emerging and star firms patenting in new areas such as AI and blockchain technologies. In Chapter 1 we demonstrate the potential use of this data by providing a preliminary analysis on the evolution of domestic linkages and international linkages. Much of the literature in international business have been concerned about the consequences of international connectedness and domestic disconnectedness (Lorenzen, Mudambi, and Schotter, 2020). Since domestic disconnectedness does have the potential of energizing populist backlash against multinational firms (Lorenzen, Mudambi, and Schotter, 2020). In our analysis of the evolution of domestic linkages and international linkages to both global and non-global cities, we observe that there has been a decline in unweighted (representing the extensiveness of innovative ties) domestic linkages following the financial crisis, but a steady increase in international unweighted and patent-weighted (representing the intensiveness of innovative ties) linkages to both global and non-global cities. Specifically, developing economies e.g., cities in Africa rely more on international linkages for innovation process, as they pursue formation of ties with global cities mostly in North America and

Asia. While cities in developed economies in Asia pursue the formation of ties with both international global and non-global cities, and North America cities pursue the formation of ties only with international global cities. The limitation of this preliminary analysis is the fact that we can only observe linkages between approximately 1200 cities in the database, hence the analysis and subsequent analysis in other chapters are biased in favor of cities more likely to form ties with the cities we have. Secondly, the quality of the database is impacted by ambiguities in names of inventors and applicants which is a major issue in patent database. Since the retrieval of missing addresses for inventors and applicants rely on looking into similar names in subsequent filings within the database, the similarity measure can only address minor cases of ambiguities. Therefore, future work on improving the quality of this database can focus on improving disambiguation algorithms (such as Morrison, Riccaboni, and Pammolli, 2017a, Li, Wei, and Wang, 2015, and Cuxac and Bonvallot, 2013) which would enhance the missing address retrieval process.

In Chapter 2 we investigate how to forecast the economic complexity and competitiveness of global cities using machine learning models. The literature on economic complexity has proposed that economic development involves accumulating diversified capabilities rather than increasing specialization in few technological areas. The principle is that most complex technologies that are needed by most people are highly leveraged and require various expertise. Cities with diversified capabilities do have the pre-requisite know-how to produce them (Balland and Rigby, 2022). The growth of these technological capabilities at the city-level hinges on the different specialization of economic agents, such that the society knows better because of differences in capabilities of its economic agents (Balland and Rigby, 2022). This notion of economic capabilities of cities is distilled in the economic complexity index capturing the productive capabilities of cities. While there are various strands of the economic complexity metrics (e.g., "Economic Complexity Index" (ECI) (Hidalgo et al., 2007), "Fitness and Complexity" (FC) (Tacchella et al., 2012) and their variations) that has been applied to measure capabilities of countries, we extend this application to the measurement of capabilities in cities by using recent

economic complexity methodology called the generalized economic complexity (GENEPY) index (Sciarra et al., 2020) which is a novel approach that reconciles both the ECI and FC using a multidimensional framework, that leverages the strengths of both methods. This way of measuring capabilities of cities adds to the toolbox of economics, as it approaches the problem using dimensionality reduction techniques that preserves information than aggregations typically used in economics (Balland and Rigby, 2022). Also, the economic complexity approach provides a new way of studying technology differently from a consequential perspective in which technology is seen as a shift parameter in measures of total factor of productivity (Balland and Rigby, 2022). By providing the notion of technology space, we can capture information of exact technologies needed to upgrade to any given technology - this can be useful in better understanding development outcomes of cities. Measuring capabilities is not a forward looking process, as economic complexity does not tell us what happens in the future in terms of whether the city becomes competitive (i.e., specialize) in a given technological area. The creation of strong institutions, industrial districts are some areas in which cities allow for a creative space that can be useful to firms and individuals, which in turn can enhance the competitiveness of these cities. Following the Ricardian comparative advantage, a competitive city should be patenting more than its fair share in a given technological area. To forecast the possibility of this happening in the future, we can rely on the historical information of the comparative advantage of cities to uncover the relationship between technologies. Combining this with an understanding of the technological portfolio in the city can help us know what technology the city would specialize in the future. The network science literature provides ways to think about the relationships between technologies, which are mostly from a linear perspective such as co-occurrence and taxonomies (Pugliese et al., 2019; Engelsman and Raan, 1991) which are problematic as technological associations can be higher-order (not pairwise) and non-linear. Methodologically, in line with recent works Tacchella et al., 2021 (forecasting competitiveness of countries using trade data) and (Straccamore and Zaccaria, 2021) (forecasting technological entry of firms using patent

data), we contribute to the discussion on forecasting the competitiveness structure of cities in technological areas by showing a better predictive machine learning model when compared to the time-delayed co-occurrence model proposed in Pugliese et al., 2019, and does not leverage the autocorrelation structure in the revealed technological advantage matrix. The performance of the machine learning model is better than results from similar work (e.g., Straccamore and Zaccaria, 2021) using patent data. Finally, we show that using the predicted competitiveness technological structure of cities, we can provide a prediction of the future capabilities of cities (generalized economic complexity index). We observed that the predicted GENEPY vary significantly from the actual GENEPY for more complex cities, and less so for least complex cities, suggesting gradual incremental changes in capabilities at early developmental stages.

This analysis is important for planning technological pathways of cities as it is well known that urbanization growth is interwoven with the economic capabilities and technological competitiveness. As cities evolve and become more urbanized, the process of urbanization increases economics of scale of production activities within the city, boosts productivity, and increases competition. The advantage of an increase in urbanization alongside the right bureaucratic environment and infrastructural investment can foster economic growth and innovation. Therefore, being able to effectively forecast the underlying capabilities of cities being distilled in an economic complexity index can be useful knowledge for cities in different ways. First, such knowledge can serve as benchmark in strategic innovation plans and can help policymakers calibrate how much impact the city can have on other developmental indicators such as reduction in greenhouse gas emission, gender inequality, and institutional quality given the presence of favorable policy space facilitating recommendations consistent with predicted economic capabilities. The forecasting exercise follows a data-driven strategy and machine learning models that are black-box, and do not incorporate how the policy space (e.g., such as a regional policy on strengthening institutions supporting innovation) and interventions impact economic capabilities. While we show that our model does have better predictive power and can reconstruct future

economic capabilities, the model predictive power specifically for the second task (i.e., forecasting competitiveness in cities that have $RTA < 0.25$ between 2000-2008) can be improved upon by consolidating the features constructed from the patent database with other policy, economic and industrial features. Further research can be done to address these issues and specifically provide index of economic capabilities of cities that is directly predicted and not one based on the predicted revealed technological advantage of cities in different technological areas.

In Chapter 3 we investigate how cities enter new technological areas. We rely on previous work on entry (Rigby, 2015) and diversification (Balland and Boschma, 2021) into new technology domains, by examining the effect of inter-city ties on entry into new technology - since exploratory activities of cities can be affected by technological structures in other cities, and the formation of ties with players in other cities can lead to diffusion of information that can serve as pre-requisite for new technological areas (Balland and Boschma, 2021; Rigby, 2015). We find evidence that this is indeed the case - that increased inter-city ties increase the likelihood that a city will enter a new technological domain, and this effect is enhanced when the ties are located in partner cities with large pool of inventors. Also, following the theory on relatedness, we examine the effect of relatedness of a city's existing technological stock with new technology. This is because similar capabilities needed to explore a new technological area will mean that the city can easily assimilate the new technology in a more useful way, rather than spending time and resources in gathering capabilities needed for the new technology. We find that the theory of relatedness holds, as cities are more likely to enter a new technological domain if the domain is related to their existing capabilities. What we observe is that the effect of technological relatedness generally outweighs the effect of inter-city ties on entry process. This can be because not all linkages matter and the inclusion of linkages with knowledge flows not complementary with the existing capabilities in a city might downgrade the effect of these ties on entry process. Since we have seen in the work of Balland and Boschma, 2021 that complementarity of capabilties is very useful in positively moderating the effect of regional linkages on techno-

logical diversification. The limitation of the analysis done in Chapter 3 is that the focus on formation of linkages does not identify if these linkages are a result of technological acquisition i.e. when a city enters a new technology during collaboration, the applicant that owns the patent is located in the city, or knowledge spillover i.e., the applicant that owns the patent is located in the partner city. Future research work can examine the impact of inter city ties and technological relatedness on these distinct kinds of entry. Also, the analysis does not take into account the role of institutions in entry process, as strong institutions should facilitate technological entry. In measuring technological relatedness, the measurement can be improved upon given the work done in Chapter 2. Although Chapter 2 is being framed to predict competitiveness, the same analysis can be done to predict entry, in which the probability score for entry can be a good proxy for the proximity of technological capabilities of cities to new technologies. Although doing this would require more historical data for the model to meaningfully predict entry in early periods. While both analyses in Chapters 2 and 3 confirm that relatedness can facilitate competitiveness and entry into new technologies, the analysis done in Chapter 3 still does not identify if technologies in the core or periphery of the cities' technological structure are driving the entry process. Finally, since the definition for the newness of a technology in a city is city-specific and not global (i.e., not new to the world), future research can examine the entry process into technologies that are new to all cities (i.e., new to the world). Other limitations are the inherent problems in the data work which we already discussed in Chapter 1.

The findings in Chapter 3 can be useful to how we think of implementing the different innovation policies such as EU smart specialization, Chile Innova, Canada' super cluster initiative and similar initiatives. For example, to achieve a more integrated innovation system in the EU, cities can leverage capabilities in other cities that do not exist in their city. To achieve this may require some policy instruments by the EU, policies such as publicly funding projects requiring collaborations between cities competent in a given technological area and those that are not, this will provide learning opportunities that can further allow innovation in areas

new to a city or ensure that projects consistent with the technological capabilities of cities are preferentially funded when collaborations are not being pursued (Balland and Boschma, 2021; Bednarz and Broekel, 2019; Barzotto and Tomlinson, 2019). Also, this can be implemented by incorporating the global city status of a city, since non-global cities are observed to rely more on ties to other cities that are competent in the new technology. Hence the former criteria for funding can be appropriate to this class of non-global cities, and the latter to the class of global cities. Such implementation can help diffuse any perceived bias in how projects are being funded by the different innovation institutions.

The analysis in Chapters 2 and 3 rely on building the network based on dyadic relationships between cities (in the case of inter-city network) or between IPCs (in the case of technological network). In Chapter 4 we investigated the likelihood of cities to emerge significantly in transnational triadic scientific and R&D collaborations based on network structures that capture higher-order interactions (i.e., beyond dyadic relationship). The motivation is that the increase in international team formation on scientific publication and technological patent suggests the need for a mathematical structure that properly encodes higher-order interactions. Hence, our methodological contribution involves the use of hypergraph structure to provide a measure for the likelihood that a city emerges in transnational triadic collaboration in these multiplex networks (i.e., publication and patents), and identifying cases in which this is done significantly by building a null model based on hypergeometric ensemble of random graphs. The result of this analysis is that US cities play a pivotal role in international networks of higher-order structure, irrespective of the dimensions of collaborations we examined - this is the case for both the patent and publication network. Specifically, these 5 US cities (San Francisco, New York, Washington, Houston, Los Angeles) consistently emerge prominently in higher-order international collaborations across all dimensions we examine, with San Francisco being more prominent. While the methodology applied may appear complex, we demonstrate that similar analysis of triadic structure in traditional network do not yield the same result as that done on hypergraph suggesting that the hypergraph struc-

ture encodes higher-order structure differently from traditional network, hence are more appropriate for such analysis. The analysis of higher collaboration structure between cities is increasingly important as cities increasingly rely on plugging into international networks of innovators to access distant knowledge. Therefore identifying cities that are central hubs in the system of multi-body interactions can be useful to know what cities to foster collaborations with when sourcing for capabilities to enter a new technology or become competitive in a technology. Future research can examine if the results of this analysis differ across technological fields and for hyperedges of higher orders beyond the triadic case. Also, examining the clustering of cities in the hypergraph for hyperedges of all sizes can provide insight as to if a city belongs to similar technological cluster with cities that are likely to emerge in higher-order collaborative structure. Further research can also address the drawback in the null model as it only account for cities' size (i.e., number of patents or publications), and does not account for other propensities such as geographic distance, and similarity in language and culture that can bias collaboration.

The thesis examines entry process and competitiveness of cities in different technological areas, which are important activities for resilient and dynamic cities. The development of a new database in the thesis provides an avenue to address the question not just from the perspective of global cities concentrated in developed economies, but also for non-global cities in emerging economies. In response to the observation of Balland and Boschma, 2021 on the limited focus on the role of inter-regional linkages in smart specialization policies, the thesis examines the role of inter-city linkages on likelihood of the city entering new technologies, the thesis provides evidence of the importance of linkages, especially for non-global cities and cities in emerging economies in Asia for entry process (findings from Chapter 3). Pursuing collaboration is the most viable strategy for less developed cities for both entry and competitiveness - this can be seen in the findings of Chapter 2 in which predicting competitiveness based on technological relatedness can be difficult for less developed cities. For less developed cities to catch-up, they must focus on policies that incentivize collaboration with developed economies. The thesis also highlights the

emergence of rising stars like Shanghai with economic reach similar to San Francisco in the multiplex innovation network. Also, the thesis highlights the importance of a path-dependent technological strategy to be relevant not just for entering a technology but also for becoming competitive in the technological area. The process of entering more technological areas brings about varieties of knowledge that are useful to the economical complexity of the city, as economic complexity of cities relies on technological diversity - especially in complex technologies. Being economically complex can be very important in predicting income levels and determining if the city will experience rapid growth in the future (Hidalgo, 2021), hence it is a useful measure of economic development.

Appendix A

Supplementary material for Chapter 1

A.0.1 Statistics for inventor dataset

	Inventor country	*Geocoding ratio*	*Number of patents*	*Inventor country*	*Geocoding ratio*	*Number of patents*
0	Afghanistan	1.00	22	Libya	0.91	22
1	Albania	0.96	93	Liechtenstein	0.87	1576
2	Algeria	0.88	285	Lithuania	0.45	1381
3	American Samoa	0.57	7	Luxembourg	0.91	3665
4	Andorra	0.82	88	Macao	0.63	193
5	Angola	0.83	24	Madagascar	0.85	39
6	Anguilla	0.63	19	Malawi	0.89	9
7	Antarctica	0.50	2	Malaysia	0.66	12329
8	Antigua and Barbuda	0.92	36	Maldives	1.00	2
9	Argentina	0.43	5950	Mali	0.97	36
10	Armenia	0.66	343	Malta	0.86	320
11	Aruba	0.83	12	Marshall Islands	1.00	13
12	Australia	0.89	51595	Mauritania	0.96	123
13	Austria	0.89	50810	Mauritius	0.82	108
14	Azerbaijan	0.24	478	Mexico	0.54	13567
15	Bahamas	0.86	357	Micronesia, Federated States of	1.00	4
16	Bahrain	0.92	61	Moldova	0.03	3143
17	Bangladesh	0.81	575	Monaco	0.82	647
18	Barbados	0.78	232	Mongolia	0.34	116
19	Belarus	0.22	3240	Montenegro	0.53	47
20	Belgium	0.92	45780	Montserrat	1.00	3
21	Belize	0.74	55	Morocco	0.80	692
22	Benin	0.50	22	Mozambique	0.60	5
23	Bermuda	0.78	282	Myanmar	0.59	123
24	Bhutan	0.71	7	Nauru	1.00	3
25	Bolivia	0.79	69	Nepal	0.74	259
26	Bonaire, Sint Eustatius and Saba	1.00	2	Netherlands	0.89	108309
27	Bosnia and Herzegovina	0.84	205	New Caledonia	0.97	34
28	Botswana	0.77	13	New Zealand	0.88	9746
29	Bouvet Island	1.00	1	Nicaragua	0.83	35
30	Brazil	0.83	22916	Niger	0.89	84
31	British Indian Ocean Territory	1.00	3	Nigeria	0.90	276
32	Brunei Darussalam	0.72	69	Niue	1.00	3
33	Bulgaria	0.61	2960	Norfolk Island	0.50	2
34	Burkina Faso	0.77	31	North Korea	0.85	483
35	Burundi	0.35	20	North Macedonia	0.84	126
36	Cabo Verde	1.00	5	Northern Mariana Islands	1.00	6
37	Cambodia	0.63	67	Norway	0.83	19846
38	Cameroon	0.85	134	Oman	0.84	89
39	Canada	0.91	161281	Pakistan	0.83	957
40	Cayman Islands	0.70	314	Palau	1.00	2
41	Central African Republic	0.88	17	Palestine, State of	0.90	11
42	Chad	0.80	5	Panama	0.76	321

