Maastricht University

UNITED NATIONS UNIVERSITY

UNU-MERIT

Working Paper Series

#2019-051

Greentech homophily and path dependence in a large patent citation network

Önder Nomaler & Bart Verspagen

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: info@merit.unu.edu | website: http://www.merit.unu.edu

Boschstraat 24, 6211 AX Maastricht, The Netherlands
Tel: (31) (43) 388 44 00

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology
UNU-MERIT

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT to stimulate discussion on the issues raised.

Greentech homophily and path dependence in a large patent citation network

Önder Nomaler & Bart Verspagen

UNU-MERIT

This version of 16 December 2019

Abstract

We propose a method to identify the main technological trends in a very large (i.e., universal) patent citation network comprising all patented technologies. Our method builds on existing literature that implements a similar procedure, but for much smaller networks, each covering a truncated sub-network comprising only the patents of a selected technology field. The increase of the scale of the network that we analyse allows us to analyse so-called macro fields of technology (distinct technology fields related by a coherent overall goal), such as environmentally friendly technologies (Greentech). Our method extracts a so-called network of main paths (NMP). We analyse the NMP in terms of the distribution of Greentech in this network. For this purpose, we construct a number of theoretical benchmark models of trajectory formation. In these models, the ideas of homophily (Green patents citing Green patents) and path dependency (the impact of upstream Green patents in the network) play a large role. We show that a model taking into account both homophily and path dependence predicts well the number of Green patents on technological trajectories, and the number of clusters of Green patents on technological trajectories.

JEL Codes: Q55, Q54, O31, O33, O34

Keywords: patent citation networks; green technology; climate change mitigation

This work was financially supported by the Academic Research Programme of the European Patent Office (EPO). Any errors and the expressed views are solely the responsibility of the authors.

1. Introduction

In this paper, we report on a new method to extract the main technological trends from a very large patent citation network covering all technologies patented under the terms of one legal jurisdiction (the European Patent Office, EPO). We also provide an application of this method to a specific patent citation network, with the aim to investigate the distribution of so-called Greentech patents over the entire network. Greentech are patents that describe technology that, as identified by the patent office, contributes to the mitigation of greenhouse gas emissions.

We look at Greentech as a so-called macro-technology field, i.e., a set of distinct technologies that are in pursuit of a common and coherent goal (in this case, combatting climate change). While methods like ours, i.e., extracting citation paths as a summary of technological trends, have a long history in the literature, they have so far only been applied to individual and smaller technology fields. Our method brings the analysis of macro-technology fields in reach of this method of analysis.

While the introduction and description of our method is one main goal of the paper, we also provide results on the nature of Greentech. For this part of our analysis, we ask how Green patents are distributed over then entire network of main technological trends. We cover the period 1978 – 2018 and include all patents published by the European Patent Office. In this set of patents, do we see clustering of Green patents in particular neighbourhoods of the citation network? In other words, do Green patents form citation paths that mainly consist of Green patents, or are they spread out over the entire network, mixed with Brown patents without regard for the "colour"? (We adopt the term "Brown" for any patent that is not defined as Green).

These questions have important consequences for the nature of Greentech as a macro-technology field. If Green patents are strongly clustered, this implies that Greentech is a macro-field that develops according to its own internal logic, and that contributing to this macro-field requires knowledge of this internal "Green logic." If, on the other hand, Green and Brown patents are perfectly mingled in the citation network, Greentech appears more as knowledge that can be added to any technology field at any stage of its development, i.e., it occurs as a way of "Greening" a wide variety of technological developments that are not inherently Green.

In order to describe the concentration of Greens in the citation network, we first introduce and apply our method of finding the main technological trends in the total citation network. This yields what is called the Network of Main Paths (NMP). We then analyse the distribution of Green and Brown patents in the NMP. This analysis focuses on the level of individual technological trajectories, which are defined as citation paths. The NMP contains a huge number of trajectories, and our method enumerates them all to be able to provide statistics on them.

These statistics will refer to two characteristics of the paths: the number of Greens, and the number of colour-clusters (Green/Brown) that they contain. The more paths there are at the extremes of the distribution (e.g., zero Greens vs. an all-Green path; and just one colour-cluster vs. as many colour-clusters as the path length), the more concentrated the Greens and Browns are in the network. However, to be able to interpret the statistics on number of Greens and number of clusters observed in the network, we need some kind of theoretical benchmark that tells us how many Greens and how many clusters should be expected.

We provide these benchmarks in the form of three theoretical stochastic models. The first of these models does not contain any mechanism that would lead to any particular concentration of Greens and Browns. Hence, any concentration that we observe to exceed the levels predicted by this model can be interpreted as relatively concentrated. The two other models introduce two specific mechanisms that would lead to concentrations of Greens and Browns in the NMP. The first mechanism is so-called homophily, which in our context, means the tendency for patents of the same colour (Green-to-Green and Brown-to-Brown) to cite each other at higher rates than patents of different colours (Green-to-Brown and Brown-to-Green). We observe homophily in the model, and our theoretical model asks to what extent this observed level of homophily can explain the concentration of Greens and Browns.

Our third model adds path dependence as a concentration mechanism. Path dependence as we define it can be seen as a higher-order form of homophily, i.e., it considers not only whether or not the cited patent is Green, but also the colour of the nodes that lie before (upstream) of the cited patent. Thus, path dependence is the tendency of clusters of more than a single Green (or Brown) node to continue as Green (or Brown). The parameters of the theoretical benchmark model that includes both homophily and path dependence can be estimated from the data of the NMP, and then the model can be simulated to provide predictions of the number of Greens and the number of colour-clusters on a path.

The rest of this paper is structured as follows. In the next section, we outline the conceptual backgrounds of our analysis. This covers the idea of technology as a sequence of incremental changes following a breakthrough invention. This also covers the idea of main path analysis to map these sequences, leading to the idea of technological trajectories or technological paths (we trajectories and paths mostly as synonyms).

Section 3 provides a brief elaboration of the specific questions on Greentech that we already introduced above. Section 4 introduces the method that we propose to construct the NMP from the total citation network. This section also provides descriptive statistics on the NMP that we extract, both in general terms, and in terms of the specific indicators on Greentech (number of Greens on a path and number of colour-clusters). Section 5 introduces the benchmark theoretical models and the notions of homophily and path dependence. It also provides the estimations necessary to implement the path dependence mechanism. Section 6 confronts the empirical data of the NMP with the predictions of the benchmark models, i.e., this section evaluates the performance of the models. Section 7 summarises the argument and points to some options for further research.

2. Conceptual backgrounds

In first instance, our analysis is aimed at outlining the major global technology trends of the last few decades by using patent citation networks. The general idea is that patent citations indicate some form of knowledge flows, from the cited to the citing patent (Trajtenberg and Jaffe, 2002). This is based on the literature that follows Hummon and Doreain (1989), who proposed a method for analysing directed and a-cyclical networks. This is the typical network that is formed by citations, either in the scientific literature, or in patent literature. The Hummon and Doreain-based methods will identify so-called technological 'main paths.'

This has usually been done for individual technological fields, as a way to quantify more qualitative impressions from engineers or the history of technology (e.g., Mina et al., 2007; Verspagen, 2007; Liu and Lu, 2012). In the current paper, by introducing a number of improvements in computing algorithms, we are able to analyse a much larger set of patent citations that represent the entire patent literature, and hence the entire spectrum of technologies that have been subject to human invention over the last decades, rather than a single technology field. By enlarging the scope in this way, we can look at a multitude of technological trajectories, and the way that these paths interact. Our emphasis can thus shift from identifying single main paths to a network of paths covering all (patented) technologies at once.

This is particularly useful in cases where the interest lies in what can be called a macro-field of technology, which we define as a collection of distinct technology fields with a common and coherent purpose. The example of a macro technology field that we will consider is so-called 'green' technology, which we define as technologies aimed at climate change mitigation. Obviously, technology with this aim consists of a large collection of distinct technology fields, e.g., in solar energy, fuel cells, biology, nutrition, agriculture, etc. The collection of main paths that we identify in the large patent citation network that is usually only studied in small parts will enable us to see how green technology is embedded in this larger context.

The idea that patent citations can be used to map technology trends has an origin in two main ideas in the economics and management literature. One idea, originating in the management field (e.g., Levinthal, 1997; Fleming and Sorensen, 2004; Aharonson and Schilling, 2016), is that technological choice of firms can be represented as a process of recombinant search on a technological landscape, and that much of this search is local, i.e., in the immediate neighbourhood of where search was previously located. The idea of a landscape is a metaphor that portrays technological knowledge as configurations of component building blocks (e.g., Kaufmann, 1993; Kaufman et al. 2000). By changing one of the components of an existing piece of knowledge, or by combining building blocks from several pieces of knowledge, new knowledge can be created from existing knowledge. Because the pieces of knowledge are related to each other by the components that they share, distance between technologies can easily be operationalised. The metaphor of a technological landscape then arises by mapping the knowledge pieces relative to each other based on how close they are.

A central tenet of this landscape concept is that performance of technologies differs and is somehow dependent on the position of the technology in the landscape. Thus, the firm (or inventor) who searches in the landscape will find particular locations of high or low opportunity and value, corresponding to peaks and valleys in the landscape metaphor. Firms will want to occupy the high value/opportunity locations of the technological landscape, and thus will direct their search efforts towards there. As a result, technological efforts by firms will cluster in technology space (e.g. Aharonson and Schilling, 2016). A logical strategy is to use prior knowledge about where the feasible and valuable technologies are located (Stuart and Podolny, 1996; Fleming and Sorensen, 2004). Such prior knowledge accumulates from the firm's own prior research, and, to the extent that they are observable, other firms' research efforts. Prior research results are guideposts (Sahal, 1981) that help current and future research. This leads to a process of dynamic increasing returns, as firms seek out the regions of technology space that are most valuable in terms of their economic returns.

Whether prior knowledge leads to useful information about where new opportunities can be found depends on the shape of the landscape. If valleys and peaks occur in the form of smooth transitions,

prior knowledge will be useful, as it will allow the researcher to follow an upward slope, and ultimately reach a (local) peak of valuable knowledge. However, if the landscape is more "rugged", information about prior research may be less useful, i.e., when spots of high and low opportunity are found randomly and independently of each other. In Kaufman's model (Kaufmann, 1993; Levinthal, 1997), a parameter tunes the ruggedness of the landscape. Intermediate values of ruggedness imply both that clustering on the basis of prior knowledge is useful, and that the landscape contains identifiable peaks and valleys (Billinger et al., 2014).

Serendipity and basic research are ways in which search in the technological landscape may occur over larger distances. This may open up new areas of the technological landscape, which can then be explored by local search. By making a large (random) jump in the technology landscape, access to a previously unknown local peak may be gained, although this must be realised by (slowly) climbing the slope that leads to the peak. Viewed in this way, the process of technological search combines elements of randomness (which areas of the landscape are opened up) and systematic exploration by collective action of the firms that are active in a specific field (Sorensen and Fleming, 2004).

This leads to the second idea that underlies our approach based on patent citations, which is that technologies develop as 'trajectories' (Dosi, 1982) that are heavily influenced by economic opportunities. The concept of a technological trajectory is also based on local search and is compatible with the metaphor of technological landscapes, while it adds to the previous discussion the idea that sequential incremental improvements of technology will generally represent a specific and collective direction in technological space, and that this direction is heavily shaped by both technological opportunities and the economic incentives that the market provides.

Dosi's starting point is that engineers will be inclined to search in the neighbourhood of a particular set of opportunities, and that such a neighbourhood tends to be opened up by a paradigm shift that follows, for example, from basic research, or from practical experimentation. Although such a paradigm shift, in principle, opens up a number of possible trajectories, there will usually only be a selective number that will actually be realised, and this is decided on the basis of specific market circumstances.

The historical case of steam engines may serve as a brief illustration (Nuvolari and Verspagen, 2009). Although based on a common technological principle, steam engines were applied in many different economic contexts, leading to a wide variety of designs that were very much adapted to the incentives found in those contexts. In Cornish mines, where steam engines were used to pump up water from flooded mine galleries, the economic incentive was saving on expensive coal, which led to very large-scale versions of the low-pressure engine that James Watt brought to Cornwall in the late 18th century. On the other hand, in the application of steam engines to railways, such large designs were unpractical because the engine had to be mobile. As a result, a trajectory emerged of much smaller high-pressure engines that could deliver adequate power for transportation. More modern examples of trajectories that show the strong impact of cumulated incremental changes can be found in digital technologies, for example in the form of the famous Moore's law.

3. Research questions and operationalisation

The aim of our analysis will be to summarise the main technological trends of the last decades, and to investigate how the macro-technology field of green technology is embedded in these trends. We are particularly interested to find out to what extent green technology is diffused across the entire set of main technology trends, or is concentrated in a smaller number of trends.

