UCD CENTRE FOR ECONOMIC RESEARCH

WORKING PAPER SERIES

2023

Construction of a Global Knowledge Input-Output Table

Ronald B. Davies Dieter F. Kogler Guohao Yang
University College Dublin

WP23/22

October 2023

UCD SCHOOL OF ECONOMICS
UNIVERSITY COLLEGE DUBLIN
BELFIELD
DUBLIN 4

Construction of a Global Knowledge Input-Output Table*

Ronald B. Davies† Dieter F. Kogler ‡ Guohao Yang §

October 13, 2023

Abstract

This paper describes the construction of the Knowledge Input-Output (KIO) table constructed as part of the RETHINK project. Using PATSTAT data on forty years of patent data from across the globae, the KIO table provides information on the number of patent applications across ten major patenting countries and the rest of the world and across 131 technology classifications. It further provides a network of patent citations, thus indicating how patents build from existing knowledge and contribute to the construction of further innovation. In addition to describing the KIO's construction, we provide a number of stylized facts on patenting activity and the citation network. These facts illustrate the lessons that can be learned from patent citation data while also identifying potential pitfalls in their use.

JEL classification: O31, O33, R15.

Keywords: Patents; Citations; Input-Output Table; Diffusion; Innovation.

*This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061123. We thank Johannes Scheuerer for useful comments. All errors are our own.

†Corresponding author. University College Dublin. Email: ronald.davies@ucd.ie

‡University College Dublin. Email: dieter.kogler@ucd.ie

§University College Dublin. Email: guohao.yang@ucd.ie

1 Introduction

Knowledge rarely arises from a vacuum, rather it builds on what has come before. From oral histories through written word to digital repositories, innovation has always learned from the existing stock of knowledge. In this fashion, each idea borrows from those that came before it and contributes to the creation of new ideas. Thus, as with traditional input-output tables where production in one industry or country both uses inputs from others and provides inputs for further production, one can conceive of a knowledge input-output table in which ideas are linked both to their predecessors and their progeny.[1] In this paper, we describe the construction of such a Knowledge Input-Output (KIO) table that is built from patents available in the PATSTAT database. Using these data, we provide a publicly available KIO that covers the ten major innovating countries (as well as the rest of the world) across four decades with patents decomposed into 131 disaggregated technological classes. This includes both the amount of patenting activity (both patent applications and granted patents) as well as the strength of connections between country-technology-time triads via patent citations. Further, we develop a set of stylized facts on innovation showing that, despite ever increasing collaborations across borders, it remains a fairly siloed activity with connections across patents largely within countries, periods of time, and technologies. Put differently, patents primarily build off of local patents in the same technologies, suggesting that the evidence for both international spillovers and "disruptive" innovations that combine disparate technologies remain fairly limited.

Specifically, the KIO is a citation matrix that links patents from one country-technology-time triad to another. Although there are a number of assumptions that must be made during the process, such as whether to use all patent applications or just granted ones, how to allocate a given patent across countries, how to deal with multiple patent filings representing a single innovation (patent families), and more, the end result reports both the level of patenting activity for each triad and the number of citations by triad pair. These citations also have a directional interpretation where a citation between two patents represents a backward citation for the citing patent and a forward citation for the cited patent. Thus, one can interpret the KIO as a weighted directional network where each triad is a node and citations are links between them. This then forms a tool to talk about the size of innovation activity, the intensity of spillovers, and more. Our goal here is to provide such a tool alongside a set of stylized facts regarding the KIO.

Ours is not the only attempt to build such a KIO. For example, Acemoglu, Akcigit, and Kerr (2016) use just granted patents filed with the US Patent Office (USPTO) from 1975-2009. Cai, Li, and Santacreu (2022) also use USPTO data (running from 2001-2010) to construct a KIO for 19 OECD countries and 19 industrial sectors.[2] Liu and Ma (2021), meanwhile, extend their analysis to global patenting activity from 1976-2020 using information from Google Patents.[3] Combining information on a patent's inventor(s), assignee(s), and

[1]We use terms such as knowledge, idea, inventions and innovation somewhat interchangeably. We do, however, acknowledge that these words have different connotations when it comes to patenting. A patent is for an innovation, that is an invention that has generated (monetary) value. An invention is an idea with a potential monetary value.

[2]To operate at the sector level, they use a concordance between technology classifications and industries.

[3]Although PATSTAT has somewhat better coverage, the advantage of Google Patents is that it is free to

the patent office where it was filed, they fractionally apportion each patent across countries and use the earliest filing date to establish the timing of an innovation. Ayerst, et al. (2023) also use Google Patents to construct a KIO across countries and technologies, however their time period is more limited and runs from 1995 to 2015. Furthermore, they only use inventor information in allocating patents to countries. Finally, they limit the time of a link to ten years. Nevertheless, Hall, Jaffe, and Trajtenberg (2001) show that a significant number of citations tend to come after this point, suggesting some long-standing contributions can be lost.

In comparison to these, we use either assignees or inventors in apportioning patents across countries and show that the two are very similar. Further, relative to Ayerst, et al. (2023), we use a time horizon that is twice as long. Finally, the KIO presented here focuses on the ten major patenting countries (based on the total number of patent filings) in order to most clearly discuss the connections across countries. Beyond alternative data assumptions, unlike them, we offer additional detail on citations across a variety of disaggregations, thereby providing important information for understanding the patterns that can be drawn from our – or any – KIO. For example, the issue of the timing of citation arrivals and what this means for both beginning- and end-of-sample truncation has been largely overlooked. Further, we point out significant country variation in data quality – for example patents from China and Japan are often missing technology details limiting the number of their patents that can be used in a KIO. Similarly, they tend to have fewer – or even zero – backward citations, again suggestive of omissions in the data. As such, their role in the international diffusion of knowledge may be systematically understated. This is important to recognize in the analyses of, for example, Ayerst, et al. (2023) and Liu and Ma (2021), who combine patent citations with international trade data to estimate the role of diffusion in productivity growth. Since patent data for major trading partners such as China and Japan are often incomplete and understate their international connections, this can have an impact when estimating the role international trade as a channel for technological diffusion.

An important admission regarding the KIO is that it draws from PATSTAT which is a database of *patents*, not innovation itself. As described by Hall, et al. (2014) the decision of whether to patent at all is an important one since by filing for protection, a firm both incurs considerable application costs and reveals its proprietary knowledge to potential competitors. Thus, patenting is a proxy for innovation since it only captures the innovation of a subset of firms and even then only for "successful" research projects, i.e. those that generate results deemed worthy of a patent application. Further, patents work as a proxy for the outcome of the innovative process. In contrast, measures such as R&D spending or the number of scientists – inputs to the research process – are alternative measures of innovation. Nevertheless, patent data are perhaps the most widely used innovation proxy because of their availability and granularity (Jaffe and de Rassenfosse, 2017). Further, if the intent is to measure how knowledge builds on knowledge, then the omission of secretive innovations for which no patent emerges may be a fairly minor issue. Thus, given our effort to analyze innovation across countries over a significant time horizon, we therefore use the patenting data while recognizing this caveat.

The rest of the paper proceeds as follows. Section 2 describes the data and methodology

researchers. Liu and Ma (2021) provide an extended comparison, finding fairly few differences.

used to construct the KIO. Section 3 provides some stylized facts on the KIO across time, countries, and technologies. Finally, Section 4 concludes.

2 Data Set Construction

In this section, we describe the construction of the KIO table including the data used.[4] Recall that the KIO is an attempt to measure the generation of knowledge within a location, technology, and time as well as capture the ways in which one set of knowledge builds from and feeds into other sets of knowledge. With this objective in mind, the KIO described here measures P_{abc}, that is, the number of patents filed in a given country a, in technology b, during time c. Further, it measures the number of citations ($N_{abc,xyz}$) made by P_{abc} (the citing patents) from the patents created in country x, in technology y, in period z. This then represents how innovations both build off of others and contribute to the creation of more innovations. To be clear on definitions, $N_{abc,xyz}$ is the number of backward citations for abc and forward citations for xyz. Although the meaning of the entries in the KIO is fairly straightforward, there are several features of the PATSTAT data that must be understood in order to appreciate what the KIO captures (and what it does not).

