# EPO ARP Project

## *Mapping Greentech Trajectories in the Universal Network of Patent Citations*

# Final Summary Report & Papers

Önder Nomaler & Bart Verspagen

UNITED NATIONS UNIVERSITY

UNU-MERIT

In this project, that has been financially-supported by the Academic Research Program (ARP) of the European Patent Office (EPO), following the evolutionary concept of technological trajectories (Dosi, 1982[1]) and building upon the literature that proposes computational tools to identify such trajectories in patent citation networks (e.g., Hummon & Doreian 1989[2], Nomaler & Verspagen 2016[3]), our primary objective was to map the entire contents of EPO patents and patent citations in PatStat. The resulting map of trajectories which we may refer to as a 'Universal Network of Main Paths' (UNMP) can be interpreted as a (historical) cartography of all main technological trends ('trajectories') and their interactive relationships of a cumulative nature.

While such a universal cartography of all technologies altogether will allow researchers and other stakeholders to understand the main trends of progress in any selected field of technology in relation to that in all other related fields, a second pillar of our project was to provide a 'proof-of-concept' in a particular 'macro' field of technology (i.e., a set of distinct technology fields related by a coherent overall goal). Given the enormous societal relevance, the macro field of technology we opted for was the so-called 'Greentech,' as captured by the Y02 tagging scheme[4] that adds to the existing CPC classification system a new set of codes indicative of a patent's contribution to climate change mitigation.

[1] Dosi, G., (1982), "Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change", Research Policy, vol. 11(3), pp.147-162.

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

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

[4] The much-appreciated outcome of collaboration between EPO, United Nations Environmental Program (UNEP) and the International Centre on Trade and Sustainable Development (ICTSD).

The main motivation of our second pillar was to provide insights into the nature of climate change mitigating technologies. How these technologies emerge, e.g., in terms of which technologies are at its basis, whether they emerge as a coherent set of specialized technologies developing within their own internal dynamics, or whether they emerge as adaptations at the fringe of non-cleantech? We also aimed at looking at the role of path dependency in the field of cleantech and whether Greentech trajectories develop in geographical isolation, or rather as the outcome of collective international effort.

Shortly, our goal was both to contribute to a more sophisticated use of large patent databases by providing new uses for the data, and to a better understanding of the technologies that will be needed for stimulating global sustainable development. As of the end date of our project, we can summarize our achievements as follows.

1) On the basis of several improvements on existing methods, we developed a new algorithm[5] that extracts the UNMP out of any given PatStat edition.
2) In addition to the extraction of the UNMP, our algorithm also introduces a new indicator of 'significance' (e.g., value) for an individual patent on the basis of its centrality in the universal citation network.
3) The algorithm was run on the 2019 spring edition of PatStat. The resulting UNMP of the citation network has been made publicly available as a database (comma-delimited text file which can be built into a relational table under any database engine. See the Appendix 1 for a snapshot). The database contains all information on application id of the UNMP nodes (i.e., patent documents), all trajectories the node belongs to, and which position it takes on each trajectory. The database, which can be linked to PatStat by patent's unique application id (appln_id) to obtain other patent meta-data information, can be downloaded at https://dataverse.nl/dataset.xhtml?persistentId=hdl:10411/ZDCQY3.
4) The analysis of the UNMP in the particular context of Greentech (and in relation to non-Greentech which we conveniently refer to as Browntech) led to two research papers. The papers are downloadable respectively at https://www.merit.unu.edu/publications/wppdf/2019/wp2019-052.pdf and https://www.merit.unu.edu/publications/wppdf/2021/wp2021-005.pdf.
5) In a nutshell, our findings strongly indicate that that, especially for policy-related concerns, progress in Greentech cannot be understood independently of developments in non-Greentech technologies

Through the rest of this document, we provide a summary of our main findings, the conclusions and the policy implications that follow.

For the convenience of interested readers, the two research papers are also attached to the end of this document.

[5] The algorithm was implemented exclusively in Microsoft® T-SQL variant of the Structured Query Language.

## 1. Research paper #1: *Greentech homophily and path dependence in a large patent citation network*

The first paper introduces in detail our (improved) method of identifying the UNMP of all patents applied for at EPO. Our algorithm identified about 3.7 million significant trajectories[6] each connecting some subset of a total of about 2.8 million patents. The length of these trajectories varies between 2 and 28.[7] On only 18% of the trajectories we find at least one (or more) Green patent.[8] We refer to this subset (of about 664 thousand) as the 'Green trajectories' although only a fraction of them (about 44 thousand) are purely Green. That is, the majority of our Green trajectories contain some mixture of both Green and Brown patents[9] showing up in varying proportions (see Appendix 2 for a complete breakdown) and many different patterns of order of appearance. For example, in the total of 357 thousand trajectories of length 6, we find that 292 thousand are purely Brown, about 3 thousand are purely green, while there are 1,731 where three Brown patents are followed by three Green patents, 755 where three Green patents are followed by three Brown patents and 525 where we observe three Green patents with three Brown ones in other (less polarized) order of appearance.

How will these observations help us understand the extent to which we could refer to Greentech as a coherent/specialized set of technologies that develop within their own technological dynamics independently of Browntech? On the one extreme, if we observed only purely Green and purely Brown trajectories, we could speak of complete specialization. On the other extreme (and given that only 6.9% of all EPO patents are tagged as Green), if we observed that through all trajectories the propensity of a patent to cite a Green patent was 6.9% independently of the color of the citing patent, we would easily conclude that Greentech and Browntech were inseparable, mutually co-evolving domains. However, we did not observe these two extremes. The actual case was somewhere in between.

Accordingly, we ventured into the question of the 'coherence' of Greentech on the basis of the concept of 'homophily'[10] which we define as the tendency of Green patents to follow other Green patents immediately through a trajectory, and the tendency of Brown patents to follow other Brown patents immediately. We observed that only about 3% of the citations of Brown patents go to Green patents (which is even less than 6.9%, indicating strong Brown-to-Brown homophily), while about 50% of the citations of Green patents go to Green patents and the other half to Brown (thus neither Green homophily nor heterophily, but absolute indifference). Accordingly, we developed a number of statistical models incorporating the notion of homophily and tested their relative explanatory power (against the purely

[6] The original network has 9,090,460 citation links, through which one can enumerate possibly billions of trajectories. By pruning 5.5 million of these citation links (ending up with only 3,494,708), while keeping all the patents of the original network, our algorithm extracts out of the billions only the 'most significant' 3.7 million trajectories.

[7] 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.

[8] i.e., a patent that is tagged as Green by a Y02 CPC code.

[9] i.e., patents that are not tagged as Green by a Y02 CPC code.

[10] We adopt the concept, which is a form of 'preferential attachment' from the literature on 'social networks.'

stochastic model of no homophily[11]) in explaining the patterns of Green-Brown co-presence in all trajectories we identified. The homophily model was by far the winner.

The baseline homophily model was enriched further by explaining the determinants of homophily conditional on the order of appearance (of Green and Brown) on individual trajectories. Our first candidate was path dependence which we define as the color of impact of upstream (occurring before the cited patents) on whether or not a citation is made by a Green patent. We found that the higher the share of the Green patents that lie upstream of a citation, the larger was the probability that the citing patent is Green, indicating that that trajectories that start off with green patents tend to remain green and vice versa.

Furthermore, we observe that various dimensions of proximity (be it technical, sectoral, intertemporal or geographical) significantly contribute to homophily. For example, we observe that the lower the time difference between the application dates of a pair of patents that show up immediately next to each other on a trajectory, the higher the likelihood that they are both Green, or both Brown. Similarly, if two consecutive patents on a trajectory have inventors from the same (or immediately neighboring countries), and/or if they are associated with the same NACE sector[12], they are more likely to have the same color. This implies that to some extent, but only to some extent, there is sectoral and geographical coherence in which progress in Greentech is decentrally organized.

These findings imply that the macro-technology field of Greentech is characterized, at least to some extent, by a specific knowledge base of its own that does not apply in the overwhelmingly Brown parts of the UNMP. In other words, the development of Greentech has been a matter of developing and applying a specific knowledge base, rather than of "greening" Brown domains 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 follow Green patents is weaker than the tendency of Brown patents to follow 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 programs, public labs, etc. As individual firms may not be able to make these investments, there may be coordination failure that warrants public policy.

As a side note, we also conjecture that knowledge about the structure of our UNMP may also help patent offices eventually to improve the algorithms used to implement Y02 tagging, and that our new patent significance indicator may introduce new perspectives for the literature that aims at estimating 'patent value' on the basis of meta-data.

[11] i.e., the model that is based on the 6.9% citation propensity to a Green patent independently of the color of the citing patent, thus the baseline model that presumes no coherence in Greentech.

[12] In terms of the concordance information provided in PatStat.

## 2. Research paper #2: *Patent Landscaping using 'green' Technological Trajectories*

In our second paper, we shift our attention from homophily (defined in terms of the Green-Brown dichotomy and the complementarities thereof) to a detailed analysis of the patterns of heterophily. In other words, we ask, to what extent Greentech development benefits from geographical and technological (i.e., as captured by IPC codes) diversity.

Figure 1 below shows an example of such diversity on an actual trajectory that spans the time period 1979 to 2012, comprising 12 patents (4 Brown followed by 8 Green patents) as contributed by inventors from six countries in seven different (4 digit) IPC codes. Looking at the patent titles, we observe the evolution of the usage of continuous variable transmission (CVT) systems, into electric and hybrid vehicles. It is indeed well-known that CVT systems that were originally developed (around late 1950s) for vehicles with a single combustion engine (clearly, Browntech), have provided the basis for the design of more sophisticated systems (e.g., Electric Variable Transmission, e-CVT) that were able to apply power from multiple sources of actuation to one output, such as a hybrid vehicle (Greentech) which has both a combustion engine and an electric motor (and in some cases also a flywheel). This trajectory nicely captures a snapshot of this main trend.

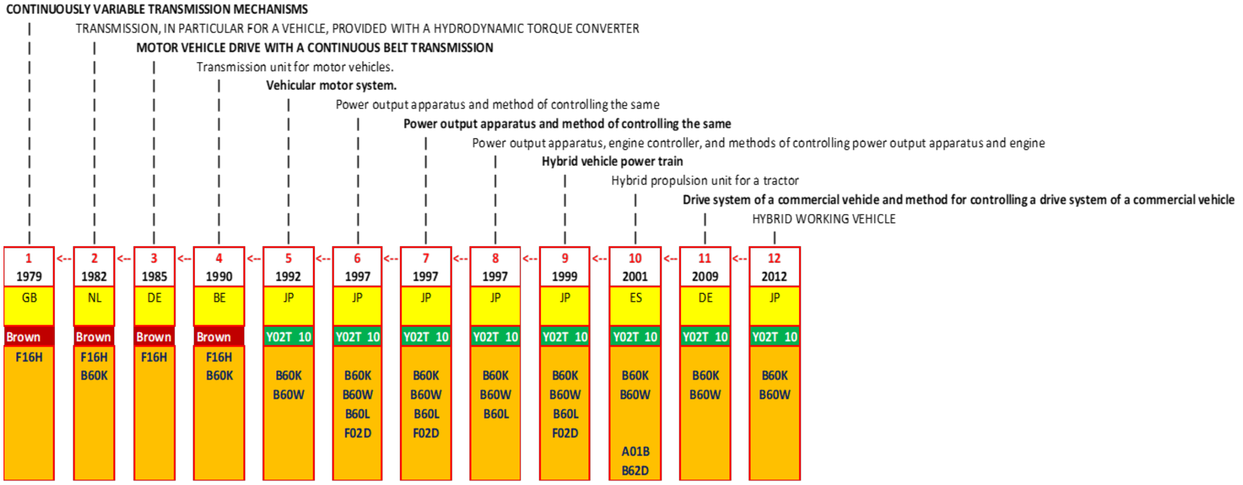

CONTINUOUSLY VARIABLE TRANSMISSION MECHANISMS
TRANSMISSION, IN PARTICULAR FOR A VEHICLE, PROVIDED WITH A HYDRODYNAMIC TORQUE CONVERTER
MOTOR VEHICLE DRIVE WITH A CONTINUOUS BELT TRANSMISSION
Transmission unit for motor vehicles.
Vehicular motor system.
Power output apparatus and method of controlling the same
Power output apparatus and method of controlling the same
Power output apparatus, engine controller, and methods of controlling power output apparatus and engine
Hybrid vehicle power train
Hybrid propulsion unit for a tractor
Drive system of a commercial vehicle and method for controlling a drive system of a commercial vehicle
HYBRID WORKING VEHICLE
1 1979 GB Brown F16H
<--
2 1982 NL Brown F16H B60K
<--
3 1985 DE Brown F16H
<--
4 1990 BE Brown F16H B60K
<--
5 1992 JP Y02T 10 B60K B60W
<--
6 1997 JP Y02T 10 B60K B60W B60L F02D
<--
7 1997 JP Y02T 10 B60K B60W B60L F02D
<--
8 1997 JP Y02T 10 B60K B60W B60L
<--
9 1999 JP Y02T 10 B60K B60W B60L F02D
<--
10 2001 ES Y02T 10 B60K B60W A01B B62D
<--
11 2009 DE Y02T 10 B60K B60W
<--
12 2012 JP Y02T 10 B60K B60W

**Figure 1. An actual example of a green trajectory. An arrow indicates a direct citation.**

In order to analyze the patterns in diversity in the entire set of Green trajectories, we build a network from the database of green technological trajectories. The network is based on 'co-occurrence' on the green trajectories. The nodes in our network are either combinations of green/brown and 4-digit IPC code, or combinations of green/brown and country of origin of the patent. In the first case, a node could, for example, be brown patents in class F16H, or green patents in class F01D. In the second case, we could have nodes like green patents from Germany, or brown patents from Japan. The networks are visualized using the LinLog method. In both cases (IPC classes or countries), we obtain sensible maps of the green technological landscape, which outline the relatedness between green technology sub-parts.

We argue that our landscaping method based on relations between technological fields that are extracted from technological trajectories fits the aim of outlining main technological trends better than methods that are merely based on individual patents or patent citation pairs. The reason is that the technological trajectories in our method are aimed at summarizing technological trends, and hence they are the most logical building blocks for mapping these trends. We look at the maps that we build as a proof-of-concept, and suggest that future patent landscaping work considers using trajectory-based metrics.

A common feature between the network based on IPC codes and the one based on countries is that brown nodes play a very important role in the network. In both cases, the nodes that are most central, are the brown nodes. This is in line with conclusions from our first paper of this project, and implies that progress in Greentech cannot be understood independently of developments in non-Greentech technologies.

In the network that uses IPC codes (see Figure 2 below), we observe a number of very broad fields that transcend Greentech as such, as well as technological areas that are clearly key to Greentech. The main examples of the first type of fields (general) are ICT and electrical, and health and medical. These are broad technological areas that serve goals that are not necessarily related to (the popularly-known domains of) Greentech, but they show up as major parts of the Greentech field in our maps.

The IPC-based map is broadly divided in one half that contains electricity-based technologies, and another half that has no direct relations to electricity. The electricity-based part includes the large ICT and electrical cluster, but also batteries electric motors and electric or hybrid mobility technologies, as well as power generation and distribution technology. In the non-electrical part of the map, the health/medical cluster is a large one, but we also find a large cluster with technologies aimed at reducing, controlling and capturing emissions and exhaust. In each cluster, we find a number of Brown technologies that occupy central positions, indicating a key role in integration.

In the geography (country) based map (see figure 3 below), we find that location is the main dividing line. This map contains three large areas. One of these contains mostly countries outside Europe, with the US and Japan as the largest nodes. The other clusters are Europe-centered. All of these clusters contain a significant number of brown nodes.

We also produced separate maps for trajectories of different lengths, and we observe a large similarity between those and the maps for all trajectories. Differences are largest for the maps based on the longest trajectories. In the geography-based map with longest trajectories, the divide changes from geography-based to brown/green-based. In other words, the major divide in the geography network of longest trajectories is between green and non-green technologies, instead of Europe-non-Europe. In the IPC-based map with longest trajectories, the two general clusters (ICT and health/medical) remain clearly visible, but a number of typical green technologies, such as electric cars and wind power, vanish from the network. These technologies have not yet accumulated the long trajectories that are found in this network.

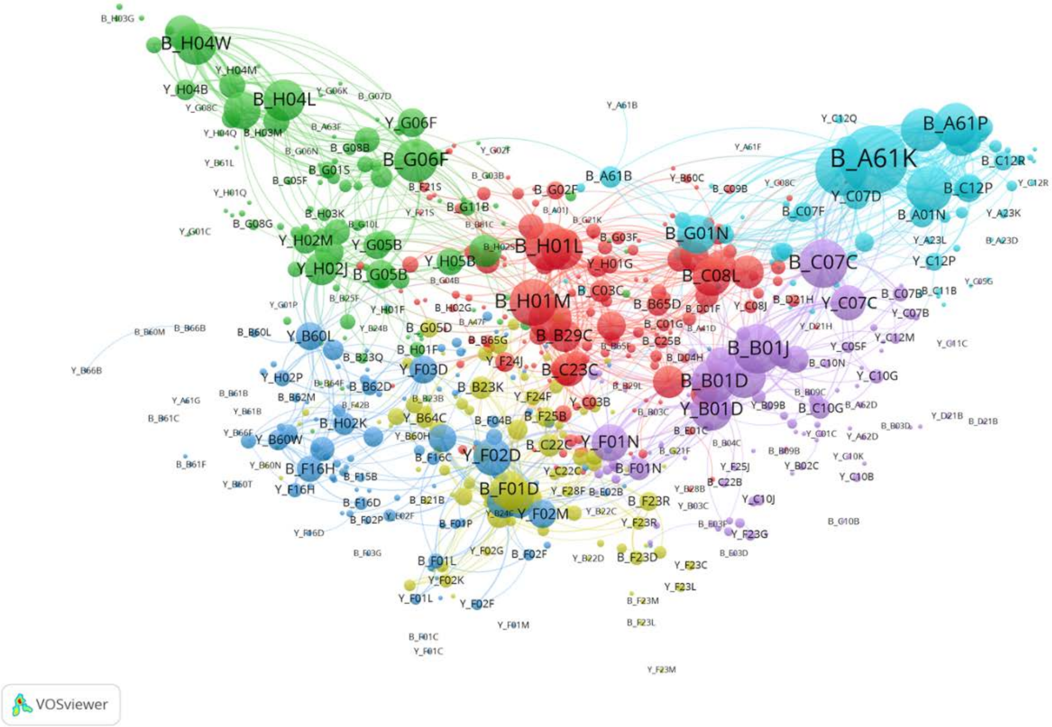


**Figure 2. Landscaping map for trajectories of all length, green/brown and IPC codes**

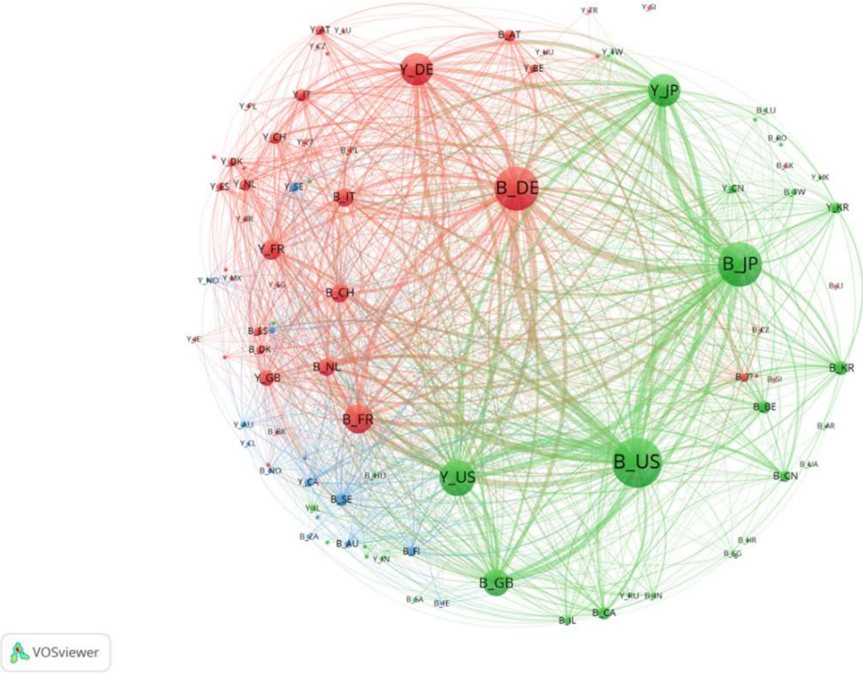
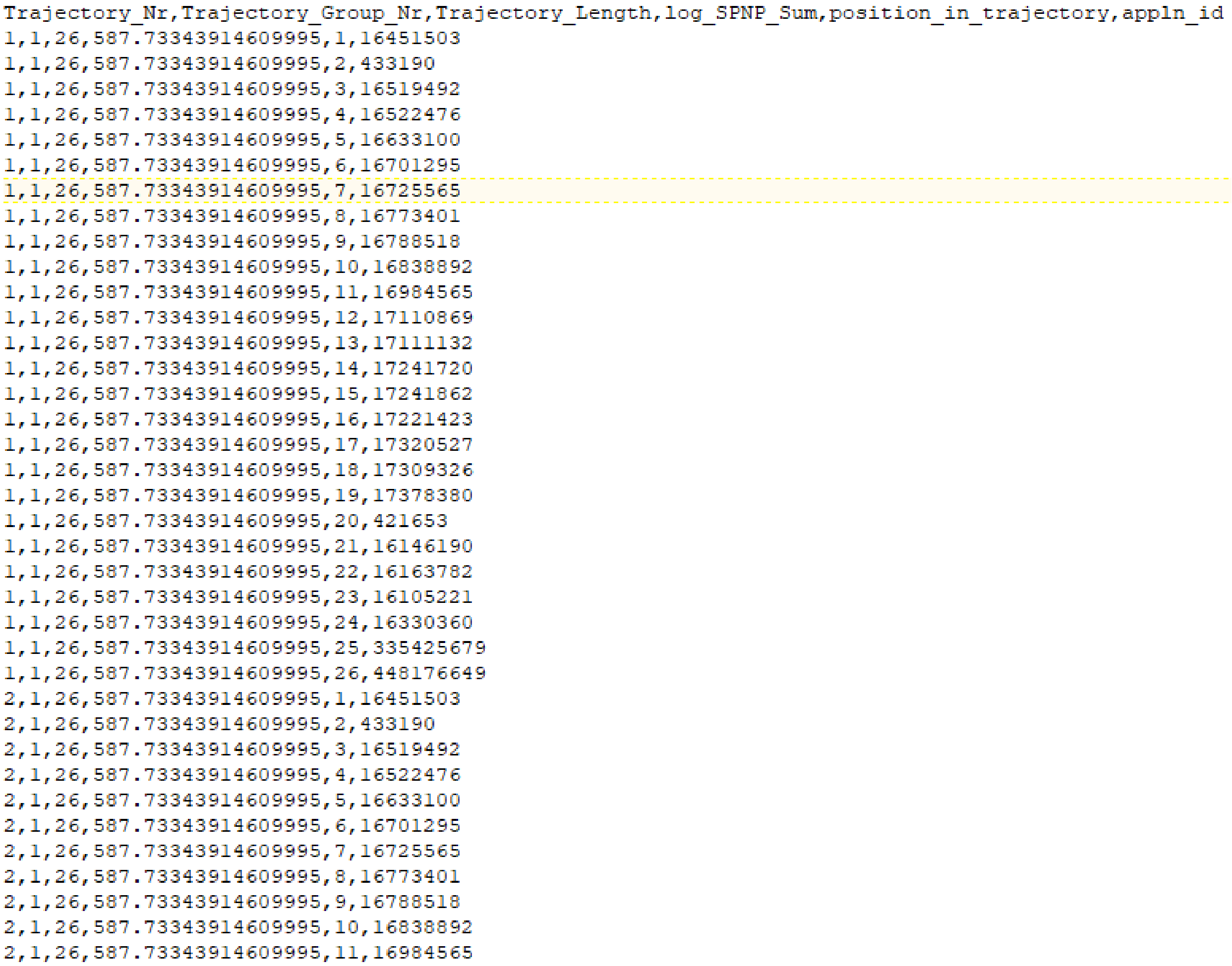


**Figure 3. Landscaping map for trajectories of all lengths, green/brown and countries**

As our analysis is mostly a proof-of-concept of the idea that trajectories are a useful unit of analysis for patent landscaping, the policy relevance of our work has a major indirect component: to the extent that patent landscaping is used to inform policymakers (e.g., innovation policy, policy on intellectual property rights), the application of our method in such studies will be one of the ways in which our method could become policy relevant.

However, there are also policy implications of the findings of our own patent landscaping exercise in green technology. First, as in our first paper, we found that non-green (brown) technology plays an important role in the green technology landscape. Policies aimed at making a green technology transition possible should therefore aim at greening non-green technologies as well as creating new and original green technology paths. Second, our landscaping maps show that large and broad technological areas such as ICT and health/medical are important sub-parts of the green technology field. Thus, a Greentech technology policy should have a broad focus, rather than only focusing on very specific Greentech areas such as electric vehicles. Finally, the geography-based maps that we produced show that Greentech technology trajectories do not develop in geographical isolation, but rather as a collective international effort. Greentech policy should therefore transcend international borders, and be based on international R&D cooperation.