Based on our above discussion, we will use a patent citation network to extract main technology trends. We define a main technological trend as a technological trajectory in the Dosi-sense, i.e., as a series of cumulative improvements to a basic design that together define a main direction of technological change. We look at local recombinant search as the main way in which firms and research organisations collectively construct these trajectories. And we operationalise the concept of a technological trajectory by drawing on the methodological tradition of Hummon and Doreian (1989). The next section will specify how we identify technological trajectories as a 'big data' variety of the original Hummon and Doreian concept of a main path in a citation network.

With patents as our smallest unit of analysis, we operationalise green technology as a specific subset of environmentally friendly patents that is aimed at greenhouse gas emission mitigation. This has the advantage that we can use the so-called Y02 tag which the major patent offices of the world assign to patents these days. The Y02 tag is in fact a technology class in the Cooperative Patent Classification (CPC) scheme. This class can be assigned to a patent document in addition to native technology classes that patent offices use, such as the US Patent Classification or the International Patent Classification. The Y02 class (the class title is 'Technologies or applications for mitigation or adaptation against climate change') is also further subdivided, for example into eight 4-digit classes that are aimed largely at specific application areas such as transport, waste agriculture etc. Using the Y02 CPC class, we classify each patent in our network as either green (having a Y02 tag) or brown (not having a Y02 tag).

This leaves the question how to operationalise the extent of diffusion (or concentration, which we consider the opposite of diffusion) across the main technology trends that we identify in the patent citation network. In order to measure diffusion, we look at the unit of individual technological trajectories, or paths, and ask how many patents on a specific path are green. Using three distinct theoretical models with varying degree of complexity, we can formulate precise statistical expectations on the number of green patents on a path with given length. Testing these expectations against what is observed in our data is the way in which we operationalise the degree of diffusion or concentration of green patents in the network of main paths.

For example, the simplest of our theoretical models predicts that 49.7% of all paths with exactly 10 patents will have no green patents at all, while a slightly more complex model predicts that 65.2% of all paths with length = 10 will have no green patents. However, in the actual data, 78.0% of all paths with length = 10 has no green patents. Thus, the second model predicts higher concentration (less diffusion) of green technology than the first model, while in the actual data, we observe more concentration than either model predicts. In the analysis below, we will also present a third model, and perform the analysis for different number of green patents (>1, >2 etc. instead of just >0) and different path lengths (up to 28), so that a complete picture of concentration of green technology in our network of main paths emerges.

In the models that we employ to predict the distribution of green technology over the main paths in our citation network, the concept of homophily will play a large role. In network analysis (e.g., McPherson et al., 2001), this refers to the idea that similarity between nodes of the network (in our

case, patents) tends to have a positive influence on the probability that a connection (in our case, a citation) exists between the nodes. In our models, we define homophily as the tendency for preferential citation, i.e., for green patents to cite green patents and brown patents to cite brown patents. We observe homophily in our network, especially in the brown-to-brown citations. Our analysis will show that incorporating homophily in the model generally increases the ability to predict the occurrence of green patents in the main paths that our analysis finds.

4. Methods – Main Path Analysis

4.1. The patent citation network

The first step in our analysis is to construct the total network of citations. This starts by extracting a citation network between PatStat application ids for which the application authority is 'EP'.[1] Citations take place between publications, while an application id may be associated with more than one publication. Thus, we consider a citation from at least one publication related to application X to at least one publication related to application Y as a citation from application X to application Y. In order to guarantee that we avoid cycles in the citation network, we consider a citation as valid only if the application date of the citing application is at least one day later than the cited application.

The citation network that is formed in this way has 2,758,196 citations linking 2,033,487 EPO patent applications. Thus, out of the 3,561,211 EPO patent applications reported in PatStat, 1,527,724 (about 43%) are not represented in the citation network, simply because these neither cite nor are cited by any other EPO patent. In order to increase coverage, the citation network is adapted in two ways, both of which add links to the network that are not actually present in the original set of intra-EPO citations.

The first extension of links in the network is aimed at capturing citations at other patent offices than the EPO. We add this to account for technological paths that are not captured exclusively by EPO patents. In this case, we look for any indirect citation linkages between EPO patents that exist between EPO patents, and add these as direct linkages in our network. For example, if EPO-application A is cited by US application B and US application B is cited by EPO application C, then we add a link from EPO application A to EPO application C in our network, even if no actual citation exists between those two EPO applications.

Our second extension deals with patent families, as documented by the DocDB families in PatStat. Patent family membership indicates a degree of similarity between the documents in the family, i.e., a family can be seen as covering a single invention by multiple patent applications. The reasons for filing more than a single application for the same invention are mostly legal. One commonly found reason is to extend coverage to multiple countries. Our exclusive focus on a single jurisdiction (EP) already implies that we do not have any family relations of this type. However, due to other legal reasons (e.g., divisionals, extensions, etc.), a DocDB family may still have more than one EPO application.

We found that treating a single family as a single invention by aggregating citations into a single link between families leads to cycles in the citation network. For example, application P and

[1] We use the 2019a edition of PatStat.

application Q could be members of the same family, but typically have different application dates. Then if patent Q cites another document with application date later than patent P, cycles will emerge easily in the aggregated citation network.

In order to avoid cycles, we deal with family membership by first ranking all EP-members of a family in terms of their application date, and then add links from the oldest EP-member to the next, and from this EP-member to the next, etc., until we reach the newest EP-member of the family. In other words, we consider a family as a technological path in itself. This procedure will prevent cycles from forming, while still recognising the similarities between inventions in a family. In this way, we have an extended patent citation network that consists of 2,771,440 patent applications (about 78% of all applications at the EPO) and 9,090,460 citations between them. This covers the period 1978 – 2018.

4.2. The network of main paths: construction

The next step in our analysis is to construct the network of main paths in the total citation network. In methodologies that draw on Hummon and Doreian (1989), the former is a systematically reduced subset of the latter, obtained by eliminating the patents and/or citations of 'lesser' importance. Thus, the network of main paths is a collection of citation chains that are representative of the most important sequences of (incremental) progress in the technology field(s) covered by the documents in the given citation network.

The first stage in constructing the network of main paths is to calculate an index of (relative) importance for each citation link in the network. These are referred to as traversal weights. Several alternative link weighing principles are proposed by Hummon and Doreian (1989) and later by Batagelj (2003). The most commonly used one is SPNP (Search Path Node Pair) which, in a nutshell, is the number of document pairs that are connected directly or indirectly by a given citation link. More formally, SPNP is the number of times a given citation link is visited if one follows through all possible upstream paths from all (direct and indirect) ancestors of the cited document (including itself) to all (direct and indirect) descendants of the citing document (including itself). We will only use SPNP in this paper.

In Hummon and Doreian (1989) and the largest part of the related literature that follows, the second stage of the method identifies a so-called main path in the network. The main path is a chain of citations that is constructed on the basis of some heuristic that aggregates the individual traversal weights of the constituent citation links of the chain. Usually, the main path is identified by a 'priority first search' algorithm, which, starting from a given start-node, follows consecutive citation links stepwise, choosing each time the next forward citation link with the highest SPNP value until hitting an end-node.[2] In case of a tie, the trajectory branches out since the algorithm separately takes each link with the highest link value and follows each emerging branch to the end.

Hummon and Doreian (1989) picked one start-node among several possible in their network, and focus on the main path that is formed by performing the priority first search algorithm from this start-node only (although they did sensitivity analysis comparing other start-nodes). If there are no

[2] A start-node is a node (patent) that does not cite any other patents. An end-node is a patent that is not cited by any other patents.

ties, this method identifies a single trajectory, the *top main path* (TMP). Verspagen (2007) starts from each start-node in the network, and constructs (based on the 'priority first search' principle) a collection of main paths that is referred to as the *network of main paths* (NMP). If the aim of the exercise is to describe the main trajectories in a specific technology field, the choice is often to focus on the TMP, because the NMP remains too large to provide a concise historical narrative.

The NMP or TMP that is generated by the priority first search algorithm consists of a subset of citations and patents of the original citation network. This is obvious for the TMP, but even the NMP generally does not cover all patents and citations. In a citation network with S start-nodes, the NMP may consist of 1 to S weakly-connected components.[3] But it is very likely that some of the individual paths in the NMP will partially overlap, leading to less components. For example, in Triulzi (2015), the largest main component of a citation network of about 114 thousand patents and about 779 thousand citations is reduced (by the procedure explained above) into a NMP of about 23.5 thousand patents (a reduction in size by about 80%) and about 22 thousand citations. This NMP consists of several weakly-connected components where the largest one consists of about 3.5 thousand patents.

As stressed by Liu et al. (2012), it is important to realise that the priority first search algorithm is a heuristic that does not guarantee a global maximum in the value of the summed SPNP over the found main path(s). This holds for the TMP as well as for any other main paths in the NMP. In other words, for any start-node, there may well be forward paths that have a higher total SPNP value than the main paths found in the priority first search algorithm. This is related to another arbitrariness identified by Liu et al. (2012): instead of starting from a start-node and implementing a forward search, one may just as well start from an end-node and search backwards. The forward search method constructs an NMP which incorporates at least one trajectory that emanates from each start-node of the original network, although only a subset of the end-nodes of the original network will make it to the NMP. With the backward search, all end-nodes of the original network, but only a subset of the start-nodes will end up in the NMP. Furthermore, the local (priority first) backward search might yield a rather different set of trajectories than the local (priority first) forward search, including a different TMP.

Our methodological innovation is threefold. First, we propose to substitute the usual priority first forward search heuristic by an alternative that combines both forward and backward search to maximise the (log-)sum of SPNP between all combinations of start-nodes and end-nodes that are connected in the citation network. Second, we separate the elimination of patents and citations in the procedure of constructing the NMP. Some citations are eliminated first, leaving all patents in the NMP, and only after this do we start to prune this NMP by removing both patents and their inward and outward citations. Third and finally, while we prune the NMP, we remove entire paths (based on their log-sum of SPNP) rather than individual patents. This has the advantage that the connectedness of the NMP remains largely intact. In this way, we can prune the NMP at any desired level, from no pruning at all to only leaving the TMP.[4]

Let us now formally describe the method, which will consist of first defining and constructing an NMP, and then pruning it step-by-step to obtain the TMP. We represent a trajectory as an *ordered*

[3] In a directed network, a weakly-connected component is a subset of patents for which there exists a path from any node to any other nodes if all unidirectional links are replaced by bidirectional connections.
[4] Our TMP is identical to the one identified by Liu et al. (2012).

set FT_i^k which refers to a forward citation chain (or a sub-chain) of successively connected NFT_i^k nodes (patent documents). This (sub)chain emanates from node *i*, which we also denote as FT_i^k *(1),* and terminates at node FT_i^k *(*NFT_i^k*).* Because multiple chains may start at note *i*, we use the index *k* to identify them. For any successive pair of documents $FT_i^k(j)$ and $FT_i^k(j+1)$ (where $j \in \{1,2,\dots,NFT_i^k - 1\}$), there exist a direct citation link from the latter to the former. Note that the document pair $FT_i^k(j)$ and $FT_i^k(j+1)$ may also appear on other trajectories than just FT_i^k.

Let F_i denote the set of all NF_i forward citation chains that emanate from a given node *i.* Thus for any citation chain $FT_i^k \in F_i$, by definition $FT_i^k(1) = i$ for $\forall k \in \{1,2,\dots,NF_i\}$). Also, let $SPx(FT_i^k(j))$ denote the *SPx* value[5] of the link through which document $FT_i^k(j+1)$ cites document $FT_i^k(j)$.

To accommodate backward search, let us draw up a set of similar definitions that take the backward perspective, by replacing the letter *F* by the letter *B* in all definitions so far. Then BT_i^k represents a trajectory as an *ordered set* of successively backward connected NBT_i^k nodes where, for any $j \in \{1,2,\dots,NBT_i^k - 1\}$, the successive pair of documents in the ordered set as $BT_i^k(j)$ and $BT_i^k(j+1)$ indicates an actual direct citation link from node *j* to node *j* + 1. Finally, let *S* denote the set of all start-nodes, and *E* the set of all end-nodes of the citation network.

We are now ready to define and construct the NMP. For any node *i* in the network, we identify the particular (forward) trajectory FT_i^o that, for an aggregation rule A(·) of choice (additive, multiplicative), satisfies the condition

$$A(SPx(FT_i^o(1)),\dots,SPx(FT_i^o(NFT_i^o - 1))) \geq A(SPxFT_i^m(1),\dots,SPx(FT_i^m(NFT_i^m - 1)))$$

for $o \in \{1,2,\dots,NF_i\}$, $\forall m \in \{1,2,\dots,NF_i\}$ and $FT_i^o(NFT_i^o) \in E$ and $FT_i^m(NFT_i^m) \in E$ (i.e., only forward trajectories that terminate at end-nodes of the network are considered). This is the forward path from node *i* that maximises aggregate forward *SPx*.