2.1 Data Selection

The first question is whether to use only granted patents or all patent applications. As discussed by Davies, Hynes, and Kogler (2021), the patenting process is a highly uncertain one with approximately half of applications being granted (something found in our data as well). While the argument can be made that granted patents are "better" than unsuccessful patent applications and therefore only granted patents should be used, two significant issues arise. First, there is a well-documented home bias in granting rates with applications arising from within the jurisdictions covered by a patent office significantly more likely to be granted (see Drivas and Kaplanis (2020), Webster, et al. (2014), Webster, et al. (2007), Guellec and van Pottelsburghe (2000), among others). Thus, to use only granted patents can lead to important, if foreign, innovations being overlooked. This is particularly worrisome in light of our effort to examine cross-border knowledge spillovers. Further, as documented in Davies, Hynes, and Kogler (2021), the patenting process is a long one, with the average granted patent requiring more than five years for approval. Therefore, restricting the KIO to granted patents would introduce a significant truncation issue for more recent years. Third, even if a given patent application is not itself granted, this does not mean it cannot influence other innovations. Indeed, even ungranted patents are cited, especially during the early years post-submission. Therefore, we use all patent applications, regardless of whether or not they have been granted, in the KIO. For ease of exposition, from this point forward we use the word "patent" to cover both granted patentand ungranted patent applications. That said, in the KIO, we provide data on both the total number of patents (as in Ayerst, et al., 2023) and just those which were granted (as in Acemoglu, Akcigit, and Kerr, 2016).

These patents are drawn from the Autumn 2022 release of the PATSTAT database.[5]

[4] Replication code for construction of the KIO is available on request.

[5] This can be found at `https://www.epo.org/en/searching-for-patents/business/patstat`.

While PATSTAT contains entries dating back to 1783, reporting of patents submitted prior to 1980 is limited and contains a number of missing fields. As this would preclude them from use, we restrict ourselves to patents submitted from 1980 onwards. As discussed in detail below, this introduces some issues surrounding backward citations at the beginning of the sample, an issue which would also feature in, for example, Cai, Li, and Santacreu (2022) and Ayerst, et al. (2023). Further, PATSTAT often features a delay between the filing of a patent and its entry into the database. As such, there is a decline in the number of patents in the 2022 version of PATSTAT starting in 2020. We therefore restrict the KIO to cover from 1980 to 2019 inclusive, i.e. four decades worth of data. A similar issue would arise in the KIO of Liu and Ma (2021) who use patents into the 2020s. Given the infrequency of patenting, particularly when breaking patents down across technologies, our KIO aggregates yearly information into four decades: the 80s (1980-89), 90s (1990-99), 00s (2000-2009), and 10s (2010-2019). This also reduces the sparsity of the citation matrix.

While on the topic of truncation, it is important to recognize that forward citations for patents closer to the end of the dataset are fewer for the simple reason that more recent patents have not been around long enough to accumulate as many citations. In their examination of US patents, Hall, Jaffe, and Trajtenberg (2001) find that a patent's forward citations tend to peak around five to seven years post-filing.[6] Thus, the KIO will automatically have a tendency to understate forward citations for patents submitted in the 10s. This issue would feature in the KIOs of Ayerst, et al. (2022) and others as well. These restrictions must be noted when attempting to compare the evolution of the KIO across decades.

In addition to restricting the years of patents, we also limit ourselves to filings with the "big five" offices: the European Patent Office (EPO), the US Patent Office (USPTO), the Chinese Patent Office (CPO), the Japanese Patent Office (JPO), and the Korean Patent Office (KPO). This differs from, say, Ayerst, et al. (2023) and Liu and Ma (2021) who use a much larger set of patent offices. Although PATSTAT offers information for over 100 patent offices, reporting from smaller offices is missing essential information for our analysis. In particular, the lag between filing and entry in the database appears to be more severe for smaller offices, meaning that their inclusion exacerbates end-of-sample truncation. Finally, as will be discussed further below, even for the major Asian offices, technology information is missing with a greater frequency than in the US or European data. Thus, the added number of usable patent filings from including those offices is limited. In any case, for our purposes, leaving out these offices is arguably a minor issue for two reasons. First, the countries covered by the big five offices are responsible for the large majority of patenting activity. WIPO (2022) indicate that these offices account for 85% of global patents, meaning that we capture a large share of patents. Second, recall that the purpose of patenting is to protect one's intellectual property. Thus, if a given innovation is valuable – and therefore important when considering linkages in knowledge creation – we would expect that the owners would have sought protection in at least one of these five major markets. This is indeed the argument made by by Coelli, Moxnes, and Ulltveit-Moe (2022) in their study of patenting and exports. Therefore, we proceed using the more reliable data arising from the big five patent offices.

[6]It is worth nothing that in their data, this peak has moved closer to the time of filing, i.e. forward citations fade out sooner.

When geo-locating patents, one can use the assignee (the owner) listed on the patent or the inventors listed in PATSTAT.[7] Each has its advantages. Assignees suggest where planning surrounding the innovation and its economic value may accrue to (although within multinationals there is the potential for patent shifting for tax purposes; see Schwab and Todtenhaupt, 2021). Inventors, on the other hand, point towards where knowledge may be created. If innovation spillovers are localized, as found by Keller (2002), this may then be preferable.[8] Rather than side with one approach or the other, we use both in constructing our KIO, providing data on patenting activity and citations when using assignee countries as well as when using inventor locations. As discussed below, although inventor location suggests more multi-country patents, the overall KIO results are extremely highly correlated across the two. In the KIO produced here, in order to limit the dimensionality of the matrix, we separately record values for the ten most frequently patenting countries and aggregate the remaining countries into a "Rest of the World (ROW)" grouping. As discussed further below, these ten separate countries make up the large bulk of overall patenting activity in our data. Note that other KIOs, such as Cai, Li, and Santacreu (2022) restrict themselves to a small number of countries or, as in Ayerst, et al. (2023), drop infrequent innovators.

Finally, in terms of technology, we use the CPC codes provided by PATSTAT. These technology classifications are assigned by the patent office when an application is submitted for the purpose of searching the prior art to assess an application's novelty and thus whether it should be granted. Note that as CPC codes are updated, PATSTAT retroactively updates its information using its own concordance, ensuring that CPC codes are consistent across time. Note that we do not include the "Y" subset of codes in the KIO's construction as this miscellaneous category is largely used to identify green technologies regardless of their underlying technologies. As such, it is not a good indicator of the co-occurrence of technologies within patents. While the CPC codes reported in PATSTAT are quite disaggregated, to reduce dimensionality of the KIO, we operate at the three-digit CPC level of which there are 131. Therefore, our KIO represents 11 countries, 131 technologies, and 4 decades, for a total of 33,223,696 elements. This figure indicates the importance of our aggregations since it grows exponentially as the number of countries, technologies, or time periods increases.

2.2 Patent Families

A complicating feature of patent data is that one innovation can lead to multiple patent filings. This can occur when patent applications are made to multiple offices in order to seek protection across multiple jurisdictions, when revised versions of a failed application are submitted, and/or when applications are made to extend an existing patent. As such, were we to use each of these patents it would artificially inflate the amount of innovation occurring as well as the number of citations between innovations. To deal with this, we take advantage of the family identifier provided by PATSTAT which links individual patents which all derive from the same innovation.

[7] Or as per Liu and Ma (2021), a combination of these along with the patent office they are filed with.

[8] It should be acknowledged, however, that his estimates suggest a strong local component of spillovers in R&D spending, suggesting limited scope for international effects. As shown below, this mirrors our finding that the bulk of citations are within country.

This then introduces two issues. The first regards the allocation of a given innovation across countries, technologies, and time. If each patent p within a family f had the same set of assignees/inventors, technology, and filing dates, then deciding what to include is straightforward and would be identical to randomly choosing a family member. This, however, is not always the case. For example, filing dates often differ across family members as the owner of the patent completes paperwork for different offices. Likewise, reapplication of a rejected patent can generate a subsequent filing date. To deal with this, we use the earliest filing date within the family to establish the year of the innovation. This mirrors the approach of, for example, Ayerst, et al. (2023).