	Inventor country	*Geocoding ratio*	*Number of patents*	*Inventor country*	*Geocoding ratio*	*Number of patents*
43	Chile	0.77	2665	Papua New Guinea	0.83	6
44	China	0.95	1492568	Paraguay	0.76	34
45	Cocos (Keeling) Islands	1.00	1	Peru	0.51	535
46	Colombia	0.49	2164	Philippines	0.85	2454
47	Comoros	0.80	5	Pitcairn	1.00	5
48	Congo	0.88	51	Poland	0.52	41432
49	Congo, The Democratic Republic of the	0.78	9	Portugal	0.73	5007
50	Cook Islands	1.00	2	Puerto Rico	0.93	138
51	Costa Rica	0.73	585	Qatar	0.90	295
52	Croatia	0.71	1446	Romania	0.50	9134
53	Cuba	0.48	723	Russia	0.12	230335
54	Curaçao	0.60	5	Rwanda	0.81	16
55	Cyprus	0.86	602	Réunion	0.83	29
56	Czech Republic	0.55	12893	Saint Barthélemy	1.00	1
57	Côte d'Ivoire	0.74	42	Saint Helena, Ascension and Tristan da Cunha	0.80	15
58	Denmark	0.92	33389	Saint Kitts and Nevis	0.83	64
59	Djibouti	0.80	25	Saint Lucia	0.75	8
60	Dominica	0.82	22	Saint Pierre and Miquelon	0.50	2
61	Dominican Republic	0.61	166	Saint Vincent and the Grenadines	0.67	6
62	Ecuador	0.64	718	Samoa	0.35	37
63	Egypt	0.88	1484	San Marino	0.48	128
64	El Salvador	0.76	96	Sao Tome and Principe	0.83	6
65	Equatorial Guinea	1.00	4	Saudi Arabia	0.93	3708
66	Eritrea	0.90	30	Senegal	0.75	48
67	Estonia	0.86	1277	Serbia	0.54	1286
68	Eswatini	0.89	81	Seychelles	0.82	57
69	Ethiopia	0.82	212	Sierra Leone	0.88	95
70	Falkland Islands (Malvinas)	1.00	3	Singapore	0.83	23583
71	Faroe Islands	0.89	29	Sint Maarten (Dutch part)	1.00	1
72	Fiji	0.89	18	Slovakia	0.54	3184
73	Finland	0.96	53089	Slovenia	0.59	4858
74	France	0.94	273586	Solomon Islands	0.67	6
75	French Guiana	0.75	8	Somalia	0.92	12
76	French Polynesia	1.00	9	South Africa	0.84	8047
77	French Southern Territories	1.00	3	South Georgia and the South Sandwich Islands	0.00	1
78	Gabon	0.92	60	South Korea	0.91	1551549
79	Gambia	0.82	11	Spain	0.77	55109
80	Georgia	0.15	1702	Sri Lanka	0.87	561
81	Germany	0.94	849314	Sudan	0.80	67
82	Ghana	0.94	134	Suriname	0.67	12
83	Gibraltar	0.94	96	Sweden	0.94	74266
84	Greece	0.79	5273	Switzerland	0.90	94033
85	Greenland	1.00	10	Syrian Arab Republic	0.61	111
86	Grenada	0.78	14	Taiwan	0.66	381021
87	Guadeloupe	0.90	11	Tajikistan	0.05	168
88	Guam	0.80	5	Tanzania, United Republic of	0.91	56
89	Guatemala	0.54	172	Thailand	0.83	3156
90	Guernsey	0.71	7	Togo	0.84	13
91	Guinea	0.82	11	Tokelau	0.94	34
92	Guyana	0.95	62	Tonga	0.82	11
93	Haiti	0.85	26	Trinidad and Tobago	0.94	188
94	Holy See (Vatican City State)	0.88	27	Tunisia	0.38	1640
95	Honduras	0.40	99	Turkey	0.65	13871
96	Hong Kong	0.86	15353	Turkmenistan	0.28	18
97	Hungary	0.70	9375	Turks and Caicos Islands	0.88	41
98	Iceland	0.89	1304	Tuvalu	0.91	32
99	India	0.95	85458	Uganda	0.86	44
100	Indonesia	0.73	1764	Ukraine	0.09	27316
101	Iran	0.91	1723	United Arab Emirates	0.89	1220
102	Iraq	0.76	85	United Kingdom	0.89	235692
103	Ireland	0.87	17041	United States	0.95	2377521
104	Isle of Man	0.95	74	United States Minor Outlying Islands	1.00	5
105	Israel	0.91	60935	Uruguay	0.53	604
106	Italy	0.93	113098	Uzbekistan	0.29	267
107	Jamaica	0.92	190	Vanuatu	0.54	11
108	Japan	0.97	2577186	Venezuela, Bolivarian Republic of	0.85	779
109	Jersey	0.98	125	Viet Nam	0.61	1154
110	Jordan	0.80	632	Vietnam	1.00	12
111	Kazakhstan	0.17	977	Virgin Islands, British	0.53	824
112	Kenya	0.87	405	Virgin Islands, U.S.	0.90	10
113	Kiribati	1.00	3	Wallis and Futuna	0.80	5
114	Kuwait	0.97	540	Western Sahara	1.00	1
115	Kyrgyzstan	0.14	110	Yemen	0.60	28
116	Lao People's Democratic Republic	0.92	50	Yugoslavia	0.68	489
117	Latvia	0.44	2129	Zambia	0.95	20
118	Lebanon	0.87	658	Zimbabwe	0.86	85
119	Lesotho	1.00	3	Åland	1.00	3
120	Liberia	0.87	15	Åland Islands	1.00	1

Table 23: Patent count and geocoding ratio for inventor's country.

A.0.2 Statistics for patent office

	Patent office	*Geocoding ratio*	*Number of patents*	*Patent office*	*Geocoding ratio*	*Number of patents*
0	Armenia	0.00	3	Japan	0.93	4736176
1	Asia Pacific	0.89	6673	Kyrgyzstan	0.00	3
2	Argentina	0.73	37632	South Korea	0.89	2182840
3	Austria	0.66	10891	Kazakhstan	0.10	30
4	Australia	0.82	605733	Lithuania	0.32	1721
5	Bosnia and Herzegovina	0.00	1	Luxembourg	0.65	1253
6	Belgium	0.61	3792	Latvia	0.27	2555
7	Bulgaria	0.59	5850	Morocco	0.77	11982
8	Brazil	0.73	329346	Monaco	0.20	140
9	Belarus	0.07	14	Moldova	0.08	4449
10	Canada	0.93	593069	Montenegro	0.97	1837
11	Switzerland	0.56	16856	Macao	1.00	1
12	Chile	0.73	14257	Malta	0.40	25
13	China	0.58	5450054	Mexico	0.84	202049
14	Colombia	0.83	15837	Malaysia	0.82	33712
15	Costa Rica	0.84	7781	Nicaragua	0.88	197
16	Cuba	0.75	2158	Netherlands	0.61	34190
17	Cyprus	0.51	406	Norway	0.87	59652
18	Czech Republic	0.61	20858	New Zealand	0.89	54869
19	Germany	0.92	929881	Africa Intellectual Patent Organization	0.81	1735
20	Denmark	0.78	5964	Panama	0.91	1826
21	Dominican Republic	0.76	3392	Peru	0.92	16370
22	Algeria	0.82	486	Philippines	0.94	6105
23	Eurasian Patent Organization	0.85	25112	Poland	0.55	66176
24	Ecuador	0.79	7584	Portugal	0.41	3350
25	Estonia	0.95	2238	Romania	0.25	10404
26	Egypt	0.76	4423	Serbia	0.79	6727
27	None	0.00	1	Russia	0.28	435065
28	European Patent Organization	0.98	1932285	Saudi Arabia	0.79	1985
29	Spain	0.50	37016	Sweden	0.67	34387
30	Finland	0.98	28382	Singapore	0.88	68354
31	France	0.89	238418	Slovenia	0.59	7484
32	United Kingdom	0.77	177788	Slovakia	0.67	6996
33	Gulf Cooperation Council	0.84	281	San Marino	0.21	608
34	Georgia	0.47	2933	El Salvador	0.93	1149
35	Greece	0.17	7120	Thailand	1.00	4
36	Guatemala	0.75	3810	Tajikistan	0.01	218
37	Hong Kong	0.82	82264	Tunisia	0.74	6952
38	Honduras	0.73	1225	Turkey	0.25	9341
39	Croatia	0.64	6642	Taiwan	0.72	650384
40	Hungary	0.77	23680	Ukraine	0.32	43890
41	Indonesia	0.78	1041	United States	0.95	4690940
42	Ireland	0.54	5039	Uruguay	0.79	10744
43	Israel	0.50	7234	Uzbekistan	0.33	6
44	India	0.95	27391	Vietnam	1.00	42
45	Iceland	0.90	3084	World Intellectual Patent Organization	0.86	2376495
46	Italy	0.50	133076	Yugoslavia	0.64	3898
47	Jordan	0.70	799	South Africa	0.83	90168

Table 24: Patent count and geocoding ratio for each patent office in the inventor dataset.

A.0.3 Statistics for Applicant dataset

	Applicant country	*Geocoding ratio*	*Number of patents*	*Applicant country*	*Geocoding ratio*	*Number of patents*
0	Japan	0.99	4134549.0	Paraguay	0.30	30.0
1	United States	0.96	2088362.0	Lao People's Democratic Republic	1.00	30.0
2	China	0.97	1633457.0	Afghanistan	0.86	28.0
3	South Korea	0.85	1401675.0	Kyrgyzstan	0.25	28.0
4	Germany	0.88	798078.0	Antigua and Barbuda	0.89	28.0
5	Taiwan	0.50	353885.0	Guernsey	0.93	27.0
6	France	0.89	277307.0	Uzbekistan	0.52	27.0
7	United Kingdom	0.91	161745.0	Aruba	0.69	26.0
8	Russia	0.10	155603.0	Albania	0.88	26.0
9	Netherlands	0.91	119769.0	Vanuatu	0.96	25.0
10	Switzerland	0.90	118989.0	Montserrat	1.00	25.0
11	Canada	0.84	109215.0	Gabon	0.92	24.0
12	Italy	0.86	97557.0	Botswana	0.68	22.0
13	Sweden	0.87	83738.0	Dominica	0.91	22.0
14	Brazil	0.23	60568.0	Cabo Verde	0.81	21.0
15	Finland	0.96	53588.0	Nigeria	0.95	21.0
16	Austria	0.77	41916.0	Madagascar	0.65	20.0
17	Spain	0.69	41432.0	NY	1.00	20.0
18	Israel	0.91	37094.0	Cameroon	0.53	19.0
19	Australia	0.94	36124.0	Papua New Guinea	0.89	19.0
20	Poland	0.29	35976.0	SU	0.63	19.0
21	Belgium	0.84	32170.0	Senegal	0.58	19.0
22	Denmark	0.92	28464.0	Faroe Islands	1.00	19.0
23	India	0.96	26275.0	Trinidad and Tobago	0.89	18.0
24	Singapore	0.91	23351.0	Mali	0.71	17.0
25	Ukraine	0.02	21105.0	Bangladesh	1.00	17.0
26	Hong Kong	0.66	19931.0	Réunion	0.76	17.0
27	Norway	0.80	15976.0	Montenegro	0.29	17.0
28	Ireland	0.93	15103.0	Bolivia	0.75	16.0
29	Turkey	0.58	10189.0	New Caledonia	1.00	16.0
30	Cayman Islands	0.81	10129.0	IB	1.00	15.0
31	Virgin Islands, British	0.80	9527.0	Nicaragua	0.93	15.0
32	Luxembourg	0.90	9052.0	Liberia	1.00	14.0
33	Czech Republic	0.37	8998.0	Holy See (Vatican City State)	0.86	14.0
34	Mexico	0.37	8494.0	SW	1.00	13.0
35	Hungary	0.35	8138.0	Zimbabwe	0.77	13.0
36	Argentina	0.09	8062.0	TX	1.00	12.0
37	South Africa	0.84	7472.0	Eritrea	1.00	12.0
38	New Zealand	0.94	6836.0	Tuvalu	0.92	12.0
39	Romania	0.23	6430.0	Sudan	0.83	12.0
40	Bermuda	0.93	5983.0	Central African Republic	0.92	12.0
41	Malaysia	0.55	5608.0	Syrian Arab Republic	0.33	12.0
42	Liechtenstein	0.88	4872.0	Tanzania	1.00	11.0
43	Barbados	0.96	4489.0	French Polynesia	0.45	11.0
44	Slovenia	0.54	3901.0	American Samoa	0.36	11.0
45	Portugal	0.63	3677.0	FL	1.00	11.0
46	Saudi Arabia	0.96	3562.0	Djibouti	1.00	10.0
47	Greece	0.66	2515.0	Macedonia	0.50	10.0
48	Moldova	0.01	2476.0	Tokelau	1.00	10.0
49	Chile	0.62	2186.0	Burundi	0.50	10.0
50	Morocco	0.12	2036.0	Myanmar	1.00	10.0
51	Bulgaria	0.48	1931.0	Grenada	0.89	9.0
52	Slovakia	0.37	1811.0	Congo	0.62	8.0
53	Latvia	0.40	1608.0	NJ	1.00	8.0
54	Cyprus	0.82	1556.0	Congo D.R.C	1.00	8.0
55	Belarus	0.06	1552.0	Guam	0.88	8.0
56	DD	0.93	1449.0	Yemen	1.00	8.0
57	Colombia	0.40	1439.0	Ghana	0.88	8.0
58	AN	0.86	1397.0	Sao Tome and Principe	0.75	8.0
59	Tunisia	0.96	1257.0	Saint Lucia	1.00	8.0
60	Iceland	0.92	1095.0	Palau	1.00	7.0
61	Malta	0.91	955.0	Benin	0.29	7.0
62	Thailand	0.80	944.0	Comoros	0.83	6.0
63	Lithuania	0.39	943.0	Guadeloupe	0.83	6.0
64	Panama	0.68	856.0	Guinea	0.83	6.0
65	Philippines	0.84	843.0	British Indian Ocean Territory	1.00	6.0
66	Estonia	0.88	788.0	Kiribati	1.00	6.0
67	Bahamas	0.86	777.0	Zambia	0.83	6.0
68	Croatia	0.69	770.0	OH	1.00	6.0
69	United Arab Emirates	0.89	714.0	Nepal	1.00	5.0
70	Cuba	0.43	638.0	CT	1.00	5.0
71	Serbia	0.11	591.0	Rwanda	0.80	5.0
72	Mauritius	0.88	554.0	Fiji	1.00	5.0
73	Uruguay	0.28	533.0	Ethiopia	0.80	5.0
74	Ecuador	0.64	507.0	OA	0.60	5.0
75	Monaco	0.72	473.0	Suriname	0.75	4.0

	Applicant country	*Geocoding ratio*	*Number of patents*	*Applicant country*	*Geocoding ratio*	*Number of patents*
76	Seychelles	0.68	446.0	Tonga	0.75	4.0
77	Puerto Rico	0.88	425.0	Uganda	0.75	4.0
78	Kazakhstan	0.26	408.0	Equatorial Guinea	0.50	4.0
79	Egypt	0.56	405.0	WA	1.00	4.0
80	Costa Rica	0.55	374.0	Somalia	0.75	4.0
81	Samoa	0.48	367.0	Solomon Islands	0.75	4.0
82	Gibraltar	0.85	347.0	Libya	0.50	4.0
83	San Marino	0.49	341.0	Iraq	0.25	4.0
84	Indonesia	0.76	332.0	Saint Helena, Ascension and Tristan da Cunha	1.00	4.0
85	North Korea	0.86	320.0	Mauritania	0.25	4.0
86	Peru	0.26	299.0	Cambodia	1.00	4.0
87	UK	1.00	290.0	Saint Barthélemy	1.00	3.0
88	Venezuela	0.83	261.0	Cocos (Keeling) Islands	1.00	3.0
89	Jordan	0.74	249.0	Angola	1.00	3.0
90	Iran	0.90	237.0	Chad	0.67	3.0
91	Brunei Darussalam	0.67	222.0	Haiti	0.33	3.0
92	Georgia	0.37	196.0	Turkmenistan	0.33	3.0
93	Isle of Man	0.96	194.0	Burkina Faso	0.67	3.0
94	Macao	0.55	179.0	MI	1.00	3.0
95	Azerbaijan	0.25	171.0	EA	0.50	2.0
96	Belize	0.67	165.0	Maldives	1.00	2.0
97	Qatar	0.93	165.0	GC	1.00	2.0
98	YU	0.24	159.0	Bhutan	1.00	2.0
99	Dominican Republic	0.49	152.0	TU	1.00	2.0
100	Saint Kitts and Nevis	0.89	150.0	BC	1.00	2.0
101	Kuwait	0.91	130.0	UT	1.00	2.0
102	UN	1.00	124.0	Northern Mariana Islands	1.00	1.0
103	Virgin Islands, U.S.	0.85	124.0	Wallis and Futuna	1.00	1.0
104	Jersey	0.97	115.0	OX	1.00	1.0
105	Vietnam	0.76	112.0	XN	1.00	1.0
106	Cook Islands	0.94	104.0	YS	1.00	1.0
107	Kenya	0.91	101.0	EM	1.00	1.0
108	Honduras	0.09	96.0	PD	1.00	1.0
109	Sri Lanka	0.91	89.0	PO	1.00	1.0
110	Lebanon	0.65	86.0	DC	1.00	1.0
111	Tajikistan	0.00	86.0	F1	1.00	1.0
112	EP	0.91	82.0	Mozambique	1.00	1.0
113	AP	0.75	76.0	Palestine	1.00	1.0
114	Andorra	0.72	76.0	NM	1.00	1.0
115	Niger	0.94	69.0	FX	1.00	1.0
116	Pakistan	0.90	68.0	Guinea-Bissau	0.00	1.0
117	Saint Vincent and the Grenadines	0.89	63.0	Malawi	0.00	1.0
118	Algeria	0.30	63.0	JR	1.00	1.0
119	Turks and Caicos Islands	0.83	60.0	BU	0.00	1.0
120	Curaçao	0.95	58.0	Nauru	1.00	1.0
121	CS	0.72	53.0	HI	1.00	1.0
122	Anguilla	0.75	51.0	Heard Island and McDonald Islands	1.00	1.0
123	Bouvet Island	0.98	50.0	Greenland	0.00	1.0
124	Mongolia	0.59	49.0	Christmas Island	1.00	1.0
125	Swaziland	0.96	47.0	TA	1.00	1.0
126	Armenia	0.59	44.0	Gambia	1.00	1.0
127	Côte d'Ivoire	0.71	38.0	Falkland Islands (Malvinas)	1.00	1.0
128	Marshall Islands	0.82	38.0	SF	1.00	1.0
129	Sierra Leone	0.84	37.0	LL	1.00	1.0
130	El Salvador	0.34	35.0	LX	1.00	1.0
131	Oman	0.91	35.0	Lesotho	1.00	1.0
132	Bosnia and Herzegovina	0.75	32.0	French Guiana	1.00	1.0
133	Bahrain	0.81	31.0	BX	1.00	1.0
134	Guatemala	0.68	31.0	Pitcairn	1.00	1.0
135	Jamaica	0.97	31.0	NB	1.00	1.0

Table 25: Patent count and geocoding ratio for Applicants' country.

Year	**2000**	**2001**	**2002**	**2003**	**2004**
Geocoding ratio	85.8	83.2	84.5	82.9	85.8
Year	**2005**	**2006**	**2007**	**2008**	**2009**
Geocoding ratio	87.3	88.1	87.3	86.9	86.9
Year	**2010**	**2011**	**2012**	**2013**	**2014**
Geocoding ratio	88.6	89.3	80.9	56.4	48.3

Table 26: The yearly geocoding ratio of applicant dataset.