## Appendix1: A snapshot of the UNMP dataset we make publicly available as a comma-delimited text file.



```
Trajectory_Nr,Trajectory_Group_Nr,Trajectory_Length,log_SPNP_Sum,position_in_trajectory,appln_id
1,1,26,587.73343914609995,1,16451503
1,1,26,587.73343914609995,2,433190
1,1,26,587.73343914609995,3,16519492
1,1,26,587.73343914609995,4,16522476
1,1,26,587.73343914609995,5,16633100
1,1,26,587.73343914609995,6,16701295
1,1,26,587.73343914609995,7,16725565
1,1,26,587.73343914609995,8,16773401
1,1,26,587.73343914609995,9,16788518
1,1,26,587.73343914609995,10,16838892
1,1,26,587.73343914609995,11,16984565
1,1,26,587.73343914609995,12,17110869
1,1,26,587.73343914609995,13,17111132
1,1,26,587.73343914609995,14,17241720
1,1,26,587.73343914609995,15,17241862
1,1,26,587.73343914609995,16,17221423
1,1,26,587.73343914609995,17,17320527
1,1,26,587.73343914609995,18,17309326
1,1,26,587.73343914609995,19,17378380
1,1,26,587.73343914609995,20,421653
1,1,26,587.73343914609995,21,16146190
1,1,26,587.73343914609995,22,16163782
1,1,26,587.73343914609995,23,16105221
1,1,26,587.73343914609995,24,16330360
1,1,26,587.73343914609995,25,335425679
1,1,26,587.73343914609995,26,448176649
2,1,26,587.73343914609995,1,16451503
2,1,26,587.73343914609995,2,433190
2,1,26,587.73343914609995,3,16519492
2,1,26,587.73343914609995,4,16522476
2,1,26,587.73343914609995,5,16633100
2,1,26,587.73343914609995,6,16701295
2,1,26,587.73343914609995,7,16725565
2,1,26,587.73343914609995,8,16773401
2,1,26,587.73343914609995,9,16788518
2,1,26,587.73343914609995,10,16838892
2,1,26,587.73343914609995,11,16984565
```

## Appendix2: Number of trajectories (out of a total of 3,710,269) per trajectory length (on rows) and the number of Greentech patents on trajectory.

| | Number of Green Patents on Trajectory | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trjectory Length | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | Row Totals |
| 2 | 471,785 | 37,837 | **16,518** | | | | | | | | | | | | | | | | | 526,140 |
| 3 | 407,494 | 36,206 | 14,028 | **9,339** | | | | | | | | | | | | | | | | 467,067 |
| 4 | 359,042 | 35,967 | 14,075 | 7,718 | **5,986** | | | | | | | | | | | | | | | 422,788 |
| 5 | 320,743 | 34,233 | 13,888 | 7,542 | 5,092 | **4,331** | | | | | | | | | | | | | | 385,829 |
| 6 | 291,513 | 32,599 | 13,895 | 7,559 | 4,764 | 3,726 | **2,983** | | | | | | | | | | | | | 357,039 |
| 7 | 254,068 | 30,311 | 12,281 | 6,929 | 4,506 | 3,249 | 2,596 | **2,008** | | | | | | | | | | | | 315,948 |
| 8 | 220,604 | 26,605 | 11,275 | 6,158 | 4,143 | 2,997 | 2,364 | 1,679 | **1,163** | | | | | | | | | | | 276,988 |
| 9 | 182,459 | 24,131 | 8,975 | 5,053 | 3,227 | 2,667 | 1,904 | 1,285 | 1,028 | **730** | | | | | | | | | | 231,459 |
| 10 | 149,688 | 20,815 | 7,309 | 4,275 | 2,835 | 2,145 | 1,721 | 1,071 | 783 | 775 | **563** | | | | | | | | | 191,980 |
| 11 | 116,095 | 17,501 | 5,313 | 3,039 | 2,196 | 1,813 | 1,504 | 1,321 | 796 | 553 | 443 | **190** | | | | | | | | 150,764 |
| 12 | 84,742 | 13,881 | 4,312 | 2,151 | 1,882 | 1,230 | 970 | 924 | 792 | 637 | 278 | 219 | **87** | | | | | | | 112,105 |
| 13 | 58,791 | 10,760 | 3,194 | 1,716 | 1,251 | 686 | 679 | 515 | 565 | 523 | 492 | 206 | 93 | **26** | | | | | | 79,497 |
| 14 | 39,694 | 9,306 | 2,328 | 1,300 | 756 | 556 | 460 | 384 | 375 | 265 | 284 | 294 | 143 | 34 | **8** | | | | | 56,187 |
| 15 | 28,398 | 7,576 | 1,954 | 705 | 357 | 351 | 355 | 222 | 161 | 166 | 229 | 179 | 190 | 78 | 7 | **3** | | | | 40,931 |
| 16 | 19,350 | 5,832 | 1,404 | 436 | 236 | 161 | 214 | 196 | 189 | 115 | 92 | 103 | 117 | 156 | 33 | 2 | | | | 28,636 |
| 17 | 15,725 | 4,895 | 830 | 276 | 130 | 124 | 119 | 112 | 108 | 50 | 41 | 58 | 51 | 83 | 140 | 21 | | | | 22,763 |
| 18 | 10,475 | 3,861 | 671 | 230 | 82 | 62 | 85 | 48 | 21 | 20 | 17 | 31 | 18 | 31 | 60 | 97 | 6 | | | 15,815 |
| 19 | 6,870 | 3,865 | 293 | 119 | 74 | 16 | 36 | 28 | 27 | 15 | 7 | 26 | 13 | 11 | 15 | 23 | 13 | | | 11,451 |
| 20 | 3,865 | 2,886 | 168 | 58 | 34 | 14 | 10 | 20 | 76 | 15 | 19 | 27 | 8 | 5 | 3 | 4 | 2 | 3 | 1 | 7,218 |
| 21 | 1,981 | 1,733 | 117 | 32 | 27 | 17 | 1 | 2 | 3 | 5 | 20 | 16 | 9 | 5 | 1 | 4 | 3 | 2 | 1 | 3,979 |
| 22 | 1,170 | 1,184 | 32 | 4 | 17 | 5 | 3 | | 1 | | 6 | 11 | 7 | 3 | | | | 10 | 5 | 2,458 |
| 23 | 677 | 914 | 11 | 18 | 6 | | | | | 12 | 4 | 11 | 9 | 4 | | | | | | 1,666 |
| 24 | 361 | 602 | 6 | | 2 | | | | | | | 4 | 1 | 3 | 3 | 2 | | | | 984 |
| 25 | 193 | 180 | 2 | | | | | | | | | 12 | | 1 | 8 | 1 | | | | 397 |
| 26 | 52 | 50 | | | | | | | | | | 3 | | | 1 | | | | | 106 |
| 27 | 18 | 42 | | | | | | | | | | | | | | | | | | 60 |
| 28 | | 14 | | | | | | | | | | | | | | | | | | 14 |
| Column Totals | 3,045,853 | 363,786 | 132,879 | 64,657 | 37,603 | 24,150 | 16,004 | 9,815 | 6,088 | 3,881 | 2,495 | 1,390 | 746 | 440 | 279 | 157 | 24 | 15 | 7 | 3,710,269 |

## EPO ARP Project

*Mapping Greentech Trajectories in the Universal Network of Patent Citations*

# RESEARCH PAPER #1

# Greentech homophily and path dependence in a large patent citation network

Önder Nomaler & Bart Verspagen

UNITED NATIONS UNIVERSITY

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 analyze allows us to analyze 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 analyze 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; technological trajectories; main path analysis; green technology; climate change mitigation

*This work was financially supported by the Academic Research Program 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 neighborhoods 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 “color”? (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 analyze 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 color-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 color-cluster vs. as many color-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 color (Green-to-Green and Brown-to-Brown) to cite each other at higher rates than patents of different colors (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 color 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 color-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 color-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 summarizes 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 Doreian (1989), who proposed a method for analyzing 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 Doreian-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 analyze 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 neighborhood 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 operationalized. 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 realized 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 neighborhood of a particular set of opportunities, and that such a neighborhood 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 realized, 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 operationalization

The aim of our analysis will be to summarize 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 organizations collectively construct these trajectories. And we operationalize 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 operationalize 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 operationalize 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 operationalize 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 or 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 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.

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

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 recognizing 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

[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.

start-nodes). If there are no 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 realize 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 maximize 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

[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.

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 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 $j \in \{1,2,\ldots,NFT_i^k - 1\}$), 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 $\forall k \in \{1,2,\ldots,NF_i\}$). 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 $j \in \{1,2,\ldots,NBT_i^k - 1\}$, 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

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

for $o \in \{1,2,\ldots,NF_i\}$, $\forall m \in \{1,2,\ldots,NF_i\}$ 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 maximizes aggregate forward *SPx*.

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

[4] Our TMP is identical to the one that would be identified by Liu et al. (2012).

[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.

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

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*.

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 traveled, 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 ≥ *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

[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%.

instead multiplicative aggregation, and identified trajectories on the basis of the maximization of $\sum_{j=1}^{NFT_i^o-1} log(SPx(FT_i^o(j)))$. This log-sum maximization 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 characterized 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 labeled 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.

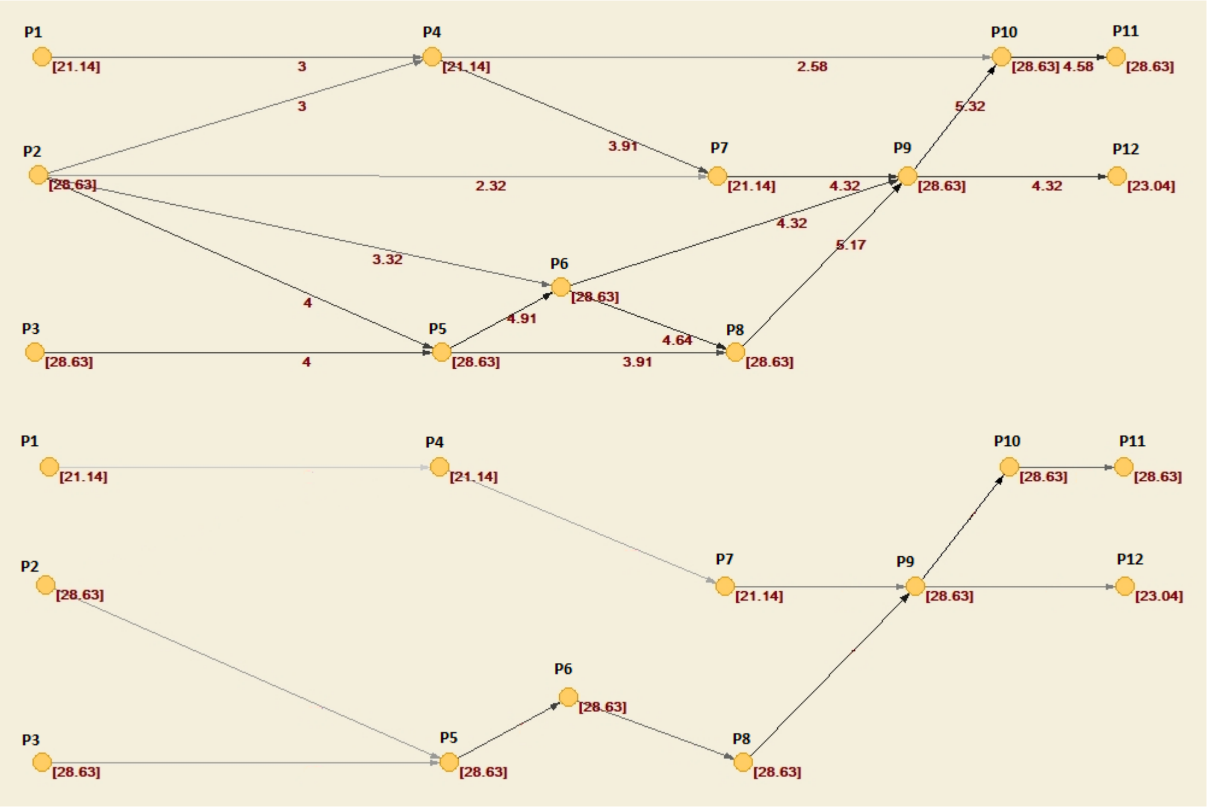


*Figure 1. Example network, total citation network (top panel) and extracted NMP (bottom panel)*

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 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.

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 descriptives 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 stabilizes 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.

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 characterize each path by two main characteristics: the number of Greens on the path, and the number of color-clusters on the path. In defining color-clusters, we simply look at subsequent nodes of the same color, and consider them as a cluster. For example, the path G →B → B → G →G has 3 clusters (G, B → B and G → G).

[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).

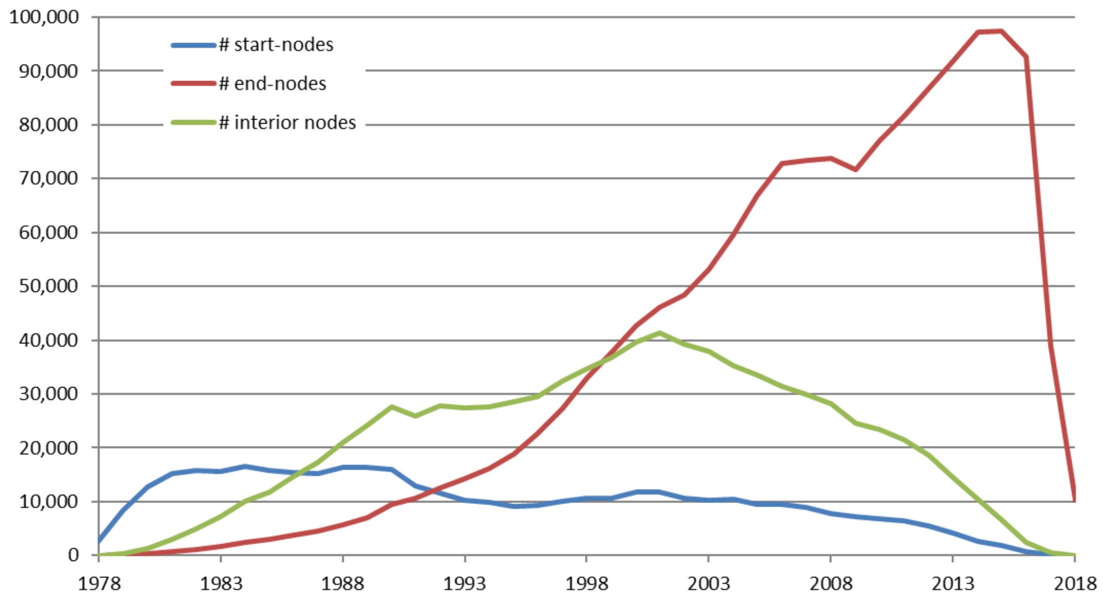


*Figure 2. Number of nodes by type, un-pruned NMP*

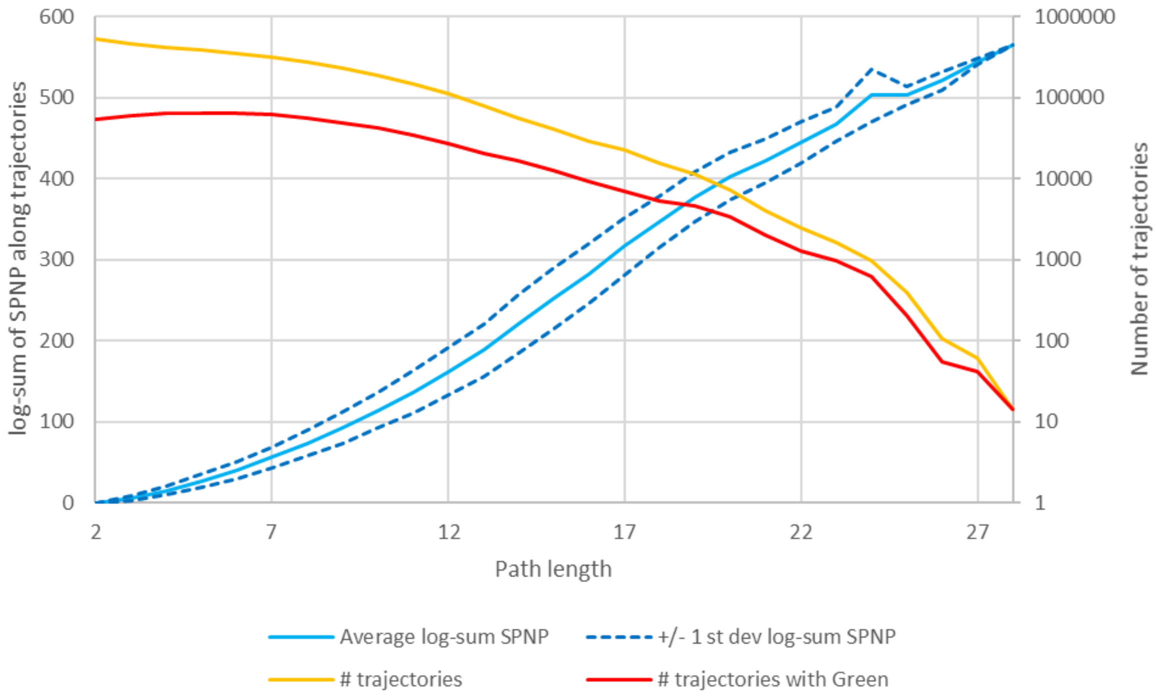


*Figure 3. Number of trajectories and long-sum of SPNP by path length*

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.

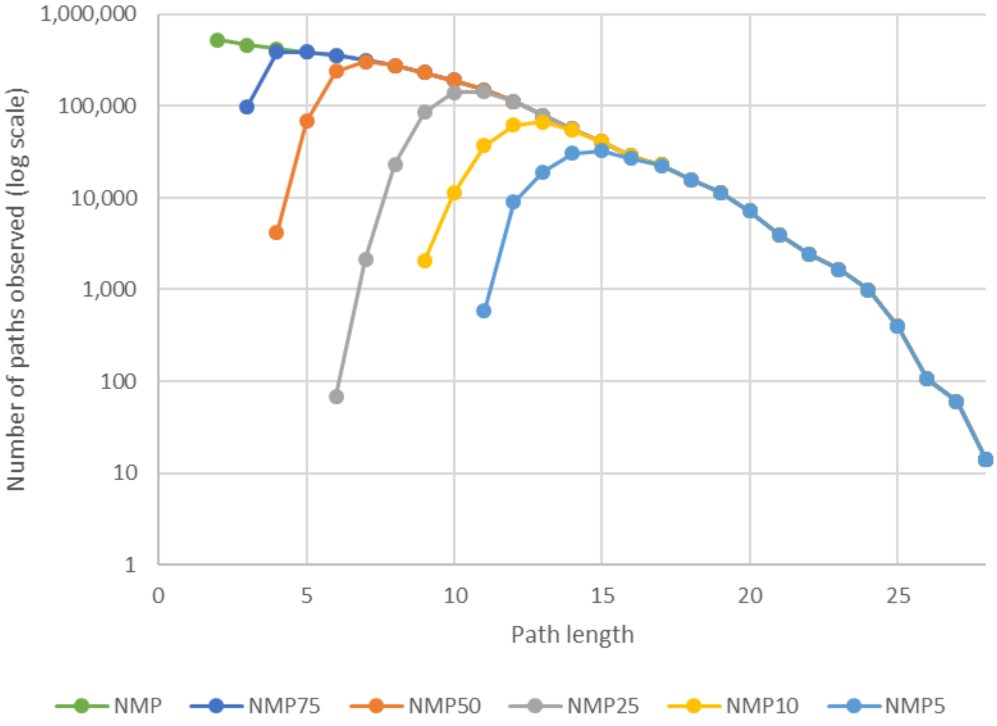


*Figure 4. Number observed paths by length, NMP and pruned versions of it*

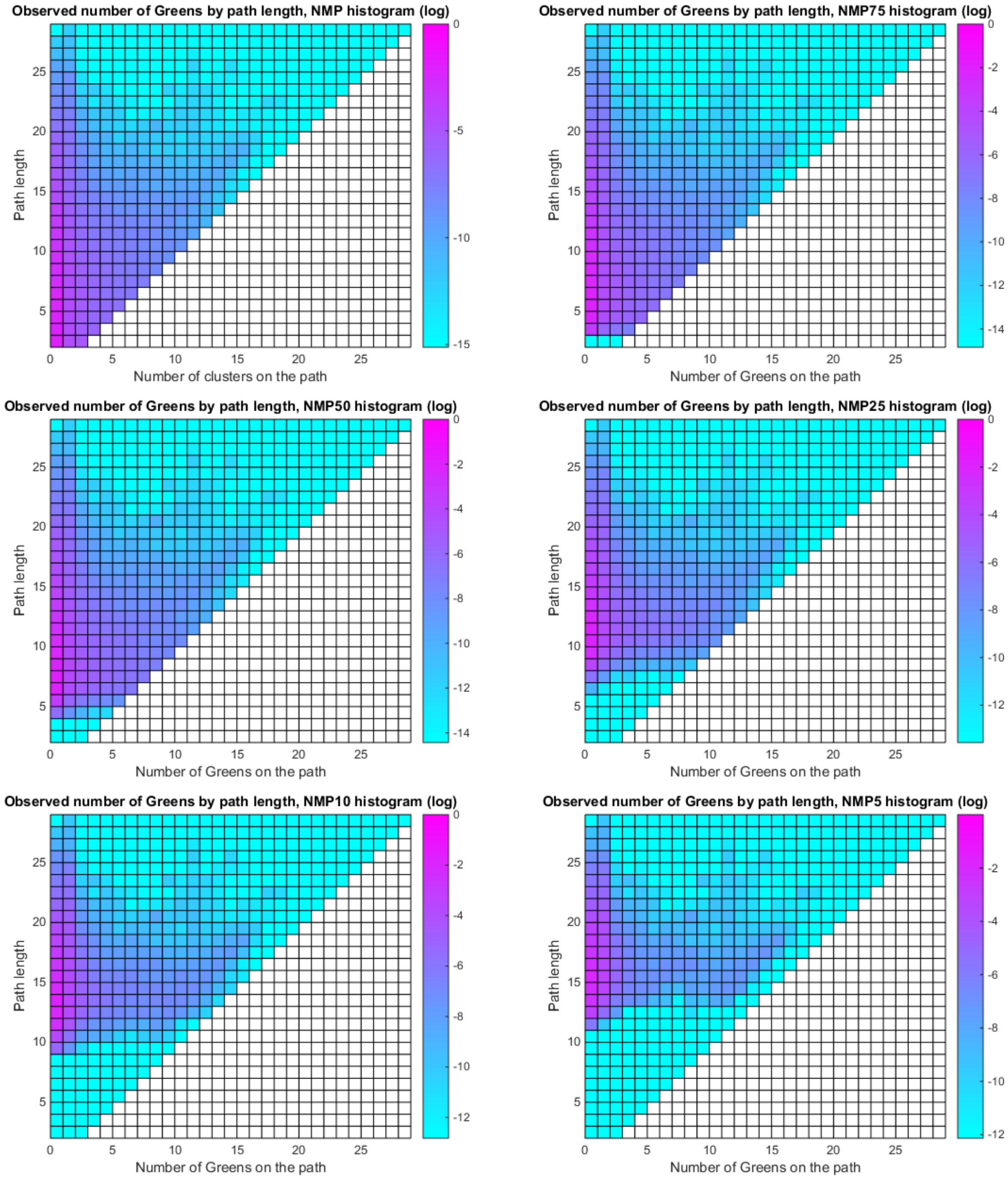


*Figure 5. Histograms of observed numbers of Greens on a path (log), by path length, NMP and pruned NMPs*

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 color 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 color 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 levels 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 color-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 neighbors 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 color-clusters, and the relative abundance of middle-long paths with relatively many Greens.

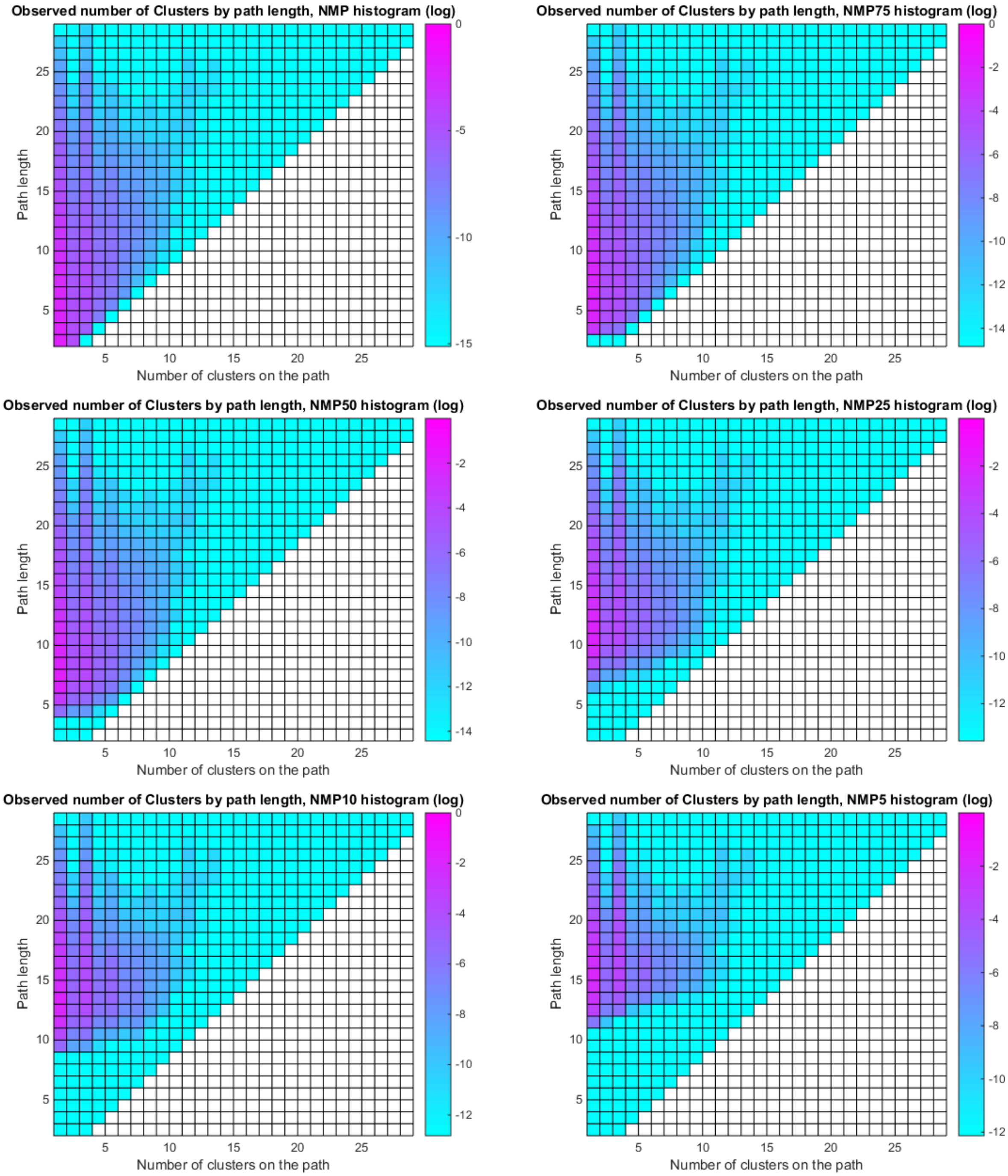


*Figure 6. Histograms of observed numbers of clusters on a path (log), by path length, NMP and pruned NMPs*

## 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 color 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 color 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 color 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 color 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 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 $\bar{p}_{G \rightarrow G}$, and $\bar{p}_{B \rightarrow G}$, and compare this to the probabilities of the Homophily model ($p_{B \rightarrow G}$ and $p_{G \rightarrow G}$).[9]

In the sample where the cited patent is Brown, the probability in the Homophily model ($p_{B \rightarrow G}$) is 0.040, while we find $\bar{p}_{B \rightarrow G}$ = 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 $\bar{p}_{G \rightarrow G}$ = 0.338, while $p_{G \rightarrow G}$ = 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

[8] We did an estimation excluding all citation pairs with cited start-nodes, and this yields very similar results.
[9] The probabilities in the Homophily model can also be estimated in a logit model, by using a model with only a constant.