Additionally, although infrequent, family members can differ in assignees/inventors and/or technologies. When filings are across multiple offices, this can potentially happen when assignees/inventors are added to mitigate home bias or when different patent offices assign different technologies to the patent. Similarly, subsequent filings stemming from, say, reapplications may add assignees or technologies. Finally, there is always the possibility of data entry errors when creating PATSTAT. One way of dealing with this would be to choose a single patent to represent the entire family. This, however, runs the risk of omitting key information. For example, if a given technology is listed on all but one family member, then it may be missed when randomly choosing. Likewise, if all technology information is missing for the chosen member, then the patent would be omitted from the KIO entirely. If such omissions are common for particular countries (as is true with China and Japan as described below), this runs the risk of under-representing innovations from those locations. With this in mind, we use all assignees/inventors and technologies included across all family members in the fractional apportionment process described momentarily.[9] Note that for individuals listed only on a subset of family members, this ensures they receive positive, if discounted, representation. In any case, in an alternative approach we used a random selection from the family members sharing the earliest filing date rather than all family members.[10] The resulting KIO was virtually identical to the one presented.

The second issue with multiple filings regards citations. This has two implications. First, different family members can share a given backward citation, i.e. they all cite the patent. More generally, there are cases where differing members of one family cite different members of another. Such instances can occur when, for example, the EPO member of family 1 cites the EPO member of family 2 whereas the USPTO member of family 1 cites the USPTO member of family 2. In such cases, there are multiple citations across families even though there is just one actual connection between them. Counting each of these separate citations would exaggerate the number of forward/backward citations (and thus importance in the citation network). With these issues in mind, we count at most one citation per family pair.[11] Furthermore, we drop within-family citations from the sample. The second citation possibility with multiple family members is within-family citations as can occur if a reapplication includes a citation of an earlier, failed version. Because these self-citations violate

[9]Note that if a patent is sold to a new assignee, that assignee is only included if it is listed on a patent within the family (as can occur if it purchases a innovation with a failed application and then makes a new one).

[10]12.7% of families have multiple patents which share this date.

[11]It is somewhat unclear how this has been done in the construction of other KIOs.

the spirit of our notion of knowledge spillovers, we omit these when constructing the KIO.[12]

Thus, our KIO is built from patent families, where the timing of the family is the earliest patent filing across members and we use all assignees, inventors, and technologies in apportioning the family across countries and technologies. To simplify exposition, from this point forward, when we use the word "patent" we are speaking about the family to which it belongs.

Although the timing of a patent c is set by the earliest filing date within the family, meaning that it is attributed to a single year, this is not the case for location or technology. We therefore follow the practice of fractional apportionment in which a given patent $p \in f$ is allocated across countries according to the share of assignees/inventors from that country. Specifically, the share of a patent p attributed to location a is $\frac{l_a}{\sum_i l_i}$ where l_i is the total number of either assignees or inventors from country i reported across members of family f. Likewise, where j_b is the number of most disaggregated CPC codes in three-digit CPC b, the share of the patent attributed to technology b is $\frac{j_b}{\sum_k j_k}$. Therefore, each patent p has a share $s^p_{ab} = \frac{l_a}{\sum_i l_i} \frac{j_b}{\sum_k j_k}$ allocated to country-CPC ab. Similarly, a citation from p_{abc} to q_{xyz} is allocated across ab, xy according to $s^p_{ab} s^q_{xy}$. To arrive at the entries for the KIO, we then sum up across patents within each country-CPC-decade triad.

2.3 Data Specifics

In this section, we provide details on the KIO construction at each stage of the process. This is intended to provide as much insight into the data which underpins the KIO. For ease of exposition, we focus on the KIO using assignee location, followed by a comparison of this to the inventor location version.

As noted above, we use PATSTAT data from the five major patent offices running from 1980-2019 which covers 48,132,094 patents (granted and otherwise). Some of these are multi-filings so that we are in practice operating with 36,277,112 unique patents/families. For 749,503 families, no filing date information is available for any family member. These are therefore dropped from the sample. Turning to the location data, we have at least one assignee with a reported country for 32,797,618 families. Those without any such information are dropped from the sample. Note that when using assignee, cross-country families are rare with only 683,742 families crossing borders. To place families in technological space, we require at least one CPC code (outside of the "Y" category) for some member of the family. Relative to the location data, CPC information is more scarce, with only 21,570,304 families having a technology code listed. This difference is primarily driven by those families located in Japan (roughly 67% of missing CPC records) and China (approximately 29%). Of patents where technology codes are available, 14,907,133 are contained within a single three-digit CPC code.

Thus, combining the data leaves us with 20,054,614 patents with country, technology, and time information. Of these, approximately 53.0% were granted.[13] By way of comparison to other KIOs, Ayerst, et al. (2023) report they have the necessary country, technology, and time data for 18.9 million patent families. Liu and Ma (2021) find approximately 11.7 million

[12]To our knowledge this was not done in the other KIOs.

[13]That is, 53.0% of families contain a member which is a granted patent by one of the big five offices.

patent families in Google Patents from 1985-2014, although it is not clear how many of these have technological information. Finally, Acemoglu, Akigit, and Kerr (2016) and Cai, Li, and Santacreu (2022), who only use filings with the USPTO, have on the order of 1.8 million patents each. Thus, despite the fact that we focus on just the five major patent offices, our number of usable patents exceeds those used elsewhere.

Turning to citations, we begin with 245,896,329 citations. However, these include both duplicate citations across families and self-citations within families.[14] Dropping these leaves us with 162,176,770 unique citations across families. Merging this with our patent information reduces this to a total of 116,953,488 citations for which we have all citing and cited family information. Although the other KIOs do not generally report the number of citations for which they have the necessary data, Cai, Li, and Santacreu (2022) indicate that they have approximately 10 million usable citations.

Within our data we have full information for 6,508,491 patents who make no listed citations. One could take this to mean that these families cite nothing. In practice, however, it is more likely that the patents they actually cite are not in the PATSTAT database. Over half of these are due to patents from China. Japan and Korea making up another 35.8% of them. This again points out that even among the main patent offices, incomplete entries happen. It is also worth noting that the issue of no backward citations is slightly higher for patents filed in the early 80s, again suggestive of the start-of-sample truncation issue discussed above. While we leave these patents in the totals describing the volume of innovation, they do not add backward citations to the KIO.

Across patents, there is significant variation in citing behaviour. First, Figure 1 illustrates the distribution for backward citations.[15] As noted above, there are a fair number of patents that list no backward citations. Beyond that, there is a long tail with some patents citing a great number of others. Similarly, Figure 2 shows the number of forward patents received by a patent for which we have all necessary information.[16] Here, two patterns emerge. First, there is a significant spike at zero citations. Unlike backward citations, this is due both to the fact that some patents simply do not get cited and an end-of-sample truncation. Second, there is a spike again at the end of the distribution. Together, these mean that many patents do not get cited, most that are cited receive only a few citations, and a handful of patents are highly influential.

Thus, the KIO provides data on the number of patents (both all and just granted) for 11 regions, 131 technologies, for 4 decades, or 5,764 country-CPC-decade triads. Further, it provides cross-family citation counts for the 33,223,696 country-CPC-decade pairs. This number of cells illustrates why we chose to aggregate less innovative countries into the RoW category in order to reduce the dimensionality of the KIO. Finally, even a casual examination of the KIO shows that it is quite sparse. Although only 3.6% of country-technology-decades report no patenting at all, 85.9% of citation cells are zero.

The above discussion focused on the case where assignee location was used in allocating patents to countries. Alternatively, we can use inventor location. Doing so leaves us with 20,204,938 patents with all required information of which 52.8% were granted. Likewise, this

[14]Self-citations account for 1,267,075 citations.

[15]Note that while this requires all information for the citing patent, it does not do so for the cited patent which is obvious when there are no forward citations.

[16]Again, this does not require that all identifying information exist for the citing patent.

allows us to use 117,601,976 citations. Overall, the two versions of the KIO are extremely similar, with a correlation of individual variables across them exceeding 0.999. The only notable difference is that the inventor approach yields far more cross-border families with 2,325,562 having inventors from multiple countries. The primary feature this influences is the sparsity of the matrix since more countries have an inventor (if not an assignee) in a given technology. Nevertheless, as this is a minority of patents and even in most of those the majority of inventors come from one country, even this does not greatly affect the KIO.