A.0.4 Python codes

```
1
2  ###This script is applied to the dataset repetitively.
3
4  def collect_coordinate(data):
5      '''data: subset of the data with coordinate'''
6
7      data['person_name'] = data['person_name'].astype(str)
8
9      mapp = defaultdict(dict)
10
11     for row in tqdm(data.itertuples()):
12         mapp[row[1+data.columns.get_loc('docdb_family_id')]][row
       [1+data.columns.get_loc('person_name')]][row[1+data.columns.
       get_loc('appln_auth')]] = row[1+data.columns.get_loc('
       coordinate')]
13
14     return mapp
15
16
17 def deterministic_imputation(mapp,x):
18     '''data: subset of the data without coordinate'''
19
20     try:
21         for i in ['US','EP','WO']:
22             if i in mapp[x['docdb_family_id']][x['person_name']]:
23                 return mapp[x['docdb_family_id']][x['person_name'
       ]][i]
24                 continue
25             else:
26                 otheroffice = random.sample(list(mapp[x['
       docdb_family_id']][x['person_name']]))
27                 return mapp[x['docdb_family_id']][x['person_name'
       ]][otheroffice]
28     except:
29         pass
30
31
32 def probabilistic_imputation(mapp,x):
33     '''data: subset of the data without coordinate
34     This adjust for slight differences in string.
35     '''
36
37     try:
38         match_name = difflib.get_close_matches(x['person_name'],
       list(mapp[x['docdb_family_id']].keys()),n=1,cutoff=0.7)  ###
       this line can also be adjusted for different ordering of names
39         if len(match_name) == 1:  ###if you are adjusting for
       order of names this should change to match_name[0]==x['
```

```
    person_name']
40          for i in ['US','EP','WO']:
41              if i in mapp[x['docdb_family_id']][x['person_name']]:
42                  return mapp[x['docdb_family_id']][x['person_name'
    ]][i]
43                  continue
44              else:
45                  otheroffice = random.sample(list(mapp[x['
    docdb_family_id']][x['person_name']]))
46                  return mapp[x['docdb_family_id']][x['person_name'
    ]][otheroffice]
47      except:
48          pass
```

Listing A.1: Python script for coordinate and address imputation

	Country	Domestic linkages (2010-2014)	Ranking (2010-2014)	Domestic linkages (2005-2009)	Ranking (2005-2009)	Domestic linkages (2000-2004)	Ranking (2000-2004)	Domestic linkages (% change)	Change in ranking
0	South Korea	7487.20	1	4469.64	1	2722.38	1	175.02	0
1	Taiwan	3925.80	2	2466.90	2	1147.70	3	242.05	1
2	Japan	1258.51	3	1415.26	3	1395.15	2	-9.79	-1
3	Israel	1202.20	4	741.00	6	519.00	6	131.63	2
4	United States	884.31	5	854.78	5	786.09	4	12.49	-1
5	China	788.60	6	273.07	8	86.70	21	809.52	15
6	Germany	772.19	7	962.78	4	650.28	5	18.74	-2
7	India	543.85	8	235.37	14	87.54	20	521.24	12
8	Singapore	518.40	9	330.00	7	307.40	7	68.64	-2
9	Canada	286.45	10	272.87	9	216.13	9	32.53	-1
10	Switzerland	273.58	11	264.66	10	194.94	11	40.34	0
11	Sweden	272.51	12	248.53	13	191.16	12	42.55	0
12	Belgium	264.94	13	258.92	11	203.61	10	30.11	-3
13	Finland	233.57	14	258.20	12	247.42	8	-5.60	-6
14	Netherlands	195.07	15	221.71	15	182.85	13	6.68	-2
15	Denmark	182.60	16	167.15	17	137.95	15	32.36	-1
16	Austria	170.40	17	187.90	16	123.80	18	37.64	1
17	Hong Kong	163.40	18	125.20	20	158.80	14	2.89	-4
18	France	160.83	19	163.69	18	127.80	17	25.84	-2
19	Ireland	131.48	20	95.60	21	90.44	19	45.37	-1

Table 27: Ranking of countries based on domestic linkages. The domestic linkage reported is the average weighted domestic linkage in the given period.

	Continent	Domestic linkages (2010-2014)	Ranking (2010-2014)	Domestic linkages (2005-2009)	Ranking (2005-2009)	Domestic linkages (2000-2004)	Ranking (2000-2004)	Domestic linkages (% change)	Change in ranking
0	Asia	1980.03	1	1465.68	1	1145.01	1	72.92	0
1	North America	710.91	2	691.07	2	649.25	2	9.49	0
2	Europe	205.56	3	242.40	3	187.80	3	9.46	0
3	Oceania	84.98	4	83.15	4	78.54	4	8.20	0
4	Africa	27.41	5	19.52	5	16.57	5	65.37	0
5	South America	17.71	6	14.06	6	9.31	6	90.25	0

Table 28: Ranking of continent based on domestic linkages. The domestic linkage reported is the average weighted domestic linkage in the given period.

	Country	Intl. global city linkages (2010-2014)	Ranking (2010-2014)	Intl. global city linkages (2005-2009)	Ranking (2005-2009)	Intl. global city linkages (2000-2004)	Ranking (2000-2004)	Intl. global city linkages (% change)	Change in ranking
0	Taiwan	1727.00	1	1024.20	1	381.80	2	352.33	1
1	Israel	1404.00	2	779.00	2	518.80	1	170.62	-1
2	Singapore	611.80	3	361.80	3	301.80	3	102.71	0
3	India	393.37	4	182.51	5	74.17	12	430.35	8
4	China	368.12	5	215.42	4	77.56	10	374.58	5
5	Hong Kong	200.80	6	141.20	7	183.60	4	9.36	-2
6	Switzerland	156.84	7	141.30	6	99.42	5	57.75	-2
7	Canada	138.99	8	125.80	8	96.40	6	44.18	-2
8	United States	136.21	9	103.07	12	90.05	7	51.26	-2
9	Belgium	119.87	10	109.27	9	83.78	9	43.07	-1
10	Malaysia	111.30	11	71.10	18	41.40	23	168.84	12
11	New Zealand	108.60	12	54.20	22	41.40	22	162.31	10
12	Denmark	107.65	13	97.20	13	74.35	11	44.78	-2
13	Thailand	100.80	14	44.40	26	38.60	24	161.13	10
14	Austria	99.43	15	106.16	10	68.23	14	45.72	-1
15	Sweden	96.01	16	73.56	16	54.33	18	76.71	2
16	South Korea	87.49	17	74.82	14	36.61	26	138.93	9
17	Ireland	84.20	18	65.64	19	69.32	13	21.46	-5
18	Finland	79.05	19	72.14	17	53.65	19	47.33	0
19	Japan	74.16	20	54.33	21	86.39	8	-14.15	-12

Table 29: Ranking of countries based on international global city linkages. The international global city linkage reported is the average international global city linkage in the given period.

	Continent	Intl. global city linkages (2010-2014)	Ranking (2010-2014)	Intl. global city linkages (2005-2009)	Ranking (2005-2009)	Intl. global city linkages (2000-2004)	Ranking (2000-2004)	Intl. global city linkages (% change)	Change in ranking
0	Asia	168.92	1	105.32	1	76.98	2	119.42	1
1	North America	119.07	2	92.72	2	81.70	1	45.73	-1
2	Oceania	57.88	3	50.15	3	44.85	3	29.05	0
3	Europe	42.26	4	40.96	4	35.65	4	18.54	0
4	Africa	24.67	5	18.23	5	10.93	5	125.57	0
5	South America	13.25	6	11.85	6	7.25	6	82.69	0

Table 30: Ranking of continents based on international global city linkages. The international global city linkage reported is the average international global city linkage in the given period.

	Countries	Intl. non-global city linkages (2010-2014)	Ranking (2010-2014)	Intl. non-global city linkages (2005-2009)	Ranking (2005-2009)	Intl. non-global city linkages (2000-2004)	Ranking (2000-2004)	Intl. non-global city linkages (% change)	Change in ranking
0	Israel	509.60	1	282.20	1	199.40	1	155.56	0
1	Singapore	258.20	2	171.40	2	142.40	2	81.32	0
2	Taiwan	242.30	3	165.70	3	109.50	3	121.27	0
3	India	154.45	4	68.40	7	30.08	21	413.39	17
4	Luxembourg	113.60	5	68.20	8	56.60	10	100.70	5
5	Switzerland	111.58	6	133.58	4	93.82	4	18.92	-2
6	Belgium	108.20	7	110.29	6	83.25	5	29.96	-2
7	China	105.58	8	56.62	12	31.88	20	231.14	12
8	Austria	90.40	9	115.83	5	57.46	9	57.30	0
9	Canada	77.25	10	62.64	10	82.65	6	-6.53	-4
10	Denmark	76.00	11	67.05	9	55.55	11	36.81	0
11	United States	71.92	12	61.77	11	69.08	7	4.11	-5
12	Japan	58.10	13	38.88	18	68.91	8	-15.69	-5
13	Ireland	51.28	14	34.40	22	33.64	19	52.43	5
14	Malaysia	50.30	15	22.80	28	13.50	31	272.59	16
15	Germany	50.20	16	56.50	13	38.92	15	29.00	-1
16	New Zealand	47.50	17	33.40	23	29.10	22	63.23	5
17	Sweden	46.96	18	44.21	15	36.36	16	29.14	-2
18	Netherlands	42.71	19	41.45	16	33.91	18	25.96	-1
19	Russia	41.40	20	50.80	14	39.20	13	5.61	-7

Table 31: Ranking of countries based on international non-global city linkages. The international non-global city linkage reported is the average international non-global city linkage in the given period.

	Continents	Intl. non-global city linkages (2010-2014)	Ranking (2010-2014)	Intl. non-global city linkages (2005-2009)	Ranking (2005-2009)	Intl. non-global city linkages (2000-2004)	Ranking (2000-2004)	Intl. non-global city linkages (% change)	Change in ranking
0	Asia	67.36	1	42.90	2	49.88	2	35.04	1
1	North America	63.31	2	54.42	1	63.55	1	-0.36	-1
2	Oceania	29.92	3	29.97	4	27.16	3	10.16	0
3	Europe	28.55	4	30.47	3	26.31	4	8.53	0
4	Africa	11.76	5	9.44	5	7.51	5	56.54	0
5	South America	7.59	6	6.93	6	5.06	6	49.87	0

Table 32: Ranking of continents based on international non-global city linkages. The international non-global city linkage reported is the average international non-global city linkage in the given period.

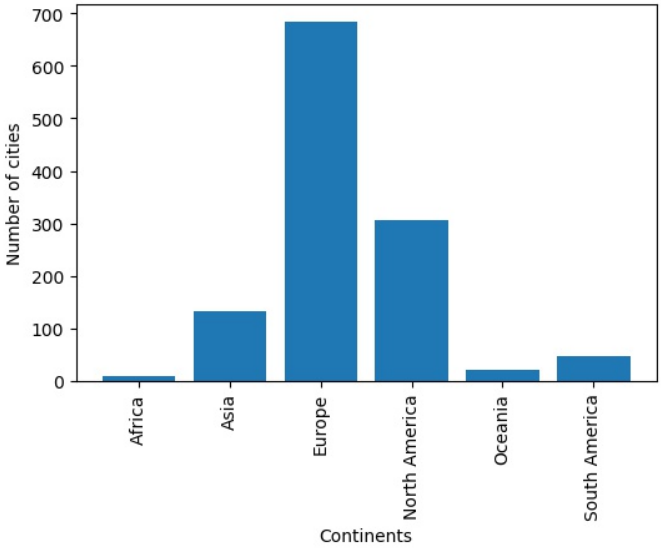


Figure 31: The number of cities in Africa, Asia, Europe, North America, Oceania, and South America is 8, 133, 684, 307, 20, 48 respectively.

Appendix B

Supplementary material for Chapter 2

B.0.1 Generalized economic complexity index of cities in specific technologies

The generalized economic complexity index for cities reported in Figure 14 is based on the patenting activities across all technological areas. However, cities can have different levels of complexity in different technological field. To determine the GENEPY index of cities in different technological fields we apply the GENEPY algorithm to the competitiveness matrix at the level of 4-digit IPCs to obtain $GENEPY_i$ - the GENEPY index for cities, and $GENEPY_j$ - the GENEPY index for each of the 4-digit IPCs. Let k be the 4-digit IPCs belonging to a technology field K - for example 23 of the 637 IPCs belong to the technological field *Chemical engineering*, then we have that the generalized economic complexity index of cities in a technological field K will be given as:

$$GENEPY_{iK} = \sum_{k \in K} GENEPY_k M_{ik} \tag{B.1}$$

Computing the complexity of cities in different technological field this way as opposed to applying the algorithm to subset of the competitiveness matrix corresponding to a technological field allows us to make use of

rich information in the dataset. For example, following the approach of applying the algorithm to a sub-matrix, trying to estimate the complexity of cities in *Pharmaceuticals* will result in applying the algorithm to a sub-matrix of size 150×2, since there are only two 4-digit IPCs (i.e., *A61K* and *A61P*) in Pharmaceuticals. Such submatrix do have limited information and the solution of the eigenvalue problem would be less meaningful.

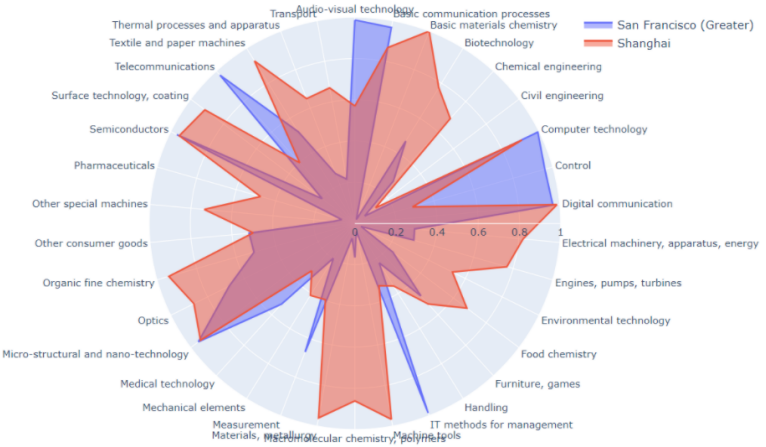


Figure 32: Cities' complexity in different technologies: San Francisco and Shanghai differ in their technology field capabilities as seen in their normalized technological complexity rank in 2014. This measure ranges from 0 (i.e., the lowest performer in the technology field) to 1 (i.e., global leader in the technology field). This is computed by summing the complexities of IPCs (within the technology field) that the city patent in competitively.

Figure 32 shows a spiderplot comparing the complexity of San Francisco and Shanghai in the 34 technological fields. The complexity of cities in each technological field are normalized such that the values range between 0 and 1, where 1 captures what city is the global leader in that technology field and 0 captures what city is the least performer in the technological field. Generally, both Shanghai are global leaders in some technological fields e.g., Shanghai is a global leader in Basic materials chemistry and Digital communication, while San Francisco is a global leader in Audio Visual technology, Computer technology, and IT methods

for management. We also find that Shanghai is ahead of San Francisco in more technological fields (24) than San Francisco is ahead of Shanghai (8), these technological fields include Basic materials chemistry, Biotechnology, Chemical engineering, Electrical machinery, Environmental technology, and Pharmaceuticals. We provide the top 5 global leaders in each technological field in Table 33.

	Audio-visual technology	Basic communication processes	Basic materials chemistry	Biotechnology	Chemical engineering
1	Taipei	San Diego	Shanghai	Lyon	Seoul
2	Malmo	Phoenix	Philadelphia (Greater)	TelAviv	Copenhagen
3	San Francisco (Greater)	Munich	New York (Greater)	Philadelphia (Greater)	Mannheim-Ludwigshafen
4	Taichung	Bangalore	Houston	Paris	Milan
5	Nuremberg	Taipei	Mannheim-Ludwigshafen	Sacramento	Dusseldorf
	Civil engineering	**Computer technology**	**Control**	**Digital communication**	**Electrical machinery, apparatus, energy**
1	Seoul	Beijing	Seattle	Hong Kong	Milwaukee
2	Amsterdam	San Francisco (Greater)	Seoul	London	Guangzhou
3	Los Angeles (Greater)	Bangalore	Madrid	Shanghai	Seoul
4	Milan	San Diego	London	Seattle	Stuttgart
5	Milwaukee	Stuttgart	Munich	San Juan	Nuremberg
	Engines, pumps, turbines	**Environmental technology**	**Food chemistry**	**Furniture, games**	**Handling**
1	Nuremberg	Seoul	Seoul	Los Angeles (Greater)	Amsterdam
2	Paris	Helsinki	Amsterdam	New York (Greater)	Hamburg
3	Toronto	Barcelona	Copenhagen	London	Taichung
4	Munich	Oslo	Minneapolis	Miami (Greater)	Milwaukee
5	Dusseldorf	Hamburg	Philadelphia (Greater)	Atlanta	Salt Lake
	IT methods for management	**Machine tools**	**Macromolecular chemistry, polymers**	**Materials, metallurgy**	**Measurement**
1	Phoenix	Seoul	Philadelphia (Greater)	Brussels	Vancouver
2	Chennai	Zurich	Lyon	Frankfurt am Main	Seoul
3	San Francisco (Greater)	Stuttgart	Houston	Shenyang	Zurich
4	Greater Melbourne	Dresden	Atlanta	Tokyo	Toulouse
5	San Antonio	Taichung	Basel	Dusseldorf	Helsinki
	Mechanical elements	**Medical technology**	**Micro-structural and nano-technology**	**Optics**	**Organic fine chemistry**
1	Stuttgart	Salt Lake	Phoenix	Tokyo	Philadelphia (Greater)
2	Toronto	Pittsburgh	Dresden	Minneapolis	Basel
3	Karlsruhe	Los Angeles (Greater)	Grenoble	Munich	Budapest
4	Chicago	Minneapolis	Singapore	Seoul	Washington (Greater)
5	Detroit (Greater)	Greater Sydney	Montreal	Taipei	Boston
	Other consumer goods	**Other special machines**	**Pharmaceuticals**	**Semiconductors**	**Surface technology, coating**
1	Seoul	Chicago	Zurich	Seoul	Dusseldorf
2	New York (Greater)	Seoul	Kuala Lumpur	Tokyo	Tokyo
3	London	Washington (Greater)	Geneva	Phoenix	Frankfurt am Main
4	Los Angeles (Greater)	Atlanta	Greater Adelaide	Dresden	Nuremberg
5	Atlanta	Minneapolis	Greater Brisbane	Fukuoka	Santiago
	Telecommunications	**Textile and paper machines**	**Thermal processes and apparatus**	**Transport**	
1	Ottawa	Seoul	Stuttgart	Paris	
2	San Diego	Munich	Seoul	Seoul	
3	Edinburgh	Minneapolis	Dusseldorf	Munich	
4	Washington (Greater)	Chicago	Karlsruhe	Berlin	
5	San Francisco (Greater)	Istanbul	Vienna	Toronto	

Table 33: Top global leaders in each technological field

B.0.2 Relationship between GENEPY index of cities and cities' F1-score of Random Forest model used in forecasting the activation of new technologies

The F1 score captures how well the Random Forest model forecast competitiveness in completely new products specifically for IPCs where cities have constantly been less competitive between 2000-2008 - i.e., the cities constantly had an RCA less than 0.25 in the IPC. We find that unlike the forecast done on the full matrix, we have a lower F1 score, and there is no systematic relationship between how well we can make a forecast and the current economic complexity of the city.