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.

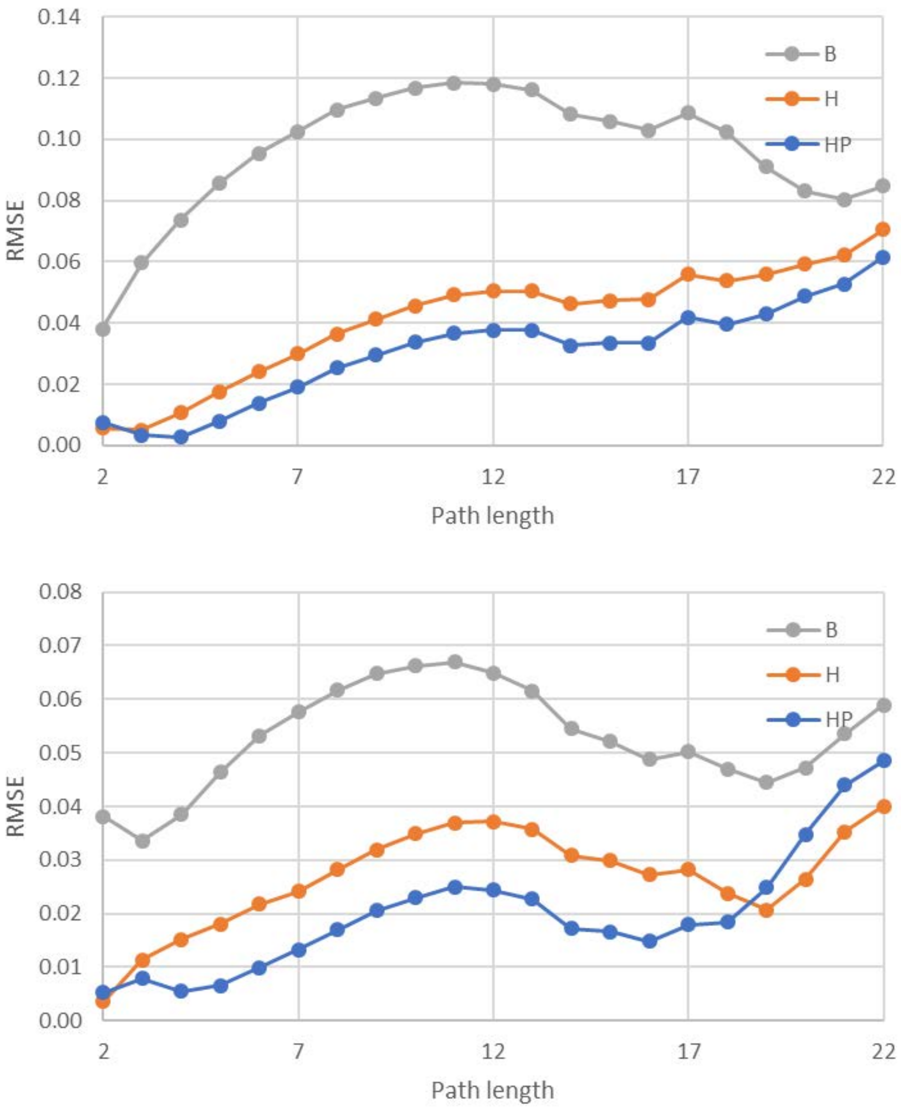


*Figure 7. Root mean squared error for predicted number of Greens (top panel) and predicted number of color-clusters (bottom panel)*

In order to undertake the comparison between actual data and predicted frequencies, we standardize 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 color-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.

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 righthand side. The three benchmark models are presented top-to-bottom. It is important to note that, as indicated by the color 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.

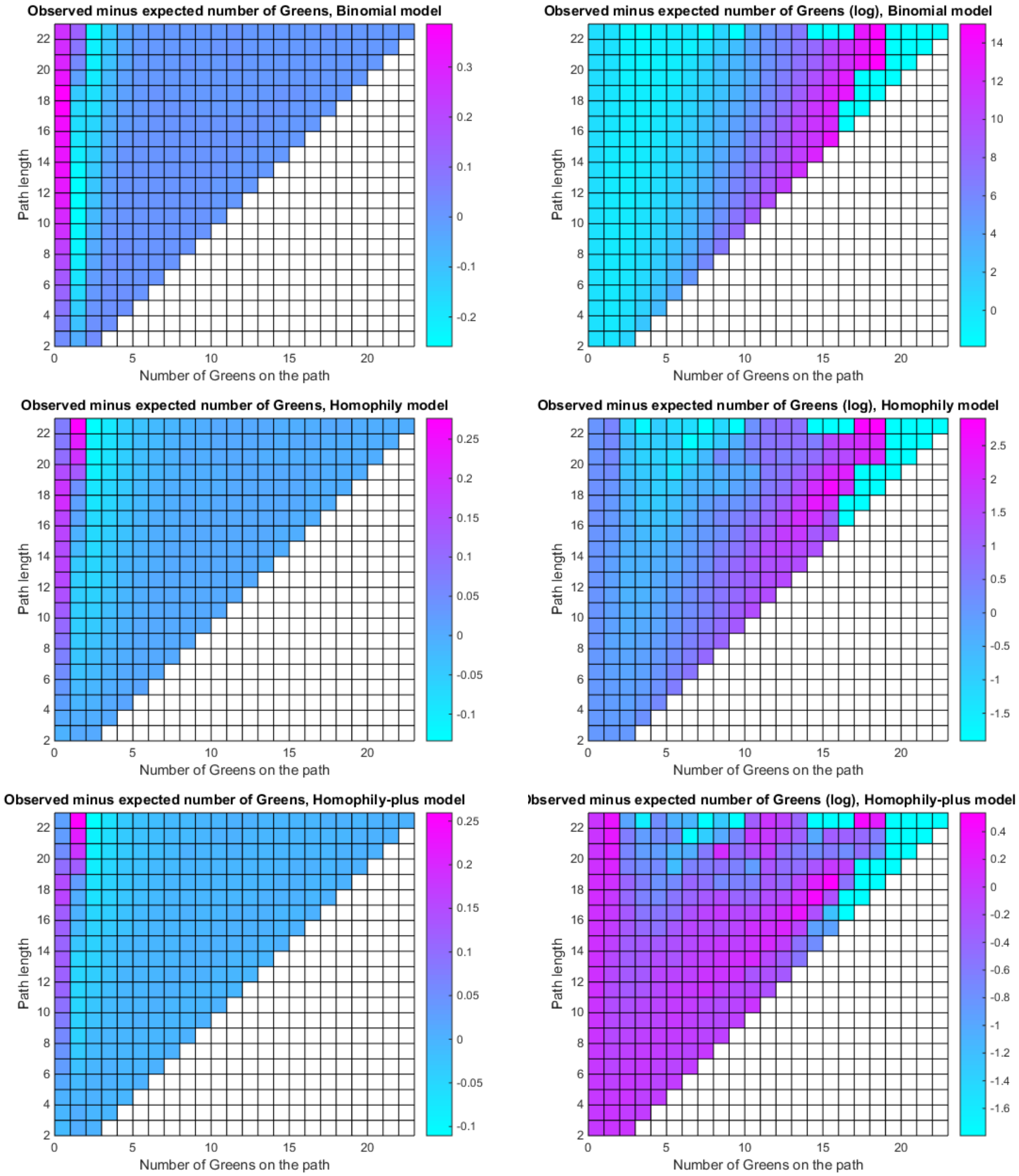


*Figure 8. Observed minus expected number of Greens, by path length and by benchmark model*

The log-difference figures on the righthand 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 –∞ (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.

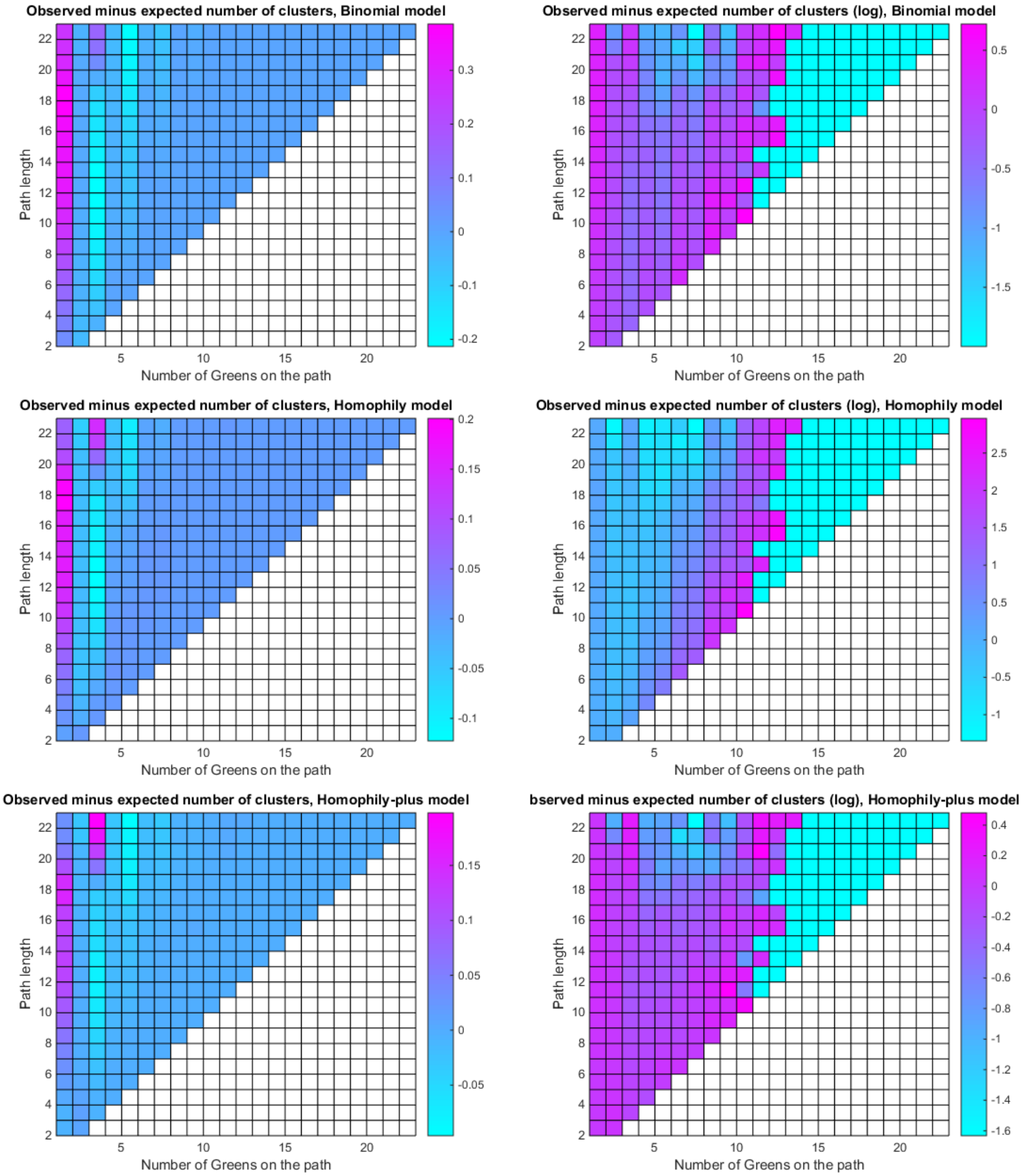


*Figure 9. Observed minus expected color-clusters, by path length and by benchmark model*

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 9 provides the comparison for observed minus predicted number of color-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 color-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.

### 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 endogenizes a part of the observed homophily in the pure homophily model, this extended model potentially endogenizes 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 neighbors. 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 labeled "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.

*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 |
| Neighboring 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 |
| Neighboring 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 |

For Brown-to-Green citations, we must keep in mind that this type of citation has a high degree of homophily ($p_{B\rightarrow B}$ = 0.96 or $p_{B\rightarrow G}$ = 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).

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 characterized 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 color-clusters. These findings are based on a theoretical model that predicts the statistical distribution of the number of Greens and the number of color-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 color 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 characterized, 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 programs, 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.

## EPO ARP Project

## *Mapping Greentech Trajectories in the Universal Network of Patent Citations*

# RESEARCH PAPER #2

# Patent Landscaping using ‘green’ Technological Trajectories

Önder Nomaler & Bart Verspagen

UNITED NATIONS UNIVERSITY

UNU-MERIT

This version of January 2021

## Abstract

We present a number of green technology patent landscaping exercises, based on a method that we developed earlier to identify the main technological trends in a very large (i.e., universal) patent citation network comprising all patented technologies. This method extracts a so-called network of main paths, where we interpret each path as a technological trajectory in the sense of Dosi (1982). We use co-occurrence on the technological trajectories as the main metric to build a network of technological relations, with green/non-green, the technology class (4-digit IPC classes) and geographical location (countries) as the main dimensions along which we observe green technology. The technology landscaping exercise visualizes these networks. In this way, we draw a detailed map of green technologies (along with the particular non-green technologies that contribute thereto or benefit therefrom), in which we find both very broad and general areas (such as ICT or medical and health), and specific green technologies, such as batteries, wind power and electric vehicles. In the geography- based map, we find specific European and non-European areas. In all our landscaping maps, non-green technologies play a large role, indicating that sectoral and geographical progress in greentech cannot be fully understood independently of developments in particular fields of non-greentech technologies.

**Keywords**: green technology; technological trajectories; patent citations; main path analysis; patent landscaping

**JEL codes**: O31, O33, Q55

*This work was financially supported by the Academic Research Program 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 patent landscaping exercise for green technology. Whereas the mapping of technological trends on the basis of patent data is usually performed using metrics calculated at the level of individual patents or patent citation pairs, we argue that it may be more appropriate to use a metric that is itself defined at the level of a technological "trend". For this reason, we build our landscaping exercise on our earlier work on technological trajectories (Nomaler and Verspagen, 2019).

A technological trajectory is perceived as a main technological trend that takes shape over time, and which consists of cumulative, and often incremental, inventions influenced by the economic and social environment. Based on Hummon and Doreian (1989), a growing literature has been established that uses patent citations to find technological trajectories, or main paths as they are also called in this literature. This literature focuses on specific technological (or scientific) fields that are defined *a priori* (e.g., fuel cells, or digital network communication). Nomaler and Verspagen (2019) propose a method that extends the Hummon and Doreian-based methods to extracting main technological trends from a very large patent citation network covering all technologies patented under the terms of one legal jurisdiction (i.e., EPO).

This enables the application of the Hummon and Doreian-based methods to so-called macro-technology fields, i.e., sets of distinct technologies that are in pursuit of a common and coherent goal. Like in our previous paper, the macro-technology field that we are interested in here is green technology, which we define as technologies aimed at combatting climate change. Whereas the concentration of green patents on technological trajectories was the topic of our previous paper, this paper looks at how the green technological trajectories that we find can be used to landscape green technology.

We understand landscaping as an impressionistic method that uses relations between patents, in our case patents occurring on specific technological trajectories, to extract and describe main technological trends and relationship between technological (sub-)fields. In our case, the landscaping exercise will consist mostly of visualization of the technological relationships in our database of green technological trajectories. The use of technological trajectories as the basic unit of analysis in the landscaping exercise is a key element of our analysis. In our view, this is the most appropriate way of proceeding if the aim is, as in our case, to map the main technological trends in a macro-field. Technological trajectories are aimed at capturing these main trends, whereas individual patents capture individual inventions, and patent citation pairs capture bilateral relationships between inventions.

Our previous analysis has already shown, among other things, that non-green (we will adopt the term brown for this) patents are an important part of green trajectories (which we define as a trajectory with at least one green patent on it). Thus, while there is some degree of green clustering, almost no green technological trajectories develop without a brown influence. It seems that progress in greentech cannot be fully understood independently of developments in particular fields of non-greentech technologies. Among other things, our landscaping exercise is expected to bring this out, i.e., to show the role of non-green technology in greentech.

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, i.e., it introduces the idea of technological trajectories as paths on metaphorical technological landscapes. This section also introduces the idea of main path analysis to map these technological trajectories or paths (we use trajectories and paths mostly as synonyms). Section 3 provides a brief, non-technical overview of our method to find technological trajectories in (very) large patent citation datasets. It also describes the construction of the basic dataset on which we apply this method, as well as a brief overview of the so-called network of main paths that results from the trajectory-extracting exercise, including a brief overview of the green content of this network of main paths.

Section 4 provides the main part of our analysis, i.e., the landscaping exercises. This section starts with a sub-section that briefly introduces the idea of patent landscaping, and the general nature of the metrics that it uses. This sub-section also argues in some more detail why we propose to calculate these metrics on technological trajectories, rather than on individual patents or patent citation pairs. Sub-section 4.1. concludes with a detailed exposition of the metrics and visualization methods that we use. The next sub-sections in section 4, i.e., 4.2 and 4.3, provide our main results in the form of maps containing the landscaping results. In section 4.2, we do this using a perspective of technology classes (4-digit IPC) and the green/brown distinction. Section 4.3 uses geographical (country) perspective combined with the green/brown distinction. Section 5 summarizes the argument and provides the main conclusions.

## 2. Technological trajectories and main path analysis

Dosi (1982) introduced the idea of technological paradigms and technological trajectories. In a nutshell, his proposal is that technological trends are both influenced by the general context of scientific and technological knowledge (the paradigm) and by economic opportunities and restrictions (trajectories). Dosi's starting point is that engineers will tend to search for technological solutions in the neighborhood of a particular set of opportunities, and that such a neighborhood 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 realized, and this is decided on the basis of specific societal circumstances, including economic markets.

Thus, the idea of a technological trajectory is based on local search. It also gives rise to the idea of a technological landscape, which 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 operationalized. The metaphor of a technological landscape then boils down to arranging the knowledge pieces relative to each other based on how close they are.

A central tenet of the concept of a technological landscape is that performance of technologies differs and is somehow dependent on the position of the technology in the landscape. Thus, the inventor who searches 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. A technological trajectory emerges from this search process, and is a specific and collective path through the technological landscape. This path is heavily shaped by both technological opportunities and the economic incentives that the economic environment (the market) provides.

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 unusable 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.

Dosi’s conceptualization of technological trajectories is broadly compatible with work in the management field (e.g., Levinthal, 1997; Fleming and Sorensen, 2004; Aharonson and Schilling, 2016). These authors propose 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 neighborhood of where search was previously located. As a result, technological efforts by firms will cluster in technology space (e.g., Aharonson and

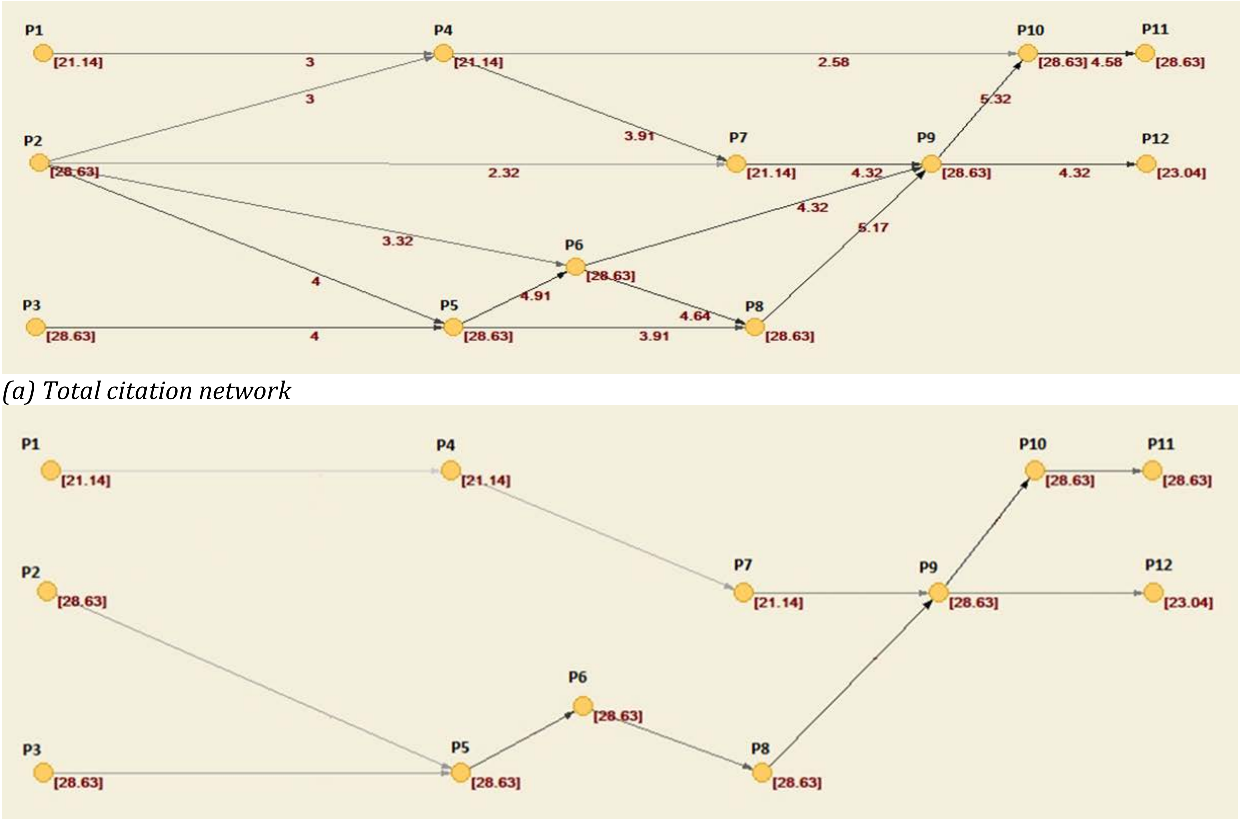
Schilling, 2016). A logical strategy is to use prior knowledge about where the feasible and valuable technologies are located (Stuart and Podolny, 1996; Fleming and Sorensen, 2004). Such prior knowledge accumulates from the firm's own prior research, and, to the extent that they are observable, other 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 realized 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).

Our analysis uses patent citations to identify technological trajectories. The legal IPR framework holds that cited patents (identified by the applicant and/or the patent examiner) constitute the prior 'state-of-the-art' against which the novelty of the citing patent's constituent claims are to be assessed by the patent examiner respectively. Thus, a citation indicates a close (directed) relation from the cited to the citing patent, in the sense of an 'inventive step' (be it marginal, incremental, or substantial) or as maintained by various scholars of innovation, an indication of some sort of 'knowledge flow' (Trajtenberg and Jaffe, 2002). This idea has been key to a growing strand of literature that operationalizes the notion of a technological trajectory in terms of an unbroken chain of citations, thus an intertemporally-ordered set of patents where there exists a citation link between each pair of immediately subsequent patents.

Let us illustrate this by a stylized example (Nomaler and Verspagen, 2019). The top panel of Figure 1 depicts a toy patent citation network as a graph. The network has 12 patents, depicted as nodes which are labeled P1 – P12. Edges (lines) indicate a citation, and the direction of the arrow to an innovative step taken by the citing patent (say P4) over the cited one (each of P1 and P2.) That is, as of its examination date, the claims in patents P1 and P2 (but not P3) were deemed by the examiner as the state of the art against which the novelty of P4 is to be assessed. P1, P2 and P3 cite no other patents in this network, thus are referred to as 'start-nodes', while P11 and P12, which are cited by no other patent of the network, are 'end-nodes'. Any chain of subsequent citation links that connect a start-node to an end-node (i.e., by a walk in the direction of the arrows) is a full path.



*(a) Total citation network*

*(b) Reduced network (Network of Main Paths)*

**Figure 1. A stylized example network (a) and its algorithmically reduced form (b)**

For example, the connected citation chain that connects P1 to P11 via patents P4 and P10 is one path. Clearly this chain describes a cumulative process (of technical progress) that begins with P1 and adds 3 subsequent inventive steps by P4, P10 and P11, each updating the state-of-the-art. However, this trajectory is neither the only one that connects P1 to P11 (there is also the one that goes via P4-P7-P9-P10), nor the only one that leads to P11. In fact, there are 11 ways in which a walk that emanates from the start-nodes P2 or P3 may end up at P11. Similarly, there are 11 different ways to reach P12 (the one and only other end-node of the network) starting from either of the three start-nodes, which, also including P1-P4-P10-P11, make a total of 23 trajectories that can be enumerated in this simple stylized network.