3 Stylized Facts

In this section, our goal is to provide some basic insights into the KIO.[17] As the purpose of freely providing the KIO to other researchers as part of the RETHINK project, we do not attempt to provide a thorough discussion of every facet. Instead, here we discuss the patterns within the KIO in broad strokes both to highlight features of its construction and identify stylized facts, both of which can form a springboard for future research. Finally, we wish to be clear that this is a descriptive exercise. We are not making any claims on the underlying drivers of these patterns. Indeed, this is the type of future research we hope the KIO can support.

In what follows, two basic approaches are taken. One seeks to provide summary statistics on the number of patents and citations. The other exploits the bilateral nature of the citation data. Fundamentally, the KIO acts as an adjacency matrix for a weighted, directional graph, that is, it describes a network where each node is a country-technology-decade and the citations from a citing node to a cited node act as the weighted, directional edge between them. For more background on the use of networks in Economics, we point the reader to the excellent contributions of Stachurski and Sargent (2022) and Jackson (2010).

3.1 Time

We begin our discussion by aggregating across countries and technologies to consider just the time dimension of the KIO. Figure 3 illustrates the number of patents – both all and only granted ones – by year. While the number of patents gradually increased up to 2015, there was a significant jump around that time. This is in large part driven by a dramatic increase in Chinese patents found in PATSTAT. After 2020, however, there is a marked reversal driven by the end-of-sample truncation in PATSTAT. This same pattern holds when looking just at granted patents, although the end-of-sample truncation begins somewhat earlier due to the time it takes for an application to be granted (if it ever is).

In Figure 4, we turn to the time trends in citations where the year in the figure corresponds to the year of the citing patent for backward citations and the year of the cited patent for forward citations. In line with the rise in the number of patents, there is a rise in backward citations over time. This does, however, fall off at the end of the sample, potentially due to end-of-sample truncation where some cited patents do not yet appear in PATSTAT. Forward citations, meanwhile, peak in 2001 and then decline even as the number of patents grows.

[17] Unless noted otherwise, all discussion in this section is based on the assignee version of the KIO. Given the similarities, the results from the inventor version are essentially identical.

This has two driving forces. First, it reflects the analysis of US patents by Hall, Jaffe, and Trajtenberg (2001) which found that the bulk of forward citations come some years after the patents filing. Second, even if those patents are being cited, it may be that the citing patents do not yet appear in PATSTAT.

In order to put these absolute changes into perspective, Figure 5 plots the average number of backward and forward citations per year.[18] Beginning with the backward citations, we see an increase in the number of citations per patent up to 2005. This is the result of the fact that only patents from 1980 onwards are used in the KIO, meaning that early patents may have missing backward citations (or their citations suffer from incomplete data). From 2005 to 2015, the average number of backward citations holds constant before starting to decline. This latter decline is the result of two things. First, it come from the same end-of-sample truncation found in Figure 3. Second, there an is increase in the number of Chinese and Japanese patents which tend to have missing technology information, meaning that they enter the denominator without contributing to the numerator of this average. Turning to the forward citations, we see an increase in the average, reaching a peak in the mid-1990s. After this, the average number of forward citations declines. This is has three causes. First, there is again the Hall, Jaffee, and Trajtenberg (2001) observation that it takes time before a given patent tends to be the bulk of its citations. Second, end-of-sample truncation means that some patents citing those in the KIO have yet to enter PATSTAT. Finally, just as Asian patents report fewer citations, they are themselves less likely to be cited, so that their increased presence pulls this average down.

As noted above, the bilateral citation data create an adjacency matrix. In Figure 6 we illustrate the citation network when aggregating to the decade level. In this, the relative size of the four nodes represents the number of patents from that decade. The colour, meanwhile, represents the relative share of forward citations received within the KIO. The relative width of the edges illustrates the share of forward citations made by the originating node (with the arrow pointing towards the cited node).

This figure illustrates three things. First, as one would expect based on Figure 4, although the number of patents increases as one moves forward in time, the share of forward citations is greatest in the 00s. This is potentially because, although numerous, patents from the 10s have yet to reach their full bloom of citations. Second, within-decade citations make up 42.2% of all citations. Third, as sounds obvious, cited decades tend to come after citing decades. This seems logical since one can only cite what already exists. That said, there are a small number of citations where patents cite something from a future decade. These, however, are infrequent and arise from three situations. First, patents that are in practice concurrent may have different filing dates due to variations in patent office processing speeds (i.e. both innovations were created around the same time but the citing application was processed before the new year whereas the cited patent was processed when the patent office returned to work). This can lead their reported dates to differ. Another cause of such "future citations" is our use of families. When later patents in the family cite patents that the earliest one did not, such a future citation can occur. That said, others are almost certainly entry errors in PATSTAT (such as the one patent that cited another which did not

[18]Note that these use all patents in the denominator, regardless of whether they themselves cite or are cited.

come out for another 39 years). Nevertheless, the tendency to cite only concurrent and past patents contributes to the overall sparsity of the KIO.

3.2 Countries

Next we aggregate across time and technologies and focus exclusively on the country dimension of the KIO. In Table 3 we provide totals on the number of patents (both all and just granted ones), the number of backward and forward citations, and the average number of citations each country's patents receive. Several lessons can be learned from this table.

First, China makes up the largest share of patents, followed by the US, Japan, and Korea. This is not true, however, when considering just granted patents where China falls to a virtual tie with Korea. Second, the US is an outlier in terms of the number of citations, both forward and backward, a pattern also noted by Ayerst, et al. (2023). Looking to the average number of citations, we see potential cause of this – American cite often and, as shown in a moment, tend to cite themselves. This, combined with the large number of US patents results in a large number of US citations. The other English-speaking nations also have rather high average citations. The Asian nations, however, tend to have few citations, particularly the Chinese. Again, part of this is largely driven by missing data. Finally, note that the Rest of the World makes up only 5.2% of patents, 6.3% of forward citations, and 7.5% of backward citations. Thus, while some information is lost by combining other countries into this catch-all category, we retain the bulk of cross-country variation while keeping the KIO a manageable scope.

Another interesting takeaway from Table 3 is the difference between the average forward and backward citations. While some countries such as the US and China have similar averages across the two, the British, Japanese, and to a lesser extent the Canadian patents have more forward citations that backward ones. This suggests that they tend to "contribute" more as inputs in the KIO compared to their outputs. The reverse is true for Switzerland and the RoW.

Figure 7 again uses the bilateral citation information to portray the KIO as a network. As before, node size indicates the relative number of patents and the colour indicates the relative share of forward citations. Edges, meanwhile, indicate the number of citations with the arrow pointing from the cited to the citing node. One feature of this figure is the role of within-country citations. These make up 56.1% of citations. As noted above, this is particularly relevant when considering US citations although this is admittedly somewhat difficult to see.

In order to focus on cross-country citations, Figure 8 omits the within-country citations when illustrating the network. This highlights the role of the US both as an input to other nation's patents and as a country that builds from theirs. In particular, the US-Japanese relationship is sizeable. China, on the other hand, has relatively weak international linkages despite its large number of patents. This suggests that, despite being a significant player in global value chains for goods production, it lags in terms of its significance in the global chain of knowledge production.

3.3 Technologies

In this section, we again aggregate by time and country to look just at the patterns across technological fields. We further aggregate to the one-digit technology code in order to simplify discussion. In Table 4, we list the number of patents, granted patents, forward and backward citations, and the average number of citations per patent. From this, several features are observed. First, there are far more patents in Electricity and Physics, with Operations and Transport in a nearby third place. As might be expected, the same holds for their citation counts. as compared to the other broad classes. Such differences are less pronounced when looking to average citations, although the averages is highly correlated with the total number of patents. Another notable feature of Table 4 is that, unlike when considering citations across countries, there is little difference between the forward and backward averages. This is unsurprising given the large share of within-class citations, meaning that within each technology, most of its forward citations are also its backward citations (since both the citing and cited patent are in the same class).