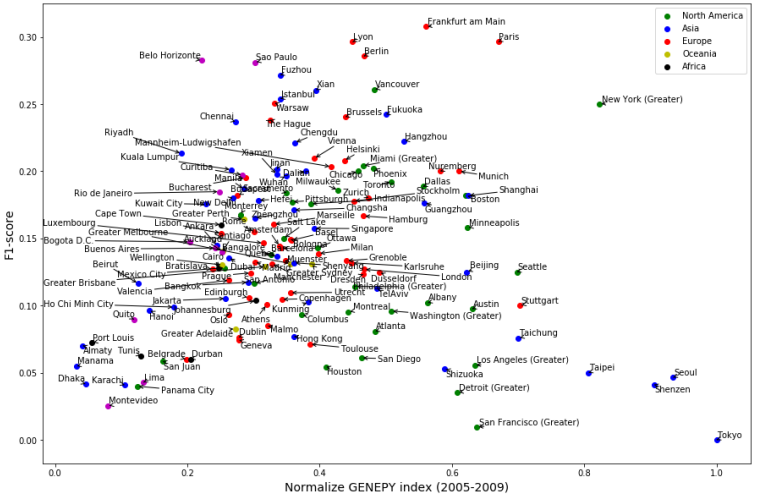


Figure 33: Generalize economic complexity vs Prediction performance (activation of new technologies) On the x-axis is the Generalized economic complexity index of cities computed between period 2005-2009, and on the y-axis is the F1-score that captures the performance of the Random Forest model in forecasting the competitiveness (activation matrix) in 2014 for each city. The correlation between the GENEPY index and the F1-score of cities is 0.06.

B.0.3 Python code for Generalized economic complexity

```
def genepy_complexity(M):

    kc = M.sum(axis=1)
    RW = M / kc[:,None]
    RW[np.isnan(RW)] = 0
    kp_1 = RW.sum(axis=0)
    den = np.matmul(kc[:,None],kp_1[None,:])

    W=M/den
    W[np.isnan(W)]=0

    genepy = defaultdict(dict)

    for i in ['country','product']:
        if i=='country':
            P = np.matmul(W,W.transpose())
        else:
```

```
            P = np.matmul(W.transpose(),W)

        np.fill_diagonal(P,0)
        eigvals, eigvecs = np.linalg.eig(P)
        idx = eigvals.argsort()[::-1]
        eigvals = eigvals[idx]
        eigvecs = eigvecs[:,idx]

        a = ((eigvals[:2]*eigvecs[:,:2]**2).sum(axis=1))**2
        b = 2*(((eigvals[:2]**2)*(eigvecs[:,:2]**2)).sum(axis=1))

        genepy[i] = a+b
    return genepy['country'], genepy['product']
```

Listing B.1: GENEPY for computing cities' technological capabilities

B.0.4 Python codes for machine learning task

```
def chunks(lst, n):
    """Yield successive n-sized chunks from lst."""
    for i in range(0, len(lst), n):
        yield lst[i:i + n]


def ML_models(X,y,X_test_sc,test,modeltype):
    if modeltype == 'RF':
        model = RandomForestClassifier()
        model.fit(X, y)
        test['predict'] = model.predict(X_test_sc)
        test['predict_prob'] = [i[int(j)] for i,j in zip(model.
    predict_proba(X_test_sc).tolist(),test['predict'].tolist())]

        return test

    elif modeltype == 'XGB':
        model = xgb.XGBClassifier( eval_metric='logloss')
        model.fit(X,y)

        test['predict'] = model.predict(X_test_sc)
        test['predict_prob'] = [i[int(j)] for i,j in zip(model.
    predict_proba(X_test_sc).tolist(),test['predict'].tolist())]

        return test

    elif modeltype == 'SVM':
        #model = make_pipeline(StandardScaler(), SVC(gamma='auto')
    )
        model = SVC()
```

```
        model.fit(X,y)

        test['predict'] = model.predict(X_test_sc)

        p = np.array(model.decision_function(X_test_sc))
        prob = np.exp(p)/np.sum(np.exp(p),axis=1, keepdims=True)

        test['predict_prob'] = prob

        return test

    elif modeltype == 'NN':
        n_inputs = X.shape[1]
        n_epochs =10
        n_batch = 5
        model = NN(n_inputs, n_units=32, dropout=0.1, l1_reg
    =0.001, activation='relu', L=3)
        model.fit(X,y,epochs=n_epochs, batch_size=n_batch, verbose
    =1)

        test['predict'] = model.predict_classes(X_test_sc)
        test['predict_prob'] = model.predict_proba(X_test_sc)

        return test


def run_models(features,target,modeltype):
    ''
    This script run each model specified
    Arg:
    features - csv file of RCAs.
    target - csv file of future competitiveness (binary).
    modeltype - a machine learning algorithm e.g. 'RF', 'XGB', '
    SVM', 'NN'
    ''
    targetdict = target.set_index(['city','year','ipc'])['mcp']
    citieschunk = list(chunks(list(features.city.unique()), 15))
    result = []
    for prod in tqdm(target.ipc.unique()):
        testpredict = []
        for chunk in citieschunk:
            test = features[features['city'].isin(chunk)]
            test['target'] = test[['city','year']].apply(lambda x:
     targetvar(x,prod,targetdict),axis=1)
            train = features[features['city'].isin(list(set(list(
    features.city.unique()))-set(chunk)))]
            train['target'] = train[['city','year']].apply(lambda
    x: targetvar(x,prod,targetdict),axis=1)

            X = train[list(train.columns)[1:-2]]
```

```
            y = train['target']
            Xtest = test[list(test.columns)[1:-2]]

            scaler = StandardScaler().fit(X)
            X_train_sc = scaler.transform(X)
            X_test_sc = scaler.transform(Xtest)

            res = ML_models(X_train_sc,y,X_test_sc,test,modeltype)
            res['ipc'] = prod
            #testpredict.append(res[['city','ipc','year','target
    ','predict']])
            testpredict.append(res[['city','ipc','year','target','
    predict','predict_prob']])

        result.append(pd.concat(testpredict,axis=0))

    return pd.concat(result,axis=0)
```

Listing B.2: Machine learning algorithm for forecasting competitiveness

Appendix C

Supplementary material for Chapter 3

Table 34 is a reference table showing the mapping between 34 technology fields and 4-digit IPC classification. These technology fields allow for the easy grouping of applications based on technology. The same technology fields are used by European Patent Office (EPO) and World Intellectual Patent Office (WIPO) for their statistics.

	Technology fields	**4-digit International Patent Classification system (IPCs)**
0	Food chemistry	C12H, A23F, A23J, A23D, C13F, C12F, C12G, A01H, A21D, A23B, A23C, A23G, A23K, A23L, C12C, C12G, C12J, C13B, C13D, C13K.
1	Other special machines	A01J, F41F, A01G, B99Z, B28D, A01C, C12L, A21B, B29K, F41C, F42B, A23N, A23P, B29C, C08J, A01B, B29D, C03B, B29B, A01D, F41H, A22B, A01F, F41J, B28C, A22C, A21C, B29L, B28B, F41A, F41B, B33Y, A01M, F41G, F42D, A01L, A01K, B02B, F42C.
2	Other consumer goods	A41D, B42C, A42B, A24D, G10G, A44C, D07B, B68C, B68B, A45D, A44B, G10H, A45F, F25D, A45C, B68F, A43B, D06F, B42B, A24B, G10D, D06N, G10C, B42D, A41B, D04D, B43M, G10F, B44D, B43L, A42C, A24F, A46B, B43K, B42F, B44C, A24C, B44F, A45B, A99Z, G10K, B44B, A41C, A43C, B68G, A62B, A41G, A41F, G10B.
3	Medical technology	A61N, A61D, H05G, A61J, A61B, G16H, A61M, A61H, A61C, A61L, A61G, A61F.
4	Furniture, games	A63J, A63D, A47K, A47B, A47C, A63C, A47G, A47F, A47D, A63G, A63B, A47L, A63H, A63F, A47H, A47J, A63K.

	Technology fields	4-digit International Patent Classification system (IPCs)
5	Machine tools	B21B, B23C, B21H, B21L, B27N, B24B, B23B, A62D, B21C, B21D, B21F, B21G, B21J, B21K, B23D, B23F, B23G, B23H, B23K, B23P, B23Q, B24C, B24D, B25B, B25C, B25D, B25F, B25G, B25H, B26B, B26D, B26F, B27B, B27C, B27D, B27F, B27G, B27H, B27J, B27K.
6	Transport	B62L, B60F, B60H, B61J, B60K, B61C, B62H, B60B, B60C, B60D, B60G, B60J, B60L, B60M, B60N, B60P, B60Q, B60R, B60S, B60T, B60V,B60W, B61B, B61D, B61F, B61G, B61H, B61K, B61L, B62B, B62C, B62D, B62K, B62M, B63B, B63C, B63G, B63H, B63J, B64B, B64C, B64D, B64F, B64G.
7	Micro-structural and nano-technology	B82Y, B81B, B81C, B82B.
8	Textile and paper machines	B41M, D04C, D21F, B31C, D21D, D21H, D04B, A41H, A43D, A46D, B31B, B31D, B31F, B41B, B41C, B41D,B41F, B41G, B41J, B41K, B41L, B41N, C14B, D01B, D01C, D01D, D01F, D01G, D01H, D02G, D02H, D02J, D03C, D03D, D03J, D04G, D04H, D05B, D05C, D06G, D06H, D06J, D06M, D06P, D06Q, D21B, D21C, D21G, D21J, D99Z.
9	Civil engineering	E01F, E05G, E06C, E04B, E01C, E21B, E02C, E01B, E01D, E01H, E02B, E02D, E02F, E03B, E03C, E03D, E03F, E04C, E04D, E04F, E04G, E04H, E05B, E05C, E05D, E05F, E06B, E21C, E21D, E21F, E99Z.
10	Thermal processes and apparatus	F24D, F24B, F22G, F22D, F28D, F24V, F24F, F22B, F23B, F23C, F23D, F23H, F23K, F23L, F23M, F23N, F23Q, F24C, F24H, F24J, F24S, F24T, F25B, F25C, F27B, F27D, F28B, F28C, F28F, F28G.
11	Measurement	G01L, G01M, G01P, G01N, G01G, G04F, G01C, G01B, G01D, G01F, G01H, G01J, G01K, G01Q, G01R, G01S, G01V, G01W, G04B, G04C, G04D, G04G, G04R, G12B, G99Z.
12	Electrical machinery, apparatus, energy	H01H, H05B, F21S, H01M, H99Z, F21K, H01F, F21H, F21L, F21V, F21W, F21Y, H01B, H01C, H01G, H01J, H01K, H01R, H01T, H02B, H02G, H02H, H02J, H02K, H02M, H02N, H02P, H02S, H05C, H05F.
13	Chemical engineering	H05H, B08B, D06B, B03C, B01J, B06B, B03D, B01B, B01D, B01F, B01L, B02C, B03B, B04B, B04C, B05B, B07B, B07C, C14C, D06C, D06L, F25J, F26B.
14	Macromolecular chemistry, polymers	C08G, C08L, C08B, C08F, C08K, C08C, C08H.
15	Basic materials chemistry	C05C, C06D, C05F, C11D, C06B, C10H, A01N, A01P, C05B, C05D, C05G, C06C, C06F, C09B, C09C, C09D, C09F, C09G, C09H, C09J, C09K, C10B, C10C, C10F, C10G, C10J, C10K, C10L, C10M, C10N, C11B, C11C, C99Z.
16	Control	G09B, G05F, G07F, G07G, G09D, G08B, G08G, G05B, G05D, G07B, G07C, G07D, G09C.
17	Audio-visual technology	H05K, H04R, G09G, G11B, H04S, G09F.
18	Engines, pumps, turbines	F02D, F04F, G21H, G21F, F02F, F02P, F03H, F01B, F01C, F01D, F01K, F01L,
		F01M, F01P, F02B, F02C, F02G, F02K, F02M, F02N, F03B, F03C, F03D, F03G, F04B, F04C, F04D, F23R, F99Z, G21B, G21C, G21D, G21G, G21J, G21K.
19	Mechanical elements	F16L, F15D, F16S, F17B, F16P, G05G, F16J, F15B, F15C, F16B, F16C, F16D, F16F, F16G, F16H, F16K, F16M, F16N, F16T, F17C, F17D.
20	Optics	G03G, G03D, H01S, G03B, G02F, G02B, G03C, G02C, G03F, G03H.
21	Handling	B67D, B65C, B66F, B65G, B65B, B25J, B65D, B65H, B66B, B66C, B66D, B67B, B67C.
22	Organic fine chemistry	C07F, C40B, C07B, C07H, A61Q, C07D, C07J, C07C.
23	Biotechnology	C12Q, C12P, C07K, C12R, C12S, C07G, C12N, C12M.
24	Environmental technology	B65F, F23G, B09B, C02F, F23J, B09C, A62C, F01N, G01T.
25	Computer technology	G06N, G06D, G06E, G06C, G06F, G06G, G06J, G06K, G06M, G06T, G10L, G11C.
26	Surface technology, coating	C25C, C25B, C30B, C23C, C23G, B05C, C23F, B05D, B32B, C23D, C25D, C25F.
27	Materials, metallurgy	C22C, C22B, B22F, B22D, B22C, C21C, C01F, C01B, C01C, C01D, C01G, C03C, C04B, C21B, C21D, C22F.
28	Basic communication processes	H03D, H03M, H03K, H03J, H03C, H03F, H03B, H03G, H03H, H03L.
29	Telecommunications	H04M, H04N, H04Q, H04K, G08C, H04H, H01Q, H01P, H04B, H04J.
30	Digital communication	H04W, H04L.
31	Semiconductors	H01L.
32	Pharmaceuticals	A61P, A61K.
33	IT methods for management	G06Q.

Table 34: Categorization of 4 digit IPCs in each technology field.

C.0.1 Descriptive for technology entry in period 2005-2014

	Mean	Std.	Min	25%	50%	75%	Max.
Inter-city ties	0.01	0.11	0.00	0.00	0.00	0.00	9.00
Technological relatedness	0.01	0.01	0.00	0.00	0.01	0.01	0.76
Distance from core IPC	1.50	0.50	0.00	1.00	1.50	2.00	3.00
Knowledge diversity	19.99	12.05	0.00	11.28	18.18	26.36	205.92
Existing IPCs	260.12	142.21	0.00	148.00	255.00	365.00	636.00
IPC growth rate	0.01	0.22	-0.83	-0.09	-0.01	0.08	10.00
City complexity	0.08	0.72	-8.84	-0.35	0.09	0.51	8.72
City's inventor density	0.28	0.64	0.00	0.03	0.10	0.28	18.59
City's share of foreign patent	0.35	0.21	0.00	0.19	0.31	0.48	1.00
City size	489.45	1777.86	0.00	52.00	157.00	427.00	117279.00
Partner size	3198.99	2386.76	0.00	1986.91	2685.13	3756.03	73823.94
Data quality	0.86	0.15	0.07	0.84	0.90	0.95	1.00

Table 35: Summary of variables for observations where entry was made between 2005-2014.

C.0.2 Descriptive for cases of no technology entry in period 2005-2014

	Mean	Std.	Min	25%	50%	75%	Max
Inter-city ties	0.00	0.07	0.00	0.00	0.00	0.00	12.33
Technological relatedness	0.01	0.01	0.00	0.00	0.01	0.01	0.39
Distance from core IPC	1.74	0.58	0.00	1.50	2.00	2.00	4.00
Knowledge diversity	16.29	13.64	0.00	5.98	13.17	23.23	89.84
Existing IPCs	167.69	151.66	0.00	28.00	129.00	278.00	636.00
IPC growth rate	-0.04	0.68	-0.95	-0.23	-0.11	0.00	10.00
City complexity	-0.06	1.05	-5.75	-0.49	0.09	0.57	4.92
City's inventor density	0.14	0.46	0.00	0.00	0.03	0.11	18.59
City's share of foreign patent	0.39	0.28	0.00	0.17	0.35	0.56	1.00
City size	175.23	960.06	0.00	5.00	28.00	114.00	91135.00
Partner size	3692.04	5445.03	0.00	1627.40	2684.69	4227.68	69414.33
Data quality	0.80	0.20	0.14	0.62	0.87	0.97	1.00

Table 36: Summary of variables for observations where there was no entry between 2005-2014.