For each of the identified paths, one can read through the (claims and/or the abstracts of the) participating patents in the order of appearance on the path, and thus form an historical narrative of the technical progress that builds incrementally on the path. Every path-narrative (23 in our stylized example) will capture some aspect of the technological trends that characterize the total network. However, due to the many patents and the citation links that are common to several paths[1], there will be strong overlap among the various narratives. Too many (partially overlapping) narratives, no matter the complementarities, is not a practical way to understand the main trends in a citation network, especially given that in actual networks of interest (comprising tens or hundreds of thousands of patents) one can identify (perhaps, uncountably) many trajectories. The main question that underlies the early work in the growing literature on 'main paths' has been whether one can identify only a single (i.e., the most 'significant') path, or a few paths, that provide the narrative that captures or highlights the essence of development through the technology landscape, summarizing the historical progress in a given technical field or application area of interest.

The exploration of this question has been based on the literature that follows Hummon and Doreian (1989), who proposed a method for analyzing 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 Doreian-based methods which will identify the technological main paths by the reduction of a complex citation network to one or few trajectories, has been seen as the operationalization of a technological trajectory (Mina et al., 2007; Verspagen, 2007).

This was initially done for individual technological fields (e.g., Mina et al., 2007; Verspagen, 2007; Liu and Lu, 2012), as a way to quantify more qualitative data from engineers or the history of technology. The early work, the findings of which have often been verified by experts in the technical field of interest, has successfully demonstrated that, despite the enormous reduction/pruning (of patents and their citations)[2], Hummon and Doreian-based

[1] For example, consider the two trajectories P2-P6-P9-P10-P11 and P3-P5-P6-P8-P9-P10-P11, the last three steps of which are identical.

[2] In terms of our stylized example in Figure 1, the original Hummon and Doreian (1989) algorithm would identify the main paths as P2-P5-P6-P8-P9-P10-P11 and P3-P5-P6-P8-P9-P10-P11 by pruning 10 of the 17 citation links and 4 (P1-P4-P7-P12) of the 12 patents.

methods can well capture the ‘main trends’ of the developments and progress through the technology landscape.

In a previous paper (Nomaler and Verspagen, 2016), we were able to analyze a much larger set of patent citations that represent the entire patent literature, and hence the entire spectrum of patented 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 are able to 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 large network of paths covering all (patented) technologies at once, providing the opportunity to address research questions that go beyond the mere provision of a historical narrative for a selected field.

In Nomaler and Verspagen (2019), we already developed this large-scale network algorithm further in order to analyze so-called ‘green’ technology, which we consider to be a macro-field of technology (a collection of distinct technology fields with a common and coherent purpose). Green technology, which we define as technologies aimed at climate change mitigation, consists of a large collection of distinct technology fields, e.g., in solar and wind energy, batteries, fuel cells, nutrition, agriculture, etc. The current paper is a follow-up in this larger project, where we offer a more visual interpretation of the network of main paths with greentech content.

## 3. Constructing the network of main paths

The network of main paths is constructed from the total network of citations, which we operationalize as a citation network between PatStat application ids for which the application authority is ‘EP’.[3] 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 that of 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 or are cited by any other EPO patent. This citation network is enhanced in two ways, both of which add links to the network that are not actually present in the original set of intra-EPO citations. First, we add technological paths that are not captured exclusively by EPO patents, by looking for any indirect citation linkages between EPO patents that exist

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

through other patent offices, 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. However, we found that treating a single family as a single invention by aggregating citations into a single link between families leads to heavy cycling in the citation network.[4] 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 sub-path in itself. This procedure will prevent cycles from forming, while still recognizing 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.

The next step in our analysis is to construct the network of main paths in the total citation network. The mathematical details of our method to do this can be found in the earlier paper (Nomaler and Verspagen, 2019), here we only provide a general description of our method and how it differs from previous methods.[5] Like in Hummon and Doreian (1989) and methods that follow that seminal paper, the network of main paths is a systematically-reduced subset of the larger network, obtained by eliminating the patents and/or citations of 'lesser significance'.

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). We choose the commonly used SPNP (Search Path Node Pair) which 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

[4] For example, application P and 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.

[5] 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 groups) 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).

itself) to all (direct and indirect) descendants of the citing document (including itself). We eventually apply a logarithmic transformation (with base 2) on the SPNP values.

For an illustration, let us revisit the stylized network example on Figure 1 (upper panel). The citation link that directly connects P1 to P4 lies on sub-trajectories that (directly or indirectly) connect one patent (P1) to upstream other patents in 8 different ways: (1) P1 to P4 (directly), (2) P1 to P7 via P4, (3) P1 to P9 via P4-P7, (4) P1 to P10 via P4, (5) once again P1 to P10 but via P4-P7-P9, (6) P1 to P11 via P4-P10, (7) once again P1 to P11 but via P4-P7-P9-10, and (8) P1 to P12 via P4-P7-P9. Thus, the SPNP value of the citation link between P1 and P4 is 8. In logarithms, $\log_2(8)$ =3, which is the value one finds on the network graph. However, the direct citation link between P6 and P9 lies on 20 sub-paths that establish (directly or indirectly) pairwise connections between individual elements of the patent set {P2, P5, P6} and the elements of the upstream set {P9, P10, P11, P12}. Thus the logarithmic SPNP value of the citation link P6-P9 is $\log_2(20)$ =4.32.[6] The most significant citation link of the network is P8-P9, which lies on 36 different trajectory segments that pairwise connect a set of 5 downstream patents {P2,P3,P5,P6,P8} to a set of 4 patents {P9, P10, P11, P12}. The individual significance of this citation link is thus $\log_2(36)$=5.17.

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 (SPNP) 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. 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.

Verspagen (2007) starts from each start-node in the network, and constructs (on the basis of the 'priority first search' principle) a collection of main paths that is referred to as the *network of main paths* (NMP). The *top main path* (TMP) is a single trajectory in the NMP, identified, for example, as the path between particular start- and end-nodes that are considered of special importance (Hummon and Doreian, 1989). As stressed by Liu et al. (2012), it is important to realize that the priority first search algorithm is a heuristic that does not guarantee a global maximum in the value of the summed (or multiplied) SPNP over the found main path(s). In other words, for any start-node, there may well be forward paths that have a higher aggregate SPNP value than the main paths found in the priority first search algorithm.

[6] Actually, the number of node pairs which are connected via P6-P9 is 3x4=12 (thus not 20), however the SPNP metric also considers the different ways in which the pairs are connected indirectly.

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-note 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 method of constructing the NMP provides three novelties. First, instead of the usual priority first forward search heuristic, we use a combination of both forward and backward search[7] to maximize the $\log_2$(sum)[8] 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 (optionally) 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.[9]

Let us go back to our stylized example in Figure 1 for further illustration. As explained before, the values indicated on the citation links are the respective SPNP values (in logarithm with base 2). The original Hummon and Doreian (1989) algorithm as well as the variants refined by Liu et al. (2012) would both identify the main paths as P2-P5-P6-P8-P9-P10-P11 and P3-P5-P6-P8-P9-P10-P11, by pruning 10 of the citation links and four (P1-P4-P7-P12) of the 12 patents.

Our algorithm, by pruning only 6 citation links and none of the patents in the full network, would instead yield the NMP depicted in the lower panel of Figure 1. It can easily be verified that the number indicated next to a patent (in square brackets) is the ($\log_2$-)sum of the of SPNP values all the citation links that are on the most significant trajectory which contains the patent in question. For example, among the $\log_2$(sum) SPNP values of all 12 possible trajectories on which P6 can be found, 28.3 is[10] the maximum value that belongs to the

[7] Emanating from each node of the network, we identify the best forward and backward sub-path and merge.

[8] This is clearly equivalent to maximizing the multiplicative product of the SPNP values.

[9] Our TMP is identical to the one identified by Liu et al. (2012).

[10] 4+4.91+4.64+5.17+5.32+4.58 corresponding to the $\log_2$(SPNP) values of the citation links P2-P5, P5-P6, P6-P8, P8-P9, P9-P10, P10-P11 respectively.

trajectories T1={P2-P5-P6-P8-P9-P10-P11} and T2={P3-P5-P6-P8-P9-P10-P11}. Of course, next to all the patents that participate in these two trajectories (i.e., P2, P3, P5, P8, P9, P10, and P11) we also find the number 28.3 in square brackets, simply because these two trajectories are also the most significant ones to which each of these other patents participate individually.

The next highest $\log_2$(sum) SPNP value we find next to a patent (P12) is 23.04. P12 is only connected directly to P9 and the most significant (downstream) sub-trajectories that lead to P9 are P2-P5-P6-P8-P9 and P3-P5-P6-P8-P9. That gives us two new (second-best) trajectories T3={P2-P5-P6-P8-P9-P12} and T4={P3-P5-P6-P8-P9-P12}. The next and the least $\log_2$(sum) SPNP value (i.e., 21.14) that we find belongs to the patents that are on the sub-trajectory P1-P4-P7. Patent P7 is only connected directly to P9 and the best trajectory that emanates (upstream) from P9 is P9-P10-P11. This gives us our final full trajectory T5={P1-P4-P7-P9-P10-P11}. The union set of T1, T2, T3, T4 and T5 is our NMP.

In case we (optionally) wanted to reduce this NMP by one step, all we would need to do would be to remove the patents that display the lowest ($\log_2$-)sum SPNP value 21.14 (i.e., P1, P4, and P7) which will effectively drop T5. Going even one step further, and eliminating P12 (with the next least $\log_2$(sum) SPNP value, 23.04) will effectively drop the trajectories T3 and T2, leaving behind only the top trajectories of the network, T1 and T2.

Furthermore, observe that T1 and T2 overlap on six of their seven patents exactly at identical positions (in terms of the order of appearance), with P2 and P3 making a tie (i.e., both have the same $\log_2$(sum) SPNP value). Accordingly, we can call the set {P2, P3, P5, P6, P8, P9, P10, P11} a 'trajectory group', where a trajectory group is a set of trajectories that share patents and have identical $\log_2$(sum) SPNP value and length.

As already stressed, we apply this method to the entire citation network for EP patents. Note that in this paper we do not opt for the (optional) patent pruning explained in the paragraph above, thus 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. In the NMP, 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 (only 14) paths of this length.

With patents as our smallest unit of analysis, we operationalize green technology as a specific subset of 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. The Y02 tag is in fact a technology class in the Cooperative Patent Classification (CPC) scheme. 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). Looking only at trajectories that contain at least one green patent, we find relatively few of them (about 660,000 of a total of 3.7

million, or about 18%).[11] The number of trajectories with some green peaks at path length 6 (about 65,000 paths), while all of the longest (length 28) trajectories have some green.

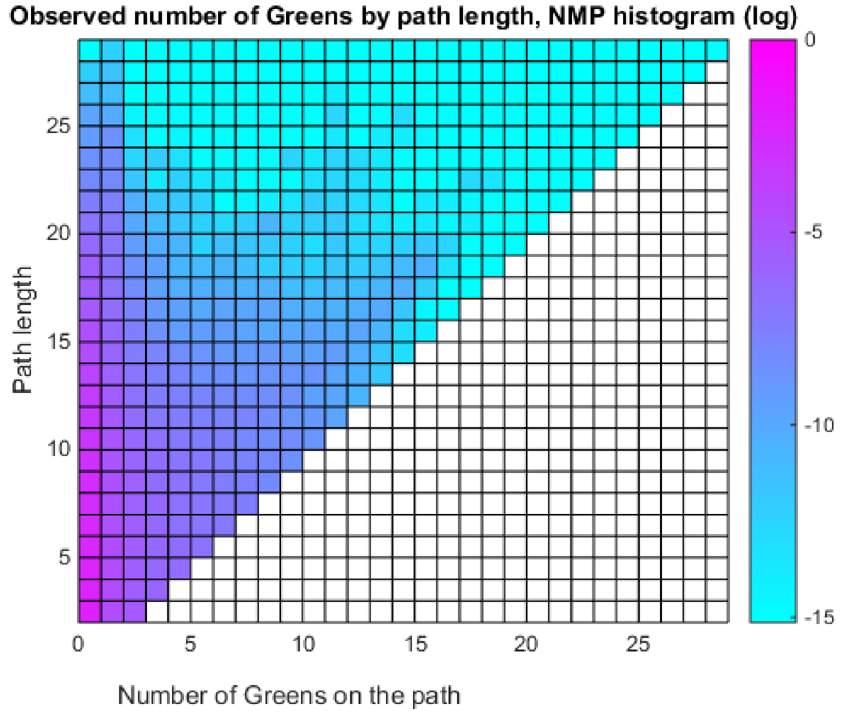


**Figure 2. Histogram of observed numbers of greens on a path (log), NMP**

In Figure 2, we enumerate all paths that are found in the NMP, and ask how many greens are found on the path. Minimum path length in the figure is 2 (there are no isolates in the NMP), and as mentioned earlier, maximum observed path length is 28. 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 color 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. 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 figure 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.

[11] In terms of 'trajectory groups': 246,551 of a total 1,262,472, or about 19.5%

## 4. Landscaping green technological trajectories

### 4.1. Conceptualizations

Patent landscaping is a set of tools that can be used to investigate and describe (recent) technological developments in a technological area, often used by research and development practitioners or policymakers to map current developments in a field (see, e.g., Bubela et al., 2013). Visualization, especially of networks, is a powerful tool used in patent landscaping to provide a quick and overall impression of relations between different parts of the knowledge base in a field (see, e.g., Federico et al., 2017).

Main paths or even networks of main paths can be visualized in a direct way, as a network of patents and citation links between them. As such, main path analysis is one of the tools in the patent landscaping toolbox. However, this direct way of visualizing the NMP is not feasible in the case of our network. Even the subset of trajectories with at least one green patent is too large to be visualized as an NMP. Therefore, we have to resort to a different way of visualizing the network. Several of such ways can be found in the existing literature (e.g., Leydesdorff et al., 2017; Yan and Luo, 2017; Kay et al., 2014).

Many of these methods are based on some form of co-occurrence. For example, if individual patents or publications can be characterized by keywords or title words, landscaping can take the form of creating and visualizing a network of such words, where the relationships between the terms (words) is defined on the basis of how often they occur together, in a patent or in a publication. One may find, for example, that the keywords “battery” and “electric vehicle” often occur together in a patent, but that these terms very rarely co-occur with terms such as “wireless communication” or “digital information”. In a network visualization, the terms “battery” and “electric vehicle” would then be mapped close to each other, pointing to a cluster of technological developments related to electric cars. Instead of co-occurrence of keywords or title words, one may also use co-citations as a measure of relatedness. For example, if two authors are often cited in a common publication, it is likely that their work is closely related.

Co-occurrence is usually measured at the level of individual patents or publications, but the resulting co-occurrence measures are often aggregated to higher-level categories, such as technology classes (in the case of patents), or journals (in the case of scientific publications). This enables the landscaper to analyze large datasets, and visualize them at an aggregation level that yields a proper impression of the trends that the landscaping exercise is after. Thus, for example, using information at the level of individual patents, one may ask which technology classes (e.g., IPC codes) co-occur often, and use this information to visualize a network of IPC codes. By observing which IPC codes are mapped close to each other, one may get an impression of in which areas the main technological developments are taking place.

In this paper, we use this kind of technological landscaping. In line with our interest in technological trajectories and using patent citations to map them, patent citations are our main source of information for the landscaping exercise. However, rather than constructing our measure of relatedness on the basis of data in patent citation pairs, we use the trajectories (paths) that are found in our NMP as the basic unit of analysis. Continuing the example of a network of IPC codes, we ask which IPC codes tend to co-occur on individual trajectories. Note that this is a deviation from existing practice, which is to measure co-occurrence at the level of individual patents (which IPC codes co-occur in patent descriptions?) or patent citations (which IPC codes co-occur  citation pairs?).

Our proposal is that defining co-occurrence at the level of technological trajectories is the preferred procedure if one is interested in mapping broad technological trends. Individual patents are indicators of inventions, a patent citation pair is an indication of a relation (including the possibility of a knowledge flow) between inventions, and trajectories (built on patents and patent citations) show the way in which individual inventions and their bilateral relations are coherent in terms of the broader trends. Thus, if the aim of the landscaping exercise is to outline the relation between broad trends of technological development, trajectories (not individual patents or patent citation pairs) are the most obvious unit for calculating co-occurrence metrics. Our analysis here can be seen as an attempt at proof-of-concept of this main idea.

Let us illustrate this with two examples. Figure 3 depicts two actual trajectories identified by our algorithm. On the first trajectory (upper panel of Figure 3) of length 12, we see four brown patents historically followed by eight green patents. All the green patents belong to the CPC category Y02T 10 (Climate Change Mitigation Technologies Related to Road Transport of Goods or Passengers). Looking at the patent titles, we observe the evolution of the usage of continuous variable transmission (CVT) systems, into electric and hybrid vehicles. It is indeed well-known that CVT systems[12] that were originally developed (around late 1950s) for vehicles with a single combustion engine (clearly, browntech), have provided the basis for the design of more sophisticated systems (e.g., Electric Variable Transmission, e-CVT) that are able to apply power from multiple sources of actuation to one output, such as a hybrid vehicle (greentech) which has both a combustion engine and an electric motor (and in some cases also a flywheel). This trajectory nicely captures that main trend.

The trajectory features inventors from 6 different countries, where Japanese inventors hold 50%, German inventors 16.67%, and four other countries each 8.33% of the 12 patents.[13]

[12] In contrasts with other transmissions that provide a limited number of gear ratios in fixed steps, a CVT system offers a continuous range of ratios, and thereby allows an engine to operate at a constant RPM while the vehicle moves at varying speeds.

[13] Throughout the analysis, we attribute patents to countries solely on the basis of location of the inventors.

Individual patents have one to (at most) four IPC codes (at 4-digit resolution) each. However, when we look beyond individual patents and consider the trajectory as a whole entity, we find 7 different IPC codes, F16H[14] showing up 4 times exclusively on brown patents, B60K[15] appearing 10 times on both green and brown patents, and the rest (B60W, B60L, F02D, A01B, and B62D) showing up (respectively on eight, four, three, one and one times) exclusively on the green patents.

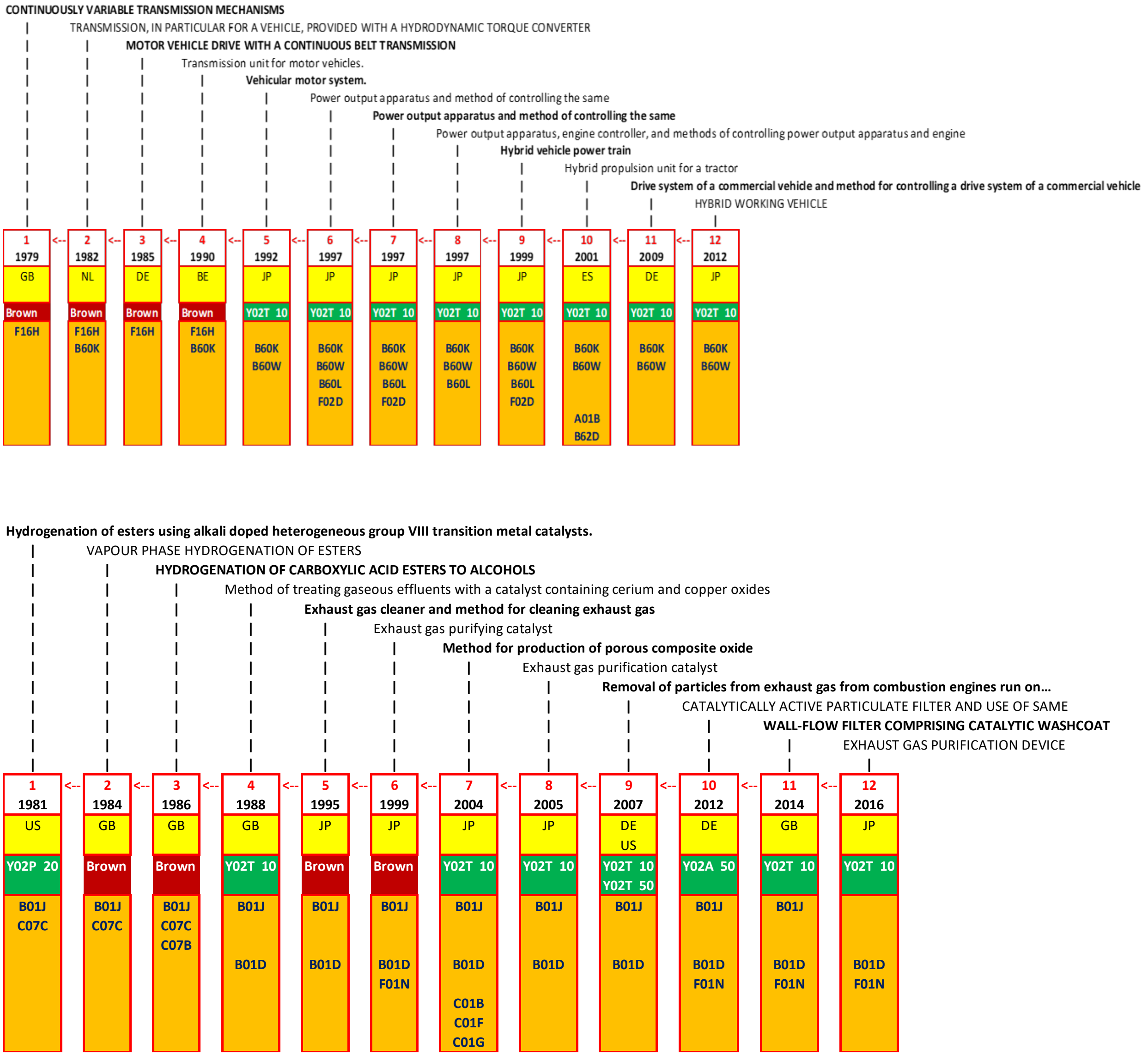

CONTINUOUSLY VARIABLE TRANSMISSION MECHANISMS
TRANSMISSION, IN PARTICULAR FOR A VEHICLE, PROVIDED WITH A HYDRODYNAMIC TORQUE CONVERTER
MOTOR VEHICLE DRIVE WITH A CONTINUOUS BELT TRANSMISSION
Transmission unit for motor vehicles.
Vehicular motor system.
Power output apparatus and method of controlling the same
Power output apparatus and method of controlling the same
Power output apparatus, engine controller, and methods of controlling power output apparatus and engine
Hybrid vehicle power train
Hybrid propulsion unit for a tractor
Drive system of a commercial vehicle and method for controlling a drive system of a commercial vehicle
HYBRID WORKING VEHICLE
1 1979 GB Brown F16H
2 1982 NL Brown F16H B60K
3 1985 DE Brown F16H
4 1990 BE Brown F16H B60K
5 1992 JP Y02T 10 B60K B60W
6 1997 JP Y02T 10 B60K B60W B60L F02D
7 1997 JP Y02T 10 B60K B60W B60L F02D
8 1997 JP Y02T 10 B60K B60W B60L
9 1999 JP Y02T 10 B60K B60W B60L F02D
10 2001 ES Y02T 10 B60K B60W A01B B62D
11 2009 DE Y02T 10 B60K B60W
12 2012 JP Y02T 10 B60K B60W
Hydrogenation of esters using alkali doped heterogeneous group VIII transition metal catalysts.
VAPOUR PHASE HYDROGENATION OF ESTERS
HYDROGENATION OF CARBOXYLIC ACID ESTERS TO ALCOHOLS
Method of treating gaseous effluents with a catalyst containing cerium and copper oxides
Exhaust gas cleaner and method for cleaning exhaust gas
Exhaust gas purifying catalyst
Method for production of porous composite oxide
Exhaust gas purification catalyst
Removal of particles from exhaust gas from combustion engines run on…
CATALYTICALLY ACTIVE PARTICULATE FILTER AND USE OF SAME
WALL-FLOW FILTER COMPRISING CATALYTIC WASHCOAT
EXHAUST GAS PURIFICATION DEVICE
1 1981 US Y02P 20 B01J C07C
2 1984 GB Brown B01J C07C
3 1986 GB Brown B01J C07C C07B
4 1988 GB Y02T 10 B01J B01D
5 1995 JP Brown B01J B01D
6 1999 JP Brown B01J B01D F01N
7 2004 JP Y02T 10 B01J B01D C01B C01F C01G
8 2005 JP Y02T 10 B01J B01D
9 2007 DE US Y02T 10 Y02T 50 B01J B01D
10 2012 DE Y02A 50 B01J B01D F01N
11 2014 GB Y02T 10 B01J B01D F01N
12 2016 JP Y02T 10 B01D F01N
<--

**Figure 3. Two actual examples of a green trajectory**

[14] Gearing.

[15] Arrangement or mounting of propulsion units or of transmissions in vehicles.

Our second example is once again an actual trajectory of length 12, and once again the trajectory comprises four brown and eight green patents. Similar to our first example, we find the brown patents showing up earlier than the green ones. However, the first and the fourth patents of the trajectory are green in this case. As in the first example, this trajectory is also related to the automotive industry, however the context of greentech is quite different (i.e., exhaust gas filtering in fossil-fuel-based vehicles). In addition to the dominance of Y02T 10 (by showing up on the majority of the green patents), we find two more green subcategories here: Y02P 20 (Technologies relating to chemical industry/ purification) and Y02A 50 (Air quality improvement or preservation, e.g., vehicle emission control…).

The IPC code B01J[16] shows up on almost all patents, B01D[17] on all of the last nine patents, C07C[18] only on the first three, F01N[19] on the last three but also on the sixth patent. The last three of the eight IPC codes that we find on this trajectory (C01B, C01F and C01G, all on chemistry of compounds) only appear on a single patent (the seventh one). With the exception of these last three IPC codes (and unlike the case with our first example), none of the IPC codes are exclusive to the brown or the green category. Another interesting matter to note regards the ninth patent of the trajectory. This patent has inventors originating from more than one country (Germany and the US) and has also been tagged by more than one green CPC code (both Y02A 10 and Y02A 50).

Following these examples, let us elaborate on our methodological basis for the definition of the network that we use for technological landscaping. This draws on the co-occurrence concept, operationalized by fractional accounting. On our second exemplar trajectory, five of the 12 patents are characterized purely by the green CPC code Y02A 10, and there is also one patent which has two green codes, one of them being Y02A 10. Fractional counting implies that this trajectory consists of 5.5 patents featuring Y02A 10, which in terms of shares is 5.5/12=45.83%. Similarly, the share of the contribution of the brown patents to the trajectory is 4/12=33.33%, that of the two Y02A 50 patents (one shared with Y02A 10) is 1.5/12=12.5%, while that of Y02A 20 patent is 1/12=8.33%. Note that these percentages add up to 100 (except for rounding), which indicates that this set of Y02 tags (including the non-Y02, or brown category) exhausts the trajectory.