In the same way, we identify the backward trajectory from any node *i* that maximises backward *SPx*, which implies finding, for each node *i* of the citation network, the particular backward trajectory BT_i^o where

$$A(SPx(BT_i^o(1)),\dots,SPx(BT_i^o(NBT_i^o - 1))) \geq A(SPxBT_i^m(1),\dots,SPx(BT_i^m(NBT_i^m - 1)))$$

for $o \in \{1,2,\dots,NB_i\}$, $\forall m \in \{1,2,\dots,NB_i\}$ and for *o* and $\forall m$, BT_i^o $(NBT_i^o) \in S$ and BT_i^m $(NBT_i^m) \in S$ (i.e., only complete trajectories that extend all the way back to a start-node of the original network are considered).

Having identified these maximum-*SPx* trajectories for all nodes of the network, we also define

$$MASPxB_i = A(SPx(BT_i^o(1)),\dots,SPx(BT_i^o(NBT_i^o - 1)))$$

$$MASPxF_i = A(SPx(FT_i^o(1)),\dots,SPx(FT_i^o(NFT_i^o - 1)))$$

These are the actual (maximum) values of aggregated *SPx* among all forward and backward trajectories from node *i*.

[5] In our analysis, we will only work with SPNP, but the method can also be applied using the alternative Hummon & Doreian citation indicators, SPLC or SPC.

We define and construct the NMP of the citation network by appending at every node *i* of the network the trajectories FT_i^o and BT_i^o. We denote this new trajectory as TT_i^o. The new path TT_i^o is, obviously, the maximum-*SPx* trajectory that goes through node *i*. Note that if node *i* is a start-node ($i \in S$), BT_i^o will be empty and $TT_i^o = FT_i^o$. Similarly, $TT_i^o = BT_i^o$ if $i \in E$ (the node is an end-node). Our NMP consists of all these appended maximum-*SPx* trajectories TT_i^o.

In this NMP, we eliminated a number of citations from the original citation network, but still all patents (nodes) of that original network are present. Thus, the NMP represents the metaphor of the most 'important' technological paths travelled, but only to the extent that this network of paths still visits all inventions that populate the landscape. In order to make the 'map' of the technological landscape a little coarser, and hence easier to interpret, we next proceed to drop also patents, and their incoming and outgoing citations, from the NMP.

To do this, we first assign every node *i* in the NMP a new indicator of significance, equal $MASPxT_i = MASPxF_i + MASPxB_i$. This is the aggregate value of *SPx* of node *i*'s maximum *SPx* trajectory TT_i^o, by which it contributed to the NMP. Having assigned all nodes with this indicator of importance, we proceed to prune the NMP by cutting the patents (and their direct forward and backward citations) with the lowest $MASPxT_i$ values. Note that by construction, this will never cut single patents, but instead the entire path TT_i^o. If we successively prune the paths TT_i^o with lowest $MASPxT_i$ value from the NMP, we will be left with a single TT_i^o. This is the TMP that is often used in other studies.[6]

Liu *et al*., (2012) propose summation as the aggregation operator. This ensures that the TMP of the citation network is the best trajectory that emanates from the start-node *s* where $MASPxF_s \geq MASPxF_i$ for all starting points $i \in S$. It also implies that backward search will not yield any trajectories with higher *SPx* sum. In Nomaler and Verspagen (2016), we chose instead multiplicative aggregation, and identified trajectories on the basis of the maximisation of $\sum_{j=1}^{NFT_i^o-1} log(SPx(FT_i^o(j)))$. This log-sum maximisation avoids the possible dominance of trajectories which might contain a few extremely high *SPx*-valued links together with many low *SPx*-valued ones, and instead gives priority to those characterised by moderately high but evenly distributed *SPx* values.

To conclude the description of our method, we will use a small example network to show how the NMP is created from the total citation network, and how the NMP can be successively pruned to yield, ultimately, the TMP. This example is displayed in Figure 1. The total citation network in the top panel has 12 patents, which are labelled P1 – P12. Arrows indicate knowledge flows or citations (knowledge flows from the cited to the citing patent). P1, P2 and P3 are start-nodes while P11 and P12 are end-nodes. The numbers attached to the arrows are log-SPNP of the citation link (these are not reproduced in the bottom panel), and the numbers in square brackets attached to the nodes (patents) are $MASPxT_i$ as explained above.

Having calculated the log-SPNP values (which works the same as it does in other studies, starting with Hummon and Doreian), our procedure to calculate the NMP drops a number of citation links from the total citation network. To see how this works, consider the citation of P4 by P10, at the top of the network diagram. This citation has log-SPNP value equal to 2.58, which is not very high, and

[6] In practice, with the large citation network that we use in the analysis below, we will not prune the NMP one-by-one, but instead at particular points of the distribution of $MASPxT_i$ values in the NMP. For example, we may prune to keep only the top-50% values of $MASPxT_i$, or the top-10%.

it lies on two paths: P1 → P4 → P10 → P11 and P2 → P4 → P10 → P11. The log-sum of SPNP is equal between those two paths: 3 + 2.58 + 4.58 = 10.16.

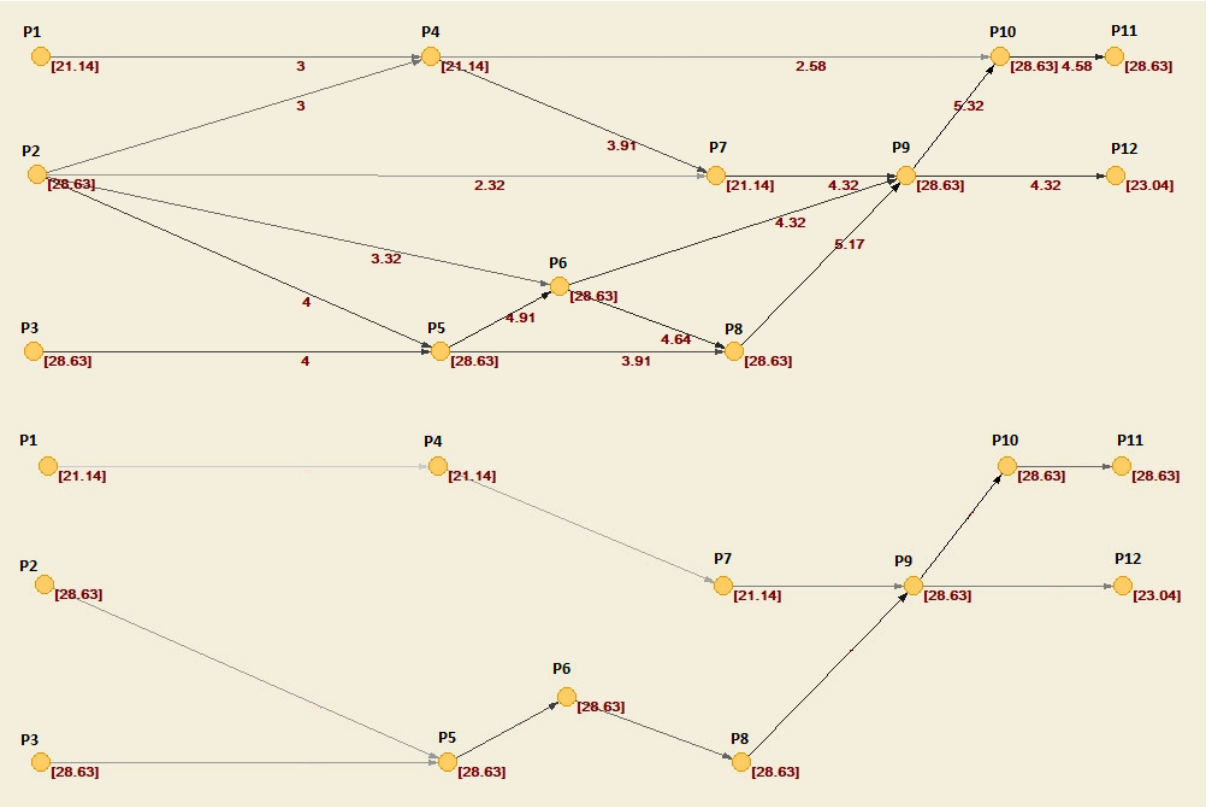


Figure 1. Example network, total citation network (top panel) and extracted NMP (bottom panel)

All patents on these two paths have alternative paths with higher SPNP. For example, P1, P2 and P4 also lie on the paths P1/P2 →P4 → P7 → P9 → P12, with total log-sum of SPNP equal to 15.55. P10 and P11 also lie on the path P3 → P5 → P6 → P9 → P10 → P11, with log-sum of SPNP equal to 23.13. Therefore, the citation P4 → P10 does not make it to the NMP, as displayed in the bottom panel.

The NMP has 5 trajectories, which our algorithm enumerates and identifies by a unique trajectory number:

T#1 (value 28.36): P2 → P5 → P6 → P8 → P9 → P10 → P11,

T#2 (value 28.36): P3 → P5 → P6 → P8 → P9 → P10 → P11,

T#3 (value 23.04): P2 → P5 → P6 → P8 → P9 → P12

T#4 (value 23.04): P3 → P5 → P6 → P8 → P9 → P12

T#5 (value 21.14): P1 → P4 → P7 → P9 → P10 → P11

Trajectories #1 and #2 have the same trajectory value (log SPNP sum) and the same length, but differ in terms of the start-node. Similarly, trajectories #3 and #4 are identical except for their start-node. Thus, we can refer to the first two trajectories as a 'trajectory group' and the third and the fourth another trajectory group. Our algorithm also enumerates trajectory groups on the basis of the following definition: A trajectory group is a set of trajectories, each with identical length and total (log) SPNP sum, and all having at least one common node (i.e., patent) exactly at the same position (order of appearance) of the trajectory.

Having constructed the NMP, it can be pruned. The first patents to be dropped, along with their inward and outward citations, would be P1, P4 and P7, as these have the lowest weight. This first cut effectively eliminates T#5 (although leaving intact P9, P10 and P11, which also participate to the more significant trajectories T#1 and T#2). Next, P12 would be dropped, eliminating trajectories #3 and #4, leaving the trajectory group formed by trajectories #1 and #2 as the only paths left, i.e., the TMP.

4.3. The network of main paths: empirical results[7]

Having defined the NMP in this way, we proceed to provide some brief descriptions of it. Note that the NMP that we constructed contains the same number of patents as in the total citation network (2,771,440), but reduces the number of citations from the original 9,090,460 to 3,494,708. Figure 2 documents the number of nodes in the NMP over time, by type of node (start-node, internal node or end-node). The number of start-nodes first rises, then stabilises and from about 1990 falls. The number of start-nodes is small as compared to the other types of nodes, except in the early period. The number of internal nodes rises slowly, peaks in 2001 and then falls slowly again. The number of end-nodes rises slowly, peaking towards the very end of the period. From about 2000 onwards, the number of end-nodes is larger than either the number of start-nodes or the number of internal nodes. This means that many of the paths in the NMP have star-like structures at the end, i.e., one final-but-one node linking to a larger number of end-nodes.

Figure 3 provides more information about the distribution of path length in the NMP and the relation between path length and log-sum of SPNP of the paths. We see that there are relatively many paths of relatively short length. Path length 2 (shortest possible) is the most frequent one (about 525,000 paths). 28 is the longest path length, but there are very few (14) paths of this length (note the log-scale for the axis of number of paths). Looking only at trajectories that contain at least one Green, we find relatively few of them (about 660,000 of a total of 3.7 million, or about 18%). The number of paths with some Green peaks at path length 6 (about 65,000 paths), while all of the longest (length 28) trajectories have some Green. The figure also shows that short paths tend to have low log-sum of SPNP, i.e., these paths would be the first one to be pruned in the procedure that was explained above. Average log-sum of SPNP rises almost linearly with path length, with a narrow standard deviation around the mean.

[7] The NMP of our citation network is available as a database (comma-delimited text file which can be built into a relational table under any database engine), and can be downloaded at https://dataverse.nl/dataset.xhtml?persistentId=hdl:10411/ZDCQY3. The database contains all information on application id of the NMP nodes (patent documents), all trajectories (and trajectory group) the node belongs to, and which position it takes on each trajectory. The database can be linked to PatStat by application id (appln_id) to obtain other patent information (such as the Green/Brown nature).

Next, we look at the phenomenon where our main interest lies: the distribution (concentration of diffusion) of the Greens and Browns on the NMP, including the pruned versions of the NMP. The basic unit of observation for this description will be individual paths in the NMP. We will enumerate all paths that are found in the NMP (or a pruned version of it), and characterise each path by two main characteristics: the number of Greens on the path, and the number of colour-clusters on the path. In defining colour-clusters, we simply look at subsequent nodes of the same colour, and consider them as a cluster. For example, the path G →B → B → G →G has 3 clusters (G, B → B and G → G).