C.0.3 Result of stratified semi-parametric cox PH model changing the threshold for new technology

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.008*	1.008*	1.008*	1.008*
		(0.00360)	(0.00422)	(0.00362)	(0.00423)
Technological relatedness		1.077**	1.077**	1.078*	1.078*
		(0.0281)	(0.0281)	(0.0317)	(0.0317)
Inter-city ties × Partner city size			1.014**		1.014**
			(0.0119)		(0.0120)
Technological relatedness × Focal city size				0.993	0.993
				(0.00557)	(0.00557)
Distance from core IPC	1.161***	1.157***	1.157***	1.158***	1.158***
	(0.00517)	(0.00530)	(0.00530)	(0.00530)	(0.00530)
Knowledge diversity	1.023**	1.023**	1.023**	1.023**	1.023**
	(0.0129)	(0.0129)	(0.0129)	(0.0129)	(0.0129)
Existing IPCs	0.951**	0.937***	0.937***	0.938***	0.938***
	(0.0147)	(0.0153)	(0.0153)	(0.0155)	(0.0155)
IPC growth rate	1.126***	1.123***	1.123***	1.123***	1.123***
	(0.0104)	(0.0105)	(0.0105)	(0.0103)	(0.0103)
City complexity	1.000	1.000	1.000	1.000	1.000
	(0.00647)	(0.00649)	(0.00648)	(0.00651)	(0.00651)
Inventor density	0.961	0.961	0.961	0.961	0.961
	(0.0234)	(0.0233)	(0.0233)	(0.0233)	(0.0233)
Share of foreign patent	0.977	0.976	0.976	0.976	0.976
	(0.0123)	(0.0125)	(0.0125)	(0.0126)	(0.0126)
Focal city size	1.069*	1.066*	1.066*	1.079*	1.079*
	(0.0475)	(0.0455)	(0.0455)	(0.0427)	(0.0427)
Partner city size	0.949***	0.951***	0.951***	0.951***	0.951***
	(0.0142)	(0.0143)	(0.0143)	(0.0143)	(0.0143)
Data quality	1.045	1.046	1.046	1.045	1.045
	(0.0193)	(0.0193)	(0.0193)	(0.0193)	(0.0193)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	707 323	707 323	707 323	707 323	707 323
Log pseudolikelihood	-391 691.75	-391 513.1	-391 513.1	-391 509.06	-391 509.06
Wald chi2	51 873.62***	52 796.06***	52 924.43***	53 123.44***	53 230.96***

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 37: Supplementary analysis. Results of stratified semi-parametric Cox PH when the threshold for an IPC to be considered new to the city is changed from 5 to 3 years.

C.0.4 Result of stratified semi-parametric cox PH model changing the measurement of inter-city ties

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.011*	1.019***	1.011*	1.019***
		(0.00464)	(0.00393)	(0.00467)	(0.00389)
Technological relatedness		1.074**	1.074**	1.074**	1.074**
		(0.0259)	(0.0258)	(0.0268)	(0.0269)
Inter-city ties × Partner city size			1.044***		1.045***
			(0.00921)		(0.00921)
Technological relatedness × Focal city size				0.999	0.998
				(0.00344)	(0.00349)
Distance from core IPC	1.139***	1.137***	1.136***	1.137***	1.137***
	(0.00572)	(0.00567)	(0.00568)	(0.00567)	(0.00567)
Knowledge diversity	1.064***	1.070***	1.070***	1.070***	1.070***
	(0.0184)	(0.0186)	(0.0187)	(0.0186)	(0.0187)
Existing IPCs	1.788***	1.740***	1.716***	1.739***	1.713***
	(0.103)	(0.101)	(0.100)	(0.102)	(0.100)
IPC growth rate	1.137***	1.130***	1.130***	1.130***	1.130***
	(0.00497)	(0.00597)	(0.00595)	(0.00590)	(0.00589)
City complexity	0.993	0.993	0.993	0.993	0.993
	(0.00729)	(0.00731)	(0.00724)	(0.00732)	(0.00726)
Inventor density	1.019	1.016	1.013	1.016	1.012
	(0.0178)	(0.0168)	(0.0160)	(0.0167)	(0.0158)
Share of foreign patent	0.983	0.982	0.982	0.982	0.982
	(0.0135)	(0.0137)	(0.0137)	(0.0137)	(0.0137)
Focal city size	1.036	1.030	1.035	1.032	1.038
	(0.0245)	(0.0252)	(0.0235)	(0.0239)	(0.0225)
Partner city size	0.945***	0.947***	0.953**	0.947***	0.953**
	(0.0148)	(0.0149)	(0.0150)	(0.0149)	(0.0150)
Data quality	1.000	1.002	1.006	1.002	1.006
	(0.0184)	(0.0185)	(0.0185)	(0.0185)	(0.0185)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	521 339	521 339	521 339	521 339	521 339
Log pseudolikelihood	-396 209.92	-396 027.31	-396 010.43	-396 027.1	-396 009.9
Wald chi2	60 277.77***	61 095.30***	61 050.34***	61 102.59***	61 062.64***

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 38: Supplementary analysis. Results of stratified semi-parametric Cox PH model when the measurement of inter-city ties is changed.

C.0.5 Hazard ratios for technological fields based on stratified semi-parametric Cox PH model

	Model 1	Model 2	Model 3	Model 4	Model 5
Basic communication processes	0.801***	0.804***	0.804***	0.804***	0.804***
	(0.0317)	(0.0318)	(0.0318)	(0.0318)	(0.0318)
Basic materials chemistry	1.739***	1.678***	1.678***	1.679***	1.679***
	(0.0541)	(0.0560)	(0.0560)	(0.0559)	(0.0559)
Biotechnology	0.302***	0.264***	0.264***	0.264***	0.264***
	(0.0144)	(0.0172)	(0.0172)	(0.0177)	(0.0177)
Chemical engineering	2.277***	2.282***	2.282***	2.281***	2.282***
	(0.0711)	(0.0715)	(0.0715)	(0.0715)	(0.0715)
Civil engineering	2.247***	2.266***	2.266***	2.266***	2.266***
	(0.0710)	(0.0720)	(0.0720)	(0.0720)	(0.0720)
Computer technology	1.155***	1.149***	1.149***	1.148***	1.148***
	(0.0466)	(0.0469)	(0.0469)	(0.0469)	(0.0469)
Control	0.914*	0.916*	0.916*	0.916*	0.916*
	(0.0346)	(0.0348)	(0.0348)	(0.0348)	(0.0348)
Digital communication	0.488***	0.466***	0.466***	0.466***	0.466***
	(0.0205)	(0.0213)	(0.0213)	(0.0214)	(0.0214)
Electrical machinery, apparatus, energy	1.545***	1.559***	1.559***	1.559***	1.559***
	(0.0452)	(0.0460)	(0.0460)	(0.0460)	(0.0460)
Engines, pumps, turbines	1.373***	1.379***	1.379***	1.379***	1.379***
	(0.0438)	(0.0441)	(0.0441)	(0.0441)	(0.0441)
Environmental technology	2.013***	1.992***	1.991***	1.992***	1.991***
	(0.0692)	(0.0692)	(0.0692)	(0.0692)	(0.0692)
Food chemistry	2.015***	1.918***	1.918***	1.918***	1.918***
	(0.0683)	(0.0731)	(0.0732)	(0.0735)	(0.0735)
Furniture, games	1.251***	1.282***	1.282***	1.282***	1.282***
	(0.0412)	(0.0436)	(0.0436)	(0.0437)	(0.0437)
Handling	1.359***	1.382***	1.382***	1.382***	1.382***
	(0.0565)	(0.0581)	(0.0581)	(0.0581)	(0.0581)
IT methods for management	0.415***	0.414***	0.414***	0.414***	0.414***
	(0.0191)	(0.0192)	(0.0192)	(0.0192)	(0.0192)
Machine tools	3.181***	3.187***	3.187***	3.187***	3.187***
	(0.0970)	(0.0974)	(0.0974)	(0.0974)	(0.0974)
Molecular chemistry, polymers	1.341***	1.279***	1.279***	1.279***	1.279***
	(0.0444)	(0.0470)	(0.0470)	(0.0475)	(0.0475)
Materials, metallurgy	1.518***	1.507***	1.507***	1.507***	1.507***
	(0.0483)	(0.0481)	(0.0481)	(0.0481)	(0.0481)
Measurement	0.901**	0.906**	0.906**	0.906**	0.906**
	(0.0323)	(0.0328)	(0.0328)	(0.0328)	(0.0328)
Mechanical elements	1.399***	1.419***	1.419***	1.419***	1.419***
	(0.0543)	(0.0559)	(0.0559)	(0.0559)	(0.0559)
Medical technology	1.088*	1.098**	1.098**	1.098**	1.098**
	(0.0379)	(0.0385)	(0.0385)	(0.0385)	(0.0385)
Micro-structural and nano-technology	3.378***	3.377***	3.378***	3.377***	3.377***
	(0.127)	(0.127)	(0.127)	(0.127)	(0.127)
Optics	0.891***	0.888***	0.888***	0.888***	0.888***
	(0.0305)	(0.0306)	(0.0306)	(0.0306)	(0.0306)
Organic fine chemistry	0.476***	0.427***	0.427***	0.427***	0.427***
	(0.0213)	(0.0238)	(0.0238)	(0.0243)	(0.0243)
Other consumer goods	1.651***	1.660***	1.660***	1.660***	1.660***
	(0.0510)	(0.0515)	(0.0515)	(0.0515)	(0.0515)
Other special machines	1.589***	1.594***	1.594***	1.594***	1.594***
	(0.0527)	(0.0530)	(0.0530)	(0.0530)	(0.0530)
Pharmaceuticals	0.0872***	0.0610***	0.0610***	0.0607***	0.0607***
	(0.00678)	(0.00564)	(0.00564)	(0.00558)	(0.00557)
Semiconductors	1.046	1.031	1.031	1.030	1.030
	(0.0458)	(0.0456)	(0.0456)	(0.0456)	(0.0456)
Surface technology, coating	1.571***	1.573***	1.573***	1.573***	1.573***
	(0.0507)	(0.0509)	(0.0509)	(0.0509)	(0.0509)
Telecommunications	0.460***	0.453***	0.453***	0.453***	0.453***
	(0.0160)	(0.0159)	(0.0159)	(0.0160)	(0.0160)
Textile and paper machines	1.528***	1.513***	1.513***	1.513***	1.513***
	(0.0488)	(0.0486)	(0.0486)	(0.0486)	(0.0486)
Thermal process and apparatus	3.189***	3.180***	3.179***	3.180***	3.179***
	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)
Transport	1.385***	1.394***	1.393***	1.393***	1.393***
	(0.0428)	(0.0432)	(0.0431)	(0.0431)	(0.0431)

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 39: Hazard ratios for technological fields based on stratified semi-parametric Cox PH model.

C.0.6 Result for ICT vs Pharmaceuticals

This section includes a supplementary analysis in which we examined the hazard of entering an IPC belonging to an ICT field by cities whose core IPCs are in the Pharmaceutical field, and vice-versa. ICT knowledge areas are collection of IPCs in the fields: Digital communication, Telecommunications, Computer technology, and IT methods for management. The model used is the Semi-Parametric Cox model with shared frailty to control for city-specific effect.

Model 1 examines entry into Pharmaceutical fields. The model includes a bivariate explanatory variable called *ICT* that takes the value 1 if the focal city's core knowledge area is ICT and 0 otherwise. Also it includes an interaction term between ICT and inter-city ties, and an interaction term between ICT and technological relatedness. Model 1 only contains subset of observations (i.e., 604 observations) with potential entry in IPCs belonging to Pharmaceuticals. Model 2 examines entry into ICT fields. The model includes a bivariate explanatory variable called *Pharmaceuticals* that takes the value 1 if the focal city's core knowledge area is in the Pharmaceuticals and 0 otherwise, also, we include an interaction term between Pharmaceutical and inter-city ties, and another interaction term between Pharmaceuticals and technological relatedness. Model 2 contains subset of observations (i.e., 18,630 observations) with potential entries in IPCs belonging to ICT field (i.e., digital communication, telecommunication, computer technology, and IT method of management).

From Model 1 we do find that inter-city ties and technological relatedness increase and decrease the likelihood of entering a new technology in the Pharmaceutical field respectively. We do not find any evidence that having core expertise in ICT increases the likelihood of entering a Pharmaceutical field. However, the effect of technological relatedness on the likelihood of entering a new technology is positively moderated for focal cities whose core expertise is in ICT. Specifically, the average marginal effect of technological relatedness on the likelihood of entering a Pharmaceutical IPC is 1.22.

From Model 2, we find that inter-city ties and technological relatedness both increase the likelihood that a city will enter a new technology in the ICT field. However, we do not find any significant evidence that a focal city with core expertise in Pharmaceuticals will increase the likelihood of exploring an IPC in ICT. However, we find that the effect of inter-city ties and technological relatedness on the likelihood of entry is positively moderated when the city's core IPC is in the Pharmaceuticals. The average marginal effect of inter-city ties and technological relatedness on entry is 2.4 and 3.11 respectively.

	Model 1		Model 2	
Inter-city ties	1.517**	(0.194)	1.223**	(0.0222)
Technological relatedness	0.670***	(0.0792)	1.144***	(0.0176)
ICT	0.833	(0.371)		
Inter-city ties × ICT	1.929	(0.751)		
Technological relatedness × ICT	1.460*	(0.253)		
Pharmaceuticals			0.968	(0.0997)
Inter-city ties × Pharmaceuticals			1.403**	(0.2024)
Technological relatedness × Pharmaceuticals			1.231**	(0.113)
City-specific effect (θ)	3.206***	(0.5562)	1.157***	(0 .0611)
City FE	Yes		Yes	
Year FE	Yes		Yes	
Technology FE	Yes		Yes	
Number of observation	604		18630	
Log likelihood	-1720.114		-33677.998	
Wald chi2	28.66***		227.22***	

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 40: Supplementary analysis: Coefficients in Model 1 are the hazard ratio of ICT core cities exploring Pharmaceutical knowledge areas, while coefficients in Model 2 are the hazard ratio of Pharmaceuticals core cities exploring ICT knowledge areas. Control variables are not included in the model to allow for convergence of the partial likelihood estimation of the coefficients.

C.0.7 Descriptive of variables for post-entry performance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1) Technological performance	1										
(2) Collaborative entry	0.099*	1									
(3) Distance from core IPC	-0.049*	-0.055*	1								
(4) Knowledge diversity	-0.033*	-0.011*	0.098*	1							
(5) Existing IPCs	0.007*	-0.041*	0.254*	0.308*	1						
(6) IPC growth rate	0.009*	0.006*	0.001	0.005	0.069*	1					
(7) City complexity	0.001	-0.005	0.084*	0.153*	0.272*	0.014*	1				
(8) Inventor density	0.022*	0.028*	0.095*	-0.027*	0.301*	0.028*	0.057*	1			
(9) Share of foreign patent	0.018*	-0.005	-0.113*	-0.142*	-0.275*	-0.023*	-0.138*	-0.101*	1		
(10) City size	0.030*	-0.025*	0.086*	-0.001	0.372*	0.059*	0.079*	0.386*	-0.069*	1	
(11) Data quality	0.039*	0.049*	0.105*	0.233*	0.371*	0.019*	0.121*	0.106*	-0.222*	0.076*	1
Mean	2.457	0.4777	1.505	19.987	260.116	0.013	0.079	0.277	0.349	489.454	0.857
Std. Dev.	7.979	0.499	0.501	12.048	142.212	0.224	0.715	0.637	0.211	1777.856	0.145
Min.	0	0	0	0	2	-0.833	-8.838	0	0	5	0.07
Max.	583	1	3	205.920	636	10	8.718	18.594	1	117279	1

Notes: * indicates significance at 5% level.

Table 41: Descriptives and correlations of variables in the post-entry performance analysis.

C.0.8 Results for post-entry performance

	Model 1	Model 2
Collaborative entry		1.727***
		(0.0360)
Distance from core IPC	0.879***	0.882***
	(0.00835)	(0.00839)
Knowledge diversity	0.964	0.931
	(0.0626)	(0.0583)
Existing IPCs	0.997	0.977
	(0.209)	(0.197)
IPC growth rate	1.099***	1.101***
	(0.0139)	(0.0132)
City complexity	1.006	1.003
	(0.0107)	(0.0103)
Inventor density	0.949	0.944
	(0.0552)	(0.0612)
Share of foreign patent	1.087**	1.081**
	(0.0298)	(0.0290)
City size	1.004	1.026
	(0.102)	(0.135)
Data quality	1.022	1.015
	(0.0413)	(0.0400)
Constant	2.268***	2.195***
	(0.0287)	(0.0276)
Over-dispersion parameter (α)	2.268	2.194
	(0.0287)	(0.0275)
City FE	Yes	Yes
Technology FE	Yes	Yes
Year FE	Yes	Yes
Log pseudolikelihood	-139224.16	-127639.7
Number of observations	102 457	102 457

Reported coefficients are incidence rate ratio and standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 42: Post entry technological performance: The role of collaboration on the performance of cities in the new knowledge after exploration.

We examined the post-entry technological performance of the city if it enters the new technology domain collaboratively. Some research have linked R&D collaboration to regional innovation performance (e.g., Lecocq and Van Looy, 2009), in which the signal of technological performance is either the development of new technologies or the enhancement of existing technologies or processes. Some of these studies found evidence that exploratory collaboration that focuses on basic research and development leads to an integrated product development path (Lecocq and Van Looy, 2009). These studies show that economic actors in a city can have different goals when engaging in collaborative effort. For example, an economic actor that is forming ties at the point of entering a new technology domain does this to discover potential of new opportunities and to learn, while when collaboration is done in already existing technology areas, this may be geared towards allowing the economic actor make the most of existing capabilities. These different objectives would imply that different ties would specifically be useful for one objective as opposed to the other. For example, a collaborative effort with a research institution could be more important when the goal of an economic actor is to learn at the point of entering a new technology domain. Generally, collaboration comes with a coordination cost between economic actors, and this cost can significantly increase with an increase in distance (e.g., international collaborations) between economic actors, hence economic actors tend to collaborate when the opportunities and potential performance in the new technology exceeds the coordination cost. However, international collaborations bring added benefit reflected in the higher than expected citations received by international publications and often time are characteristics of "big science" with larger impact.

In analyzing the impact of collaboration at the point of entering a new technology domain we rely on patent indicators. There are different measures of technological performance that can be used some of which include patent count per cities subsequent upon entering the new technology, which captures the amount of technological activity (Lecocq and Van Looy, 2009) or the "citation-weighted number of patent applications in a city over a fixed period of years subsequent upon entry" (Leten, Belderbos,

and Looy, 2016). We adopted the latter in measuring the technological performance of cities upon entering a new technology. We calculated the number of citations a patent received over a fixed four-year window. Then for all the patents a city produced three years after entering the new technology, we take the average citation. This is done in consistence with other studies on technological performance. To measure "collaborative entry", we simply identify if the patent invented by an economic actor in a city at the point of entering the new technology was done jointly with economic actors located in another city. The explanatory variable takes the value 1 if the patent was done jointly and 0 otherwise. We identify that of the 103,049 entries observed, only 47.7% were done through collaborations between inventors in different cities. We also have that on average cities received 2.46 citations upon entering an IPC, and the most number of citations received was 583 in an IPC belonging to the field of Electrical machinery, apparatus, energy.