In terms of the respective contributions of countries, due to the co-presence of Germany (DE) and US in the ninth patent (i.e., a fractional contribution), the contribution of these two countries to the trajectory (also considering their exclusive appearance on the 10th and the

[16] Catalysis or colloid chemistry.

[17] Chemical separation.

[18] Acyclic or carbocyclic compounds.

[19] Gas-flow silencers or exhaust apparatus for internal-combustion engines.

1st patent) is 1.5/12=12.5%, while that of the United Kingdom (GB) is 4/12=33.33% and that of Japan is 5/12=41.67%. Again, these percentages add up to 100.

**Table 1. Co-occurrence of various types on example trajectory (on lower panel of Figure 3)**

OCCURRENCES

| | |
|---|---|
| Y02P 20 | 8.33% |
| Y02A 50 | 12.50% |
| Brown | 33.33% |
| Y02T 10 | 45.83% |
| | |
| DE | 12.50% |
| US | 12.50% |
| GB | 33.33% |
| JP | 41.67% |

CO-OCCURRENCES

| | Y02P 20 | Y02A 50 | Brown | Y02T 10 |
|---|---|---|---|---|
| Y02P 20 | 0.69% | **2.08%** | **5.56%** | **7.64%** |
| Y02A 50 | | 1.56% | **8.33%** | **11.46%** |
| Brown | | | 11.11% | **30.56%** |
| Y02T 10 | | | | 21.01% |

| | DE | US | GB | JP |
|---|---|---|---|---|
| DE | 0.69% | **2.08%** | **5.56%** | **7.64%** |
| US | | 1.56% | **8.33%** | **11.46%** |
| GB | | | 11.11% | **30.56%** |
| JP | | | | 21.01% |

Table 1 documents these occurrences and the implied co-occurrences for the second example trajectory. The complete picture of the contribution shares of green CPC categories (at 8-digit resolution)[20] and countries are reproduced in the left column, labelled as occurrences. But our ultimate interest is clearly not in these occurrence shares, but in the extent to which all possible pairs of these occurrence categories co-occur on the trajectory, that is, the breakdown of the occurrences to co-occurrences. We label this as fractional co-occurrence, which is equal to the multiplicative product of the respective occurrence shares of each pair. For instance, in the bottom part of Table 1, the total contribution of GB is 33.33% and that of Japan is 41.67%. Thus, 41.67% of Great Britain's 33.33% contribution to the trajectory (41.67%x33.33%=20.83%) is attributable to its co-occurrence with Japan and similarly 33.33% of Japan's 41.67% contribution to the trajectory (20.83%, once again) is attributable to its co-occurrence with the United Kingdom (GB). Clearly, these co-occurrence shares are symmetrical (the JP-GB co-occurrence value is equal to the GB-JP one). Hence, we collapse the above- and below-diagonal values into just once value and report it above the diagonal. Thus, the joint contribution of the GB-JP pair is calculated as 2x20.83%=41.66%. At the same time, the contribution of the co-occurrence of GB with itself amounts to 33.33%x33.33%=11.11% and that of Japan with itself is 41.67%x41.67%=21.01%.

To put this formally, on a given trajectory $t$ that contains $N_t$ patents, and given $M$ categories of a selected taxonomy (such as inventor countries or IPC codes) the fractionally-counted co-occurrence contribution of a pair of categories indexed as integers is

[20] In the PatStat convention for IPC and CPC codes, there are 8 characters (including white spaces) that appear before the '/' in the full code.

$$c_{ij}^{t} = \frac{o_i^t o_j^t}{N_t^2} \tag{1}$$

where $o_i$ and $o_j$ are respectively the fractional number of patents where category $i$ and category $j$ can be found on the trajectory. As already noted, the matrix of $c$ values is symmetric around the diagonal, (that is, for $i \neq j$, $c_{ij} = c_{ji}$), thus we can as well express the co-occurrence matrix in terms of a diagonal + upper diagonal form as

$$c_{ij}^{t} = \begin{cases} 2\frac{o_i^t o_j^t}{N_t^2} \; if \; i < j \\ \left(\frac{o_i^t}{N_t}\right)^2 \; if \; i = j \end{cases} \tag{2}$$

The upper diagonal form of the co-occurrence matrices for our second exemplar trajectory are shown on the right-hand side panels of Table 2 (respectively for countries and 8-digit CPC codes including brown). Remember that the vectors on the right-hand-side panels show the occurrence shares of the respective categories (i.e., $o_i/N$ for each ). Observe that the sum of the elements of each co-occurrence matrix is exactly 100%. That is, each matrix breaks the given 'one' trajectory down to  co-occurrences,  indicating the cooccurrence between pairs of different categories (i.e., off-diagonal values).

Having broken each trajectory down to a co-occurrence matrix, the elements of which add up to unity (or 100%), the next step is to aggregate over a selected set of trajectories (or trajectory groups) $ST$ into a large matrix whose rows and columns enumerate all the categories (i.e., countries, IPCs, etc.) that appear on the selected trajectories. In formal terms, the task is to construct the matrix according to either specification given by equations (1) or (2), i.e., either a symmetric form or a diagonal + upper diagonal form

$$AC_{ij}^{ST} = \sum\nolimits_{\forall t \in ST} c_{ij}^{t} \tag{3}$$

In the remainder of the paper, we will aggregate over trajectory groups, rather than individual trajectories, although we will still write the shorter term "trajectories" instead of the longer "trajectory groups" when we refer to the co-occurrence results and their usage in the analysis below. There are 246,551 (green) trajectory groups in the database, and our analysis starts by doing the co-occurrence calculations as explained so far on each of these. In terms of aggregation of co-occurrence over the trajectory groups, let us use a complete example, where we use only 4-digit Y02 (green) CPC codes (plus the non-Y02 or class). The upper panel of Table 2 shows an actual matrix[21] where $ST$ is the set of all 246,551 green trajectory groups in our green NMP. The last column shows the row sums of the matrix, which is obviously the occurrence of the category.

[21] In the diagonal + upper diagonal form according to equation (2).

**Table 2. Co-occurrences of 4_digit green CPC codes with each other and with brown.**

**Top panel: raw frequencies, middle panel: normalized frequencies (i.e., association strength), bottom panel: frequencies normalized excluding diagonal elements.**

| | Brown | Y02A | Y02B | Y02C | Y02D | Y02E | Y02P | Y02T | Y02W | Total Occur | Total occur only with others |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Brown | 124849 | 8125 | 6699 | 848 | 5465 | 13454 | 18050 | 18761 | 4159 | 162629 | 37780 |
| Y02A | | 2532 | 201 | 74 | 3 | 214 | 268 | 802 | 186 | 7468 | 4937 |
| Y02B | | | 3127 | 2 | 79 | 985 | 180 | 126 | 19 | 7273 | 4146 |
| Y02C | | | | 389 | 0 | 109 | 196 | 100 | 15 | 1061 | 672 |
| Y02D | | | | | 1880 | 20 | 20 | 16 | 0 | 4682 | 2802 |
| Y02E | | | | | | 9690 | 1888 | 833 | 279 | 18580 | 8891 |
| Y02P | | | | | | | 6426 | 384 | 542 | 17190 | 10764 |
| Y02T | | | | | | | | 12104 | 23 | 22627 | 10522 |
| Y02W | | | | | | | | | 2429 | 5041 | 2612 |

| | Brown | Y02A | Y02B | Y02C | Y02D | Y02E | Y02P | Y02T | Y02W |
|---|---|---|---|---|---|---|---|---|---|
| Brown | 1.16 | 1.65 | 1.40 | 1.21 | 1.77 | 1.10 | 1.59 | 1.26 | 1.25 |
| Y02A | | 11.19 | 0.91 | 2.31 | 0.02 | 0.38 | 0.52 | 1.17 | 1.22 |
| Y02B | | | 14.58 | 0.08 | 0.57 | 1.80 | 0.35 | 0.19 | 0.13 |
| Y02C | | | | 85.24 | 0.00 | 1.36 | 2.64 | 1.03 | 0.70 |
| Y02D | | | | | 21.15 | 0.06 | 0.06 | 0.04 | 0.00 |
| Y02E | | | | | | 6.92 | 1.46 | 0.49 | 0.73 |
| Y02P | | | | | | | 5.36 | 0.24 | 1.54 |
| Y02T | | | | | | | | 5.83 | 0.05 |
| Y02W | | | | | | | | | 23.57 |

| | Brown | Y02A | Y02B | Y02C | Y02D | Y02E | Y02P | Y02T | Y02W |
|---|---|---|---|---|---|---|---|---|---|
| Brown | | 3.62 | 3.56 | 2.78 | 4.29 | 3.33 | 3.69 | 3.92 | 3.50 |
| Y02A | | | 0.81 | 1.86 | 0.02 | 0.41 | 0.42 | 1.28 | 1.20 |
| Y02B | | | | 0.07 | 0.57 | 2.22 | 0.33 | 0.24 | 0.15 |
| Y02C | | | | | 0.00 | 1.52 | 2.25 | 1.17 | 0.72 |
| Y02D | | | | | | 0.07 | 0.06 | 0.04 | 0.01 |
| Y02E | | | | | | | 1.64 | 0.74 | 1.00 |
| Y02P | | | | | | | | 0.28 | 1.60 |
| Y02T | | | | | | | | | 0.07 |
| Y02W | | | | | | | | | |

Note that, thanks to fractional counting within each trajectory group (which makes the co-occurrence matrix of the trajectory group add up to one), the matrix elements add up to 246,551, which is the number of trajectory groups considered. [22]

The aggregated co-occurrence matrix shows that the majority of potential co-occurrences have been actually realized, but to different extents. An intuitive assessment of the relative extents, even with such a small number of categories, is not straightforward, especially because the categories are quite heterogenous in their total occurrences (e.g., there are significantly more brown patents in our green trajectories than green). A normalization of the matrix (against heterogeneity in the 'size' of the categories) greatly helps.

We use a highly common method of normalization, referred to as the Association method (Van Eck and Waltman, 2009) that draws on graph theory. The method divides each actual co-occurrence by the 'expected' (given the total occurrences) co-occurrence. In terms of a diagonal + upper-diagonal representation:

$$Assoc_{ij}^{ST} = \begin{cases} \frac{2AC_{ij}^{ST}}{E\left(AC_{ij}^{ST}\right)}, if\ i < j \\ \frac{AC_{ij}^{ST}}{E\left(AC_{ij}^{ST}\right)}, if\ i = j \end{cases} \tag{4}$$

with $E\left(AC_{ij}^{ST}\right) = \frac{O_i O_j}{2O}$, (5)

where $O_i = \sum_{\forall j} AC_{ij}^{ST}$, $O_j = \sum_{\forall i} AC_{ij}^{ST}$, and $O = \sum_{\forall i} \sum_{\forall j} AC_{ij}^{ST}$ (6)

Equation (5) gives expected co-occurrence in a scenario where existing total occurrences were randomly redistributed among the co-occurrence categories, while keeping the total occurrence of each category as it is. Equation (4) divides actual co-occurrence by this expected value, so that high (low) values will indicate stronger (weaker) than expected co-occurrence.

The association-normalized matrix of the 4-digit level CPC co-occurrences (i.e., top panel of Table 2 as computed according to equations 4 to 6) is shown in the middle panel of the same table. Here we clearly observe strong homophily. That is, the association of a category with itself is extremely strong (stronger than expected). For example, the actual co-occurrence of Y02C with itself is about 85 times higher than its expected value in a random network. As an example of non-homophily, the co-occurrence of brown and Y02A patents is less strong, but

[22] A trajectory group is the union set of a number (say $n$) trajectories where each constituent trajectory (all of the same length $L$) is an ordered set of $L$ patents. In our fractional accounting scheme, the weight of each patent of each contributing trajectory is $1/nL$. This way, the contribution of each patent that is common to all $n$ contributing trajectories to the trajectory group adds up to $1/L$, while that of an uncommon one that shows up at the kth position is $1/mL$, where $m$ is the number of all patents that co-occupy the kth position along the trajectory group.

still higher than one (i.e., higher than expected). In fact, actual co-occurrences of all categories with brown are more than expected (i.e., association values larger than unity). The other outstanding co-occurrences are between the pairs Y02A-Y02C, Y02B-Y02E, Y02C-Y02P, and Y02P-Y02W.

In this paper, vis-à-vis Nomaler & Verspagen (2019), our interest shifts from homophily to heterophily. That is, we focus here on the interaction between different categories. This means that, as is common practice in graph theory (especially when visualizing and clustering of co-occurrence networks), we ignore the diagonal values of the co-occurrence matrix (i.e., the co-occurrence of a category with itself). This implies replacing equation (5) with

$$O_i = \sum_{\forall j \neq i} AC_{ij}^{ST}, O_j = \sum_{\forall i \neq j} AC_{ij}^{ST}, \text{ and}^{23} \; O = \sum_{\forall i} \sum_{\forall j \neq i} AC_{ij}^{ST} \qquad (5a)$$

In terms of our co-occurrence network in the top panel of Table1, this approach (which we adopt throughout the rest of the paper) yields the association matrix shown on the bottom panel of Table 2. In qualitative terms, what this matrix suggests is quite similar to that suggested by the matrix in the middle panel (where co-occurrence of a category with itself is also accounted for), but the matrix that ignores the diagonal elements indicates an even stronger association between the brown patents and all the patents of all green subcategories.

The rest of this paper will draw on network graph representations of these co-occurrence matrices. Indeed, a co-occurrence matrix is a weighted network graph, where the nodes of the networks are the bibliometric categories (such as countries, IPC codes etc.) that are associated to individual patents and the weighted edges between pairs of nodes correspond to the co-occurrence of the pairs in trajectories. For visualization, we will use a combination of network graph layout and clustering methods. In particular, we use the so-called LinLog visualization method (Noack, 2007, 2009), combined with the modularity clustering technique proposed by Newman (2004). LinLog visualization and modularity clustering are closely related methods (see Noack 2009), i.e., they are based on the same principles and ideas. Modularity clustering provides a way of dividing the network nodes into communities (groups of nodes), based on the idea that co-occurrence linkages should be more frequent and stronger within the communities than between communities. In our maps, we will show these groups (clusters, or communities) by using colors. Because of the relatedness of the clustering and visualization methods, the nodes within a cluster will tend to appear close to each other in the map, i.e., the clusters will tend to appear as entities of their own. However,

[23] In our example on Table 2, the alternative vector of the occurrence figures $O_i$, that exclude the diagonal values of the co-occurrence matrix are indicated on the rightmost column of the top panel of the table.

because we reduce a highly-dimensional dataset to a 2-dimensional visualization and a limited number of clusters (typically 5-10), some of the clusters will overlap in the map.

As a preview, we provide in Figure 4 the network graph of the co-occurrence network given in Table 2 before. Observe that the brown category, which is closely associated with the other, green, categories is placed (by the layout algorithm) at a central location and the distance between pairs of nodes closely reflect the extent to which they are mutually associated (as indicated by the bottom panel of Table 2). For example, node Y02D which is the least associated with the rest of the nodes (other than brown) appears furthest from the rest of the nodes. On the basis of mutual association, the algorithm assigns nodes Y02C, Y02P, Y02E and brown to a single cluster, while each of the rest are assigned to a singleton cluster of its own.

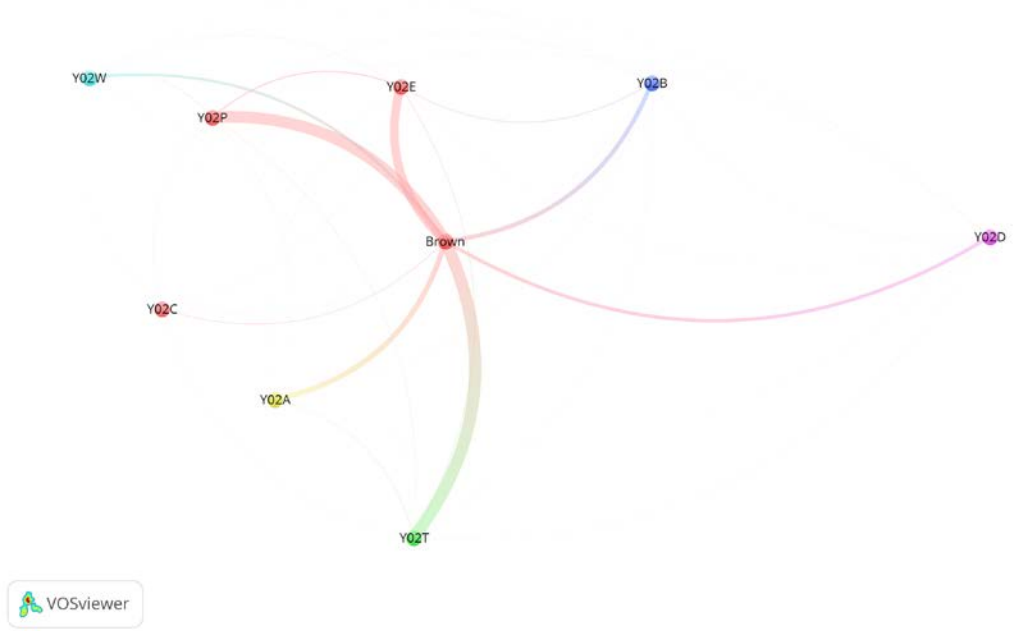


**Figure 4. Co-occurrence matrix in Table 2 visualized and clustered**

The next step is to define our networks in terms of what constitutes them, or, in other words, what the nodes are. The simplest choice would be to consider separately a network of countries (about 55 nodes), another network of green CPC codes (as in Figure 4, at 4-digit resolution), and another for IPC codes (at 4-digit resolution which comprises about 650 different categories). As convenient as it is (i.e., relatively small number of nodes), this straightforward approach would devoid our analysis from an important source of information that can only be harvested by also considering the cross co-occurrence of the various types of categories on the patents.

Let us revisit our example trajectory shown in the upper panel of Figure 3. As discussed, the IPC code F16H shows up exclusively on brown patents, B60K appears both as green and brown, and the other IPC codes found on the trajectory (B60W, B60L, F02D, A01B, and B62D) appear exclusively on green patents. Similarly, only the brown patents of this trajectory have inventors from the Netherlands, Belgium or the UK, while inventors from Japan and Spain exclusively contribute to the green patents, and German inventors contribute both to a green and a brown patent. Such valuable information can only be captured by considering various crossings between the basic categories.

**Table 3. Example of occurrence and co-occurrence of composite characteristics on a trajectory (that appears on the upper panel of Figure 3)**

| | |
|---|---|
| Br_BE | 8.33% |
| Br_DE | 8.33% |
| Br_GB | 8.33% |
| Br_NL | 8.33% |
| Gr_DE | 8.33% |
| Gr_ES | 8.33% |
| Gr_JP | 50.00% |

| | Br_BE | Br_DE | Br_GB | Br_NL | Gr_DE | Gr_ES | Gr_JP |
|---|---|---|---|---|---|---|---|
| Br_BE | 0.69% | **1.39%** | **1.39%** | **1.39%** | **1.39%** | **1.39%** | **8.33%** |
| Br_DE | | 0.69% | **1.39%** | **1.39%** | **1.39%** | **1.39%** | **8.33%** |
| Br_GB | | | 0.69% | **1.39%** | **1.39%** | **1.39%** | **8.33%** |
| Br_NL | | | | 0.69% | **1.39%** | **1.39%** | **8.33%** |
| Gr_DE | | | | | 0.69% | **1.39%** | **8.33%** |
| Gr_ES | | | | | | 0.69% | **8.33%** |
| Gr_JP | | | | | | | 25.00% |

| | |
|---|---|
| Gr_A01B | 2.08% |
| Gr_B62D | 2.08% |
| Gr_F02D | 6.25% |
| Br_B60K | 8.33% |
| Gr_B60L | 9.03% |
| Gr_B60K | 23.61% |
| Gr_B60W | 23.61% |
| Br_F16H | 25.00% |

| | Gr_A01B | Gr_B62D | Gr_F02D | Br_B60K | Gr_B60L | Gr_B60K | Gr_B60W | Br_F16H |
|---|---|---|---|---|---|---|---|---|
| Gr_A01B | 0.04% | **0.09%** | **0.26%** | **0.35%** | **0.38%** | **0.98%** | **0.98%** | **1.04%** |
| Gr_B62D | | 0.04% | **0.26%** | **0.35%** | **0.38%** | **0.98%** | **0.98%** | **1.04%** |
| Gr_F02D | | | 0.39% | **1.04%** | **1.13%** | **2.95%** | **2.95%** | **3.13%** |
| Br_B60K | | | | 0.69% | **1.50%** | **3.94%** | **3.94%** | **4.17%** |
| Gr_B60L | | | | | 0.82% | **4.26%** | **4.26%** | **4.51%** |
| Gr_B60K | | | | | | 5.57% | **11.15%** | **11.81%** |
| Gr_B60W | | | | | | | 5.57% | **11.81%** |
| Br_F16H | | | | | | | | 6.25% |

We opt to define this in two different ways, each of which is a combination of the (non-)green nature of a patent, and one other characteristic, which is either the IPC class, or the geographical location of the patent. Thus, we will define each patent in terms of two 2-dimensional characteristic sets: (brown/green, 4-digit IPC class) and (brown/green, country of inventor). This yields characteristics such as B_B60K (B for brown and B60K as the IPC class), Y_B60W (Y for Y02, i.e., a green patent) and B60W as the IPC class, B_DE (DE for a German patent) or Y_US (for a US patent). We will draw separate networks for each of the two characteristic sets (based on either IPC classes or countries). Let us illustrate this further. Table 3 enumerates all the composite characteristics that are found on the trajectory depicted earlier on the upper panel of Figure 3. The crossing between 6 different participating countries and the green/brown divide implies 12 composite characteristics, however only seven of these are realized on the trajectory. Similarly, out of the 14 possible crossings between the green/brown divide and the seven 4-digit IPC codes found on the trajectory, only eight combinations are realized. The computation of occurrence shares and the co-occurrence shares are based on the same principles of fractional counting.

Last but not the least, let us explain our filtering strategy, which is a common practice in network visualization and clustering. We find that quite a few of the composite characteristics network nodes that we identify on our trajectories appear on only a few trajectories and therefore many of the co-occurrences are very small numbers. We consider these small numbers as noise and in order to avoid cluttering of the network graphs, we eliminate any node whose total co-occurrence with others amounts to less than 20 trajectories and also any co-occurrence number that amounts to less than one full trajectory. Although these specific thresholds are arbitrary, their exact values do not influence the main results in our analysis, while applying no filtering does clutter our maps significantly. If none of the co-occurrences of a node with the other nodes satisfied the latter criteria (i.e., if the node becomes an isolate), that node was also eliminated from the network. In the resulting dataset, there are 709 elements in the (brown/green, 4-digit IPC class) set (423 of these are brown, the other 286 green), and 93 in the (brown/green, country of location) set (50 brown, 43 green).

### 4.2. Results: IPC classes

We will start with the landscaping of the dataset with green/brown and IPC (4-digit) combinations. In Figure 5, we show the distribution of association values of the edges (links) in this network. The figure is based on the set of all trajectories with at least one green patent. In this figure, both axes are on a log scale, and therefore the approximate linearity (for most of the range) of the relationship suggests a power law distribution. This implies a high degree of skewness, with just a few edges that have very high association value, and many edges with small association. The vertical segment of the distribution on the right (i.e., for low values of association) is likely due to the thresholding (filtering) that we applied in building the network, which has led to the cutting of edges with small weight.

Next, in Figure 6, we document some characteristics of the nodes in the network. Here we look both at total node weight (the variable *O* as specified in the previous sub-section), and at the diversification of nodes in terms of the distribution of their association with other nodes, i.e., the edge values in the network. The diversification variable is defined as the inverse of the Herfindahl index of the association values with other nodes. In other words, for every node *j*, we first calculate $/O_j$ for all partner nodes , then calculate the sum of squares of these (this is the Herfindahl index), and finally take the inverse. Because there are 708 potential partner nodes in the network, the maximum value for the inverse Herfindahl is 708. This value would correspond to an entirely equal distribution of association values over the

708 partner nodes. The smaller the inverse Herfindahl of a particular node is, the more unequal the distribution of association values is.

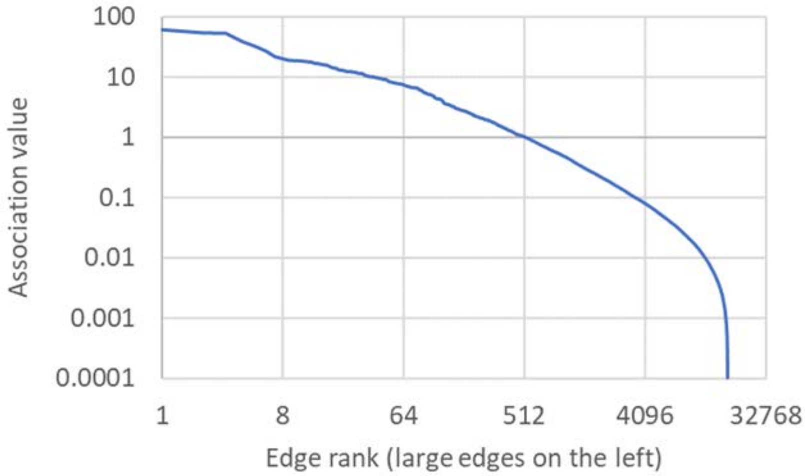


**Figure 5. Distribution of association values in the green/brown and IPC network, all green trajectories**

The horizontal axis of the figure displays the rank of the value $O$ for the node, the vertical axis the diversification value (inverse Herfindahl). Green and brown nodes are indicated by different colors. The overall relationship is again one indication skewness, with the few largest nodes (in terms of the variable $O$) being most diversified. The maximum diversification value observed in the figure is about 99, which is significantly smaller than the theoretical maximum of 708. This implies that even the most diversified nodes have limited diversification. Diversification decays rapidly when nodes become smaller (in terms of the variable $O$). This relationship is pretty similar between green and brown nodes, although the fitted power law relations (dotted lines in the graph) show a slightly steeper slope for the set of green nodes.