In Figure 9 we once again present the KIO as a network. To aid in examination, this aggregates the KIO from three to one-digit technology codes. As seems plausible, within-technology citations are the bulk of citations, making up 70.8% of citations. To focus attention on the cross-class citations, Figure 10 drops the within-class citations. Here, the most notable feature is that technology classes G and H rely heavily on one another.

4 Conclusion

As is well accepted, knowledge begets knowledge. To understand this phenomenon, and its relationship to global value chains for goods and other economic outcomes, it is necessary to develop a tool to describe the inter-relations of innovation across time, borders, and technological classes. In this paper, we describe one such method – the use of patent and citation data to create a Knowledge Input-Output (KIO) table. We do so for forty years of data on global patenting activity, providing details for the ten most active innovators from 1980-2019 across 131 technological classes. We also provide some stylized facts regarding innovative activity and knowledge flows as embodied by patent data. It is our hope that this KIO works in much the same way as the phenomenon it studies, namely that it builds on the existing work on innovation and serves as a springboard for further research.

References

Acemoglu, D., U. Akcigit, and W.R. Kerr (2016). Innovation Network. Proceedings of the National Academy of Sciences of the United States of America, 113 (41) 11483-11488.

Ayerst, S., F. Ibrahim, G. MacKenzie, and S. Rachapalli (2023). Trade and Diffusion of Embodied Technology: An Empirical Analysis. Journal of Monetary Economics, 137, 128-145.

Cai, J., N. Li, and A.M. Santacreu (2022) Knowledge diffusion, trade and innovation across countries and sectors. American Economic Journal: Macroeconomics, 14(1), 104–145.

Coelli, F., A. Moxnes, and K. Ulltveit-Moe. (2022). Better, Faster, Stronger: Global Innovation and Trade Liberalization. The Review of Economics and Statistics, 104(2), 205-216.

Davies, R.B., R. Hynes, and D.F. Kogler. 2021. Patent Boxes and the Success Rate of Applications. Mimeo.

Drivas, K. and I. Kaplanis (2020). The Role of International Collaborations in Securing the Patent Grant. MPRA Paper No. 99520.

Guellec, D. and B. van Pottelsberghe (2000). Applications, Grants and the Value of Patent. Economics Letters, 69, 109-114.

Hall, B.H., C. Helmers, M. Rogers and V. Sena (2014). The Choice between Formal and Informal Intellectual Property: a Review. Journal of Economic Literature, 52, 375-423.

Hall, B.H., A. Jaffe, and M. Trajtenberg (2001). The NBER Patent Citations Data File: Lessons, Insights, and Methodolgogical Tools. NBER Working Paper No. 8498.

Jackson, M. (2010) Social and Economic Networks. Princeton: Princeton University Press.

Jaffe, A. and G. de Rassenfosse (2017). Patent Citation Data in Social Science Research: Overview and Best Practices. Journal of the Association of Information Science and Technology, 68(6), 1360-1374.

Keller, W. (2002). Geographic Localization of International Technology Diffusion. American Economic Review, 92(1), 120-142.

Liu, E. and S. Ma (2021) Innovation Networks and R&D Allocation. NBER Working Paper No. 29607.s

Schwab, T. and Todtenhaupt, M. (2021). Thinking Outside the Box: The Cross-Border Effect of Tax Cuts on R&D. Journal of Public Economics, 204.

Stachurski, J. and T. Sargent (2022). Economic Networks: Theory and Computation. `https://doi.org/10.48550/arXiv.2203.11972`.

Webster, E., P.H. Jensen, and A. Palangkaraya (2014). Patent Examination Outcomes and the National Treatment Principle. Rand Journal of Economics, 45(2), 449-469.

Webster, E., A. Palangkaraya, and P.H. Jensen (2007). Characteristics of International Patent Application Outcomes. Economics Letters, 95(3), 362-368.

World Intellectual Property Organization. 2022. WIPO IP Facts and Figures 2022. Geneva: World Intellectual Property Organization.