To estimate the effect of "collaborative entry" on post-entry technological performance we used a count model because the dependent variable is a count variable, and unlike other regression models, the model accounts for the non-negativity and discreteness of the dependent variable. Specifically, we applied the Negative Binomial regression model with fixed effect, which accounts for over-dispersion in the dependent variable (as the standard deviation of technological performance is 7.98 which exceeds the mean 2.46). Correlations in error terms arising from unobserved heterogeneity in cities are accounted for by clustering standard errors at the level of cities. Table 42 shows the empirical results for technological performance. The coefficients in the table are incidence rate ratio. Hence, the coefficients are interpreted as the proportional change in the performance of a city in a new technology when the city's patent in the new technology at the point of entry was done in collaboration with other cities, as opposed to patenting non-collaboratively. If the coefficient is greater than 1 we have a positive impact, if it is less than 1 we have a negative impact. Model 1 has only the control variables, while in Model 2, in addition to the control variables in Model 1, we have the bi-variate variable of interest i.e. "collaborative entry." From Model 2 in Table 42

we see a significant positive effect of collaborative entry on technological performance. Specifically, collaborating at the point of entering a new technology domain as opposed to not collaborating at the point of entry increase the city's performance in the new technology domain by 72.7%. From Model 1 in Table 42 we see that entering a new technology domain that is far away from the city's core IPC decreases the performance of the city in the technological area. Also, cities experience increased performance entering more attractive IPCs. Finally, we capture the effect of multinational corporations by examining the share of the city's foreign patents, we have that a standard deviation increase in the share of foreign patent results in an 8.1% increase in the city's performance upon entering a new technology domain.

C.0.9 Results of stratified semi-parametric Cox PH for a sample of cities in Europe only

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.005*	1.016***	1.006*	1.016***
		(0.00254)	(0.00237)	(0.00255)	(0.00237)
Technological relatedness		1.055**	1.055**	1.055***	1.055***
		(0.0187)	(0.0187)	(0.0169)	(0.0170)
Inter-city ties × Partner city size			1.032***		1.032***
			(0.00473)		(0.00464)
Technological relatedness × Focal city size				1.008*	1.008*
				(0.00358)	(0.00358)
Distance from core IPC	1.130***	1.129***	1.128***	1.128***	1.128***
	(0.00749)	(0.00746)	(0.00746)	(0.00746)	(0.00745)
Knowledge diversity	1.080***	1.082***	1.082***	1.082***	1.082***
	(0.0228)	(0.0231)	(0.0231)	(0.0231)	(0.0231)
Existing IPCs	1.749***	1.698***	1.697***	1.707***	1.707***
	(0.139)	(0.136)	(0.136)	(0.136)	(0.136)
IPC growth rate	1.124***	1.119***	1.119***	1.118***	1.118***
	(0.00683)	(0.00744)	(0.00745)	(0.00747)	(0.00748)
City complexity	0.985	0.986	0.986	0.986	0.986
	(0.0108)	(0.0108)	(0.0108)	(0.0108)	(0.0108)
Inventor density	1.010	1.009	1.010	1.010	1.010
	(0.0201)	(0.0199)	(0.0200)	(0.0200)	(0.0201)
Share of foreign patent	1.018	1.019	1.019	1.018	1.018
	(0.0183)	(0.0184)	(0.0184)	(0.0184)	(0.0184)
Focal city size	1.092*	1.097**	1.096**	1.095**	1.094*
	(0.0387)	(0.0390)	(0.0390)	(0.0384)	(0.0384)
Partner city size	0.944**	0.945**	0.946*	0.945**	0.946**
	(0.0203)	(0.0204)	(0.0205)	(0.0204)	(0.0204)
Data quality	1.019	1.026	1.026	1.026	1.026
	(0.0244)	(0.0245)	(0.0245)	(0.0245)	(0.0245)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	329 468	329 468	329 468	329 468	329 468
Log pseudolikelihood	-227 829.24	-227 751.01	-227 745.66	-227 741.79	-227 736.5
Wald chi2	39 706.45***	40 268.84***	40 285.24***	40 433.91***	40 453.72***

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 43: Supplementary analysis. Results of stratified semi-parametric Cox PH for a sample of cities in Europe only.

C.0.10 Results of stratified semi-parametric Cox PH for a sample of cities in North America only

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		0.994	0.976	0.993	0.977
		(0.00511)	(0.0157)	(0.00508)	(0.0157)
Technological relatedness		1.143***	1.143***	1.152***	1.152***
		(0.0124)	(0.0124)	(0.0132)	(0.0132)
Inter-city ties × Partner city size			0.969		0.970
			(0.0213)		(0.0212)
Technological relatedness × Focal city size				0.991	0.991
				(0.00399)	(0.00401)
Distance from core IPC	1.133***	1.128***	1.128***	1.128***	1.128***
	(0.0110)	(0.0108)	(0.0108)	(0.0108)	(0.0108)
Knowledge diversity	1.078*	1.086*	1.086*	1.086*	1.086*
	(0.0399)	(0.0406)	(0.0406)	(0.0406)	(0.0406)
Existing IPCs	1.619***	1.599**	1.603**	1.572**	1.577**
	(0.234)	(0.233)	(0.234)	(0.230)	(0.230)
IPC growth rate	1.158***	1.144***	1.144***	1.144***	1.145***
	(0.00899)	(0.0112)	(0.0111)	(0.0110)	(0.0110)
City complexity	1.013	1.018	1.019	1.016	1.017
	(0.0120)	(0.0120)	(0.0120)	(0.0120)	(0.0119)
Inventor density	1.107*	1.110*	1.111*	1.109*	1.109*
	(0.0512)	(0.0524)	(0.0524)	(0.0523)	(0.0523)
Share of foreign patent	1.014	1.012	1.012	1.012	1.012
	(0.0304)	(0.0293)	(0.0293)	(0.0292)	(0.0293)
Focal city size	1.036	1.038	1.039	1.057	1.057
	(0.0596)	(0.0626)	(0.0626)	(0.0650)	(0.0651)
Partner city size	0.954*	0.955*	0.954*	0.955*	0.954*
	(0.0200)	(0.0198)	(0.0198)	(0.0198)	(0.0198)
Data quality	1.125	1.061	1.061	1.058	1.058
	(0.127)	(0.116)	(0.116)	(0.115)	(0.115)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	116 920	116 920	116 920	116 920	116 920
Log pseudolikelihood	-95 274.886	-95 198.904	-95 197.976	-95 196.097	-95 195.221
Wald chi2	20 526.21***	21 352.32***	21 474.57***	21 494.26***	21 616.36***

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 44: Supplementary analysis. Results of stratified semi-parametric Cox PH for a sample of cities in North America only.

C.0.11 Results of stratified semi-parametric Cox PH for a sample of cities in Asia only

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.020**	1.027**	1.020**	1.028**
		(0.00746)	(0.00938)	(0.00752)	(0.00941)
Technological relatedness		1.201***	1.201***	1.220***	1.219***
		(0.0258)	(0.0258)	(0.0270)	(0.0270)
Inter-city ties × Partner city size			1.017		1.016
			(0.0107)		(0.0107)
Technological relatedness × Focal city size				0.969***	0.969***
				(0.00887)	(0.00884)
Distance from core IPC	1.193***	1.181***	1.182***	1.181***	1.181***
	(0.0128)	(0.0128)	(0.0128)	(0.0128)	(0.0128)
Knowledge diversity	0.987	1.000	1.000	1.003	1.003
	(0.0485)	(0.0495)	(0.0496)	(0.0497)	(0.0497)
Existing IPCs	2.038***	2.075***	2.074***	2.068***	2.067***
	(0.202)	(0.211)	(0.210)	(0.211)	(0.211)
IPC growth rate	1.153***	1.140***	1.140***	1.139***	1.139***
	(0.0144)	(0.0153)	(0.0153)	(0.0154)	(0.0154)
City complexity	0.972	0.964	0.964	0.964	0.964
	(0.0173)	(0.0182)	(0.0182)	(0.0186)	(0.0186)
Inventor density	1.065	1.046	1.047	1.026	1.027
	(0.0361)	(0.0354)	(0.0354)	(0.0419)	(0.0418)
Share of foreign patent	0.882***	0.874***	0.874***	0.872***	0.872***
	(0.0299)	(0.0301)	(0.0301)	(0.0302)	(0.0302)
Focal city size	0.991	0.981	0.982	1.063	1.064
	(0.0520)	(0.0488)	(0.0492)	(0.0542)	(0.0543)
Partner city size	0.949	0.954	0.954	0.953	0.954
	(0.0346)	(0.0359)	(0.0361)	(0.0362)	(0.0363)
Data quality	1.007	0.998	0.998	0.998	0.998
	(0.0286)	(0.0288)	(0.0288)	(0.0289)	(0.0289)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	42 347	42 347	42 347	42 347	42 347
Log pseudolikelihood	-56 084.435	-56 010.613	-56 010.387	-55 998.242	-55 998.022
Wald chi2	10 674.44***	11 736.90***	12 326.55***	12 480.83***	13 007.81***

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 45: Supplementary analysis. Results of stratified semi-parametric Cox PH for a sample of cities in Asia only.

C.0.12 Results of stratified semi-parametric Cox PH for Global cities

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.002	1.008	1.001	1.008
		(0.00519)	(0.0102)	(0.00524)	(0.0102)
Technological relatedness		1.141***	1.141***	1.160***	1.160***
		(0.0174)	(0.0174)	(0.0181)	(0.0181)
Inter-city ties × Partner city size			1.009		1.010
			(0.0127)		(0.0125)
Technological relatedness × Focal city size				0.973**	0.973***
				(0.00803)	(0.00802)
Distance from core IPC	1.142***	1.135***	1.135***	1.134***	1.135***
	(0.0117)	(0.0115)	(0.0115)	(0.0115)	(0.0115)
Knowledge diversity	0.929	0.936	0.936	0.939	0.939
	(0.0393)	(0.0406)	(0.0406)	(0.0409)	(0.0409)
Existing IPCs	1.611***	1.664***	1.664***	1.645***	1.645***
	(0.151)	(0.162)	(0.162)	(0.160)	(0.160)
IPC growth rate	1.132***	1.120***	1.120***	1.121***	1.120***
	(0.0120)	(0.0142)	(0.0142)	(0.0141)	(0.0141)
City complexity	1.009	1.009	1.009	1.007	1.007
	(0.0135)	(0.0139)	(0.0140)	(0.0141)	(0.0141)
Inventor density	1.036	1.032	1.031	1.016	1.015
	(0.0444)	(0.0440)	(0.0443)	(0.0442)	(0.0445)
Share of foreign patent	0.954	0.950*	0.950*	0.950*	0.950*
	(0.0243)	(0.0245)	(0.0245)	(0.0245)	(0.0245)
Focal city size	0.982	0.981	0.982	1.047	1.048
	(0.0516)	(0.0508)	(0.0510)	(0.0565)	(0.0567)
Partner city size	0.951**	0.950**	0.950**	0.950**	0.950*
	(0.0184)	(0.0188)	(0.0188)	(0.0189)	(0.0189)
Data quality	1.002	0.988	0.988	0.986	0.986
	(0.0323)	(0.0317)	(0.0317)	(0.0316)	(0.0316)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	54 731	54 731	54 731	54 731	54 731

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 46: Supplementary analysis. Results of stratified semi-parametric Cox PH for a sample of 168 global cities. These global cities are alpha and beta cities identified in the Globalization and World Cities (GaWC) project.

C.0.13 Results of stratified semi-parametric Cox PH for non-global cities

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.005*	1.008**	1.005*	1.008**
		(0.00258)	(0.00243)	(0.00259)	(0.00244)
Technological relatedness		1.066**	1.066**	1.065**	1.065**
		(0.0227)	(0.0227)	(0.0209)	(0.0209)
Inter-city ties × Partner city size			1.011***		1.011***
			(0.00300)		(0.00300)
Technological relatedness × Focal city size				1.007	1.007
				(0.00375)	(0.00375)
Distance from core IPC	1.139***	1.138***	1.138***	1.137***	1.137***
	(0.00641)	(0.00636)	(0.00636)	(0.00637)	(0.00637)
Knowledge diversity	1.083***	1.087***	1.088***	1.088***	1.089***
	(0.0197)	(0.0199)	(0.0200)	(0.0200)	(0.0200)
Existing IPCs	1.830***	1.777***	1.777***	1.788***	1.788***
	(0.120)	(0.118)	(0.118)	(0.119)	(0.119)
IPC growth rate	1.137***	1.131***	1.131***	1.130***	1.130***
	(0.00536)	(0.00611)	(0.00611)	(0.00610)	(0.00610)
City complexity	0.985	0.986	0.986	0.986	0.986
	(0.00896)	(0.00894)	(0.00894)	(0.00893)	(0.00893)
Inventor density	0.998	0.996	0.996	0.994	0.994
	(0.0131)	(0.0128)	(0.0128)	(0.0126)	(0.0126)
Share of foreign patent	1.001	1.001	1.001	1.001	1.001
	(0.0154)	(0.0155)	(0.0155)	(0.0155)	(0.0155)
Focal city size	1.111***	1.114***	1.113***	1.110***	1.110***
	(0.0269)	(0.0264)	(0.0264)	(0.0258)	(0.0259)
Partner city size	0.944***	0.946**	0.946**	0.946**	0.946**
	(0.0164)	(0.0164)	(0.0164)	(0.0164)	(0.0164)
Data quality	1.014	1.019	1.019	1.019	1.019
	(0.0236)	(0.0237)	(0.0237)	(0.0237)	(0.0237)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	466 608	466 608	466 608	466 608	466 608

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 47: Supplementary analysis. Results of stratified semi-parametric Cox PH for a sample of non-global cities.

C.0.14 Results of stratified semi-parametric Cox PH using global city status as a moderator of inter-city ties

The non-global city variable takes the value 1 if the city is classified as being non-global and 0 if it is classified as being global.

	Model 1	Model 2	Model 3	Model 4	Model 5
Inter-city ties		1.005*	1.001	1.004*	1.001
		(0.00209)	(0.00271)	(0.00210)	(0.00272)
Technological relatedness		1.074**	1.074**	1.074**	1.074**
		(0.0253)	(0.0253)	(0.0264)	(0.0264)
Inter-city ties × Non global city status			1.005		1.006*
			(0.00398)		(0.00398)
Technological relatedness × Focal city size				1.007	1.007
				(0.00375)	(0.00375)
Distance from core IPC	1.139***	1.136***	1.137***	1.137***	1.137***
	(0.00572)	(0.00570)	(0.00570)	(0.00570)	(0.00570)
Knowledge diversity	1.064***	1.064***	1.064***	1.064***	1.064***
	(0.0184)	(0.0186)	(0.0186)	(0.0186)	(0.0186)
Existing IPCs	1.788***	1.736***	1.735***	1.734***	1.733***
	(0.103)	(0.101)	(0.101)	(0.102)	(0.101)
IPC growth rate	1.137***	1.130***	1.130***	1.130***	1.130***
	(0.00497)	(0.00593)	(0.00593)	(0.00587)	(0.00587)
City complexity	0.993	0.994	0.994	0.994	0.994
	(0.00729)	(0.00726)	(0.00727)	(0.00727)	(0.00728)
Inventor density	1.019	1.012	1.013	1.012	1.012
	(0.0178)	(0.0162)	(0.0162)	(0.0161)	(0.0160)
Share of foreign patent	1.001	1.001	1.001	1.001	1.001
	(0.0135)	(0.0137)	(0.0137)	(0.0137)	(0.0137)
Focal city size	1.036	1.037	1.038	1.040	1.041
	(0.0245)	(0.0238)	(0.0240)	(0.0229)	(0.0229)
Non-global city	0.944*	0.946*	0.946*	0.946*	0.946*
	(0.0164)	(0.0164)	(0.0164)	(0.0164)	(0.0164)
Data quality	1.000	1.008	1.008	1.008	1.008
	(0.0184)	(0.0184)	(0.0184)	(0.0184)	(0.0184)
City FE	Yes	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Number of observations	521 339	521 339	521 339	521 339	521 339

Notes: (i) Coefficients are exponentiated (ii) Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 48: Supplementary analysis. Results of stratified semi-parametric Cox PH using global city status as moderator of inter-city ties.

Partner size	Relative hazard	P > |z|
-1	-0.0014	0.127
0	0.0024	0.000
1	0.0058	0.000
2	0.0088	0.000
3	0.0116	0.000
4	0.0142	0.000
5	0.0173	0.000
6	0.0239	0.001
7	0.0514	0.001
8	0.2069	0.003
9	1.1675	0.645
10	7.2500	0.780
11	46.0784	0.856
12	294.3229	0.868
13	1879.51	0.878
14	11978.2	0.886

Table 49: Average marginal effect of inter-city ties: Relative hazard estimated at different points of the mean-centered partner size.

Bibliography

Adams, J. (2012). "Collaborations: The rise of research networks". In: *Nature, 490(7420), 335–336*.

Adams, J.D., G.C. Black, J.R. Clemmons, and P.E. Stephan (2005). "Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999". In: *Research Policy, 34, 259-285.*

Agrawal, A., D. Kapur, and J. McHale (2008). "How do spatial and social proximity influence knowledge flows? Evidence from patent data". In: *Journal of urban economics, 64(2), pp.258-269*.

Ahuja, G. and C. Morris Lampert (2001). "Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions". In: *Strategic management journal, 22(6-7), pp.521-543*.

Akira Goto, Kazuyuki Motohashi (2007). "Construction of a Japanese Patent Database and a First Look at Japanese Patenting Activities". In: *Research Policy, 36 (9), 1431-1442*.

Alcácer, J. and M. Zhao (2016). "Zooming in: A practical manual for identifying geographic clusters". In: *Strategic Management Journal, 37(1), pp.10-21*.