Thus, before having actually visualized the network, we already have the impression that this network is one in which a small set of nodes plays a very central role in keeping the network together. The nodes with large weights are also the ones with large connectivity, and without these nodes the network would quickly fall apart. Table 4 looks at which these nodes are, as it shows the top-15 nodes in terms of the node weights $O$. The table documents the top-15 nodes for a number of categories of trajectory lengths. This starts, on the left-hand side, with very short trajectories (2 – 4 patents), and goes up to 15 – 28 patents length. The rightmost column with codes documents the ranking for trajectories of any length. The short description in the last column refers to the IPC code only, and only to the 'all lengths' column, and is an informal summary (by the authors) of the full title of the class. Note that one of these descriptions (for B01J) occurs twice, because that code appears in the table both as brown and green.

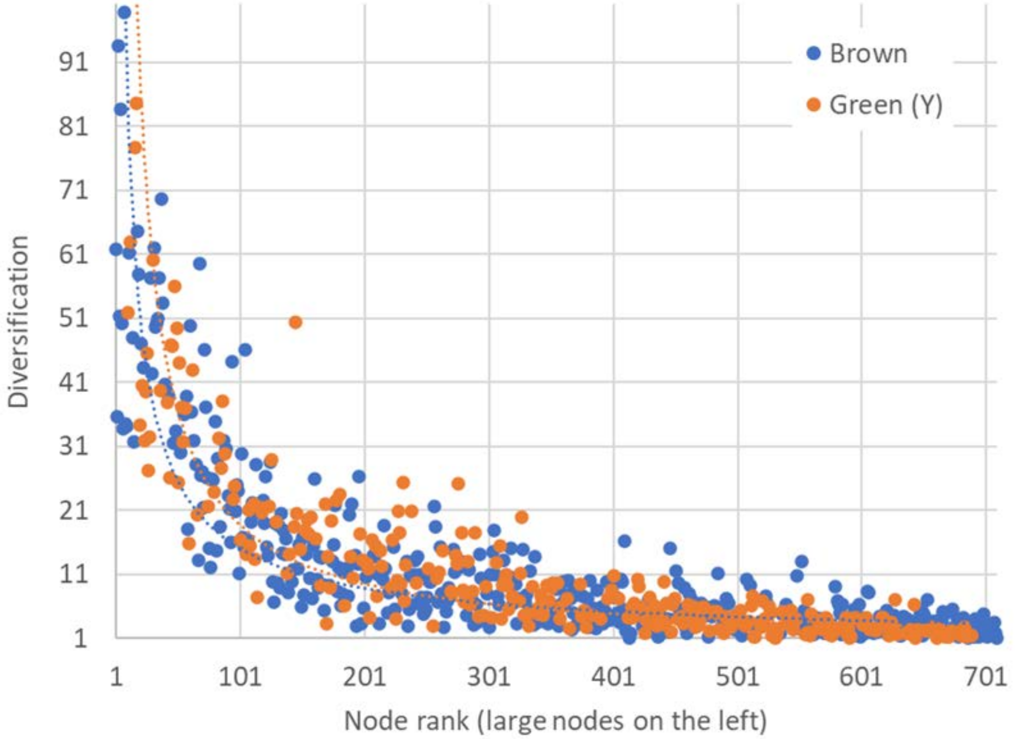


**Figure 6. Diversification of association values of individual nodes in the network, green/brown and IPC**

One interesting aspect of the rankings in the table is that for all trajectory lengths longer than 4 patents, as well as for the set of all trajectories, the brown labels dominate the top of the lists. For these cases, the first green (Y) label appears at rank 9 (Y_B01J for length 15-28). For all lengths, only two green (Y) labels appear in the top-15. Thus, except for very small trajectory lengths, brown technology plays a dominant role in the networks that we will use for the landscaping exercise.

Another interesting feature of the rankings is that many of the labels appear often across the columns for different trajectory length. There is a total of 29 labels in the five columns (which together have 75 positions. The top label (B_A61K) is common to all five columns, and there are five more labels that appear in all columns (B_B01J, B_C07C, B_C07D, B_H01M, Y_B01J), while there are 13 labels that appear in only one column.

**Table 4. Node labels of the top-15 nodes in terms of total node weight, for various trajectory length, brown/green and IPC codes**

| Rank | Len 2-4 | Len 5-8 | Len 9-14 | Len 15-28 | All Len | Short description of the IPC code (All length) |
|---|---|---|---|---|---|---|
| 1 | B_A61K | B_A61K | B_A61K | B_A61K | B_A61K | Medical preparations |
| 2 | B_C07C | B_B01J | B_C07D | B_C12N | B_C07D | Heterocyclic compounds |
| 3 | Y_H01M | B_C07C | B_H01L | B_C07D | B_B01J | Chemical or physical processes |
| 4 | B_B01D | B_B01D | B_H04W | B_C07K | B_C07C | Acyclic or carbocyclic compounds |
| 5 | Y_H01L | B_H01L | B_H01M | B_A61P | B_H01L | Semiconductor devices |
| 6 | B_B01J | B_G06F | B_C12N | B_H01M | B_H01M | Conversion of chemical to electrical energy |
| 7 | Y_B01D | B_C07D | B_C07K | B_C12P | B_C12N | Microorganisms or enzymes |
| 8 | Y_C07C | B_H04L | B_B01J | B_C07C | B_B01D | Separation |
| 9 | B_H01M | B_H01M | B_C08L | Y_B01J | B_A61P | Therapeutic activity of chemicals |
| 10 | Y_B01J | B_F01D | B_H04L | B_B01J | B_C07K | Peptides |
| 11 | Y_A61K | Y_B01J | B_A61P | Y_F01N | Y_H01M | Conversion of chemical to electrical energy |
| 12 | B_G06F | Y_B01D | B_H04B | B_C12Q | B_G06F | Digital data processing |
| 13 | Y_H02J | Y_H01M | B_C07C | B_A61Q | Y_B01J | Chemical or physical processes |
| 14 | B_H01L | Y_F02D | Y_B01J | Y_H01M | B_H04L | Transmission of digital information |
| 15 | B_C07D | B_A61P | B_G06F | B_C12R | B_H04W | Wireless communication networks |

The landscaping map that we produced is in Figure 7.[24] The size of the nodes is proportional to node weight *O*. The colors of the nodes in this map represent the clusters that are found using the modularity method. There are six clusters in the graph, ranging in size from 178 nodes (the red cluster in the center) to 77 nodes (the cyan cluster on the right-top). Exact cluster configurations for all landscape maps are documented in the appendix. Each of these clusters is built around a number of the top-nodes in Table 4. For example, the large red cluster includes nodes B_H01L (semiconductor devices), B_H01M and Y_H01M (both conversion of chemical to electrical energy), while the top-node B_A61K is found in the smaller cyan cluster.

The areas of the map contain clearly identifiable green sub-technologies. In the center of the map, in the red cluster, and roughly spanning the are closely surrounding the two large nodes B_H01L (semiconductor devices) and B_H01M (conversion of chemical to electrical energy), we find battery technology. This area includes both the brown and green versions of H01M and H01L.

[24] We use VOS viewer to produce all landscaping maps in this paper, always in LinLog/Modularity mode. The map in this figure is produced with attraction = 4, repulsion = 2 and resolution = 0.7.

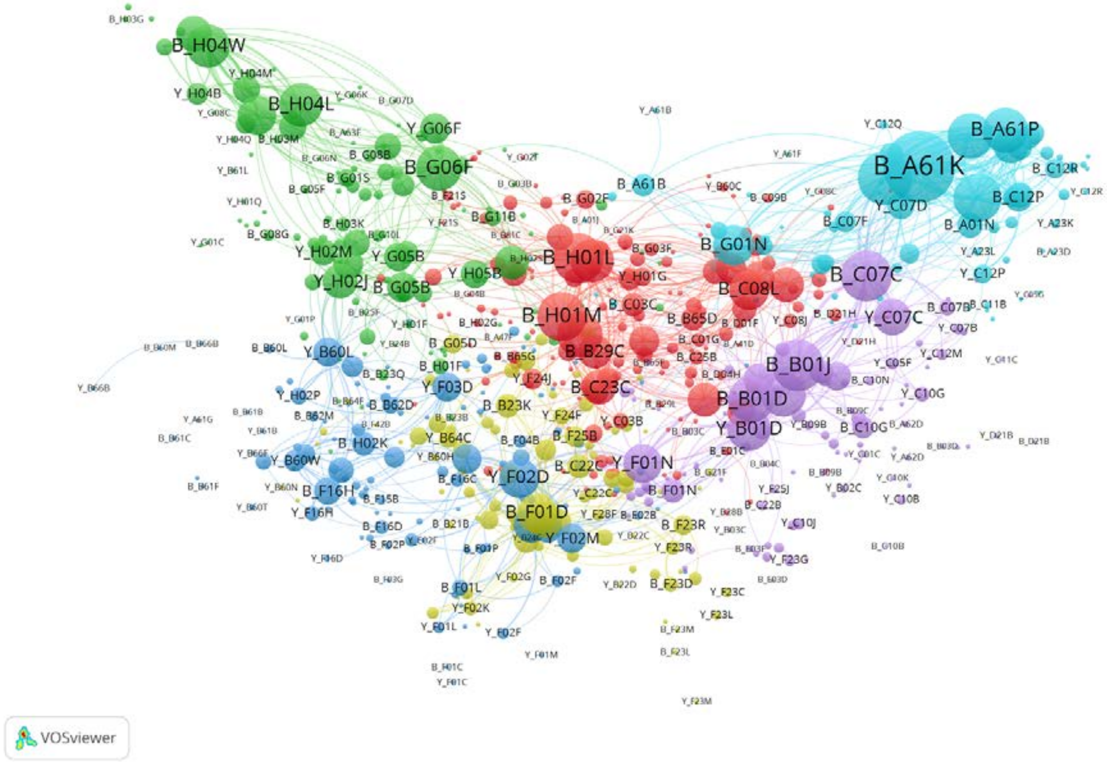


**Figure 7. Landscaping map for trajectories of all length, green/brown and IPC codes**

On the top left, the green cluster spans ICT and electrical technologies, with three large sub-areas: wireless communication (at the top right extreme of the cluster, with IPC codes H04W, wireless communication, and H04L, transmission of digital information, both in green and brown versions), computing (the core is IPC code G06F both in green and brown versions), and electrical power distribution in the lower part of the green cluster (with codes H02J, distributing power; H02M, power conversion, and G05B, control systems, all in green and brown versions).

The green cluster with its large emphasis on electronics and electricity fits in with the other (lower) parts on the left-hand side of the graph, as most of these have a clear link to electric power. The top area in the light blue cluster with codes Y_B60L (electric vehicles) and Y_H02P (electric motors) is led by green nodes (code Y), and clearly corresponds to electric mobility. The area below that, with large nodes Y_B60W (hybrid vehicles), B_H02K (dynamos) and B_F16H (gearing) can either be seen as a closely related sub-part, or part of the larger electric mobility group. In between the electric mobility area and the battery area that we pointed to above, we find a wind power area, with the central blue node Y_F03D (wind motors).

The area with green and blue nodes in the center-bottom part of the graph is focused on (combustion) engines and turbines. The central nodes here are B_F01D (steam turbines), Y_F02D (control of combustion engines) and Y_F02M (fuel supply of combustion engines). The right-hand side of the graph contains two large groups. One of these is the purple cluster, which is built around the reduction, capture and control of emissions. Central nodes in this cluster are B_B01J (chemical and physical processes), B_B01D and Y_B01D (separation), B_F01N and Y_F01N (exhaust apparatus), and B_C07C and Y_C07C (acyclic or carbocyclic compounds).

Finally, the cyan cluster on the top right-hand side is a medical and health cluster. The central nodes here are B_A61K (medical preparations) and B_A61P (therapeutic activity of chemicals), as well as B_C12N (microorganisms or enzymes), B_A01N (biocides). Interestingly, in this cluster, a far majority of the large nodes are brown.

We also produced similar landscaping maps for the individual categories of trajectory lengths. Most of these (all categories shorter than 15 patents) are provided in the appendix. These maps show a large degree of similarity to the above map for all trajectory lengths. In other words, the general configuration between greentech sub-fields that we observe above, is generally also appropriate for specific trajectory lengths. We do document, however, the map for trajectory lengths 15 – 28 patents, in Figure 8.[25] This map has 73 nodes and four clusters, which vary in size from 31 nodes (the central red cluster, which corresponds to the largest cluster in the previous map) to 10 nodes (the yellow cluster).

This map also shows similarities to the previous map. In particular, some of the previous areas are still recognizable, e.g., the ICT cluster (wireless communication and computing), which is now yellow/green and found in the upper-right hand side corner; the health medicine cluster, which is now green and found in the upper-right hand side corner; the cluster on reduction, capture and control of emissions, which is now light blue and found at the bottom of the graph, and batteries (to the right of the center in the red cluster). However, a number of other prominent green technologies are no longer found in this graph, e.g., the electric mobility cluster(s) and wind power. The IPC codes (green or brown) corresponding to the core of these fields are no longer in the network. This means that these technologies are not represented in the longest and most cumulative trajectories of the database.

[25] This map was produced with settings attraction = 4, repulsion = 2, resolution = 0.7.

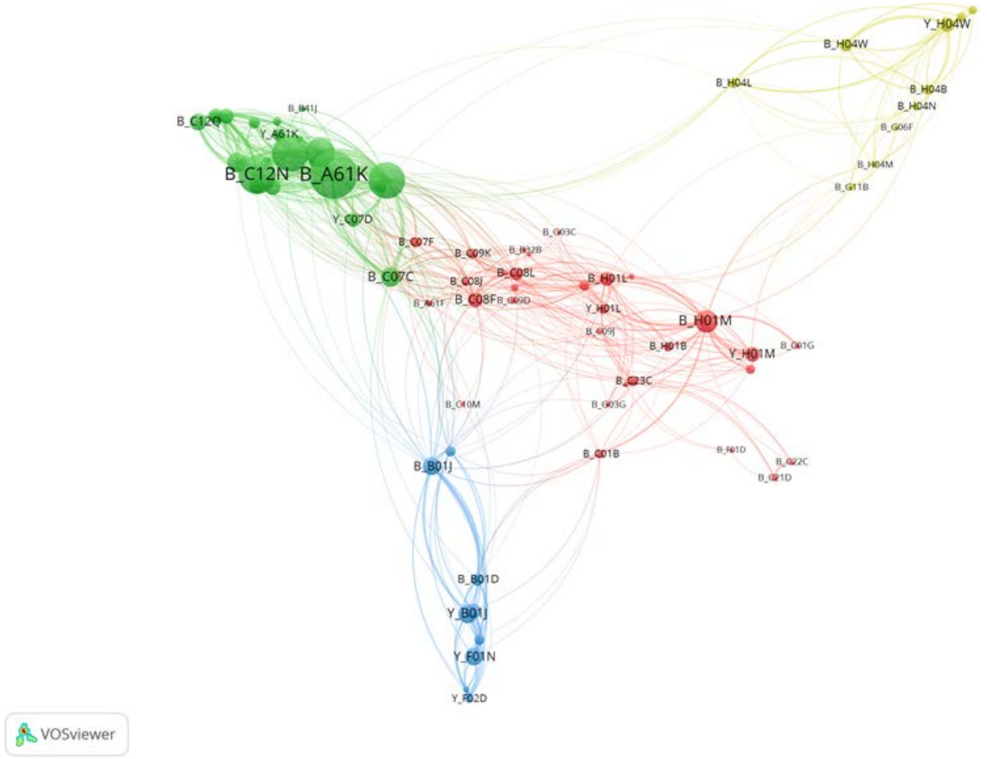


**Figure 8. Landscaping map for trajectories of length 15-28, green/brown and IPC codes**

### 4.3. Results: countries

We now move to describe the results of the landscaping exercise in the dataset with green/brown and country codes. Like before, we first look at some general properties of the distribution of node weight and edge weight in this network. Figure 9 describes the distribution of edge weight. As in the previous case, this is a skewed distribution that resembles a power law, except at the right end, where we find the small edge values (again, this is likely the result of applying a cutoff value for co-occurrences.

Figure 10 describes the relationship between node weight and diversification. Again, this is similar to the previous case, with the large nodes being the most diversified ones, and a power law being a reasonably good fit. Diversification peaks at around 87, which is much closer to the theoretical maximum (92) than before. Thus, nodes in the country network can be much more diversified (relatively speaking) than nodes in the IPC codes network. Like before, the distribution for green nodes resembles that for brown nodes.

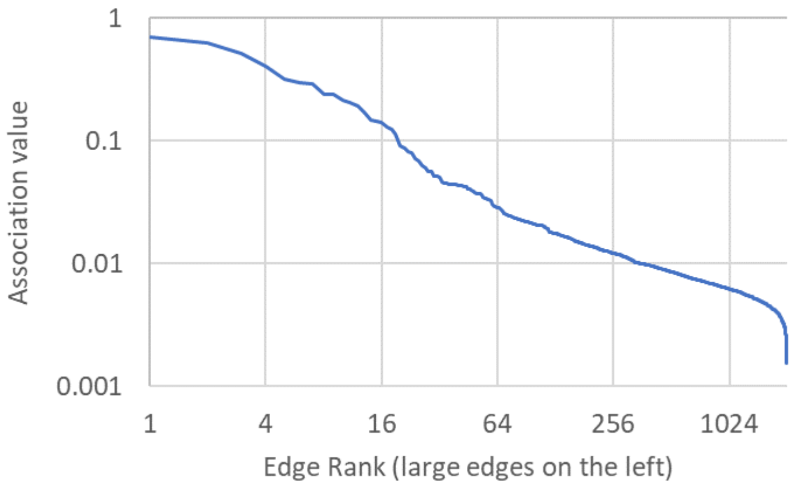


**Figure 9. Distribution of association values in the green/brown and country network, all green trajectories**

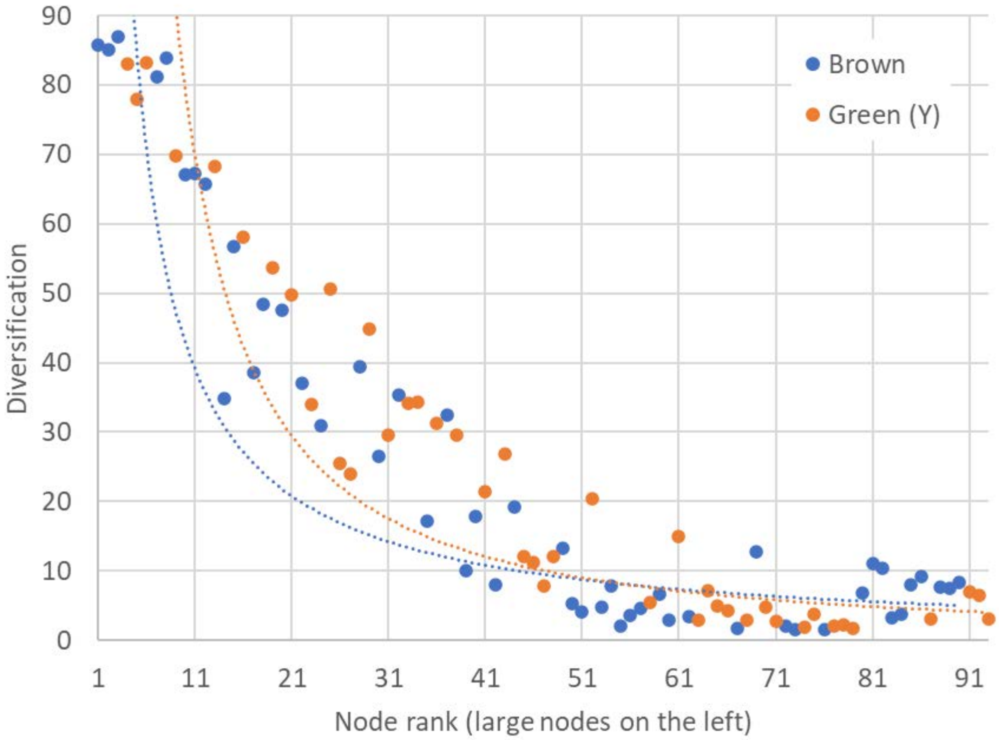


**Figure 10. Diversification of association values of individual nodes in the network, green/brown and countries**

Table 5 shows the top-15 nodes in terms of node weight *O*, again for various network lengths. Brown/US tops the table in each of the columns, with brown/Japan in second place for the longer trajectories, as well as the all-lengths network, and brown/Germany in second place for shorter trajectory lengths. Thus, we also see brown nodes dominating the top of the node weight distributions, as was the case before. However, in this case, the first green nodes start appearing from rank 3 and 4. In positions 1 – 5, no other countries than the US, Germany and Japan appear, with the UK (listed as GB) as the first other country appearing. The code B_?? indicates an unknown country. There are a total of 20 labels occupying the 75 positions in the table, with 11 labels occurring the maximum of five times. Three of those 11 labels are green (Y_US, Y_DE and Y_JP), the other eight are brown (B_US, B_DE, B_JP, B_FR, B_CH, B_GB, B_IT and B_NL).

**Table 5. Node labels of the top-15 nodes in terms of total node weight, for various trajectory lengths, brown/green and countries**

| Rank | Len 2-4 | Len 5-8 | Len 9-14 | Len 15-28 | All lengths |
|---|---|---|---|---|---|
| 1 | B_US | B_US | B_US | B_US | B_US |
| 2 | B_DE | B_DE | B_JP | B_JP | B_JP |
| 3 | Y_US | B_JP | B_DE | B_DE | B_DE |
| 4 | Y_DE | Y_US | Y_US | B_FR | Y_US |
| 5 | B_JP | Y_DE | Y_JP | Y_JP | Y_DE |
| 6 | Y_JP | Y_JP | B_GB | B_GB | Y_JP |
| 7 | B_FR | B_FR | B_FR | Y_US | B_FR |
| 8 | Y_FR | B_GB | Y_DE | B_BE | B_GB |
| 9 | B_GB | Y_FR | B_CH | Y_DE | Y_FR |
| 10 | Y_GB | B_IT | B_NL | B_CH | B_CH |
| 11 | B_IT | B_CH | B_IT | B_CA | B_IT |
| 12 | B_NL | B_NL | B_KR | B_NL | B_NL |
| 13 | B_CH | Y_GB | Y_FR | B_KR | Y_GB |
| 14 | Y_IT | B_KR | B_BE | B_?? | B_KR |
| 15 | Y_NL | B_SE | Y_GB | B_IT | B_SE |

The landscaping map for green/brown and countries is in Figure 11.[26] There are three clusters of nodes, which are clearly organized geographically. The largest cluster is the red one (44 nodes), which consists of relatively advanced European countries. Germany, France, Italy, the UK and the Netherlands comprise the largest nodes in this cluster. This cluster, like the other two, has a fairly even representation of brown and green nodes, although we notice that most often the two versions of an individual country are in the same cluster.

The second largest cluster has 35 nodes and is the green cluster. This cluster is dominated by non-European countries, although there are some European nodes in it as well. The US and Japan supply the largest nodes to this cluster, with B_GB also being a large node. Note that while B_GB is in the green cluster, Y_GB (the green version) is in the first cluster, thus making the UK one of the relatively few countries that are split over two clusters. Asian countries also make up a large part of the second, green cluster.

The third and final cluster is colored blue, and it has only 14 nodes. The largest node in this cluster is B_SE, which just makes it to the list in the last column of Table 5. Thus, this is a cluster with relatively small and peripheral nodes, which is also indicated by the position of the cluster in the graph.

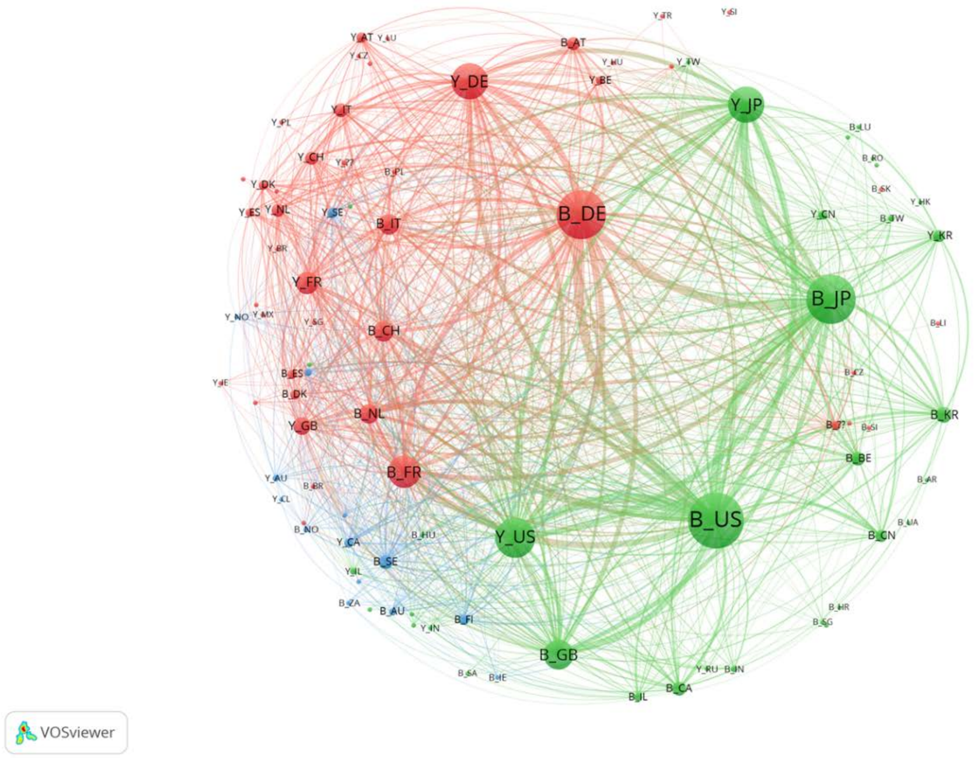


**Figure 11. Landscaping map for trajectories of all lengths, green/brown and countries**

[26] The settings fort his graph are attraction 3, repulsion 1, resolution 1.