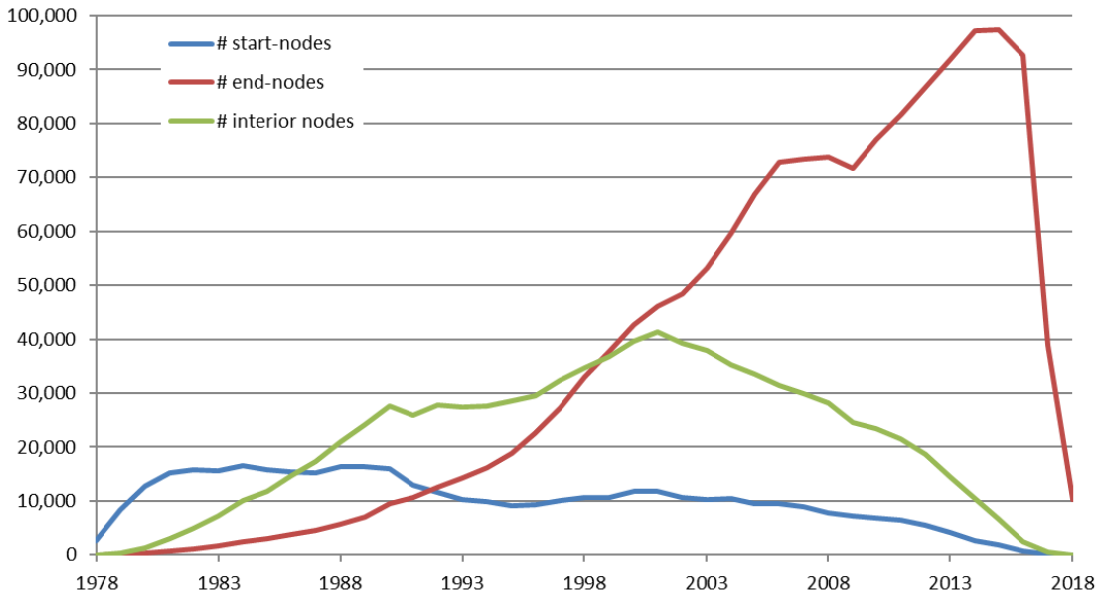


Figure 2. Number of nodes by type, un-pruned NMP

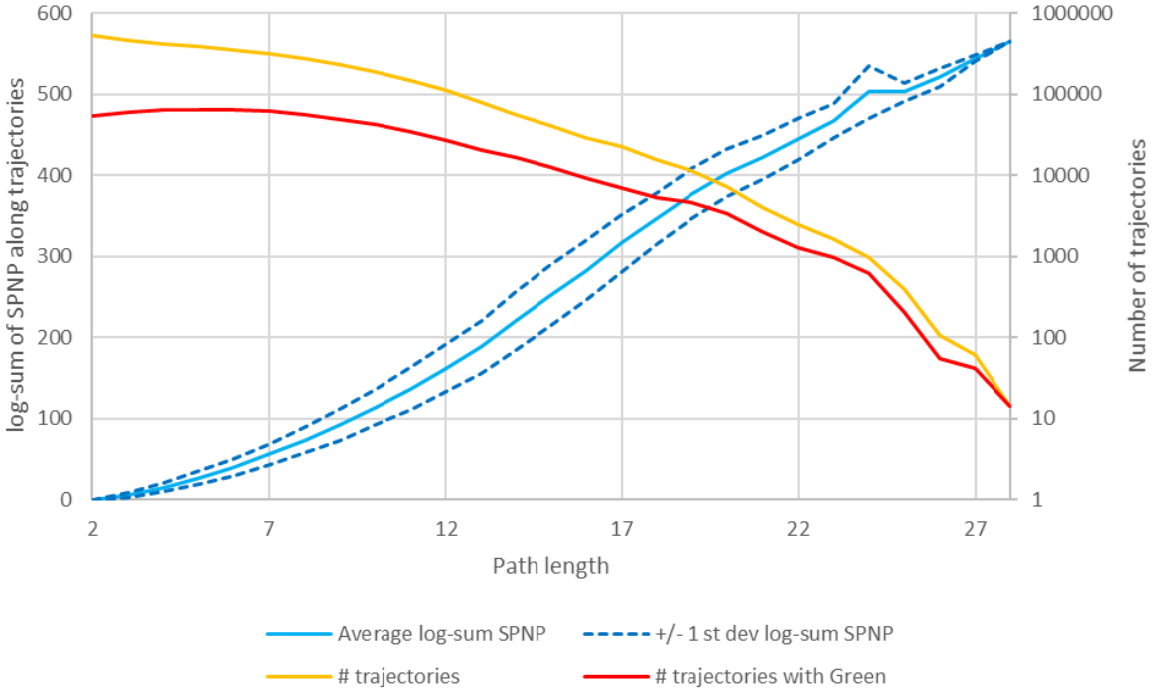


Figure 3. Number of trajectories and long-sum of SPNP by path length

Obviously, the possible number of Greens and the number of clusters on a path depend on the length of the path. Therefore, we will perform our analysis for each observed path length in the NMP. As implied by our method, there will be no isolates in the NMP, and hence minimum observed path length in the (non-pruned) NMP is 2. Figure 4 shows the observed frequency of paths by length in the NMP and five pruned version of it. Pruning has been done by percentile of the node weights as defined above, and the label indicates how much of the full NMP is kept. For example, NMP75 refers to a network in which the bottom 25 percentile nodes (and their citations) have been removed from the NMP (the largest pruned NMP), while NMP5 drops the bottom 95 percentile nodes (smallest pruned NMP).

The line for the full NMP is the same as in Figure 3, where short paths (length 2) are most frequent, and every longer path length shows a lower observed frequency. In line with what is expected on the basis of the SPNP line in Figure 3, pruning this network removes mostly the short paths, because these are the paths with low log-sum of SPNP. In the NMP75, all paths of length 2 and some of length have disappeared, while paths of length 4 and longer remain (almost) as frequent as in the NMP. This process repeats itself with further pruning until in the NMP5, the shortest path length is at 11. This implies that looking at longer path lengths in the (unpruned) NMP is a good approximation of the actual pruning process.

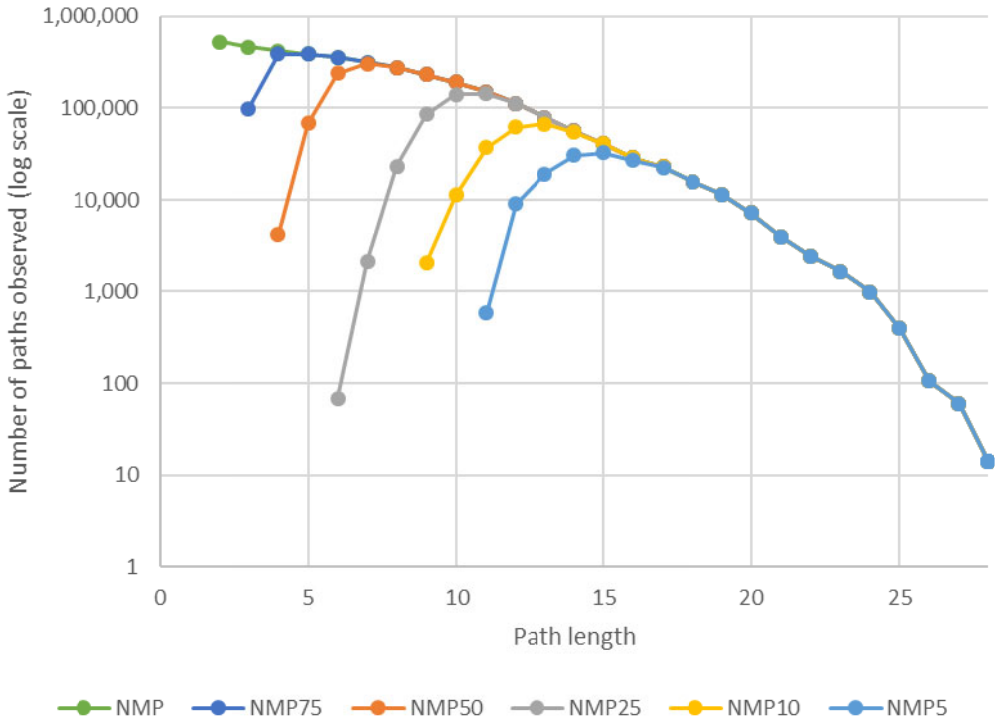
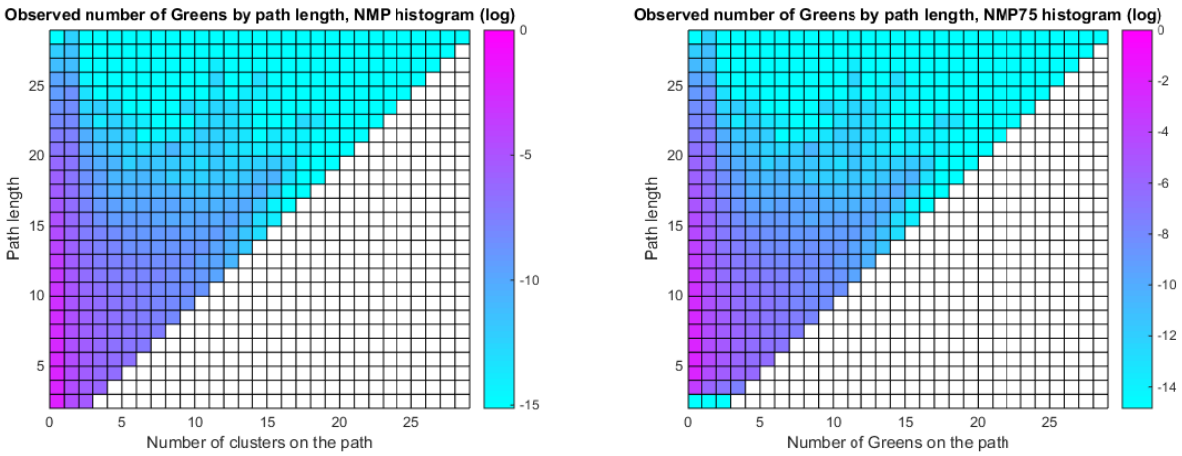


Figure 4. Number observed paths by length, NMP and pruned versions of it



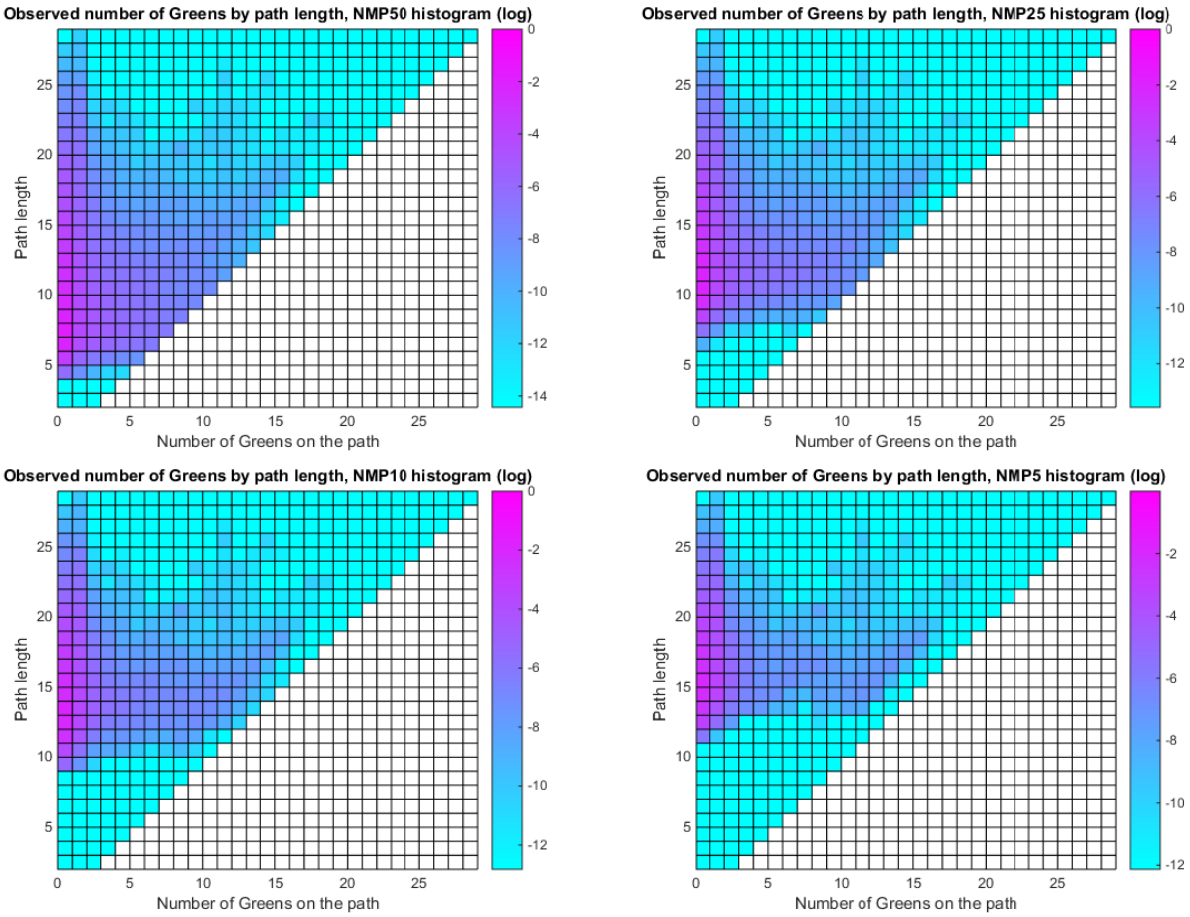


Figure 5. Histograms of observed numbers of Greens on a path (log), by path length, NMP and pruned NMPs