Alderson, A.S. and J. Beckfield (2004). "Power and position in the world city system". In: *American Journal of sociology, 109(4), pp.811-851*.

Angelou, K., M. Maragakis, K. Kosmidis, and P. Argyrakis (2020a). "Dynamics of regional multilinks in research innovation temporal networks". In: *EPL (Europhysics Letters), 130(2), p.28001*.

— (2020b). "Triangular research and innovation collaborations in the European area". In: *arXiv preprint arXiv:2011.04553*.

Arruda, G.F. de, G. Petri, and Y. Moreno (2020). "Social contagion models on hypergraphs". In: *Physical Review Research, 2(2), p.023032*.

Arruda, G.F. de, M. Tizzani, and Y. Moreno (2021). "Phase transitions and stability of dynamical processes on hypergraphs". In: *Communications Physics, 4(1), pp.1-9*.

Asheim, B. T. and L. Coenen (2006). "Contextualising regional innovation systems in a globalising learning economy: On knowledge bases and institutional frameworks". In: *The Journal of Technology Transfer, 31(1), 163-173*.

Asheim, B.R.T. (1996). "Industrial districts as "learning regions": a condition for prosperity". In: *European planning studies, 4(4), pp.379-400*.

Atkinson, A.B. and J.E. Stiglitz (1969). "A new view of technological change". In: *The Economic Journal, 79(315), pp.573-578*.

Bala Subrahmanya, M.H. (2017). "How did Bangalore emerge as a global hub of tech start-ups in India? Entrepreneurial ecosystem—Evolution, structure and role". In: *Journal of Developmental Entrepreneurship, 22(01), p.1750006*.

Balland P.A., Broekel T. Diodato D. Giuliani E. Hausmann R. O'Clery N. and D. Rigby (2022). "The new paradigm of economic complexity". In: *Research Policy, 51(3), p.104450*.

Balland, P. A., R. Boschma, J. Crespo, and D. L. Rigby (2018). "Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification". In: *Reg. Stud. 53, 1252–1268*.

Balland, P.A. and R. Boschma (2021). "Complementary interregional linkages and Smart Specialisation: an empirical study on European regions". In: *Regional Studies, pp.1-12*.

Balland, P.A., R. Boschma, J. Crespo, and D.L. Rigby (2017). "Relatedness, knowledge complexity and technological opportunities of regions: A framework for smart specialisation". In: *Papers in Evolutionary Economic Geography, 17*.

— (2019). "Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification". In: *Regional Studies, 53(9), pp.1252-1268.*

Balland, P.A., C. Jara-Figueroa, S.G. Petralia, M.P. Steijn, D.L. Rigby, and C.A. Hidalgo (2020). "Complex economic activities concentrate in large cities". In: *Nature Human Behaviour, 4(3), pp.248-254*.

Balsmeier, B., A. Chavosh, G.C. Li, G. Fierro, K. Johnson, A. Kaulagi, D. O'Reagan, B. Yeh, and L. Fleming (2015). "Automated disambiguation of US patent grants and applications". In: *Unpublished working paper, Fung Institute for Engineering Leadership*.

Barjak, F. and S. Robinson (2008). "International collaboration, mobility and team diversity in the life sciences: Impact on research performance". In: *Social Geography, 3(1), 23–36*.

Barzotto M., Corradini C. Fai F.M. Labory S. and P.R. Tomlinson (2019). "Enhancing innovative capabilities in lagging regions: an extra-regional collaborative approach to RIS3". In: *Cambridge Journal of Regions, Economy and Society, 12(2), pp.213-232*.

Basile, R., G. Cicerone, and L. Iapadre (2019). "Economic complexity and regional labor productivity distribution: evidence from Italy". In: *Rev. Reg. Stud. 47, 201–219*.

Bassens, D., B. Derudder, and F. Witlox (2011). "A multiple-perspectives construct of the American global city". In: *Urban Studies, 45(1), 3–28*.

Bathelt, H., A. Malmberg, and P. Maskell (2004). "Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation". In: *Progress in human geography, 28(1), pp.31-56*.

Beaverstock, J.V., R.G. Smith, and P.J. Taylor (1999). "A roster of world cities". In: *cities, 16(6), pp.445-458*.

Becattini, G. (2017). "The Marshallian industrial district as a socio-economic notion". In: *Revue d'economie industrielle, (157), pp.13-32*.

Bednarz, M. and T. Broekel (2019). "The relationship of policy induced R&D networks and inter-regional knowledge diffusion". In: *Journal of Evolutionary Economics, 29(5), pp.1459-1481*.

Belderbos, R., F. Benoit, S. Edet, G.H. Lee, and M. Riccaboni (2021). "Global cities' cross-border innovation networks. A chapter contribution in the book Cross-Border Innovation in a Changing World: Players, Places and Policies, edited by D. Castellani, A. Perri, V. Scalera, A. Zanfei". In: *Oxford University Press.*

Belderbos, R., H. S. Du, and A. Goerzen (2017). "Global cities, connectivity, and the location choice of MNC regional headquarters". In: *Journal of Management Studies, 54(8), 1271–1302*.

Belderbos, R., D. Faems, B. Leten, and B.V. Looy (2010). "Technological activities and their impact on the financial performance of the firm: Exploitation and exploration within and between firms". In: *Journal of Product Innovation Management, 27(6), pp.869-882*.

Ben Saâd, M. and G. Assoumou-Ella (2019). "Economic complexity and gender inequality in education: an empirical study". In: *Econ. Bull. 39, 321–334*.

Besedeš, T. (2008). "A search cost perspective on formation and duration of trade". In: *Review of International Economics, 16(5), pp.835-849*.

Bettencourt, L.M., J. Lobo, and D. Strumsky (2007). "Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size". In: *Research policy, 36(1), pp.107-120*.

Blondel, V.D., J.L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment, 2008(10), p.P10008*.

Bornmann, L., C. Wagner, and L. Leydesdorff (2015). "BRICS countries and scientific excellence: A bibliometric analysis of most frequently cited papers". In: *Journal of the Association for Information Science and Technology, 66(7), 1507–1513.*

Boschken, H. L. (2008). "A multiple-perspectives construct of the American global city". In: *Urban Studies, 45(1), 3–28*.

Boschma, R. (2004). "Competitiveness of regions from an evolutionary perspective". In: *Regional studies, 38(9), pp.1001-1014*.

Boschma, R., P.A. Balland, and D.F. Kogler (2015). "Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010". In: *Industrial and corporate change, 24(1), pp.223-250*.

Boschma, R. and K. Frenken (2011). "Technological relatedness and regional branching". In: *In: Bathelt, H. Feldman, MP, Kogler DF (eds), Dynamic geographies of knowledge creation and innovation*.

Boschma, R., A. Minondo, and M. Navarro (2013). "Related variety and regional growth in Spain". In: *Papers in Regional Science, 91(2), pp.241-256*.

Boschma, R. A. and S. Iammarino (2009). "Related variety, trade linkages and regional growth". In: *Economic Geography, 85(3):289–311*.

Boser, B.E., I.M. Guyon, and V.N. Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152)*.

Boughorbel S., Jarray F. and M. El-Anbari (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". In: *PloS one, 12(6), p.e0177678*.

Breschi, S. and C. Lenzi (2016). "Co-invention networks and inventive productivity in US cities". In: *Journal of Urban Economics, 92, pp.66-75.*

Breschi, S. and F. Lissoni (2001). "Localised knowledge spillovers vs. innovative milieux: Knowledge "tacitness" reconsidered". In: *Papers in regional science, 80(3), pp.255-273*.

Breschi, S., F. Lissoni, and F. Malerba (2003). "Knowledge-relatedness in firm technological diversification". In: *Research policy, 32(1), pp.69-87*.

Burt, R.S. (2004). "Structural holes and good ideas". In: *American journal of sociology, 110(2), pp.349-399*.
Camagni, R. (1991). "Innovation networks: spatial perspectives". In: *Belhaven-Pinter*.
Cantwell, J. and G. Vertova (2004). "Historical evolution of technological diversification". In: *Research Policy, 33(3), pp.511-529*.
Carletti, T., F. Battiston, G. Cencetti, and D. Fanelli (2020). "Random walks on hypergraphs". In: *Physical Review E 101, 022308*.
Carletti, T., D. Fanelli, and S. Nicoletti (2020). "Dynamical systems on hypergraphs". In: *Journal of Physics: Complexity, 1(3), p.035006*.
Casiraghi, G. and V. Nanumyan (2018). "Generalised hypergeometric ensembles of random graphs: The configuration model as an urn problem". In: *arXiv preprint arXiv:1810.06495.*
Casiraghi, G., V. Nanumyan, I. Scholtes, and F. Schweitzer (2017). "From relational data to graphs: Inferring significant links using generalized hypergeometric ensembles". In: *In International Conference on Social Informatics (pp. 111-120). Springer, Cham.*
Castellani, D. and K. Lavoratori (2019). "Location of R&D abroad: An analysis on global cities". In: *In Relocation of Economic Activity (pp. 145-162)*.
Castells, M. (2002). "Local and global: cities in the network society". In: *Tijdschrift voor Economische en Sociale Geografie 93(5), 548-558*.
Catini, R., D. Karamshuk, O. Penner, and M. Riccaboni (2015). "Identifying geographic clusters: A network analytic approach". In: *Research policy* 44.9, pp. 1749–1762.
Cerina, F., A. Chessa, F. Pammolli, and M. Riccaboni (2014). "Network communities within and across borders". In: *Scientific Reports* 4.1, pp. 1–7.
Chakravarty, D., A. Goerzen, M. Musteen, and M. Ahsan (2021). "Global cities: A multi-disciplinary review and research agenda". In: *Journal of World Business, 56(3), p.101182*.
Chen, T. and C. Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794)*.
Cheng, Y. and R. LeGates (2018). "China's hybrid global city region pathway: Evidence from the Yangtze River Delta". In: *Cities, 77, pp.81-91*.
Chessa, A., A. Morescalchi, F. Pammolli, O. Penner, A.M. Petersen, and M. Riccaboni (2013). "Is Europe evolving toward an integrated research area?" In: *Science, 339(6120), pp.650-651.*

Choi, S. (2012). "Core-periphery, new clusters, or rising stars? international scientific collaboration among "advanced" countries in the era of globalization". In: *Scientometrics, 90(1), 25–41.*

Cohen, R.B. (2018). "The new international division of labor, multinational corporations and urban hierarchy (pp. 287-315)". In: *Routledge*.

Cohen, W.M. and D.A. Levinthal (1990). "Absorptive capacity: A new perspective on learning and innovation". In: *Administrative Science Quarterly, 35(1), 128–152*.

Colombelli, A., J. Krafft, and F. Quatraro (2014). "The emergence of new technology-based sectors in European regions: A proximity-based analysis of nanotechnology". In: *Research Policy, 43(10), pp.1681-1696*.

Cortes, C. and V. Vapnik (1995). "Support-vector networks". In: *Machine learning, 20(3), pp.273-297*.

Cox, D. R. (1972). "Regression Models and Life-Tables (with Discussion)". In: *Journal of the Royal Statistical Society, Series B: Methodological 34: 187–220*.

Crespi G., Geuna A. and L.J. Nesta (2006). "Labour mobility of academic inventors: career decision and knowledge transfer". In: *In A study of inventors (pp. 65-119). Deutscher Universitätsverlag, Wiesbaden*.

Cuxac P., Lamirel J.C. and V. Bonvallot (2013). "Efficient supervised and semi-supervised approaches for affiliations disambiguation". In: *Scientometrics, 97(1), pp.47-58*.

De Rassenfosse, G., H. Dernis, D. Guellec, L. Picci, and B.V.P. de la Potterie (2013). "The worldwide count of priority patents: A new indicator of inventive activity". In: *Research Policy, 42(3), pp.720-737*.

De Rassenfosse, G., J. Kozak, and F. Seliger (2019). "Geocoding of worldwide patent data". In: *Scientific data, 6(1), pp.1-15*.

De Rassenfosse, G. and F. Seliger (2021). "Imputation of missing information in worldwide patent data". In: *Data in Brief, 34, p.106615*.

Derudder, B. and P. Taylor (2005). "The cliquishness of world cities". In: *Global Networks, 5(1), pp.71-91*.

Derudder, B. and F. Witlox (2008). "Mapping world city networks through airline flows: context, relevance, and problems". In: *Journal of Transport Geography, 16(5), pp.305-312*.

Derudder, B., F. Witlox, and P.J. Taylor (2013). "US cities in the world city network: Comparing their positions using global origins and destinations of airline passengers". In: *Urban Geography, 28(1), pp.74-91*.

Dijkstra, L. and H. Poelman (2012). "Cities in Europe: the new OECD-EC definition". In: *Regional focus, 1(2012), pp.1-13*.

Dijkstra, L., H. Poelman, and P. Veneri (2019). "The EU-OECD definition of a functional urban area". In: *OECD regional Development working papers*.

Edquist, C. and B.A. Lundvall (1993). "Comparing the Danish and Swedish systems of innovation". In: *National innovation systems: A comparative analysis, pp.265-298*.

Engelsman, E.C. and A.F.J. van Raan (1991). "Mapping of Technology, A First Exploration of Knowledge Diffusion Amongst Fields of Technology". In: *Policy Studies on Technology and Economy (BTE) Series, No. 15. The Hague*.

— (1992). "A patent-based cartography of technology". In: *Research Policy 23, 1–26*.

Essletzbichler, J. (2015). "Relatedness, industrial branching and technological cohesion in US metropolitan areas". In: *Regional Studies, 49(5), pp.752-766*.

Estrada, E. and G.J. Ross (2018). "Centralities in simplicial complexes: Applications to protein interaction networks". In: *Journal of theoretical biology, 438, pp.46-60*.

Frenken, K., J. Hoekman, S. Kok, R. Ponds, F. van Oort, and J. van Vliet (2009). "Death of distance in science? A gravity approach to research collaboration". In: *In Innovation networks (pp. 43-57). Springer Berlin Heidelberg*.

Frenken, K., F. Van Oort, and T. Verburg (2007). "Related variety, unrelated variety and regional economic growth". In: *Regional studies, 41(5), pp.685-697*.

Friedman, J.H. (2002). "Stochastic gradient boosting". In: *Computational statistics & data analysis, 38(4), pp.367-378*.

Friedmann, J. (1986). "The world city hypothesis". In: *Development and Change, 17, 69–83*.

Fritz, B. S. L. and R. A. Manduca (2019). "The economic complexity of US metropolitan areas". In: *Preprint at arXiv https://arxiv.org/abs/1901.08112*.

Furman, J.L., M.E. Porter, and S. Stern (2002). "The determinants of national innovative capacity". In: *Research policy, 31(6), pp.899-933*.

Gertler, M.S. (2003). "Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there)". In: *Journal of economic geography, 3(1), pp.75-99*.

Glaeser, E.L. and J.D. Gottlieb (2009). "The wealth of cities: Agglomeration economies and spatial equilibrium in the United States". In: *Journal of economic literature, 47(4), pp.983-1028*.

Goerzen, A., C. Asmussen, and B. Nielsen (2013). "Global cities and multinational enterprise location strategy". In: *Journal of International Business Studies, 44(5): 427–450*.

Google Geocoding API. Accessed: 2020-10-18.

Griliches, Z. (1990). "Patent statistics as economic indicators: a survey (No. w3301)". In: *National Bureau of Economic Research*.

Grilli, J., G. Barabás, M.J. Michalska-Smith, and S. Allesina (2017). "Higher-order interactions stabilize dynamics in competitive network models". In: *Nature, 548(7666), pp.210-213*.

Grossman, G.M. and E. Helpman (1991). "Innovation and growth in the global economy". In: *MIT press*.

Güneri, B. and A. Y. Yalta (2020). "Does economic complexity reduce output volatility in developing countries?" In: *Bull. Econ. Res. https://doi.org /10.1111/boer.12257*.

Hannigan, T. J., M. Cano-Kollmann, and R. Mudambi (2015). "Thriving innovation amidst manufacturing decline: The Detroit auto cluster and resilience of local knowledge production". In: *Industrial and Corporate Change, 24(3): 613–634*.

Hausmann, R., C.A. Hidalgo, S. Bustos, M. Coscia, and A. Simoes (2014). "The atlas of economic complexity: Mapping paths to prosperity". In: *MIT Press*.

Hausmann, R., J. Hwang, and D. Rodrik (2007). "What you export matters". In: *Journal of economic growth, 12(1), pp.1-25*.

He, T. (2008). "International scientific collaboration of China with the G7 countries". In: *Scientometrics, 80(3), 571–582.*

Helmers, C. (2019). "Choose the neighbor before the house: Agglomeration externalities in a UK science park". In: *Journal of economic geography, 19(1), pp.31-55*.

Henderson, J. and M. eds. Castells (1987). "Global restructuring and territorial development". In: *(p. 110). London: Sage*.

Hennemann, S., D. Rybski, and I. Liefner (2012). "The myth of global science collaboration: Collaboration patterns in epistemic communities". In: *Journal of Informetrics, 6(2), pp.217-225*.

Hess, W. and M. Persson (2010). "The Duration of Trade Revisited: Continuous-Time vs. Discrete-Time Hazards". In: *IFN Working Paper, No. 829, Research Institute of Industrial Economics (IFN), Stockholm*.

Hidalgo, C.A. (2021). "Economic complexity theory and applications". In: *Nature Reviews Physics, 3(2), pp.92-113*.

Hidalgo, C.A. and R. Hausmann (2009). "The building blocks of economic complexity". In: *Proceedings of the national academy of sciences, 106(26), pp.10570-10575*.

Hidalgo, C.A., B. Klinger, A.L. Barabási, and R. Hausmann (2007). "The product space conditions the development of nations". In: *Science, 317(5837), pp.482-487*.

Hoekman, J., K. Frenken, and R. J. Tijssen (2010). "Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe". In: *Research Policy, 39(5), 662–673.*

Hoisl, K. (2007). "Tracing mobile inventors—the causality between inventor mobility and inventor productivity". In: *In A study of inventors (pp. 65-119). Deutscher Universitätsverlag, Wiesbaden*.

Jaffe, A.B. (1989). "Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers". In: *Research Policy, 18(2), pp.87-97*.