Figure 1: Distribution of Backward Citations

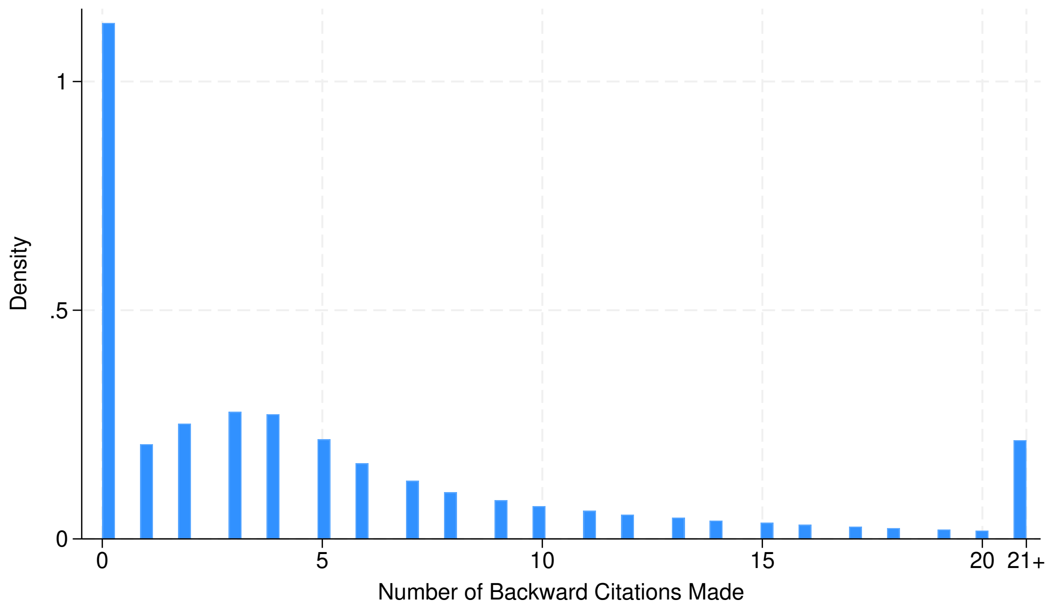


Figure 2: Distribution of Forward Citations

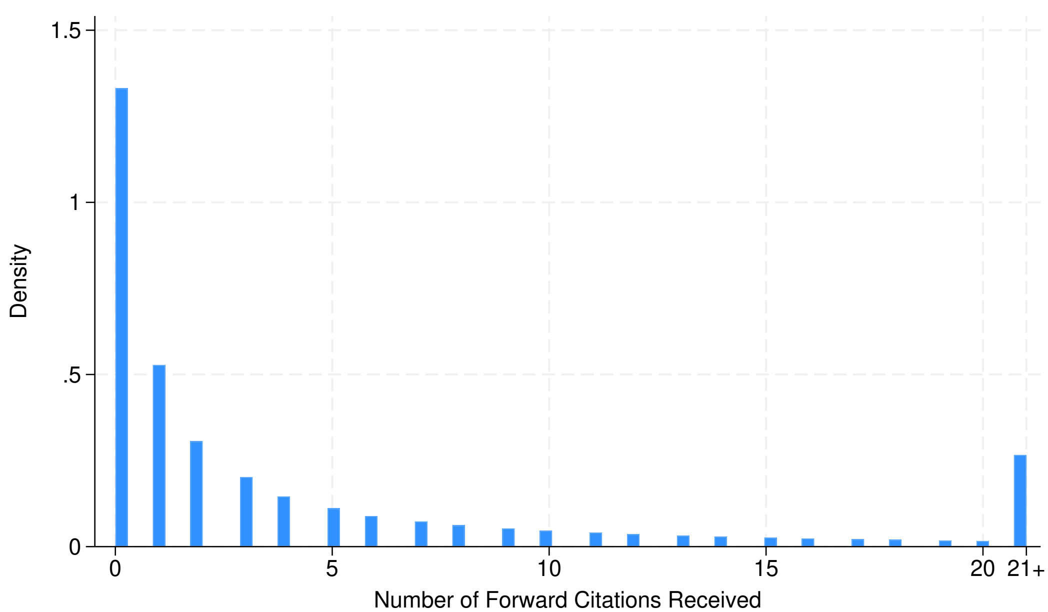


Figure 3: Patents Filed by Year

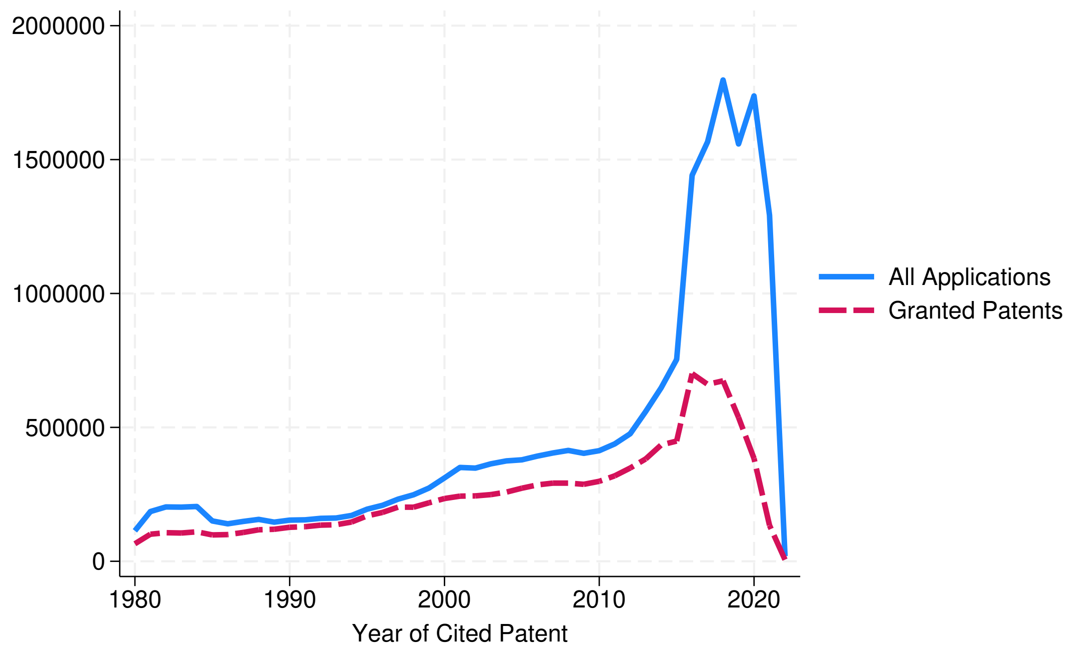
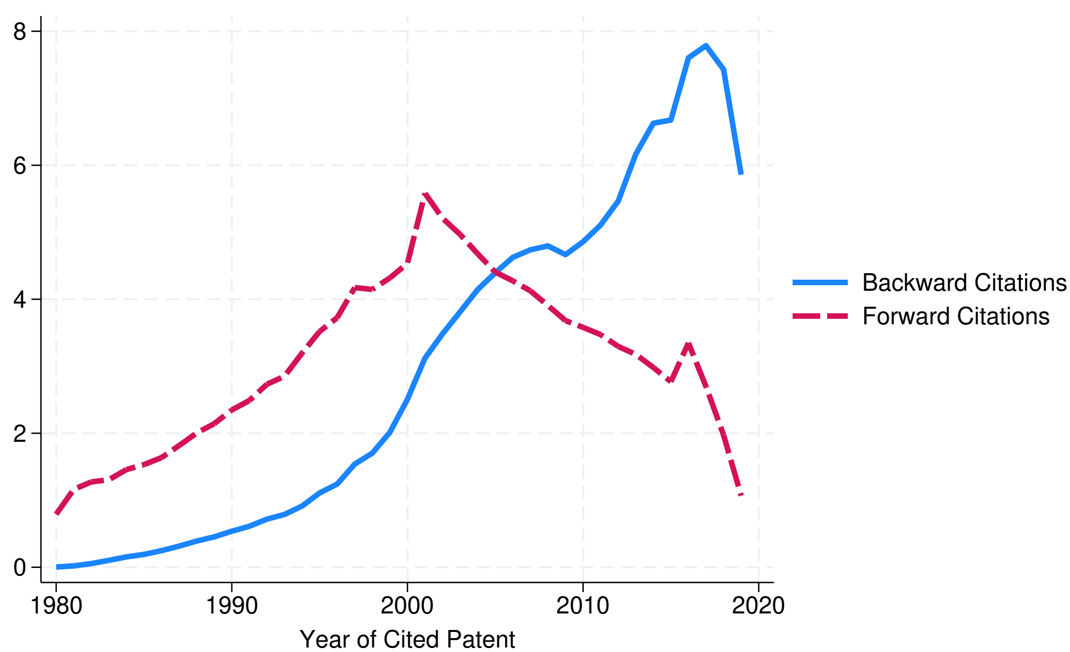
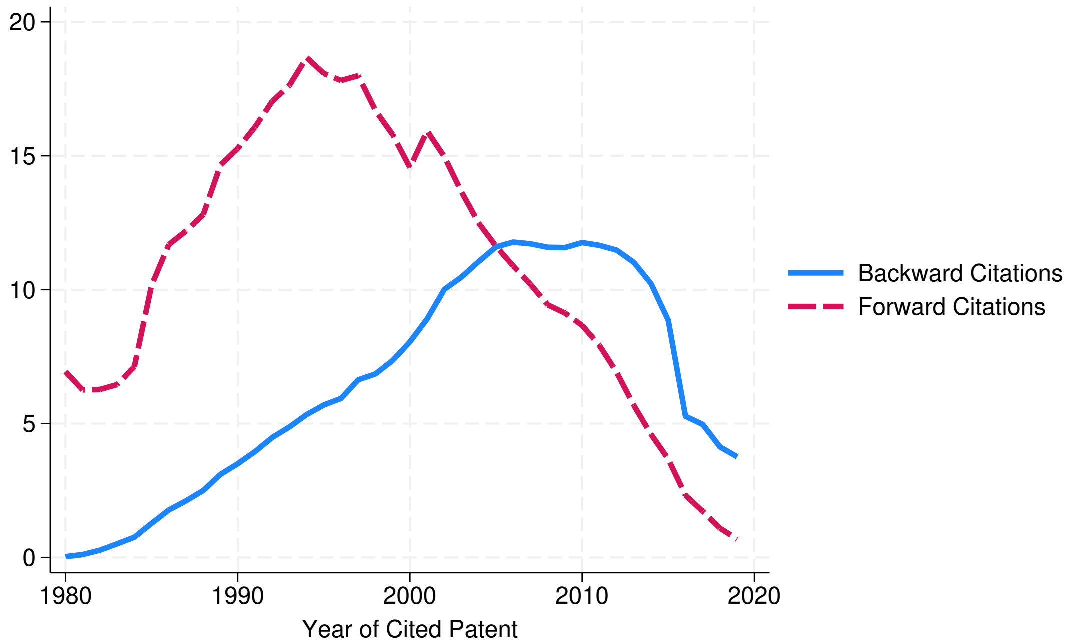


Figure 4: Citations by Year



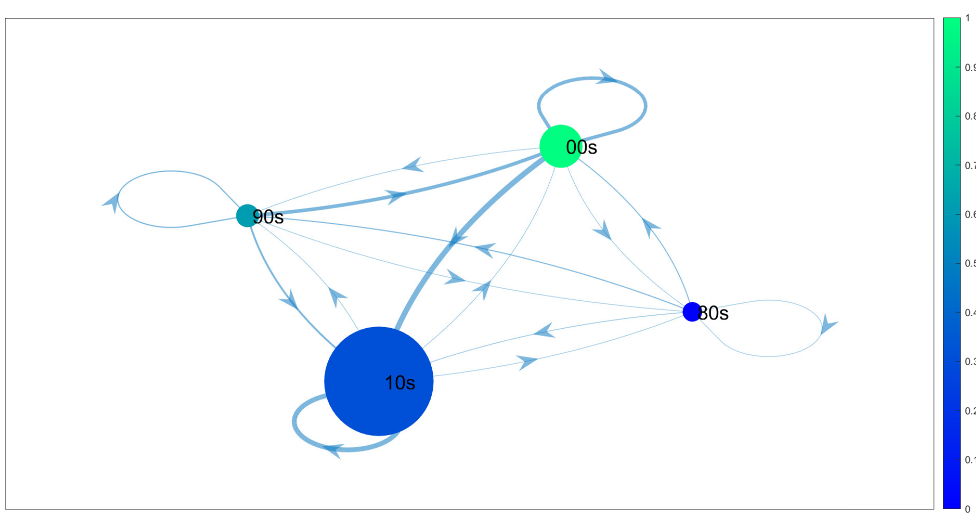
Notes: Year corresponds to year of filing for citing/cited patent.