Figure 5 shows the distribution of the number of Greens on a path in the NMP and its pruned versions, by path length. Path length is on the vertical axis, so that each horizontal row represents paths of identical length. The horizontal axis of each figure displays the number of Greens on a path, and the colour shading indicates the relative frequency in the network. These frequencies are the log of the share of a particular path type in the entire network. For example, the colour for the cell with path length 3 and number of Greens 1 indicates the relative frequency (log) of paths of length 3 with one green in the network. White cells indicate impossible combinations (number of Greens larger than the path length), and the lightest shade (cyan) indicates cells with zero observed cases (for example, we observe no purely Green paths of length 28).

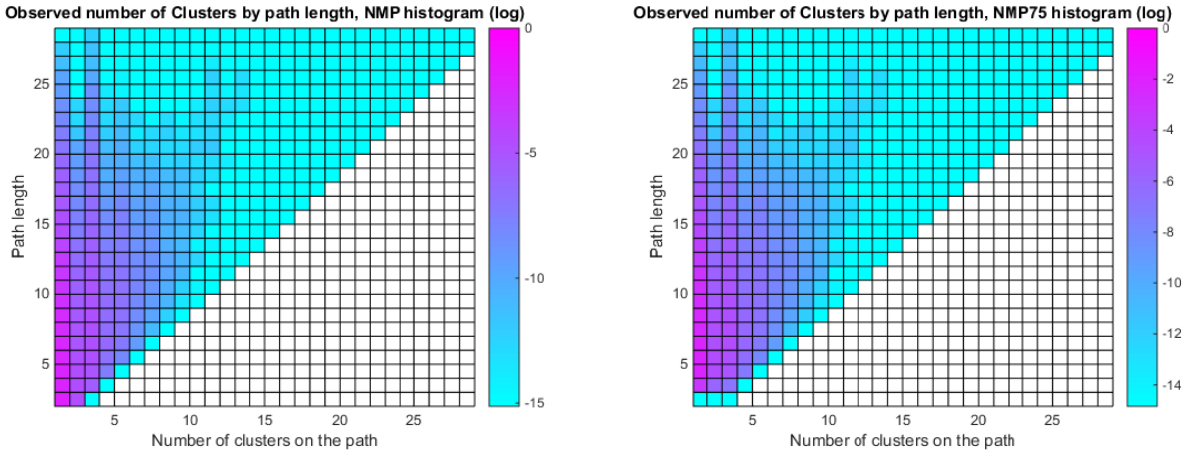
The subfigures indicate different pruning level of the NMP. The lower-left corner of each subfigure disappears when the NMP is pruned more (as in Figure 4). Each of the subfigures shows a strong concentration of paths with zero Greens or just one Green. Paths of length (about) 5 – 15 are most often found to contain relatively large numbers of Greens. Longer paths mostly occur with only one or no Green at all.

Figure 6 shows the same type of histogram, but for the number of colour-clusters. This shows a very similar picture, with a large concentration of paths that have just one or a few clusters. These

are mostly paths with very few Greens, i.e., all-Brown paths (one cluster) or paths with just one Green (either two or three clusters, depending on whether the Green occurs internal to the path). One difference that we observe between the two histograms is in the near-diagonal area for long paths, which exclusively has zeros for the cluster histogram, but some paths in the number of Greens histogram.

Figure 6 also shows that an uneven number of clusters occurs more often than an even number. For example, the cells for one or three clusters show higher frequencies than their neighbours for two and four clusters. This is expected, especially for longer path lengths. For example, for a path of length 5 with just 1 Green to have an even number of clusters (2), the Green must be either a start-node or an end-node, which is an *a priori* probability of 2/5. On the other hand, if the Green is internal to the path (the larger probability equal to 3/5), there will be an odd number of clusters (3).

This is a good illustration of the fact that we need a benchmark to interpret the histograms. This benchmark should guide us in judging whether the observed frequencies in these histograms are more or less frequent than what can be expected on the basis of the benchmark. Our next section will introduce three benchmark models, all based in probability theory. The task we set for these benchmark models is to try to predict the particular distribution of Greens that are observed in Figure 5 and Figure 6. This means that the models must be able to explain, among other things, the relative abundance of paths with few Greens (0 or 1) and few colour-clusters, and the relative abundance of middle-long paths with relatively many Greens.



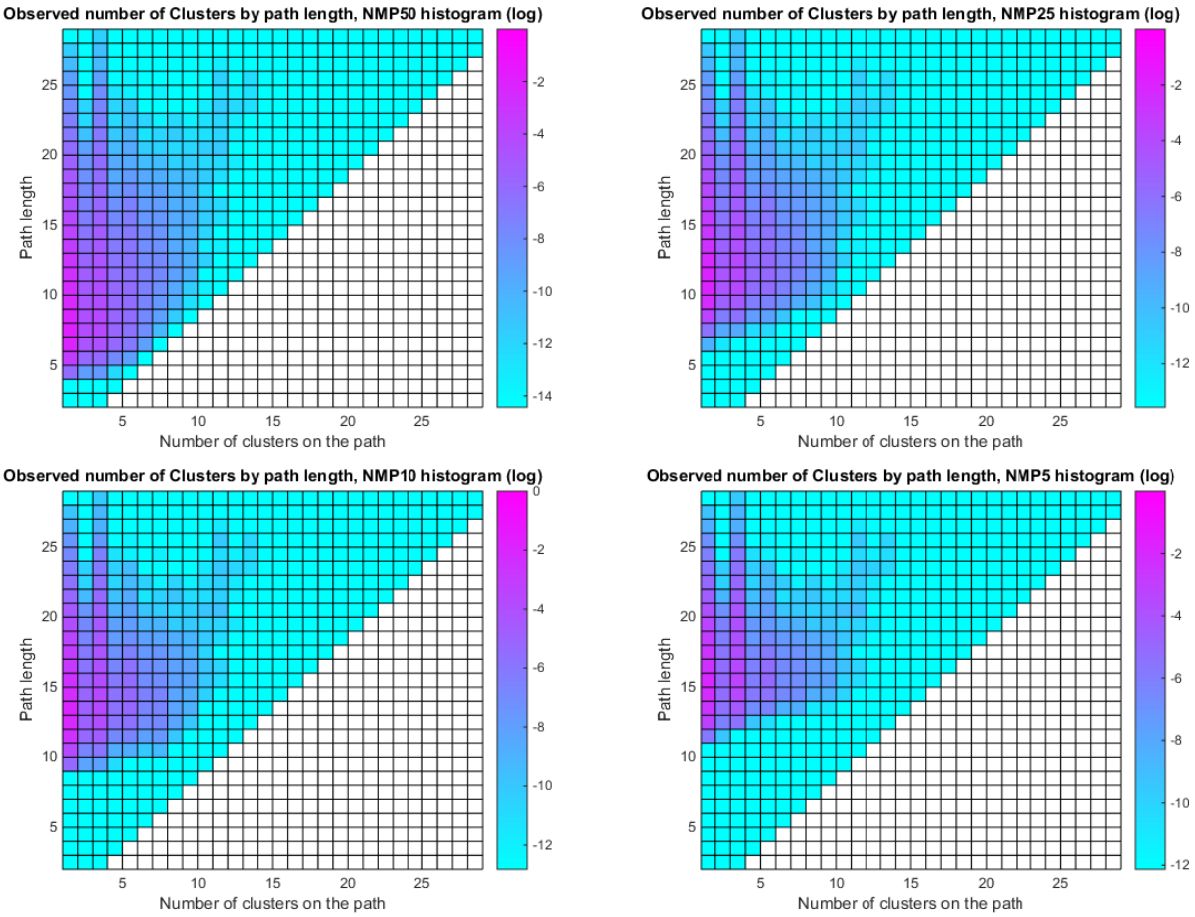


Figure 6. Histograms of observed numbers of clusters on a path (log), by path length, NMP and pruned NMPs

5. Benchmark models for trajectory formation

Do the empirical characteristics of the NMP, at different levels of pruning, represent any substantial level of concentration of green patents? This is the question that we now turn to. In order to answer it, we need a benchmark to compare the empirical data against. We will provide a number of those benchmarks, in the form of theoretical models of trajectory formation.

All models that we present will take trajectory length as given, i.e., we will use the models to generate predictions for trajectories of a specified length, and then compare these predictions to the empirically observed trajectories of the same length. The three models that we use have various degrees of Greentech concentration built into their assumptions. The comparison of the model predictions with actual data will therefore enable us to assess empirical concentration of Greentech. We will now present the three models in turn.

5.1. The Binomial model

The first benchmark model is based on the Binomial distribution. This model assumes no particularly strong concentration of Greentech across the NMP. It is built based on two probabilities, which we denote p_{GS} and p_G. p_{GS} is the probability that the start-node of the trajectory is a Green, and p_G is the probability that a non-start-node is a Green. These probabilities are observed in the NMP, and we use these observed probabilities to construct the Binomial benchmark model.

The prediction of the Binomial model for the number of Greens on a trajectory of given length can be derived directly from the analytical expression for the Binomial distribution. However, this is not possible for the prediction of the number of clusters. Therefore, our implementation of the binomial model enumerates all possible trajectories of a given length in terms of their Green/Brown content. For example, for trajectory length 3, the possible trajectories are G_B_B; B_G_B; B_B_G; G_G_B; G_B_G; B_G_G; G_G_G; B_B_B. Each one of those possibilities has an easy-to-calculate probability, for example the probability of G_G_B is equal to p_{GS} x p_G x (1 – p_G). The probability of n Greens on a trajectory of length 3 is then the sum of these probabilities over all possible trajectories with n Greens (e.g., the probability to find exactly one Green on a trajectory of length 3 is the sum of probabilities of the first three possibilities enumerated above). Similarly, the probability of m clusters on a trajectory of length 3 is the sum over all possibilities that yield m clusters (e.g., the probability to find exactly one cluster on a trajectory of length 3 is the sum of probabilities of the last two possibilities enumerated above).

5.2. The Homophily model

Our next benchmark model is an elaboration of the Binomial model that assumes some degree of concentration of the Greens and Browns in the network by the mechanism of homophily. This model assumes the same probability p_{GS} for a start-node to be a Green, but it differentiates the probability for any non-start-node to be green, depending on what colour the cited node has. Thus, we distinguish $p_{G\to G}$ and $p_{B\to G}$, which are, respectively, the probability that a node is a Green conditional on the previous node being Green, and the probability that a node is a Green conditional on the previous node being Brown. Again, these probabilities are observed in the empirical data of the NMP.

The logic of calculating the expected number of Greens or number of clusters is the same in the Homophily model as in the Binomial model, i.e., we enumerate the options and aggregate probabilities. But the outcomes are rather different between the two models. For example, for trajectory length 3, the Binomial model predicts that 17.6% of all trajectories will have exactly one Green, while the Homophily model gives 7.2%. Similarly, the Homophily model predicts that 6.7% of those trajectories will have exactly 2 clusters, while the Binomial model gives 12.6%.

These differences arise from the difference between $p_{G\to G}$ and $p_{B\to G}$. We observe a low value for $p_{B\to G}$ (0.040), while the value for $p_{G\to G}$ is much higher (0.491). These numbers imply a high degree of homophily for the Brown patents ($p_{B\to B}$ = 1 – $p_{B\to G}$ = 0.960) but less so between the Greens ($p_{G\to G}$ = 0.491). Interpreting these numbers loosely, we can say that Brown patents have a strong preference for citing other Brown patents, whereas Green patents are more or less indifferent

between citing other Green patents or citing Brown patents. This implies that the concentration levels that are observed in the Homophily model are mostly due to the Brown homophily.