Like before, we document the landscaping maps for trajectory lengths 2 – 14 in the appendix, and move to the map for the longest trajectories (15 – 28) in Figure 12. This map has two clusters, of 17 and 12 nodes (i.e., 29 nodes in total). The small cluster from the previous graph has disappeared from the graph, these nodes do not appear in the long trajectories often enough to be included. Interestingly, the clustering of this network has switched from geographical to the green/brown distinction. The largest cluster (red) has 16 brown nodes and one green node (Y_IT), the other cluster (green) has nine green nodes and three brown nodes. Thus, in the longest and most cumulative trajectories, the green/brown distinction is a more useful way of organizing the technology landscape than geography. This switch from geography to green/brown already happens in the network for length 9-15, as can be seen in the appendix.

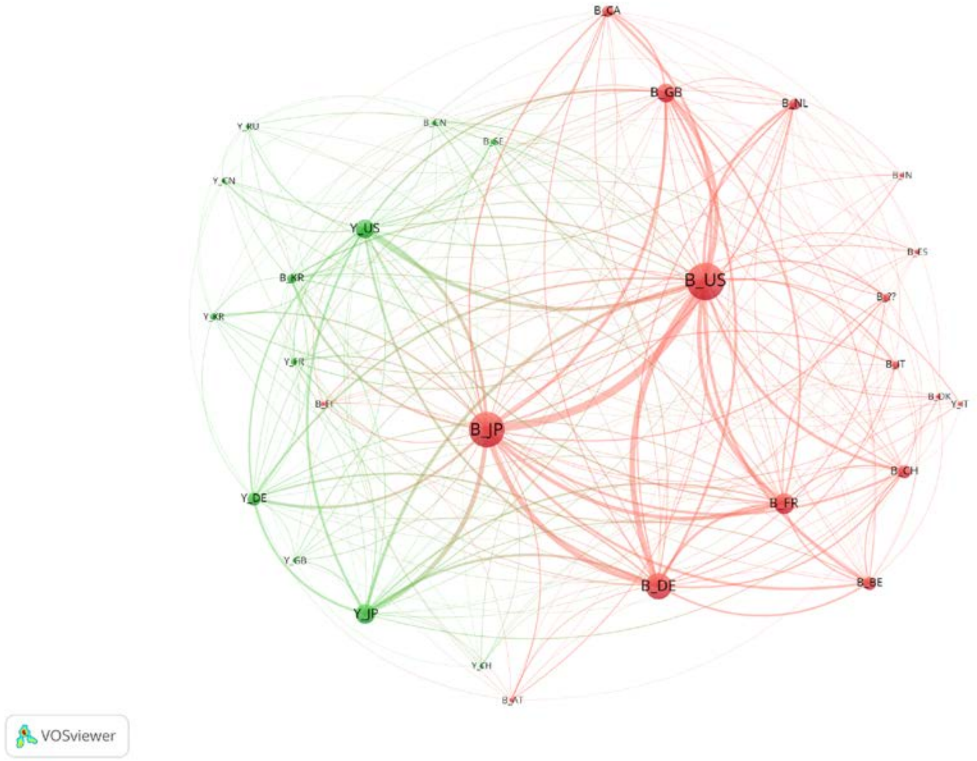


**Figure 12. Landscaping map for trajectories of length 15-28, green/brown and countries**

## 5. Summary and conclusions

In a previous paper (Nomaler and Verspagen, 2019), we introduced a method that uses a (very) large patent citation network to extract a collection of technological trajectories (which are represented by chains of citations) that are aimed at describing the global main technological trends over the last decades. Using this method, we extract a so-called network of main paths (NMP), which consists of overlapping paths that represent the trajectories that represent large technology flows, from the entire set of EPO patents. We characterized each patent on the NMP as either green (contributing to the mitigation of greenhouse gas emissions) or brown (non-green), and then selected the set of trajectories that contain at least one green patent.

In this paper, we use the set of green trajectories to perform a patent landscaping exercise that aims to map the main developments in the greentech field. To this end, we build a network from the database of green technological trajectories. The network is based on co-occurrence on the green trajectories. The nodes in our network are either combinations of green/brown and 4-digit IPC code, or combinations of green/brown and country of origin of the patent. In the first case, a node could, for example, be brown patents in class F16H, or green patents in class F01D. In the second case, we could have nodes like green patents from Germany, or brown patents from Japan. The networks are visualized using the LinLog method. In both cases (IPC classes or countries), we obtain sensible maps of the green technological landscape, which outline the relatedness between green technology sub-parts.

We argue that our landscaping method based on relations between technological fields that are extracted from technological trajectories fits the aim of outlining main technological trends better than methods that are based on individual patents or patent citation pairs. The reason is that the technological trajectories in our method are aimed at summarizing technological trends, and hence they are the most logical building blocks for mapping these trends. We look at the maps that we build as a proof-of-concept, and suggest that future patent landscaping work considers using trajectory-based metrics (our data are publicly available to support such work, see footnote 5).

A common feature between the network based on IPC codes and the one based on countries is that brown nodes play a very important role in the network. In both cases, the nodes that have the highest weight, are brown nodes. This is in line with conclusions from our earlier analysis (Nomaler and Verspagen, 2019), and implies that progress in greentech cannot be understood independently of developments in non-greentech technologies. We also find that both in the country and IPC network, the distribution of node and vertex weight is very skewed, with a few nodes or edges being responsible for the main part of co-occurrence. The large nodes are also the ones that are broadly connected, i.e., they keep the network together and link different sub-areas of the landscape maps.

In the network that uses IPC codes, we observe a number of very broad fields that transcend greentech as such, as well as technological areas that are clearly key to greentech. The main examples of the first type of fields (general) are ICT and electrical, and health and medical. These re broad technological areas that serve goals that are not necessarily related to greentech, but they show up as major parts of the greentech field in our maps.

The IPC-based map is broadly divided in one half that contains electricity-based technologies, and another half that has no direct relations to electricity. The electricity-based part includes the large ICT and electrical cluster, but also batteries electric motors and electric or hybrid mobility technologies, as well as power generation and distribution technology. In the non-electrical part of the map, the health/medical cluster is a large one, but we also find a large cluster with technologies aimed at reducing, controlling and capturing emissions and exhaust.

In the geography (country) based map, we find that location is the main dividing line. This map contains three large areas. One of these contains mostly countries outside Europe, with the US and Japan as the largest nodes. The other clusters are Europe-centered. All of these clusters contain a significant number of brown nodes.

We also produced separate maps for trajectories of different lengths, and we observe a large similarity between those and the maps for all trajectories. Differences are largest for the maps based on the longest trajectories. In the geography-based map with longest trajectories, the divide changes from geography-based to brown/green-based. In other words, the major divide in the geography network of longest trajectories is between green and non-green technologies, instead of Europe-non-Europe. In the IPC-based map with longest trajectories, the two general clusters (ICT and health/medical) remain clearly visible, but a number of typical green technologies, such as electric cars and wind power, vanish from the network. These technologies have not yet accumulated the long trajectories that are found in this network.

As our analysis is mostly a proof-of-concept of the idea that trajectories are a useful unit of analysis for patent landscaping, the policy relevance of our work has a major indirect component: to the extent that patent landscaping is used to inform policymakers (e.g., innovation policy, policy on intellectual property rights), the application of our method in such studies will be one of the ways in which our method could become policy relevant. However, there are also policy implications of the findings of our own patent landscaping exercise in green technology. First, as in our previous study, we found that non-green (brown) technology plays an important role in the green technology landscape. Policies aimed at making a green technology transition possible should therefore aim at greening non-green technologies as well as creating new and original green technology paths. Second, our landscaping maps show that large and broad technological areas such as ICT and health/medical are important sub-parts of the green technology field. Thus, a greentech technology policy should have a broad focus, rather than only focusing on very specific greentech areas such as electric vehicles. Finally, the geography-based maps that we produced show that greentech technology trajectories do not develop in geographical isolation, but rather as a collective international effort. Greentech policy should therefore transcend international borders, and be based on international R&D cooperation.

## 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.

Bubela, T., Gold, E., Graff, G. et al. (2013), Patent landscaping for life sciences innovation: toward consistent and transparent practices. Nature Biotechnology, vol. 31, 202–206

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

Federico, P., Heimerl, F., Koch, S. and S. Miksch (2017). A Survey on Visual Approaches for Analyzing Scientific Literature and Patents, IEEE Transactions on Visualization and Computer Graphics, vol. 23, pp. 2179-2198

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.

Kay, L., Newman, N., Youtie, J., Porter, A.L. and Rafols, I. (2014), Patent Overlay Mapping: Visualizing Technological Distance. Journal of the Association for Information Science and Technology, vol. 65: 2432-2443.

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

Leydesdorff, L., Kogler, D.F. and B. Yan (2017). Mapping Patent Classifications: Portfolio and Statistical Analysis, and the Comparison of Strengths and Weaknesses, Scientometrics, vol. 112: 1573-1591

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.

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.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69, 066133.

Noack, A. (2007). Energy models for graph clustering. Journal of Graph Algorithms and Applications, 11(2), 453–480.

Noack, A. (2009). Modularity clustering is force-directed layout. Physical Review E, 79, 026102

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

Nomaler, Ö. & B. Verspagen, 2019, greentech homophily and path dependence in a large patent citation network, UNU-MERIT working paper #2019-051

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

Van Eck, N.J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology, 60(8), 1635–1651.

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.

Yan, B. and Luo, J. (2017), Measuring technological distance for patent mapping. Journal of the Association for Information Science and Technology, 68: 423-437.

## Appendix. Additional landscaping maps

All maps use the same parameters as the corresponding map for all trajectory lengths in the main text.

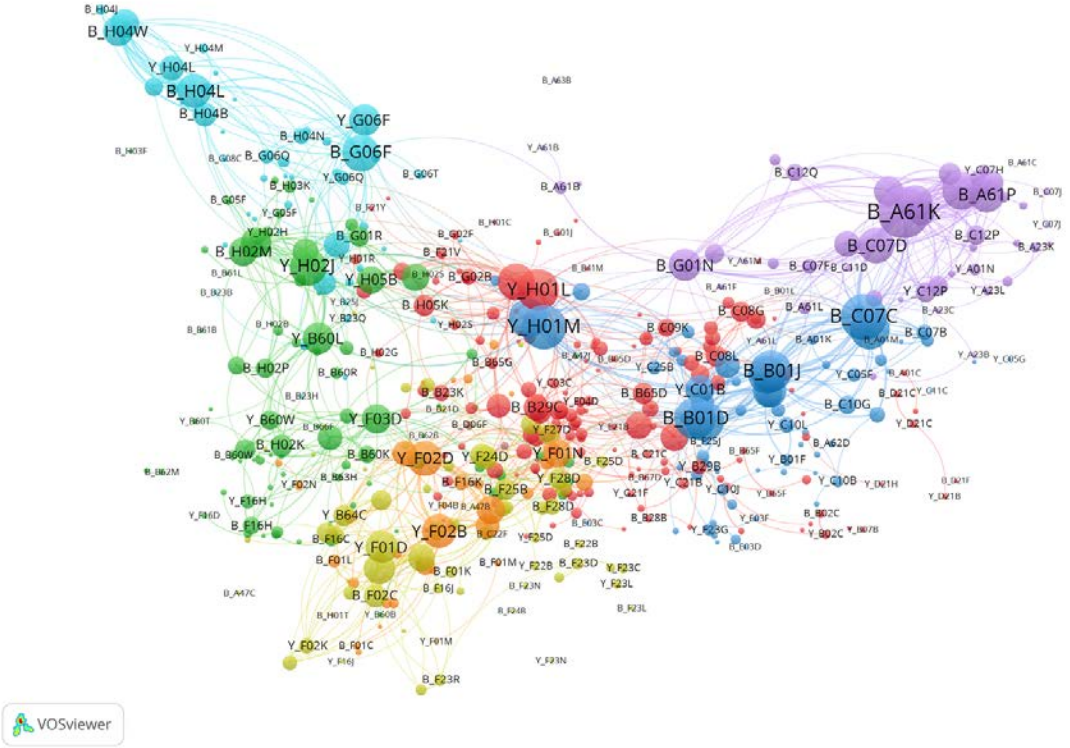


**Figure A1. Landscaping map for trajectories of length 2-4, green/brown and IPC codes**

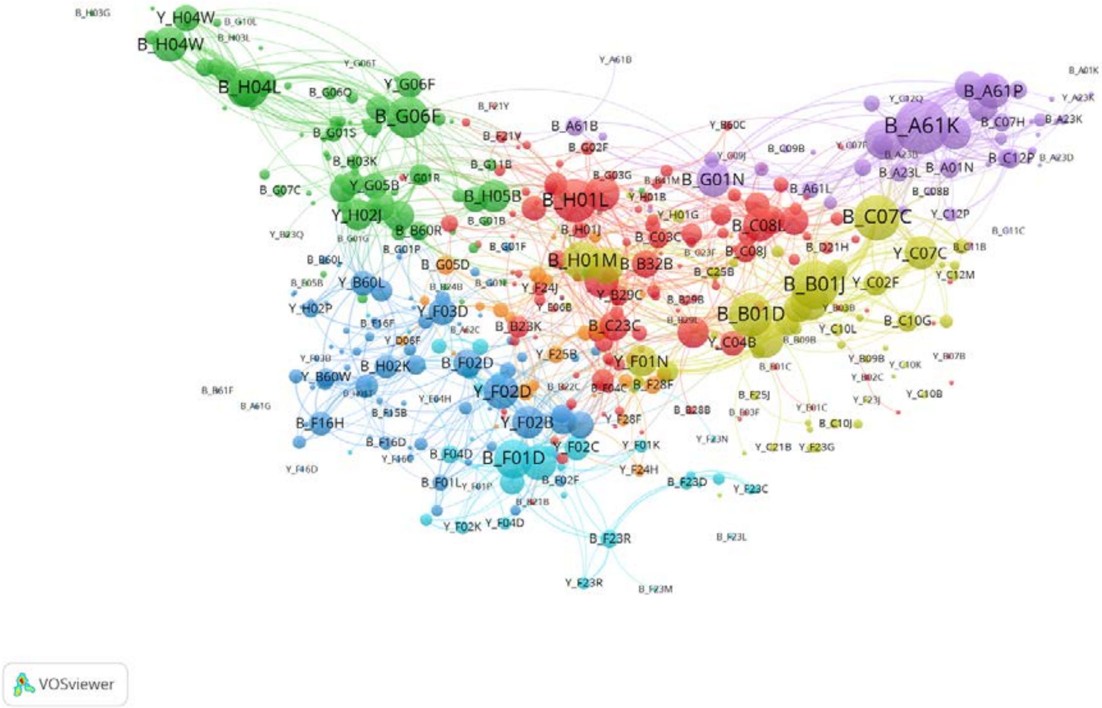


**Figure A2. Landscaping map for trajectories of length 5-8, green/brown and IPC codes**

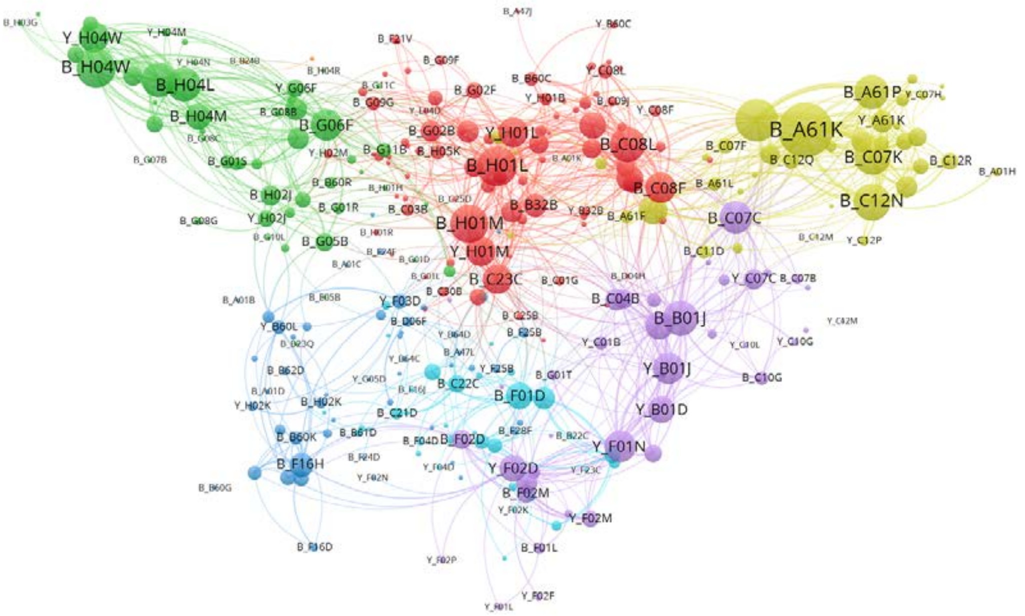


**Figure A3. Landscaping map for trajectories of length 9-14, green/brown and IPC codes**

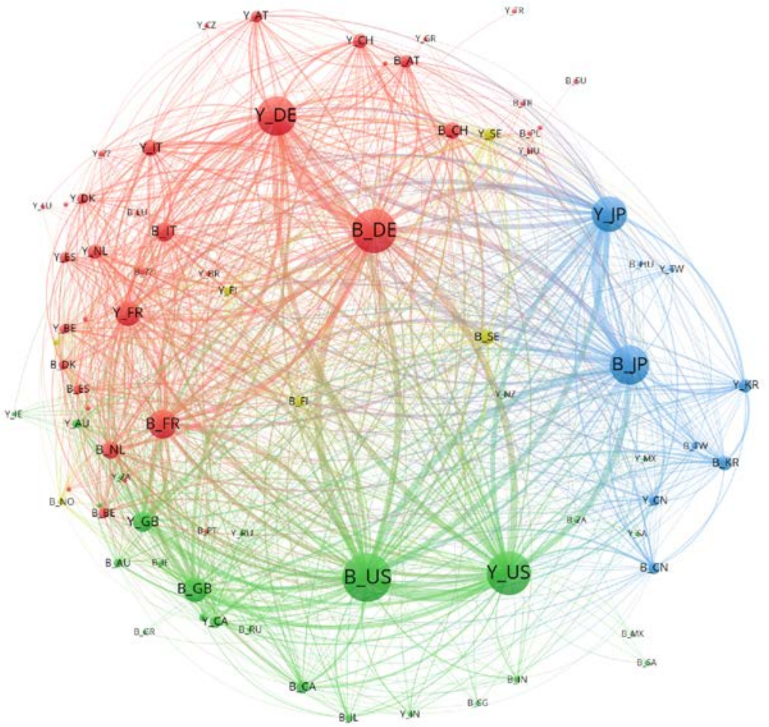
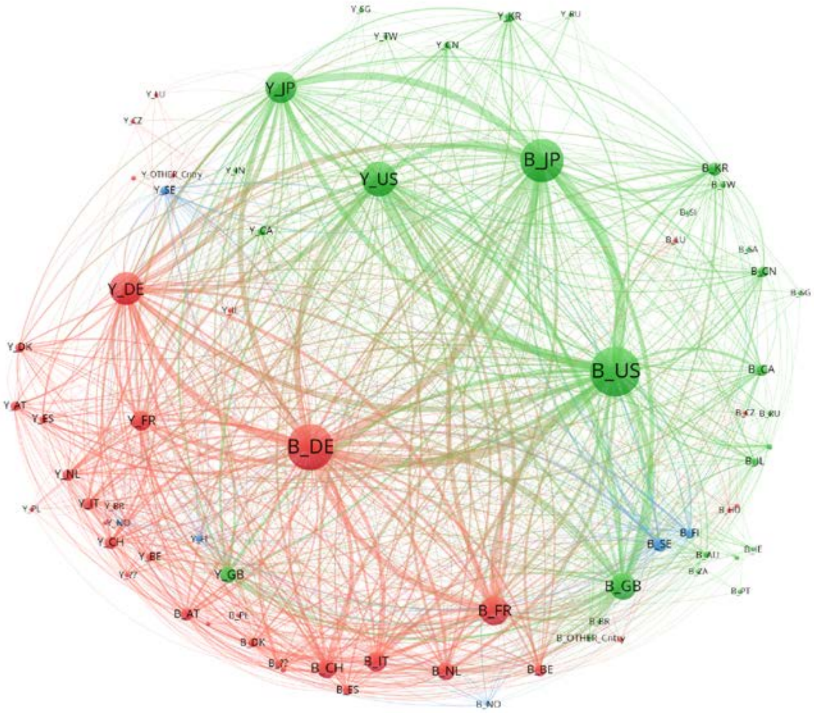
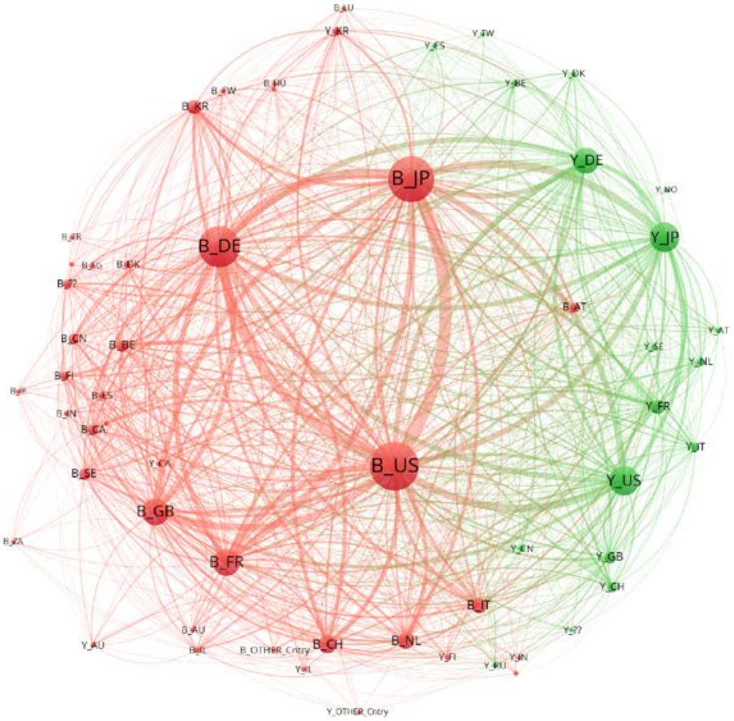


**Figure A4. Landscaping map for trajectories of length 2-4, green/brown and countries**



VOSviewer

**Figure A5. Landscaping map for trajectories of length 5-8, green/brown and countries**



VOSviewer

**Figure A6. Landscaping map for trajectories of length 9-14, green/brown and countries**

**Table A1. Cluster membership of brown/green & IPC combinations, by trajectory length**

| | Length | | | | | | Length | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 | | All | 2-4 | 5-8 | 9-14 | 15-28 |
| B_A01G | 1 | 1 | 1 | - | - | B_C03C | 1 | 1 | 1 | 1 | - |
| B_A41D | 1 | - | - | - | - | B_C04B | 1 | 1 | 1 | 5 | 3 |
| B_A43B | 1 | - | 1 | - | - | B_C08C | 1 | - | 1 | - | - |
| B_A44B | 1 | - | - | - | - | B_C08F | 1 | 1 | 1 | 1 | 1 |
| B_A47G | 1 | - | - | - | - | B_C08G | 1 | 1 | 1 | 1 | 1 |
| B_A62C | 1 | 4 | 1 | - | - | B_C08J | 1 | 1 | 1 | 1 | 1 |
| B_B05B | 1 | 1 | 1 | 1 | - | B_C08K | 1 | 1 | 1 | 1 | 1 |
| B_B05C | 1 | 1 | 1 | - | - | B_C08L | 1 | 1 | 1 | 1 | 1 |
| B_B05D | 1 | 1 | 1 | 1 | - | B_C09B | 1 | 1 | 5 | 1 | - |
| B_B24D | 1 | - | - | - | - | B_C09C | 1 | 1 | 1 | 1 | - |
| B_B26D | 1 | 1 | 2 | - | - | B_C09D | 1 | 1 | 1 | 1 | 1 |
| B_B26F | 1 | - | - | - | - | B_C09J | 1 | 1 | 1 | 1 | 1 |
| B_B27N | 1 | - | 1 | - | - | B_C09K | 1 | 1 | 1 | 1 | 1 |
| B_B28B | 1 | 1 | 1 | - | - | B_C23C | 1 | 1 | 1 | 1 | 1 |
| B_B28C | 1 | - | - | - | - | B_C23F | 1 | 1 | 1 | - | - |
| B_B29B | 1 | 1 | 1 | 1 | - | B_C23G | 1 | - | - | - | - |
| B_B29C | 1 | 1 | 1 | 1 | - | B_C25B | 1 | 3 | 4 | 1 | - |
| B_B29D | 1 | 1 | 1 | 1 | - | B_C25D | 1 | 1 | 1 | 1 | - |
| B_B29K | 1 | 1 | 1 | 1 | - | B_C30B | 1 | 1 | 1 | 1 | - |
| B_B29L | 1 | - | 1 | 1 | - | B_D01D | 1 | - | - | - | - |
| B_B31B | 1 | - | - | - | - | B_D01F | 1 | 1 | 1 | 1 | - |
| B_B32B | 1 | 1 | 1 | 1 | 1 | B_D01G | 1 | - | - | - | - |
| B_B33Y | 1 | - | - | - | - | B_D01H | 1 | - | 2 | - | - |
| B_B41C | 1 | - | - | - | - | B_D02G | 1 | - | - | - | - |
| B_B41M | 1 | 1 | 1 | 1 | - | B_D03D | 1 | - | 1 | - | - |
| B_B41N | 1 | - | - | - | - | B_D04B | 1 | - | 1 | - | - |
| B_B44C | 1 | - | - | - | - | B_D04H | 1 | 1 | 1 | 1 | - |
| B_B60C | 1 | 1 | 1 | 1 | - | B_D06M | 1 | 1 | 1 | 4 | - |
| B_B60J | 1 | 2 | 1 | 1 | - | B_D06N | 1 | - | - | - | - |
| B_B64G | 1 | 4 | - | - | - | B_D06P | 1 | - | - | - | - |
| B_B65B | 1 | 1 | 1 | 1 | - | B_D21H | 1 | 1 | 1 | 1 | - |
| B_B65D | 1 | 1 | 1 | 1 | - | B_E01C | 1 | 1 | 1 | - | - |
| B_B65G | 1 | 1 | 2 | 1 | - | B_E01F | 1 | 1 | 1 | - | - |
| B_B65H | 1 | 1 | 2 | 1 | - | B_E04B | 1 | 1 | 1 | 1 | - |
| B_B81B | 1 | - | - | - | - | B_E04C | 1 | 1 | 1 | - | - |
| B_B81C | 1 | - | - | - | - | B_E04D | 1 | 1 | 1 | - | - |
| B_B82Y | 1 | - | - | - | - | B_E04F | 1 | 1 | 1 | - | - |
| B_C01G | 1 | 3 | 4 | 1 | 1 | B_E04G | 1 | 1 | 1 | - | - |
| B_C03B | 1 | 1 | 1 | 1 | - | B_E06B | 1 | 1 | 1 | 1 | - |

Note: a dash (-) indicates that a node is not in the network.