Figure 5: Average Citations by Year



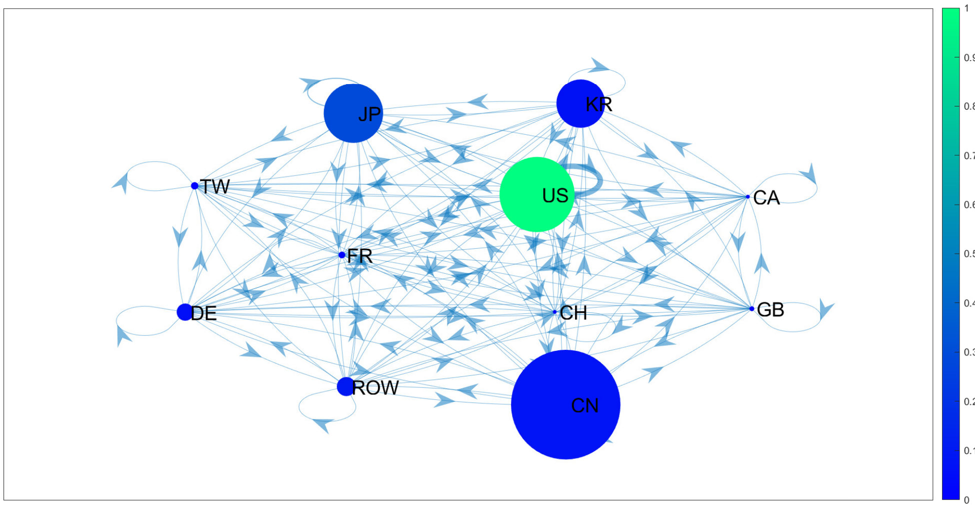
Notes: Year corresponds to year of filing for citing/cited patent.

Figure 6: Citation Network Across Decades



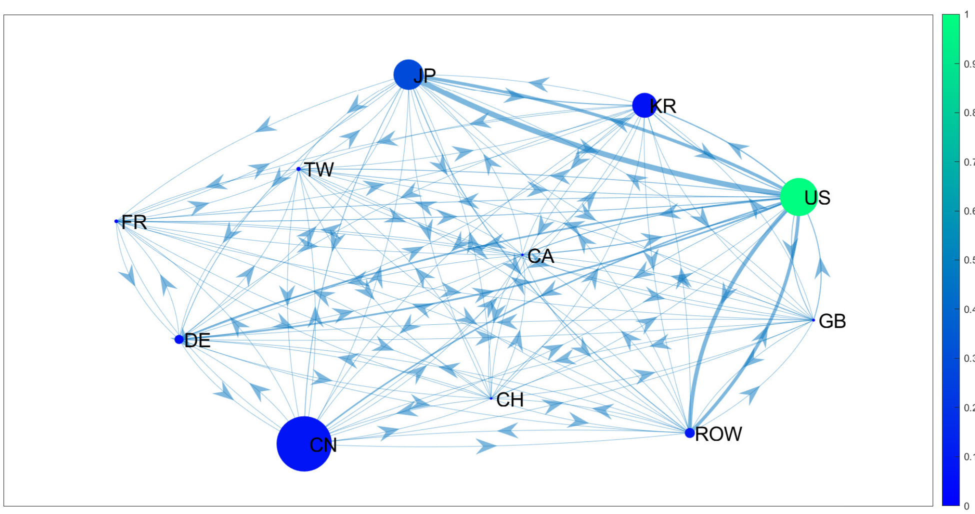
Notes: Size and colour of nodes indicate the number of citing patents. Size of edges indicate number of citations. Arrows run from the citing node to the cited node.

Figure 7: Citation Network Across Countries



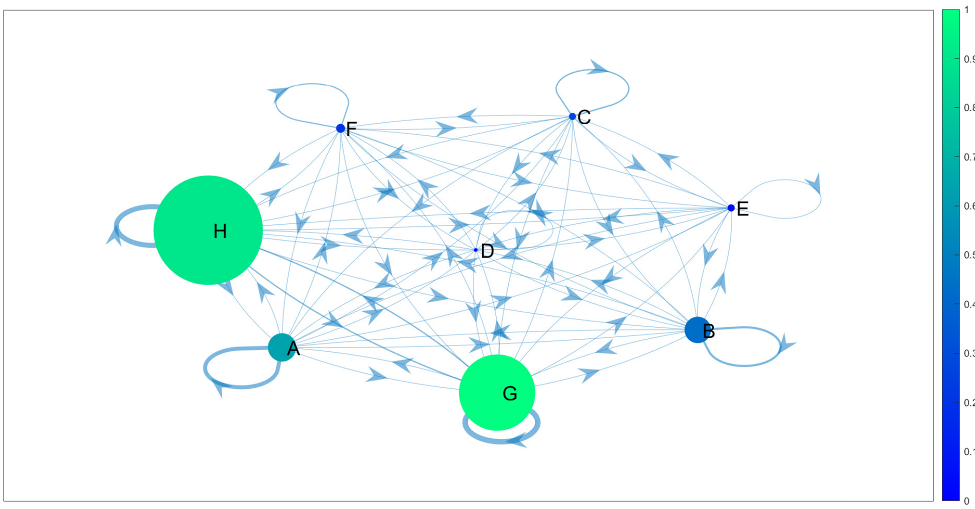
Notes: Size and colour of nodes indicate the number of citing patents. Size of edges indicate number of citations. Arrows run from the citing node to the cited node.

Figure 8: Citation Network Across Countries Omitting Within-Country Citations



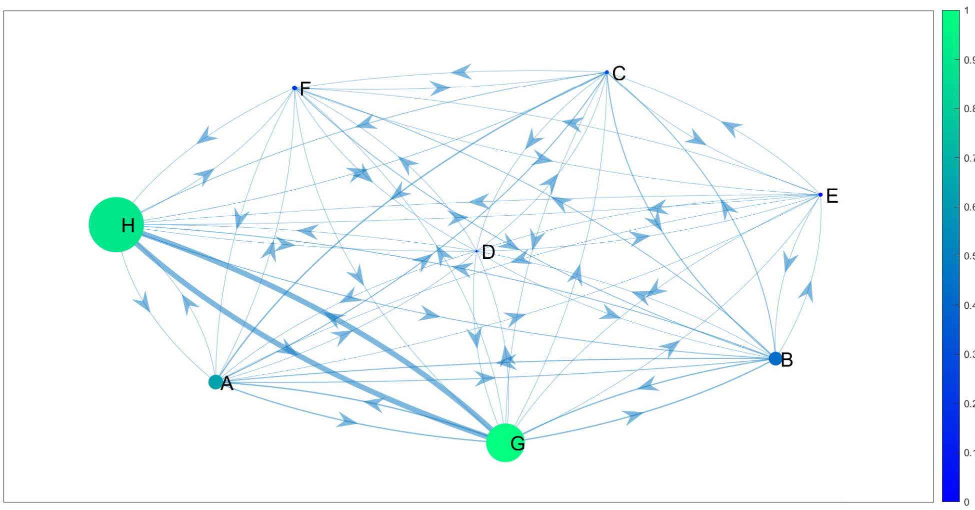
Notes: Size and colour of nodes indicate the number of citing patents. Size of edges indicate number of citations. Arrows run from the citing node to the cited node.

Figure 9: Citation Network Across Technology Classes



Notes: Size and colour of nodes indicate the number of citing patents. Size of edges indicate number of citations. Arrows run from the citing node to the cited node.

Figure 10: Citation Network Across Technology Classes Omitting Within-Class Citations



Notes: Size and colour of nodes indicate the number of citing patents. Size of edges indicate number of citations. Arrows run from the citing node to the cited node.