5.3. The Homophily-plus-Path dependence model

Our last model again extends the previous one by assuming an additional mechanism that will likely lead to concentration. It assumes that the probability of the citing patent being a Green depends not only on the colour of the cited patent (as in the Homophily model), but also on the patents that lie before the cited patent (if any). To measure this, we count all Green patents upstream from the cited patent (i.e., the cited patent is not included in this count), and express this as a fraction of the number of upstream patents. This is called the *path dependence indicator*. For example, when considering the colour of the fifth patent following after G_B_G_G, we calculate the path dependence indicator as 2/3 (2 Greens in a total of 3 upstream patents).

In the Homophily-plus-Path dependence (HP) model, we assume that the citation probability is homophilic and path dependent, i.e., we assume $p_{G\to G} = \bar{p}_{G\to G} + a_G D$ and $p_{B\to G} = \bar{p}_{B\to G} + a_B D$, where D is the path dependence indicator as defined above, and a_G, a_B, $\bar{p}_{G\to G}$, and $\bar{p}_{B\to G}$ are parameters that must be estimated econometrically from the data.

We use a logit model to obtain these estimates. This model takes the binary variable that a citing patent is a Green patent (1 if that is the case, 0 otherwise) as the dependent variable. It has just one independent variable (in addition to a constant) and this is the path dependence indicator as explained above. We estimate this model on the sample of citation pairs that are present in the NMP, separately for the samples where the cited patent is Green and where it is Brown. For citation pairs where the cited patent is a start-node, we impute the average value of the path dependence indicator for Green or Brown patents (depending on the colour of the start-node).[8]

Table 1. Logit estimations of the parameters of path dependence model

Independent variable	*Estimate*	*Standard Error (significance)*
Sample with cited patent Green		
Path dependence	1.738	0.014 (***)
Constant	-0.669	0.007 (***)
Sample with cited patent Brown		
Path dependence	2.988	0.014 (***)
Constant	-3.327	0.003 (***)

Table 1 provides the logit estimates. We see that the path dependence variable is highly significant in both samples, and so is the constant. These estimated values are not very meaningful in themselves, as they need to be combined with the path dependence indicator values, and then

[8] We did an estimation excluding all citation pairs with cited start-nodes, and this yields very similar results.

transformed to estimates of the actual probability. To obtain a rough indication of the importance of path dependence in forming trajectories, we can calculate the implied probability under the assumption of path dependence = 0, which gives us $\bar{p}_{G\to G}$, and $\bar{p}_{B\to G}$, and compare this to the probabilities of the Homophily model ($p_{B\to G}$ and $p_{G\to G}$).[9]

In the sample where the cited patent is Brown, the probability in the Homophily model ($p_{B\to G}$) is 0.040, while we find $\bar{p}_{B\to G}$ = 0.035. Thus, on average, path dependence contributes about (0.040 – 0.035)/0.040 ≈ 16% of the "baseline" probability in the Homophily model. For the sample of Green cited patents, we find $\bar{p}_{G\to G}$ = 0.338, while $p_{G\to G}$ = 0.493. Here the difference ≈ 31%. Thus, indirect homophily in the form of path dependence explains a substantial part of the baseline homophily, especially for Green-to-Green citations.

6. Clustering and concentration in the NMP

We are now able to compare the nature of the actually observed paths in the NMP to the expected number of paths in the three benchmark models. The results are documented only for paths up to length 22, because the expected frequencies must be derived computationally, and this takes very long for longer path lengths. Also, the number of observed long paths is very low, so that the statistical comparison that we are after is hard for long paths. To save space, we do not distinguish between pruned versions of the NMP, as we already know that by and large we may achieve this by looking at longer paths.

In order to undertake the comparison between actual data and predicted frequencies, we standardise the predicted probabilities and the observed shares to unity for each path length, i.e., for each path length, we compare the expected and observed shares of paths with zero Greens, one Green, etc. in all paths of the specified length. The differences between observed and predicted are then either expressed simply as the difference, or as the difference of their logs. This distinction is made because all benchmark models predict a relative abundance of paths with few (1 – 3) Greens or colour-clusters, while paths with a high number of Greens or high number of clusters are very improbable (and infrequent). As a result of this, the difference between observed and expected frequencies has very different scales between high and low number of Greens or clusters. The log or non-log versions of the difference each bring out one of these scales in a better way.

[9] The probabilities in the Homophily model can also be estimated in a logit model, by using a model with only a constant.

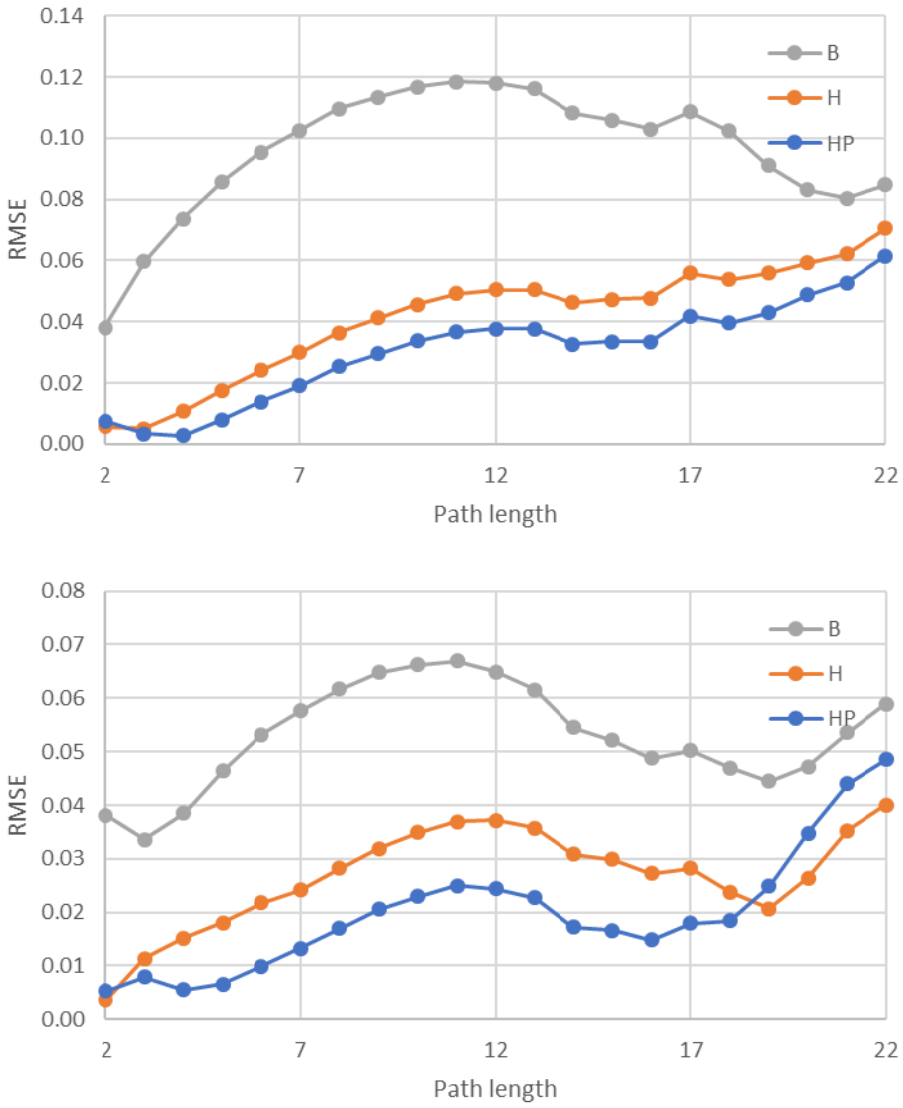


Figure 7. Root mean squared error for predicted number of Greens (top panel) and predicted number of colour-clusters (bottom panel)

We first look at a summary measure of the performance of each of the three benchmark models. This is displayed in Figure 7, which documents the root mean squared error for each path length, and for each of the three benchmark models (this is based on the non-log differences only). Several conclusions can be drawn from these figures. First, the Binomial model clearly under-performs as compared to both other models. For every path length, it predicts the number of Greens and the number of clusters worse than the two other models do. This means that the distribution of Greens, either in terms of their sheer number or in terms of their clustering on the paths of the NMP, is more concentrated than could be expected based on randomness as represented in the Binomial model. The concentration forces that are represented in the other models (homophily and path dependence) add explanatory power to the model.

Second, the HP model generally does better than the pure Homophily model, although this differs systematically with path length. For short paths (2 or 3, i.e., mainly in the bottom-25% SPNP values of the NMP), the Homophily model and the HP model perform approximately the same. For the number of Greens on a path, the HP model performs better for the entire range of paths lengths larger than 3. For the number of clusters, HP does better for paths up to length 18, after which pure Homophily does better. One may conclude from this that both concentration mechanisms, homophily and path dependence, play a significant role in predicting the concentration of Greens and Browns in the NMP.

In Figure 8, we take a more detailed view on how well the three models predict the number of Greens on a path. In this figure, we have the non-log difference on the left-hand side, and the log differences on the right-hand side. The three benchmark models are presented top-to-bottom. It is important to note that, as indicated by the colour bars with each of the figures, the scales of the differences are very different between the subfigures, especially between the log-differences, as a result of the fact that the three models have such differential levels of performance (as in Figure 7).

Focusing first on the non-log differences, we see that what dominates in this case is the prediction error for a low number of Greens. Each of the models tends to under-predict the number of paths with zero Greens, for each path length, except very short paths (2 or 3) in the case of the Homophily and HP model. The extent of under-prediction rises with path length, i.e., it is more severe for longer paths. On the other hand, the number of paths with relatively few Greens is over-predicted. This is especially the case for just one Green, although only for paths up to about 18 or 19 long. For 2, 3 or 4 Greens, over-prediction keeps occurring also for long paths. Beyond 4 or 5 Greens, the differences between observed and predicted become indistinguishable from zero on the non-log scale.

The log-difference figures on the right-hand side provide further insight into the performance of the models with respect to paths with many Greens on them. Note that for very long paths, we do not observe some of the theoretically possible values for number of Greens. These are indicated by the shade (pure cyan) that corresponds to the lowest value on the scale, to represent $-\infty$ (associated with log(0)). The impression that emerges from these plots is that for each path length, the large number of Greens is over-represented (under-predicted). But this is much less the case for the HP model than for the other two models.

As an intermediate conclusion, we may say that the Homophily and HP benchmark models predict the concentration of Greens and Browns relatively well. The actual NMP has relatively many pure Brown paths, relatively few paths with just one (or a few) Greens, and homophily and path dependence come some way towards explaining these phenomena. Thus, the Greens are somewhat concentrated on the NMP, and homophily and path dependence seem to be relevant in explaining these tendencies (even if they cannot explain it fully).