Jaffe, A.B., M. Trajtenberg, and R. Henderson (1993). "Geographic localization of knowledge spillovers as evidenced by patent citations". In: *the Quarterly journal of Economics, 108(3), pp.577-598*.

Jones, Benjamin F. (2021). "The Rise of Research Teams: Benefits and Costs in Economics". In: *Journal of Economic Perspectives, 35(2), 191-216*.

Jones, Benjamin F, Stefan Wuchty, and Brian Uzzi (2008). "Multi-university research teams: Shifting impact, geography, and stratification in science". In: *Science, 322(5905), 1259-1262*.

Jonkers, K. and L. Cruz-Castro (2013). "Research upon return: The effect of international mobility on scientific ties, production and impact". In: *Research Policy, 42(8), 1366–1377.*

Kahn, M. E., W. Sun, J. Wu, and S. Zheng (2018). "The Revealed Preference of the Chinese Communist Party Leadership: Investing in Local Economic Development versus Rewarding Social Connections". In: *National Bureau of Economic Research (NBER) http://www.nber.org/papers/w24457*.

Kerr, W.R. (2010). "Breakthrough inventions and migrating clusters of innovation". In: *Journal of urban Economics, 67(1), pp.46-60*.

Kim, K., M. Khabsa, and C.L. Giles (2016). "Inventor name disambiguation for a patent database using a random forest and DBSCAN". In: *IEEE/ACM joint conference on digital libraries (JCDL) (pp. 269-270).*

Krugman, P. (1991). "Increasing returns and economic geography". In: *Journal of political economy, 99(3), pp.483-499*.

Lambiotte, R., V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren (2008). "Geographical dispersal of mobile communication networks". In: *Physica A: Statistical Mechanics and its*

Applications 387 (21) (2008) 5317–5325. arXiv:0802.2178, doi:10.1016/j.physa.2008.05.014.

Lapatinas, A., A. Kyriakou, and A. Garas (2019). "Taxation and economic sophistication: evidence from OECD countries". In: *PLoS ONE 14, e0213498*.

Laperche, B. and E.G. Carayannis (Ed.) (2013). "Knowledge capital and small businesses". In: *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship, Springer, New York (2013), pp. 1177-1181*.

Lavie, D., U. Stettner, and M.L. Tushman (2010). "Exploration and exploitation within and across organizations". In: *Academy of Management annals, 4(1), pp.109-155*.

Lawson, C. and E. Lorenz (1999). "Collective learning, tacit knowledge and regional innovative capacity". In: *Regional studies, 33(4), pp.305-317*.

Lecocq, C. and B. Van Looy (2009). "The impact of collaboration on the technological performance of regions: time invariant or driven by life cycle dynamics? An explorative investigation of European regions in the field of biotechnology". In: *Scientometrics, 80(3), pp.845-865*.

Leten, B., R. Belderbos, and B.V. Looy (2016). "Entry and technological performance in new technology domains: Technological opportunities, technology competition and technological relatedness". In: *Journal of Management Studies, 53(8), pp.1257-1291*.

Levinthal, D.A. and J.G. March (1993). "The myopia of learning". In: *Strategic management journal, 14(S2), pp.95-112*.

Levitt, B. and J.G. March (1988). "Organizational learning". In: *Annual review of sociology, 14(1), pp.319-338*.

Li, D., Y.D. Wei, and T. Wang (2015). "Spatial and temporal evolution of urban innovation network in China". In: *Habitat International, 49, pp.484-496*.

Li, G.C., R. Lai, A. D'Amour, D.M. Doolin, Y. Sun, V.I. Torvik, Z.Y. Amy, and L. Fleming (2014). "Disambiguation and co-authorship networks of the US patent inventor database (1975–2010)". In: *Research Policy, 43(6), pp.941-955*.

Lobo, J. and D. Strumsky (2008). "Metropolitan patenting, inventor agglomeration and social networks: A tale of two effects". In: *Journal of Urban Economics, 63(3), pp.871-884*.

Lord, L.D., P. Expert, H.M. Fernandes, G. Petri, T.J. Van Hartevelt, F. Vaccarino, G. Deco, F. Turkheimer, and M.L. Kringelbach (2016). "Insights into brain architectures from the homological scaffolds of functional connectivity networks". In: *Frontiers in systems neuroscience, 10, p.85*.

Lorenzen, M. and R. Mudambi (2013). "Clusters, connectivity and catch-up: Bollywood and Bangalore in the global economy". In: *Journal of Economic Geography, 13(3), 501-534*.

Lorenzen, M., R. Mudambi, and A. Schotter (2020). "International connectedness and local disconnectedness: MNE strategy, city-regions and disruption". In: *Journal of International Business Studies, 51(8), pp.1199-1222*.

Maggioni, M. A., M. Nosvelli, and T.E. Uberti (2007). "Space versus networks in the geography of innovation: A European analysis". In: *Regional science, 86(3): 471-493*.

Malecki, E. (2002). "The economic geography of the Internet's infrastructure". In: *Economic Geography 78, 399–424*.

Malmberg, A. and P. Maskell (2002). "The elusive concept of localization economies: towards a knowledge-based theory of spatial clustering". In: *Environment and Planning A, 34(3), 429-450*.

Maraut, Stéphane, Hélène Dernis, Colin Webb, Vincenzo Spiezia, and Dominique Guellec (2008). "The OECD REGPAT database: a presentation". In:

March, J.G. (1991). "Exploration and exploitation in organizational learning". In: *Organization science, 2(1), pp.71-87*.

Masey, D. (1984). "Spatial divisions of labour: Social structures and the geography of production". In: *London: Macmillan*.

Maskell, P. and A. Malmberg (1999). "The Competitiveness of Firms and Regions: "Ubiquitification" and the Importance of Localized Learning". In: *European urban and regional studies, 6(1), pp.9-25*.

Matthiessen, C.W., A.W. Schwarz, and S. Find (2010). "World cities of scientific knowledge: systems, networks and potential dynamics. An analysis based on bibliometric indicators". In: *Urban Studies, 47(9), 1879-97*.

Maurseth, P.B. and B. Verspagen (2002). "Knowledge spillovers in Europe: a patent citations analysis". In: *Scandinavian Journal of Economics, 104(4), pp.531-545*.

Mealy, P. and D. Coyle (2019). "To Them That Hath: Economic Complexity and Local Industrial Strategy in the UK". In: *Elsevier*.

Metcalfe, J.S. (1995). "Technology systems and technology policy in an evolutionary framework". In: *Cambridge journal of economics, 19(1), pp.25-46*.

Mewes, Lars and Tom Broekel (2020). "Technological complexity and economic growth of regions". In: *Research Policy (2020): 104156*.

Miguelez, E. (2019). "Collaborative patents and the mobility of knowledge workers". In: *Technovation, 86, pp.62-74*.
Miguélez, E. and R. Moreno (2012). "Skilled labour mobility, networks and knowledge creation in regions: a panel data approach". In: *The Annals of Regional Science: 1-22*.
Misra, S., H. Li, and J. He (2019). "Machine learning for subsurface characterization". In: *Gulf Professional Publishing*.
Molloy, M., B. Reed, M. Newman, A.L. Barabási, and D.J. Watts (2011). "A critical point for random graphs with a given degree sequence". In: *In The Structure and Dynamics of Networks (pp. 240-258). Princeton University Press.*
Morescalchi, Andrea, Fabio Pammolli, Orion Penner, Alexander M Petersen, and Massimo Riccaboni (2015). "The evolution of networks of innovators within and across borders: Evidence from patent data". In: *Research Policy, 44(3), 651-668*.
Morrison, G., S.V. Buldyrev, M. Imbruno, O.A.D. Arrieta, A. Rungi, M. Riccaboni, and F. Pammolli (2017). "On economic complexity and the fitness of nations". In: *Scientific reports, 7(1), pp.1-11*.
Morrison, G., M. Riccaboni, and F. Pammolli (2017b). "Disambiguation of patent inventors and assignees using high-resolution geolocation data". In: *Scientific data, 4, p.170064*.
— (2017a). "Disambiguation of patent inventors and assignees using high-resolution geolocation data". In: *Scientific data* 4.1, pp. 1–21.
Neagu, O. and M. C. Teodoru (2019). "The relationship between economic complexity, energy consumption structure and greenhouse gas emission: heterogeneous panel evidence from the EU countries". In: *Sustainability 11, 497*.
Neal, Z.P. (2011). "From central places to network bases: A transition in the US urban hierarchy, 1900-2000". In: *City & Community, 10(1), pp.49-75*.
Neffke, F., M. Henning, and R. Boschma (2011). "How do regions diversify over time? Industry relatedness and the development of new growth paths in regions". In: *Economic geography, 87(3), pp.237-265*.
Neuhäuser, L., A. Mellor, and R. Lambiotte (2020). "Multibody interactions and nonlinear consensus dynamics on networked systems". In: *Physical Review E, 101(3), p.032310*.
Newman, M.E.J., E.B. Naim, H. Frauenfelder, and Z. Toroczkai (Eds.) (2004). "Who is the best connected scientist? A study of scientific co-authorship networks". In: *Complex networks, Springer, Berlin, pp. 337-370*.

Niosi, J., P. Saviotti, B. Bellon, and M. Crow (1993). "National systems of innovation: in search of a workable concept". In: *Technology in society, 15(2), pp.207-227*.
Nitsch, V. (2009). "Die another day: Duration in German import trade". In: *Review of World Economics, 145(1), pp.133-154*.
OECD (2012). "Redefining Urban: A New Way to Measure Metropolitan Areas". In: *OECD Publishing*.
Orellana, A. and L. Fuentes (2019). "Metropolitan Area". In: *The Wiley Blackwell Encyclopedia of Urban and Regional Studies, pp.1-9*.
Otiso, K. M., B. Derudder, D. Bassens, L. Devriendt, and F Witlox (2011). "Airline connectivity as a measure of the globalization of African cities". In: *Applied Geography, 31 (2), 609–620*.
Owen-Smith, J. and W.W. Powell (2004). "Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community". In: *Organization science, 15(1), pp.5-21*.
Parnreiter, C. (2019). "Global cities and the geographical transfer of value". In: *Urban Studies, 56(1), 81–96*.
Patel, P. and K. Pavitt (1994). "National innovation systems: why they are important, and how they might be measured and compared". In: *Economics of innovation and new technology, 3(1), pp.77-95*.
PatentView (2018). "PatentView Database". In: *https://www.patentsview.org/web/viz/comparisonscmp=all/state/numDesc/2018*.
Penrose, E. and E.T. Penrose (2009). "The Theory of the Growth of the Firm". In: *Oxford university press*.
Persson, O., W. Glänzel, and R. Danell (2008). "Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies". In: *Scientometrics, 54, 75–89.*
Petralia, S., P.A. Balland, and A. Morrison (2017). "Climbing the ladder of technological development". In: *Research Policy, 46(5), pp.956-969*.
Pezzoni, M., F. Lissoni, and G. Tarasconi (2014). "How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation". In: *Scientometrics, 101(1), pp.477-504*.
Porter, M.E. (1998). "Clusters and the new economics of competition". In: *(Vol. 76, No. 6, pp. 77-90). Boston: Harvard Business Review*.
Pugliese, E., G. Cimini, A. Patelli, A. Zaccaria, L. Pietronero, and A. Gabrielli (2019). "Unfolding the innovation system for the development of countries: coevolution of Science, Technology and Production". In: *Scientific reports, 9(1), pp.1-12*.

Pugliese, E., A. Zaccaria, and L. Pietronero (2016). "On the convergence of the Fitness-Complexity Algorithm". In: *The European Physical Journal Special Topics, 225(10), pp.1893-1911*.

Reynolds, C., M. Agrawal, I. Lee, C. Zhan, J. Li, P. Taylor, T. Mares, J. Morison, N. Angelakis, and G. Roos (2018). "A sub- national economic complexity analysis of Australia's states and territories". In: *Regional Studies, 52(5), pp.715-726*.

Rigby, D.L. (2015). "Technological relatedness and knowledge space: entry and exit of US cities from patent classes". In: *Regional Studies, 49(11), pp.1922-1937*.

Rodrik, D. (2019). "Where Are We in the Economics of Industrial Policies?" In: *Frontiers of Economics in China, 14(3), pp.329-335*.

Rozenblat, C. and D. Pumain (2005). "Firm linkages, innovation and the evolution of urban systems". In: *Cities in globalization: Practices, policies, theories, pp.130-156*.

Rutherford, J., A. Gillespie, and R. Richardson (2004). "The territoriality of pan-European telecommunications backbone networks". In: *Journal of Urban Technology 11 (3), 1–34*.

Sachini, E., K. Sioumalas-Christodoulou, and C. Chrysomallidis (2021). "COVID-19 enabled co-authoring networks: a country-case analysis". In: *Scientometrics 126, 5225–5244*.

Sasaki, Y. (2007). "The truth of the F-measure". In: *https://www. cs. odu. edu/ mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07*.

Sassen, S. (1991). "The Global City: New York, London, Tokyo." In: *Princeton University Press, Princeton*.

— (2001). "Cities in the global economy". In: *Handbook of urban studies, pp.256-272*.

Saxenian, A. (1994). "Regional networks: industrial adaptation in Silicon Valley and route 128". In: ,

Saxenian, A. and J.Y. Hsu (2001). "The Silicon Valley-Hsinchu connection: Technical communities and industrial upgrading". In: *Industrial and Corporate Change, 10, 893-920*.

Scherer, F.M. (1982). "Inter-industry technology flows in the United States". In: *Research policy, 11(4), pp.227-245*.

Schilling, M.A. and C.C. Phelps (2007). "Interfirm collaboration networks: The impact of large-scale network structure on firm innovation". In: *Management science, 53(7), pp.1113-1126*.

Scholl, T., A. Garas, and F. Schweitzer (2018). "The spatial component of R&D networks". In: *Journal of Evolutionary Economics, 28(2), pp.417-436*.

Sciarra, C., G. Chiarotti, L. Ridolfi, and F. Laio (2020). "Reconciling contrasting views on economic complexity". In: *Nature communications, 11(1), pp.1-10*.
Singh, J. (2005). "Collaborative networks as determinants of knowledge diffusion patterns". In: *Management science, 51(5), pp.756-770*.
Soja, E. (2014). "My Los Angeles: From urban restructuring to regional urbanization". In: *Los Angeles: UCLA Press*.
Song P., Zong X. Chen X. Zhao Q. and L. Guo (2021). "The Shortest Duration Constrained Hidden Markov Model: Data denoise and forecast optimization on the country-product matrix for the Fitness-Complexity Algorithm". In: *PloS one, 16(7), p.e0253845*.
Sonnenwald, D. H. (2007). "Scientific collaboration". In: *Annual review of information science and technology, 41(1), 643–681.*
Storper, M. and A.J. Scott (2009). "Rethinking human capital, creativity and urban growth". In: *Journal of economic geography, 9(2), pp.147-167*.
Straccamore M., Pietronero L. and A. Zaccaria (2021). "Which will be your firm's next technology? Comparison between machine learning and network-based algorithms". In: *arXiv preprint arXiv:2110.02004*.
Sweet, C. M. and D. S. Eterovic Maggio (2015). "Do stronger intellectual property rights increase innovation?" In: *World Dev. 66, 665–677*.
Tacchella, A., M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero (2012). "A new metrics for countries' fitness and products' complexity". In: *Scientific reports, 2(1), pp.1-7*.
Tacchella, A., A. Zaccaria, M. Miccheli, and L. Pietronero (2021). "Relatedness in the Era of Machine Learning". In: *arXiv preprint arXiv:2103.06017*.
Taylor, P. J. (2001). "Specification of the world city network". In: *Geographical Analysis, 33(2), 181–194*.
Taylor, P.J., B. Derudder, P. Saey, and F. (Eds.) Witlox (2007). "Cities in Globalization: Practices, Policies, Theories". In: *Routledge, London*.
Valero, A. and J. Van Reenen (2019). "The economic impact of universities: Evidence from across the globe". In: *Economics of Education Review, 68, pp.53-67.*
Ventura, S.L., R. Nugent, and E.R. Fuchs (2015). "Seeing the non-stars:(some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records". In: *Research Policy, 44(9), pp.1672-1701*.
Verginer, L. and M. Riccaboni (2020). "Talent goes to global cities: The world network of scientists' mobility". In: *Research Policy, 50(1), p.104127*.
— (2021). "alent goes to global cities: The world network of scientists' mobility". In: *Research policy, 50(1), p.104127*.

Vlčková, J., N. Kaspříková, and M. Vlčková (2018). "Technological relatedness, knowledge space and smart specialisation: The case of Germany". In: *Moravian Geographical Reports, 26(2), pp.95-108*.

Wagner, C.S., H.W. Park, and L. Leydesdorff (2015). "The continuing growth of global cooperation networks in research: A conundrum for national governments". In: *PLOS ONE, 10(7), e0131816.*

Wagner, C.S., T.A. Whetsell, and L. Leydesdorff (2017). "Growth of international collaboration in science: Revisiting six specialties". In: *Scientometrics, 110(3), pp.1633-1652.*

WEF (2014). "The Competitiveness of Cities: A report on the Global Agenda Council on Competitiveness". In: *Geneva: World Economic Forum*.

Wu, L., H. Zhu, H. Chen, and M.C. Roco (2019). "Comparing nanotechnology landscapes in the US and China: a patent analysis perspective". In: *Journal of Nanoparticle Research, 21(8), pp.1-20.*

Zaccaria, A., M. Cristelli, A. Tacchella, and L. Pietronero (2014). "How the taxonomy of products drives the economic development of countries". In: *PloS one, 9(12), p.e113770*.

Zaccaria, A., S. Mishra, M.Z. Cader, and L. Pietronero (2018). "Integrating services in the economic fitness approach". In: *World Bank Policy Research Working Paper, (8485)*.

Zhang F., Wang Y. and W. Liu (2020). "Science and technology resource allocation, spatial association, and regional innovation". In: *Sustainability, 12(2), p.694*.

Zheng, S., W. Sun, J. Wu, and M. E. Kahn (2016). "Urban Agglomeration and Local Economic Growth in China: The Role of New Industrial Parks". In: *Social Science Research Network (SSRN): https://papers.ssrn.com /abstract=2746711*.