Table 1: KIO Variable Definitions

Variable	Definition
cited_country	Cited Country.
cited_cpc3	Cited three digit CPC code.
cited_decade	Cited Decade; 1980 indicates patents from 180-1989, etc.
citing_country	Citing Country.
citing_cpc3	Citing three digit CPC code.
citing_decade	Citing Decade; 1980 indicates patents from 180-1989, etc.
number_of_cites_inventor	Number of backward citations made by citing country-cpc-decade; number of forward citations received by cited country-cpc-decade. Country allocation made based on inventor locations.
number_of_cites_owner	Number of backward citations made by citing country-cpc-decade; number of forward citations received by cited country-cpc-decade. Country allocation made based on assignee locations.
cited_total_app_inventor	Total number of patents in cited country-cpc-decade. Country allocation made based on inventor locations.
cited_total_grant_inventor	Total number of granted patents in cited country-cpc-decade. Country allocation made based on inventor locations.
citing_total_app_inventor	Total number of patents in cited country-cpc-decade. Country allocation made based on inventor locations.
citing_total_grant_inventor	Total number of granted patents in cited country-cpc-decade. Country allocation made based on inventor locations.
cited_total_app_owner	Total number of patents in cited country-cpc-decade. Country allocation made based on assignee locations.
cited_total_grant_owner	Total number of granted patents in cited country-cpc-decade. Country allocation made based on assignee locations.
citing_total_app_owner	Total number of patents in cited country-cpc-decade. Country allocation made based on assignee locations.
citing_total_grant_owner	Total number of granted patents in cited country-cpc-decade. Country allocation made based on assignee locations.

Table 2: Information by Country

Country	Patents	Granted	Forward Cites	Backward Cites	Avg. Forward	Avg. Backward
Canada	0.146	0.109	2.03	1.821	13.947	12.507
China	5.354	1.758	6.23	9.934	1.164	1.856
Germany	0.797	0.595	5.018	4.984	6.294	6.252
France	0.28	0.21	1.925	1.761	6.881	6.295
Japan	2.884	1.846	19.544	13.757	6.777	4.77
Korea	2.356	1.616	5.461	6.103	2.318	2.59
RoW	0.889	0.622	7.389	8.828	8.315	9.934
Switzerland	0.14	0.102	1.19	1.404	8.501	10.031
Taiwan	0.312	0.209	1.97	2.243	6.31	7.182
UK	0.19	0.13	1.937	1.647	10.182	8.656
USA	3.658	2.946	64.259	64.472	17.566	17.624

Notes: Excepting averages, numbers are in millions.

Table 3: Information by Country

Country	Patents	Granted	Forward Cites	Backward Cites	Avg. Forward	Avg. Backward
Canada	0.146	0.109	2.03	1.821	13.947	12.507
China	5.354	1.758	6.23	9.934	1.164	1.856
Germany	0.797	0.595	5.018	4.984	6.294	6.252
France	0.28	0.21	1.925	1.761	6.881	6.295
Japan	2.884	1.846	19.544	13.757	6.777	4.77
Korea	2.356	1.616	5.461	6.103	2.318	2.59
RoW	0.889	0.622	7.389	8.828	8.315	9.934
Switzerland	0.14	0.102	1.19	1.404	8.501	10.031
Taiwan	0.312	0.209	1.97	2.243	6.31	7.182
UK	0.19	0.13	1.937	1.647	10.182	8.656
USA	3.658	2.946	64.259	64.472	17.566	17.624

Notes: Excepting averages, numbers are in millions.

Table 4: Information by Technology Class

Technology	Patents	Granted	Forward Cites	Backward Cites	Avg. Forward	Avg. Backward
A: Human Necessities	2.365	1.249	20.74	21.609	3.564	3.641
B: Operations & Transport	3.005	1.723	14.304	14.36	5.605	5.595
C: Chemistry & Metallurgy	1.853	1.081	9.588	8.875	4.657	4.09
D: Textiles	0.223	0.129	0.886	0.842	3.559	3.135
E: Fixed Constructions	0.612	0.36	2.647	2.644	2.71	2.62
F: Mechanical Engineering	1.481	0.883	7.254	7.216	6.6	6.485
G: Physics	3.89	2.353	32.215	32.081	5.855	5.628
H: Electricity	3.577	2.367	29.319	29.328	7.098	6.92

Notes: Excepting averages, numbers are in millions.

UCD CENTRE FOR ECONOMIC RESEARCH – RECENT WORKING PAPERS

WP22/25 Margaret Samahita, Kevin Devereux: 'Are economics conferences gender-neutral? Evidence from Ireland' October 2022

WP22/26 Michele Gubello, Nora M. Strecker: 'Weakly Progressive: Disproportionate Higher Education Attendance and the Structure of Income Taxes' November 2022

WP22/27 David Madden: 'Mental Health in Ireland During the Covid Pandemic: Evidence from Two Longitudinal Surveys' December 2022

WP22/28 Ronald B. Davies, Zuzanna Studnicka: 'Tariff Evasion, the Trade Gap, and Structural Trade' December 2022

WP23/01 Xidong Guo: 'An Analysis of a Rural Hospital's Investment Decision under Different Payment Systems' January 2023

WP23/02 Emanuele Albarosa, Benjamin Elsner: 'Forced Migration and Social Cohesion: Evidence from the 2015/16 Mass Inflow in Germany' January 2023

WP23/03 Xidong Guo: 'Hospital Efficiency and Consultants' Private Practices: Analysing the Impact of a Voluntary Reform' February 2023

WP23/04 Tadgh Hegarty, Karl Whelan: 'Calculating The Bookmaker's Margin: Why Bets Lose More On Average Than You Are Warned' February 2023

WP23/05 Tadgh Hegarty, Karl Whelan: 'Forecasting Soccer MatchesWith Betting Odds: A Tale of Two Markets' February 2023

WP23/06 Karl Whelan: 'How Do Prediction Market Fees Affect Prices and Participants?' March 2023

WP23/07 Kevin Devereux, Zuzanna Studnicka: 'Local Labour Market Concentration and Wages in Ireland' March 2023

WP23/08 Kevin Devereux, Zuzanna Studnicka: 'When and Where do Minimum Wage Hikes Increase Hours? Evidence from Ireland' April 2023

WP23/09 Karl Whelan 'Fortune's Formula or the Road to Ruin?The Generalized Kelly CriterionWith Multiple Outcomes' April 2023

WP23/10 Manuel E. Lago: '#MeToo...or not? Do salient shocks affect gender social norms?' April 2023

WP23/11 Cormac Ó Gráda, Tyler Anbinder, Simone A. Wegge: 'Gaming theSystem: The Not-So-Poor and Savings Banks in Antebellum New York' April 2023

WP23/12 Tadgh Hegarty, Karl Whelan: 'Disagreement and Market Structure in Betting Markets: Theory and Evidence from European Soccer' May 2023

WP23/13 Tadgh Hegarty, Karl Whelan: 'Do Gamblers Understand Complex Bets? Evidence From Asian Handicap Betting on Soccer' May 2023

WP23/14 Matthew Amalitinga Abagna: 'Special Economic Zones and Local Economic Activities in Ethiopia.' May 2023

WP23/15 Gregory Corcos, Stefanie Haller: 'Importer Dynamics: Do Peers Matter?' July 2023

WP23/16 Karl Whelan: 'How Does Inside Information Affect Sports Betting Odds?' July 2023

WP23/17 Michele Gubello: 'Social Trust and the Support for Universal Basic Income' July 2023

WP23/18 Constantin Bürgi, Zuzanna Studnicka: 'Online Interactions, Trade And The Gravity Model' August 2023

WP23/19 Juan S. Morales, Margaret Samahita: 'Can Social Pressure Stifle Free Speech?' August 2023

WP23/20 Fabrice Kämpfen, Xavier Gómez-Olivé, Owen O'Donnell, Carlos Riumallo Herl: 'Effectiveness of Population-Based Hypertension Screening: A Multidimensional Regression Discontinuity Design' August 2023

WP23/21 Annette Alstadsæter, Ronald Davies, Mathieu Parenti, Farid Toubal: 'The Real Effects of Tax Havens' October 2023

UCD Centre for Economic Research
Email economics@ucd.ie