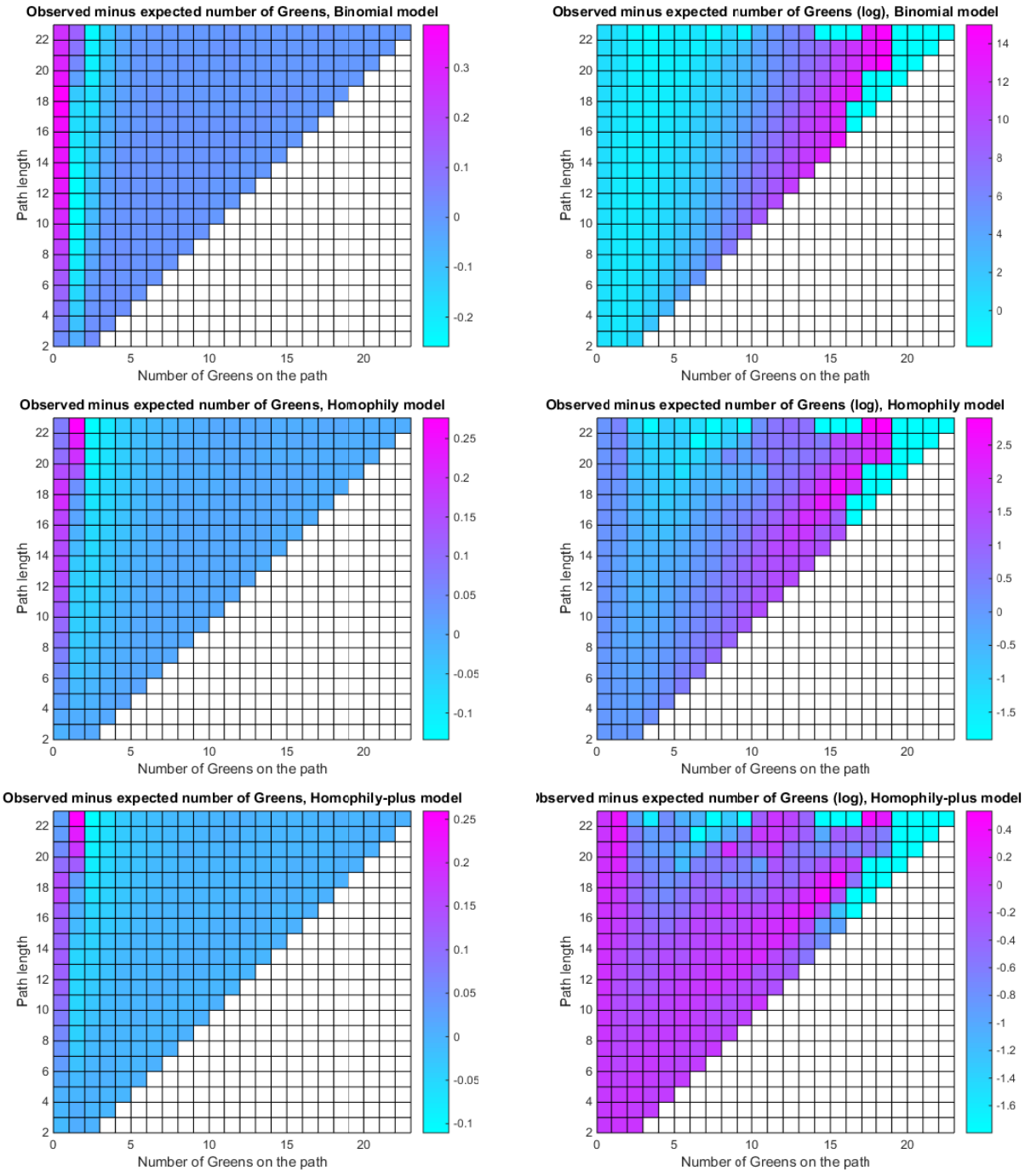


Figure 8. Observed minus expected number of Greens, by path length and by benchmark model

Figure 9 provides the comparison for observed minus predicted number of colour-clusters. This has many similarities with the previous figure, especially the under-prediction of paths with just one cluster (obviously, paths with zero Greens have just one colour-cluster). We also see under-representation in the data of paths with 3 clusters, relative to all three models, except for long path lengths. For path length 11 onwards, we see complete absence of paths with many clusters, i.e., the entire upper-right corner has zero observed paths, which implies under-representation in the

actual data (this is most obvious in the log-plots). Also, we observe relatively good performance of the homophily-plus model, at least for paths that are not very long (this is fully in line with Figure 7). Thus, our earlier conclusions on the importance of homophily and path dependence for clustering of Greens and Browns in the NMP are essentially confirmed by the results in Figure 9.

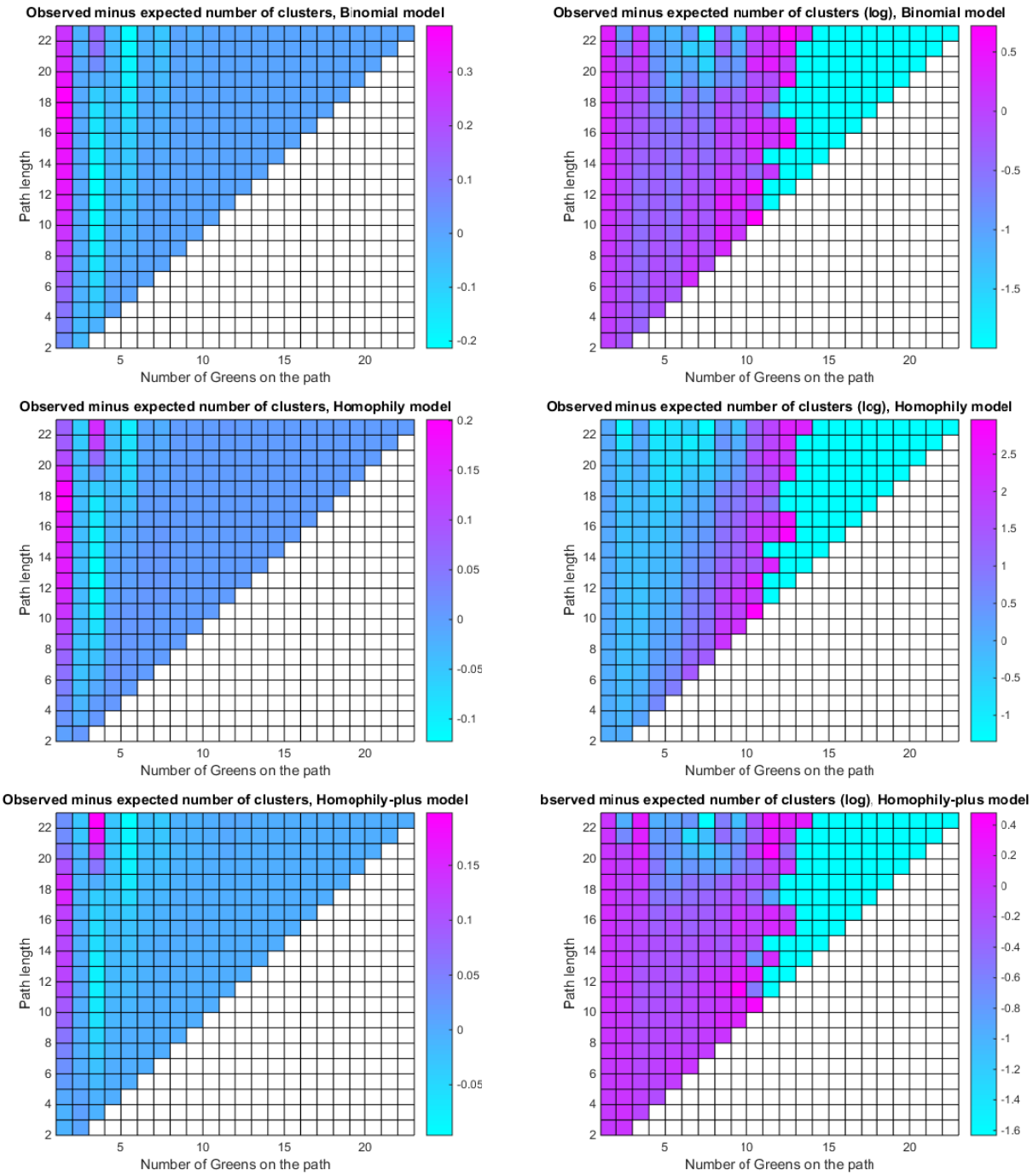


Figure 9. Observed minus expected colour-clusters, by path length and by benchmark model

6.1. What drives homophily?

As a final step in our analysis we implement an alternative regression model to the one in Table 1, by including a number of control variables drawing on the literature in innovation studies. Just as the homophily-plus-path dependency model of Table 1 endogenises a part of the observed homophily in the pure homophily model, this extended model potentially endogenises a larger part of observed homophily, because it takes into account a larger set of variables than just path dependence. Admittedly, we do not have a proper theory of homophily, so we use variables that are common in the patent citation literature (e.g., Criscuolo and Verspagen, 2008).

Our first two variables control for timing. We have the filing year of the citing patent, and the lag in years between cited and citing patent. The fling year is expressed as a fraction between zero and one, where zero indicates the year 1978, and one is 2018. The citation lag is also expressed as a fraction, with zero indicating zero years and one indicating 40 years.

In their most basic form, all other new variables that we add are defined as a binary dummy variable, although some of these may take a non-binary value due to fractional counting. Three of those variables refer to the citation type. The variable called *Negative citation* is equal to one if the citation is deemed (by the examiners) as either an X, a Y or an I type citation. All these citation types somehow prejudice the citing patent as not sufficiently novel (as compared to the cited patent). The variable *Applicant citation* is one if the citation was added by the applicant (D type citation). The last of the citation type variables (*Family-link*) is a dummy variable indicating if we added this citation as a family-relationship (see the description of our total citation network above).

The next two variables capture geography. We have one dummy that is one if the cited and citing patent are from the same country, as indicated by inventor addresses. Because inventor countries are counted fractionally, this variable generally takes non-binary values (it is bounded between zero and one, however). The other geographical variable is one if the citing and cited country are geographical neighbours. Again, this is counted fractionally, yielding values for the variable between zero and one. Finally, we have a dummy variable that indicates whether the cited and citing patent are from the same NACE sector. We use the PatStat concordance to NACE sectors, and again this is counted fractionally, yielding values between zero and one.

The estimation results of the extended model are in Table 2. Besides the parameter estimates and their significance, this table also contains three extra columns, which provide information on the impact of the variable on observed homophily. The column that is labelled “Max effect” documents the (marginal) effect that is associated to an increase of the variable from zero to one. This is evaluated taking all other variables at their sample mean. The sample mean of each variable is also documented, along with its standard deviation.

Looking at Green-to-Green citations first (top part of the table), we see that the maximum effect of the path dependence variable and the family-link both have large positive maximum effects. The citation lag and the different-NACE variables have relatively large negative effects (i.e., they decrease Green-to-Green. All of these are based on highly significant parameter estimates. Thus, belonging to the same patent family, belonging to the same NACE sector and a small citation lag seem to be the main driving factors in Green-to-Green homophily.

For Brown-to-Green citations, we must keep in mind that this type of citation has a high degree of homophily ($p_{B \to B} = 0.96$ or $p_{B \to G} = 0.040$ in the pure Homophily model). Therefore, the threshold for

contributing significantly to homophily is much lower in this case. The path dependency variable stands out with a large potential impact, but note that in this sample, the average value of path dependency is only 2.9% (vs 39.6% in the Green-to-Green sample). There are very few citation pairs in this sample with path dependency indicator equal to one, but the few that have this have a large bonus probability have a Green citing patent. Other influential variables in this sample are the filing year of the citing patent, the citation lag and different NACE sectors (all of these have a positive impact, i.e., they decrease the degree of Brown-to-Brown homophily) and the family-link (negative impact, i.e., this increases Brown-to-Brown homophily).

Table 2. Logit estimation of the parameters of the extended homophily-path dependence model

Explanatory variable	Estimate	Standard error (significance)	Max effect	Mean	Std dev
Sample with cited patent Green					
Path dependence	1.711	0.021 (***)	0.396	0.390	0.323
Filing year citing	0.039	0.032	0.010	0.728	0.209
Citation lag (years)	-0.558	0.042 (***)	-0.135	0.178	0.153
Negative citation	0.069	0.015 (***)	0.017	0.265	0.441
Applicant citation	-0.012	0.038	-0.003	0.030	0.170
Family-link	2.777	0.097 (***)	0.440	0.038	0.190
Identical country	0.080	0.015 (***)	0.020	0.373	0.473
Neighbouring countries	0.044	0.020 (**)	0.011	0.139	0.334
Different NACE sector	-1.211	0.020 (***)	-0.293	0.354	0.343
Constant	-0.213	0.027 (***)			0.447
Sample with cited patent Brown					
Path dependence	2.973	0.020 (***)	0.310	0.029	0.100
Filing year citing	0.680	0.022 (***)	0.017	0.664	0.220
Citation lag (years)	0.468	0.026 (***)	0.014	0.190	0.155
Negative citation	-0.051	0.010 (***)	-0.001	0.250	0.433
Application citation	-0.244	0.027 (***)	-0.006	0.041	0.198
Family-link	-5.583	0.299 (***)	-0.032	0.029	0.169
Identical country	-0.128	0.010 (***)	-0.003	0.377	0.474
Neighbouring countries	-0.111	0.013 (***)	-0.003	0.139	0.336
Different NACE sector	0.576	0.011 (***)	0.017	0.314	0.346
Constant	-4.129	0.016 (***)			0.016

Overall, these estimation results confirm the relevance of the path dependency mechanism as an additional factor to pure homophily. They also point to several other factors influencing homophily in the citation network, such as intra-NACE sector increases both Brown-to-Brown and Green-to-Green homophily, citation type (prejudicing novelty and applicant citations), and the timing of the citation. Geographic distance does not seem to have a large impact.

7. Conclusions

We introduced a method that uses a (very) large patent citation network to extract a collection of technological trajectories that are aimed at describing the global main technological trends over the last decades. The method yields a so-called network of main paths (NMP), which consists of overlapping paths that represent the trajectories that represent large technology flows, as represented by patent citations. We characterised each patent on the NMP as either Green (contributing to the mitigation of greenhouse gas emissions) or Brown (non-Green). We propose that the NMP and the Green/Brown representation of its nodes can be used to represent the nature of the macro-technology field of Greentech.

In terms of the content of Greentech, our main finding is that Green patents are rather concentrated in the NMP, i.e., we find relatively many paths that have either fewer Greens than expected (e.g., zero Greens, or all Brown paths), or more Greens than expected; and we find more paths with relatively few colour-clusters. These findings are based on a theoretical model that predicts the statistical distribution of the number of Greens and the number of colour-clusters over paths of a fixed length, i.e., we find a stronger concentration of Greens than this model predicts.

We also have two alternative models, which introduce two separate mechanisms that will lead to concentration of Greens. We find that these models, especially the one that includes both mechanisms, predict the data in the NMP relatively well. The concentration-mechanisms that these models include are homophily, which we define as the tendency of Green patents to cite other Green patents, and the tendency of Brown patents to cite other Brown patents; and path dependence, which we define as the colour of impact of upstream (occurring before the cited patents) on whether or not a citation is made by a Green patent. We find that the more Green patents lie upstream of a citation, the larger is the probability that the citing patent is Green.

This implies that the macro-technology field of Greentech is characterised, at least to some extent, by a specific knowledge base of its own, that does not apply in the overwhelmingly Brown parts of the NMP. In other words, the development of Greentech is a matter of developing and applying a specific knowledge base, rather than of "greening" Brown environments without specific knowledge of Greentech. To the extent that this is reflected in homophily, it is mainly the result of Brown-to-Brown homophily, which we observe to be very strong, rather than of Green-to-Green homophily, which is weaker (the tendency of Green patents to cite Green patents is weaker than the tendency of Brown patents to cite Brown patents).

The concentration of Green (and Brown) patents that results from homophily and path dependence has implications for policy makers who want to "green" the economy. It means that for green technology to emerge at a substantial scale, there needs to be investment in the green knowledge base. This will be associated with fixed costs, e.g., investment in academic study programmes, public labs, etc. As individual firms may not be able to make these investments, there may be coordination failure that warrants public policy. At a much more down-to-earth level, we imagine that knowledge about the structure of our NMP may also help patent offices to improve the algorithms used to implement Y02 tagging.

The dual purpose of the analysis in this paper was to present the method, and to apply it. With the method and the database available, applications to other (macro-)fields of technology are also possible. But our analysis also leaves open research questions in terms of Greentech. For example, we have been unable to touch upon the possibility of subdividing Greentech into more specific

fields. The Y02 tagging system that we applied also defines eight subclasses, which can provide more information about the concentration of specific types of Greentech over the NMP. This could be researched using the same type of benchmark models as we applied.

It will also be useful to investigate the explanatory factors for homophily and path dependence in Greentech citation networks in a more detailed way. Our final section provided some exploratory evidence on this matter, but it is beyond the scope of this paper to develop and test a proper theory of homophily and path dependence in citation networks.

References

Aharonson, B.S., Schilling, M.A., 2016. Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution, Research Policy, vol. 45, pp. 81–96.

Batagelj, V. (2003), Efficient Algorithms for Citation Network Analysis, mimeo, reprinted in: V. Batagelj, P. Doreian, A. Ferligoj, N. Kejzar: Understanding Large Temporal Networks and Spatial Networks. Wiley, 2014

Billinger, S. Stieglitz, N. and T.R. Schumacher, 2014, Search on Rugged Landscapes: An Experimental Study, Organization Science 25(1): 93-108.

Criscuolo, P. and B. Verspagen, 2008, 'Does it matter where patent citations come from? Inventor vs. examiner citations in European patent, Research Policy, vol. 37, pp. 1892-1908.

Dosi, G. (1982) Technological paradigms and technological trajectories. Research Policy, 11: 147–162.

Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strategic Management Journal, vol. 25, pp. 909–928.

Hummon, N.P., Doreian, P. (1989) Connectivity in a citation network: The development of DNA theory. Social Networks, 11: 39-63.

Kauffman, S.A. 1993. The origins of order: self-organization selection in evolution. Oxford University Press, Oxford, U.K.

Kauffman, S.A., Lobo, J., Macready, W.G., 2000. Optimal search on a technology landscape. Journal Economic Behavior and Organization, vol. 43, pp. 141–166.

Levinthal, D.A., 1997. Adaptation on Rugged Landscapes, Management Science, vol. 43, pp. 934-950.

Liu, J. S., Lu, L. Y. Y. (2012) An Integrated Approach for Main Path Analysis: Development of the Hirsch Index as an Example. Journal of the American Society for Information Science and Technology, 63:528-542.

McPherson, M.; Smith-Lovin, L.; Cook, J. M. (2001). "Birds of a Feather: Homophily in Social Networks". Annual Review of Sociology. 27: 415–444.

Mina, A., Ramlogan, R., Tampubolon, G., Metcalfe, J.S. (2007) Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. Research Policy, 36: 789-806.

Nomaler, Ö. & B. Verspagen, 2016, River deep, mountain high: of long run knowledge trajectories within and between innovation clusters, Journal of Economic Geography, 16, pp. 1259-1278

Nuvolari, A. Verspagen, B. (2009) Technical choice, innovation, and British steam engineering, 1800–50. Economic History Review, 62: 685-710.

Sahal, D. (1981) Patterns of Technological Innovation (Addison-Wesley).

Stuart, T.E., Podolny, J.M., 1996. Local search and the evolution of technological capabilities. Strategic Management Journal, vol. 17, pp. 21–38.

Trajtenberg, M. and A. Jaffe, 2002, Patents, Citations, and Innovations. A Window on the Knowledge Economy, Cambridge, MA: MIT Press

Triulzi, G. (2015), Looking for the right path. Technology Dynamics, Inventive Strategies and Catching-up in the Semiconductor Industry, PhD Thesis, UNU-MERIT, Maastricht, https://www.merit.unu.edu › training › theses › triulzi_giorgio

Verspagen, B. (2007) Mapping Technological Trajectories as Patent Citation Networks: a Study on the History of Fuel Cell Research, Advances in Complex Systems, vol. 10: 93-115.

The UNU-MERIT WORKING Paper Series

2019-01 *From "destructive creation" to "creative destruction": Rethinking Science, Technology and innovation in a global context* by Luc Soete
2019-02 *Do young innovative companies create more jobs? Evidence from Pakistani textile firms* by Waqar Wadho, Micheline Goedhuys and Azam Chaudhry
2019-03 *What gains and distributional implications result from trade liberalization?* by Maria Bas and Caroline Paunov
2019-04 *FDI, multinationals and structural change in developing countries* by André Pineli, Rajneesh Narula and Rene Belderbos
2019-05 *The race against the robots and the fallacy of the giant cheesecake: Immediate and imagined impacts of artificial intelligence* Wim Naudé
2019-06 *The middle-technology trap: The case of the automotive industry in Turkey* by Ibrahim Semih Akçomak and Serkan Bürken
2019-07 *The impact of a mathematics computer-assisted learning platform on students' mathematics test scores* by Marcelo Perera and Diego Aboal
2019-08 *Health insurance and self-employment transitions in Vietnam* by Nga Le, Wim Groot, Sonila M. Tomini and Florian Tomini
2019-09 *Knowledge economy and economic development in the Arab region* by Samia Mohamed Nour
2019-10 *Migration of higher education students from the North Africa region* by Samia Mohamed Nour
2019-11 *Job automation risk, economic structure and trade: a European perspective* by Neil Foster-McGregor, Önder Nomaler an Bart Verspagen
2019-12 *The breadth of preferential trade agreements and the margins of* exports by Rod Falvey and Neil Foster-McGregor
2019-13 *What a firm produces matters: diversification, coherence and performance of Indian manufacturing firms* by Giovanni Dosi, Nanditha Mathew and Emanuele Pugliese
2019-14 *Brazilian exporters and the rise of Global Value Chains: an empirical assessment* by Caio Torres Mazzi
2019-15 *How has globalisation affected the economic growth, structural change and poverty reduction linkages? Insights from international comparisons* by Aradhna Aggarwal
2019-16 *R&D, innovation and productivity* by Pierre Mohnen
2019-17 *Domestic intellectual property rights protection and exports: Accessing the credit channel* by Gideon Ndubuisi
2019-18 *The role of early-career university prestige stratification on the future academic performance of scholars* by Mario Gonzalez-Sauri and Giulia Rossello
2019-19 *The employment impact of product innovations in sub-Saharan Africa: Firm-level evidence* by Elvis Korku Avenyo, Maty Konte and Pierre Mohnen
2019-20 *Embodied and disembodied technological change: the sectoral patterns of job-creation and job-destruction* by G. Dosi, M. Piva, M. E. Virgillito and M. Vivarelli
2019-21 *Can we have growth when population is stagnant? Testing linear growth rate formulas and their cross-unit cointegration of non-scale endogenous growth models* by Thomas H.W. Ziesemer

2019-22 *Technical progress and structural change: a long-term view* by Alessandro Nuvolari and Emanuele Russo

2019-23 *No evidence of an oil curse: Natural resource abundance, capital formation and productivity* by Mueid al Raee, Denis Crombrughe and Jo Ritzen

2019-24 Far from random? The role of homophily in student supervision by Giulia Rossello and Robin Cowan

2019-25 *Semi-endogenous growth models with domestic and foreign private and public R&D linked to VECMs* by Thomas H. W. Ziesemer

2019-26 *Characterizing growth instability: new evidence on unit roots and structural breaks in long run time series* by Emanuele Russo, Neil Foster-McGregor and Bart Verspagen

2019-27 *Measuring attitudes on gender equality and domestic violence in the Arab context: The role of framing, priming and interviewer effects* by Ann-Kristin Reitmann, Micheline Goedhuys, Michael Grimm and Eleonora E. M. Nillesen

2019-28 *Imported intermediates, technological capabilities and exports: Evidence from Brazilian firm-level data* by Caio Torres Mazzi and Neil Foster-McGregor

2019-29 *Japan's productivity and GDP growth: The role of GBAORD, public and foreign R&D* by Thomas Ziesemer

2019-30 *The decline in entrepreneurship in the West: Is complexity ossifying the economy?* by Wim Naudé

2019-31 *Modern industrial policy in Latin America: Lessons from cluster development policies* by Carlo Pietrobelli

2019-32 *Testing the employment and skill impact of new technologies: A survey and some methodological issues* by Laura Barbieri, Chiara Mussida, Mariacristina Piva and Marco Vivarelli

2019-33 *The Potential for innovation in mining value chains. Evidence from Latin America* by Michiko Iizuka, Carlo Pietrobelli and Fernando Vargas

2019-34 *Enforcing higher labour standards within developing country value chains: Consequences for MNEs and informal actors in a dual economy* by Rajneesh Narula

2019-35 *A comment on the multifaceted relationship between multinational enterprises and within-country inequality* by Rajneesh Narula and Khadija van der Straaten

2019-36 *The effects of R&D subsidies and publicly performed R&D on business R&D: A survey* by Thomas H.W. Ziesemer

2019-37 *Does it pay to do novel science? The selectivity patterns in science funding* by Charles Ayoubi, Michele Pezzoni and Fabiana Visentin

2019-38 *Regulation and innovation under Industry 4.0: Case of medical/healthcare robot, HAL by Cyberdyne* by Michiko Iizuka and Yoko Ikeda

2019-39 *The future of work and its implications for social protection and the welfare state* by Franziska Gassmann and Bruno Martorano

2019-40 *Return, circular, and onward migration decisions in a knowledge society* by Amelie F. Constant

2019-41 *Mining and quality of public services: The role of local governance and decentralisation* by Maty Konte and Rose Camille Vincent

2019-42 *Corruption and tax morale in Africa* by Amadou Boly, Maty Konte and Abebe Shimeles

2019-43 *Remittances and Bribery in Africa* by Maty Konte and Gideon Ndubuisi

2019-44 *Women's Political and Reproductive Health Empowerment in Africa: A literature review* by Maty Konte, Victor Osei Kwadwo and Tatenda Zinyemba

2019-45 *The effect of public funding on scientific performance: A comparison between China and the EU* by Lili Wang, Xianwen Wang, Fredrik Niclas Piro and Niels J. Philipsen

2019-46 *Credit constraints and trade performance: Does trust-based social capital matter?* By Gideon Ndubuisi and Maty Konte

2019-47 *The impact of mission-oriented R&D on domestic and foreign private and public R&D, total factor productivity and GDP* by Thomas H.W. Ziesemer

2019-48 *Confronting the challenge of immigrant and refugee student underachievement: Policies and practices from Canada, New Zealand and the European Union* by Özge Bilgili, Louis Volante, Don A. Klinger and Melissa Siegel

2019-49 *Stuctural transformation in general equilibrium* by Alessio Moro and Carlo Valdes

2019-50 *Systemising social innovation initiatives and their regional context in Europe* by René Wintjes, Nordine Es-sadki and Ad Notten

2019-51 *Greentech homophily and path dependence in a large patent citation network* by Önder Nomaler & Bart Verspagen