### Table A1 (continued)

| | | Length | | | |
|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 |
| B_E21B | 1 | 1 | 1 | - | - |
| B_E21D | 1 | - | - | - | - |
| B_F16B | 1 | 1 | 1 | - | - |
| B_F16L | 1 | 1 | 1 | 1 | - |
| B_F21K | 1 | - | - | - | - |
| B_F21S | 1 | 1 | 1 | 1 | - |
| B_F21V | 1 | 1 | 1 | 1 | - |
| B_F21Y | 1 | 1 | 1 | - | - |
| B_G01J | 1 | 1 | 1 | - | - |
| B_G01Q | 1 | - | - | - | - |
| B_G02B | 1 | 1 | 1 | 1 | - |
| B_G02F | 1 | 1 | 1 | 1 | 1 |
| B_G03C | 1 | - | 1 | 1 | 1 |
| B_G03F | 1 | 1 | 1 | 1 | 1 |
| B_G03G | 1 | 1 | 1 | 1 | 1 |
| B_G03H | 1 | - | - | - | - |
| B_G04B | 1 | - | - | - | - |
| B_G09F | 1 | 1 | 1 | 1 | - |
| B_G21K | 1 | - | - | - | - |
| B_H01B | 1 | 1 | 1 | 1 | 1 |
| B_H01C | 1 | 1 | 2 | - | - |
| B_H01G | 1 | 3 | 4 | 1 | - |
| B_H01J | 1 | 1 | 1 | 1 | - |
| B_H01K | 1 | - | 1 | - | - |
| B_H01L | 1 | 1 | 1 | 1 | 1 |
| B_H01M | 1 | 3 | 4 | 1 | 1 |
| B_H01R | 1 | 1 | 1 | 1 | - |
| B_H01S | 1 | 1 | 1 | 1 | - |
| B_H02G | 1 | 1 | 1 | - | - |
| B_H02N | 1 | - | - | - | - |
| B_H02S | 1 | 1 | - | - | - |
| B_H05H | 1 | 1 | 1 | - | - |
| B_H05K | 1 | 1 | 1 | 1 | - |
| Y_A01G | 1 | 1 | 1 | 1 | - |
| Y_B05B | 1 | 1 | - | - | - |
| Y_B05D | 1 | - | - | - | - |
| Y_B28B | 1 | 1 | 1 | - | - |
| Y_B29B | 1 | 1 | 1 | - | - |
| Y_B29C | 1 | 1 | 1 | 1 | - |
| Y_B29D | 1 | - | 1 | - | - |
| Y_B29K | 1 | 1 | - | - | - |
| Y_B29L | 1 | - | - | - | - |
| Y_B32B | 1 | 1 | 1 | 1 | - |
| Y_B60C | 1 | 1 | 1 | 1 | - |
| Y_B60J | 1 | - | - | - | - |
| Y_B65B | 1 | - | - | - | - |
| Y_B65D | 1 | 1 | 1 | 1 | - |
| Y_B65G | 1 | 1 | - | - | - |
| Y_C03B | 1 | 1 | 1 | 1 | - |
| Y_C03C | 1 | 1 | 1 | 1 | - |
| Y_C04B | 1 | 1 | 1 | 5 | - |
| Y_C08C | 1 | - | - | - | - |
| Y_C08F | 1 | 1 | 1 | 1 | - |
| Y_C08G | 1 | 1 | 1 | 1 | - |
| Y_C08J | 1 | 1 | 1 | 1 | - |
| Y_C08K | 1 | 1 | 1 | 1 | - |
| Y_C08L | 1 | 1 | 1 | 1 | - |
| Y_C09B | 1 | - | - | - | - |
| Y_C09C | 1 | - | - | - | - |
| Y_C09D | 1 | 1 | 1 | 1 | - |
| Y_C09J | 1 | 1 | 1 | - | - |
| Y_C09K | 1 | 1 | 1 | 1 | - |
| Y_C23C | 1 | 1 | 1 | 1 | 1 |
| Y_C25B | 1 | 3 | 4 | 1 | - |
| Y_C25D | 1 | 1 | - | - | - |
| Y_C30B | 1 | 1 | 1 | - | - |
| Y_D01F | 1 | - | - | - | - |
| Y_D04H | 1 | - | - | - | - |
| Y_D21H | 1 | 1 | 1 | - | - |
| Y_E01C | 1 | 1 | 1 | - | - |
| Y_E04B | 1 | 1 | 1 | - | - |
| Y_E04C | 1 | 1 | 1 | - | - |
| Y_E04D | 1 | 1 | 1 | 1 | - |
| Y_E04F | 1 | 1 | - | - | - |
| Y_E06B | 1 | 1 | 1 | - | - |
| Y_E21B | 1 | 1 | - | - | - |
| Y_F16B | 1 | 1 | 1 | - | - |
| Y_F16L | 1 | 1 | 1 | - | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A1 (continued)**

| | | | Length | | |
|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 |
| Y_F21S | 1 | 1 | - | - | - |
| Y_F21V | 1 | 1 | - | - | - |
| Y_F21Y | 1 | - | - | - | - |
| Y_F24J | 1 | 1 | 1 | 1 | - |
| Y_F24S | 1 | 1 | 1 | - | - |
| Y_G01J | 1 | - | - | - | - |
| Y_G02B | 1 | 1 | 1 | 1 | - |
| Y_G02F | 1 | - | - | - | - |
| Y_G03F | 1 | - | - | - | - |
| Y_G09F | 1 | - | - | - | - |
| Y_G21B | 1 | 3 | - | - | - |
| Y_H01B | 1 | 1 | 1 | 1 | - |
| Y_H01G | 1 | 3 | 4 | 1 | 1 |
| Y_H01J | 1 | 1 | 1 | 1 | - |
| Y_H01K | 1 | - | - | - | - |
| Y_H01L | 1 | 1 | 1 | 1 | 1 |
| Y_H01M | 1 | 3 | 4 | 1 | 1 |
| Y_H01R | 1 | 1 | 1 | - | - |
| Y_H02G | 1 | 1 | 3 | - | - |
| Y_H02S | 1 | 1 | 1 | - | - |
| Y_H05H | 1 | - | - | - | - |
| Y_H05K | 1 | 1 | 1 | 1 | - |
| B_A01M | 2 | 5 | 5 | - | - |
| B_A21B | 2 | - | - | - | - |
| B_A47J | 2 | 2 | 2 | 1 | - |
| B_A63F | 2 | - | - | 2 | - |
| B_B23B | 2 | 6 | 2 | - | - |
| B_B23C | 2 | - | - | - | - |
| B_B23D | 2 | - | - | - | - |
| B_B23Q | 2 | 6 | 2 | 2 | - |
| B_B24B | 2 | 6 | 2 | 7 | - |
| B_B25F | 2 | - | 3 | - | - |
| B_B25J | 2 | 6 | 2 | 3 | - |
| B_B28D | 2 | 1 | - | - | - |
| B_B41F | 2 | 1 | 2 | 2 | - |
| B_B41J | 2 | 1 | 2 | 2 | 2 |
| B_B42D | 2 | - | 2 | 2 | - |
| B_B60Q | 2 | 2 | 2 | 1 | - |
| B_B60R | 2 | 2 | 2 | 2 | - |
| B_B61L | 2 | 2 | 2 | - | - |
| B_B65C | 2 | - | - | - | - |
| B_C06B | 2 | - | - | - | - |
| B_E05B | 2 | 4 | 2 | 2 | - |
| B_E05C | 2 | - | - | - | - |
| B_E05D | 2 | - | - | - | - |
| B_E05F | 2 | - | 3 | - | - |
| B_F24C | 2 | 2 | 2 | 1 | - |
| B_G01B | 2 | 6 | 2 | 2 | - |
| B_G01C | 2 | 6 | 2 | 2 | - |
| B_G01D | 2 | 6 | 3 | 2 | - |
| B_G01G | 2 | - | 2 | 3 | - |
| B_G01K | 2 | 4 | 7 | - | - |
| B_G01R | 2 | 2 | 2 | 2 | - |
| B_G01S | 2 | 6 | 2 | 2 | - |
| B_G01V | 2 | 6 | 1 | - | - |
| B_G01W | 2 | - | - | - | - |
| B_G02C | 2 | - | 1 | 7 | - |
| B_G03B | 2 | 1 | 2 | - | - |
| B_G04C | 2 | - | - | - | - |
| B_G04G | 2 | - | 2 | - | - |
| B_G05B | 2 | 6 | 2 | 2 | - |
| B_G05F | 2 | 2 | 2 | - | - |
| B_G06F | 2 | 6 | 2 | 2 | 4 |
| B_G06K | 2 | 6 | 2 | 2 | - |
| B_G06N | 2 | - | - | - | - |
| B_G06Q | 2 | 6 | 2 | 2 | - |
| B_G06T | 2 | 6 | 2 | 2 | - |
| B_G07B | 2 | - | 2 | 2 | - |
| B_G07C | 2 | 6 | 2 | 2 | - |
| B_G07D | 2 | - | - | 2 | - |
| B_G07F | 2 | 6 | 2 | 2 | - |
| B_G08B | 2 | 6 | 2 | 2 | - |
| B_G08C | 2 | 6 | 2 | 2 | - |
| B_G08G | 2 | 6 | 2 | 2 | - |
| B_G09B | 2 | - | 2 | - | - |
| B_G09G | 2 | 6 | 2 | 1 | - |
| B_G10L | 2 | - | 2 | 2 | - |
| B_G11B | 2 | 6 | 2 | 2 | 4 |

Note: a dash (-) indicates that a node is not in the network.

**Table A1 (continued)**

| | | Length | | | |
|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 |
| B_G11C | 2 | 6 | 2 | 2 | - |
| B_H01F | 2 | 2 | 3 | 6 | - |
| B_H01H | 2 | 2 | 2 | 2 | - |
| B_H01P | 2 | 6 | 2 | - | - |
| B_H01Q | 2 | 6 | 2 | 2 | - |
| B_H02B | 2 | 2 | - | - | - |
| B_H02H | 2 | 2 | 2 | 2 | - |
| B_H02J | 2 | 2 | 2 | 2 | - |
| B_H02M | 2 | 2 | 2 | 1 | - |
| B_H03B | 2 | - | - | - | - |
| B_H03F | 2 | 2 | 2 | - | - |
| B_H03G | 2 | - | 2 | 2 | - |
| B_H03H | 2 | 2 | 2 | - | - |
| B_H03J | 2 | - | 2 | 2 | - |
| B_H03K | 2 | 2 | 2 | 2 | - |
| B_H03L | 2 | - | 2 | - | - |
| B_H03M | 2 | 6 | 2 | 2 | - |
| B_H04B | 2 | 6 | 2 | 2 | 4 |
| B_H04H | 2 | - | 2 | 2 | - |
| B_H04J | 2 | 6 | 2 | 2 | - |
| B_H04L | 2 | 6 | 2 | 2 | 4 |
| B_H04M | 2 | 6 | 2 | 2 | 4 |
| B_H04N | 2 | 6 | 2 | 2 | 4 |
| B_H04Q | 2 | 6 | 2 | 2 | - |
| B_H04R | 2 | 6 | 2 | 2 | - |
| B_H04W | 2 | 6 | 2 | 2 | 4 |
| B_H05B | 2 | 2 | 2 | 1 | 1 |
| B_H05G | 2 | - | - | - | - |
| Y_A47J | 2 | - | - | - | - |
| Y_B23Q | 2 | 6 | 2 | - | - |
| Y_B24B | 2 | - | - | - | - |
| Y_B25J | 2 | 6 | 2 | - | - |
| Y_B28D | 2 | - | - | - | - |
| Y_B60R | 2 | 2 | 2 | - | - |
| Y_B61L | 2 | - | - | - | - |
| Y_F24C | 2 | - | - | - | - |
| Y_G01B | 2 | - | - | - | - |
| Y_G01C | 2 | - | - | - | - |
| Y_G01D | 2 | 6 | 2 | - | - |
| Y_G01K | 2 | - | - | - | - |
| Y_G01R | 2 | 2 | 2 | 2 | - |
| Y_G01S | 2 | 6 | 2 | 2 | - |
| Y_G01V | 2 | - | - | - | - |
| Y_G05B | 2 | 6 | 2 | 2 | - |
| Y_G05F | 2 | 2 | 2 | - | - |
| Y_G06F | 2 | 6 | 2 | 2 | - |
| Y_G06K | 2 | - | 2 | 2 | - |
| Y_G06Q | 2 | 6 | 2 | 2 | - |
| Y_G06T | 2 | - | 2 | - | - |
| Y_G07C | 2 | - | - | - | - |
| Y_G07F | 2 | - | - | - | - |
| Y_G08B | 2 | 6 | - | - | - |
| Y_G08C | 2 | - | - | - | - |
| Y_G08G | 2 | - | - | - | - |
| Y_G09G | 2 | 6 | 2 | - | - |
| Y_G11B | 2 | - | - | - | - |
| Y_G11C | 2 | - | - | - | - |
| Y_H01F | 2 | 2 | 3 | - | - |
| Y_H01H | 2 | 2 | 2 | - | - |
| Y_H01Q | 2 | - | - | - | - |
| Y_H02H | 2 | 2 | 2 | - | - |
| Y_H02J | 2 | 2 | 2 | 2 | - |
| Y_H02M | 2 | 2 | 2 | 1 | - |
| Y_H03K | 2 | 2 | - | - | - |
| Y_H04B | 2 | 6 | 2 | 2 | 4 |
| Y_H04J | 2 | - | - | - | - |
| Y_H04L | 2 | 6 | 2 | 2 | 4 |
| Y_H04M | 2 | 6 | 2 | 2 | - |
| Y_H04N | 2 | 6 | 2 | 2 | - |
| Y_H04Q | 2 | 6 | 2 | - | - |
| Y_H04W | 2 | 6 | 2 | 2 | 4 |
| Y_H05B | 2 | 2 | 2 | 1 | - |
| B_A01B | 3 | 1 | 7 | 3 | - |
| B_A01C | 3 | 1 | 7 | 3 | - |
| B_A01D | 3 | 1 | 7 | 3 | - |
| B_A01F | 3 | - | - | - | - |
| B_A61G | 3 | 2 | 3 | - | - |
| B_B60B | 3 | 2 | 3 | - | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A1 (continued)**

| | | Length | | | | | | Length | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 | | All | 2-4 | 5-8 | 9-14 | 15-28 |
| B_B60G | 3 | 2 | 3 | 3 | - | B_F02P | 3 | 7 | 3 | - | - |
| B_B60K | 3 | 2 | 3 | 3 | - | B_F03D | 3 | 2 | 3 | - | - |
| B_B60L | 3 | 2 | 3 | - | - | B_F03G | 3 | - | - | - | - |
| B_B60M | 3 | - | - | - | - | B_F04B | 3 | 7 | 3 | 3 | - |
| B_B60P | 3 | 4 | 6 | - | - | B_F04C | 3 | 7 | 7 | 3 | - |
| B_B60S | 3 | - | 3 | - | - | B_F04F | 3 | 2 | - | - | - |
| B_B60T | 3 | 2 | 3 | 3 | - | B_F15B | 3 | 2 | 3 | 3 | - |
| B_B60W | 3 | 2 | 3 | 3 | - | B_F16C | 3 | 2 | 3 | 6 | - |
| B_B61B | 3 | 2 | 3 | - | - | B_F16D | 3 | 2 | 3 | 3 | - |
| B_B61C | 3 | - | - | - | - | B_F16F | 3 | 2 | 3 | 3 | - |
| B_B61D | 3 | 2 | 3 | 6 | - | B_F16H | 3 | 2 | 3 | 3 | - |
| B_B61F | 3 | - | 3 | - | - | B_F16K | 3 | 7 | 3 | 5 | - |
| B_B62B | 3 | 2 | - | - | - | B_F16N | 3 | 2 | - | - | - |
| B_B62D | 3 | 2 | 3 | 3 | - | B_F17C | 3 | 7 | 3 | - | - |
| B_B62J | 3 | 2 | 3 | - | - | B_F17D | 3 | - | - | - | - |
| B_B62K | 3 | 2 | 3 | - | - | B_F41H | 3 | - | - | - | - |
| B_B62M | 3 | 2 | 3 | 3 | - | B_G01F | 3 | 7 | 3 | 5 | - |
| B_B63B | 3 | 2 | 3 | - | - | B_G01H | 3 | - | - | - | - |
| B_B63H | 3 | 2 | 3 | - | - | B_G01L | 3 | 7 | 3 | 1 | - |
| B_B66B | 3 | 2 | 3 | - | - | B_G01M | 3 | 7 | 3 | 5 | - |
| B_B66C | 3 | 2 | 3 | - | - | B_G01P | 3 | 7 | 3 | 1 | - |
| B_B66F | 3 | 2 | 3 | - | - | B_G05G | 3 | - | - | - | - |
| B_B67D | 3 | 1 | 3 | - | - | B_H01T | 3 | 7 | 3 | - | - |
| B_E01B | 3 | - | - | - | - | B_H02K | 3 | 2 | 3 | 3 | - |
| B_E01H | 3 | - | - | - | - | B_H02P | 3 | 2 | 3 | 3 | - |
| B_E02B | 3 | 2 | 1 | - | - | Y_A01C | 3 | 1 | - | - | - |
| B_E02D | 3 | 2 | 1 | - | - | Y_A01D | 3 | - | - | - | - |
| B_E02F | 3 | 2 | 3 | - | - | Y_A61G | 3 | - | - | - | - |
| B_E04H | 3 | 2 | 1 | - | - | Y_B60B | 3 | 2 | - | - | - |
| B_F01B | 3 | 7 | - | - | - | Y_B60K | 3 | 2 | 3 | 3 | - |
| B_F01C | 3 | 7 | - | - | - | Y_B60L | 3 | 2 | 3 | 3 | - |
| B_F01L | 3 | 7 | 3 | 5 | - | Y_B60S | 3 | - | - | - | - |
| B_F01M | 3 | 7 | 3 | - | - | Y_B60T | 3 | 2 | 3 | - | - |
| B_F01P | 3 | 7 | 3 | - | - | Y_B60W | 3 | 2 | 3 | 3 | - |
| B_F02B | 3 | 7 | 3 | 5 | - | Y_B61B | 3 | - | - | - | - |
| B_F02D | 3 | 7 | 3 | 5 | - | Y_B61C | 3 | - | - | - | - |
| B_F02F | 3 | 7 | 3 | 5 | - | Y_B61D | 3 | 2 | 3 | 6 | - |
| B_F02M | 3 | 7 | 3 | 5 | - | Y_B62D | 3 | 2 | 3 | - | - |
| B_F02N | 3 | 7 | 3 | - | - | Y_B62J | 3 | - | - | - | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A1 (continued)**

| | | | Length | | | | | | Length | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 | | All | 2-4 | 5-8 | 9-14 | 15-28 |
| Y_B62K | 3 | - | - | - | - | B_A47B | 4 | 4 | - | - | - |
| Y_B62M | 3 | 2 | 3 | - | - | B_A47C | 4 | 4 | 6 | - | - |
| Y_B63B | 3 | 2 | 3 | - | - | B_A47F | 4 | - | - | - | - |
| Y_B63H | 3 | 2 | 3 | - | - | B_A47L | 4 | 8 | 7 | 3 | - |
| Y_B66B | 3 | - | - | - | - | B_A63H | 4 | - | - | - | - |
| Y_B66C | 3 | 2 | - | - | - | B_B08B | 4 | 1 | 1 | - | - |
| Y_B66F | 3 | - | - | - | - | B_B21B | 4 | 1 | 1 | 6 | - |
| Y_E02B | 3 | 2 | 3 | - | - | B_B21C | 4 | 1 | - | - | - |
| Y_E02D | 3 | 2 | - | - | - | B_B21D | 4 | 4 | 3 | 6 | - |
| Y_E02F | 3 | 2 | - | - | - | B_B22C | 4 | 1 | 1 | 6 | - |
| Y_E04H | 3 | 2 | 3 | - | - | B_B22D | 4 | 1 | 1 | 6 | - |
| Y_F01C | 3 | 7 | - | - | - | B_B22F | 4 | 1 | 1 | 6 | - |
| Y_F01L | 3 | 7 | 3 | 5 | - | B_B23H | 4 | 6 | 6 | - | - |
| Y_F01M | 3 | 7 | - | - | - | B_B23K | 4 | 1 | 1 | 6 | - |
| Y_F01P | 3 | 7 | 3 | - | - | B_B23P | 4 | 4 | 6 | 6 | - |
| Y_F02B | 3 | 7 | 3 | 5 | 3 | B_B24C | 4 | 1 | 1 | - | - |
| Y_F02D | 3 | 7 | 3 | 5 | 3 | B_B25B | 4 | 4 | 6 | - | - |
| Y_F02F | 3 | 7 | 3 | 5 | - | B_B25C | 4 | - | - | - | - |
| Y_F02M | 3 | 7 | 3 | 5 | - | B_B60H | 4 | 4 | 7 | 3 | - |
| Y_F02N | 3 | 7 | 3 | 3 | - | B_B60N | 4 | 4 | 6 | - | - |
| Y_F02P | 3 | 7 | 3 | 5 | - | B_B64C | 4 | 4 | 6 | 3 | - |
| Y_F03B | 3 | 2 | 3 | - | - | B_B64D | 4 | 4 | 6 | 3 | - |
| Y_F03D | 3 | 2 | 3 | 3 | - | B_B64F | 4 | 4 | 6 | - | - |
| Y_F04B | 3 | 7 | - | - | - | B_C21D | 4 | 1 | 1 | 6 | 1 |
| Y_F04C | 3 | - | - | - | - | B_C22C | 4 | 1 | 1 | 6 | 1 |
| Y_F15B | 3 | 2 | - | - | - | B_C22F | 4 | 1 | 1 | 6 | - |
| Y_F16C | 3 | 2 | 3 | - | - | B_D06F | 4 | 8 | 7 | 3 | - |
| Y_F16D | 3 | 2 | 3 | - | - | B_F01D | 4 | 4 | 6 | 6 | 1 |
| Y_F16F | 3 | 2 | 3 | - | - | B_F01K | 4 | 4 | 6 | - | - |
| Y_F16H | 3 | 2 | 3 | 3 | - | B_F02C | 4 | 4 | 6 | 6 | - |
| Y_F16K | 3 | 7 | 3 | - | - | B_F02G | 4 | 4 | - | - | - |
| Y_F17C | 3 | 7 | 3 | - | - | B_F02K | 4 | 4 | 6 | - | - |
| Y_F23K | 3 | - | - | - | - | B_F04D | 4 | 4 | 6 | 6 | - |
| Y_G01F | 3 | 7 | - | - | - | B_F16J | 4 | 4 | 6 | 6 | - |
| Y_G01L | 3 | - | - | - | - | B_F22B | 4 | 4 | 6 | 6 | - |
| Y_G01M | 3 | 7 | 3 | - | - | B_F23C | 4 | 4 | 6 | 6 | - |
| Y_G01P | 3 | - | - | - | - | B_F23D | 4 | 4 | 6 | 6 | - |
| Y_H02K | 3 | 2 | 3 | 3 | - | B_F23K | 4 | - | - | - | - |
| Y_H02P | 3 | 2 | 3 | 3 | - | B_F23L | 4 | 4 | 6 | - | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A1 (continued)**

| | Length | | | | | | Length | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 | | All | 2-4 | 5-8 | 9-14 | 15-28 |
| B_F23M | 4 | - | 6 | - | - | Y_F02G | 4 | 4 | 6 | 6 | - |
| B_F23N | 4 | 4 | 6 | 6 | - | Y_F02K | 4 | 4 | 6 | 6 | - |
| B_F23R | 4 | 4 | 6 | 6 | - | Y_F03G | 4 | 4 | - | - | - |
| B_F24B | 4 | 4 | - | - | - | Y_F04D | 4 | 4 | 6 | 6 | - |
| B_F24D | 4 | 4 | 7 | 3 | - | Y_F15D | 4 | - | - | - | - |
| B_F24F | 4 | 4 | 7 | 3 | - | Y_F16J | 4 | 4 | 6 | - | - |
| B_F24H | 4 | 4 | 7 | 3 | - | Y_F22B | 4 | 4 | 6 | - | - |
| B_F25B | 4 | 4 | 7 | 3 | - | Y_F23C | 4 | 4 | 6 | 6 | - |
| B_F25D | 4 | 4 | 7 | 3 | - | Y_F23D | 4 | 4 | 6 | - | - |
| B_F28D | 4 | 4 | 7 | 3 | - | Y_F23L | 4 | 4 | 6 | - | - |
| B_F28F | 4 | 4 | 7 | 3 | - | Y_F23M | 4 | - | - | - | - |
| B_F42B | 4 | - | - | - | - | Y_F23N | 4 | 4 | 6 | - | - |
| B_G05D | 4 | 4 | 7 | 3 | - | Y_F23R | 4 | 4 | 6 | 6 | - |
| B_G10K | 4 | 7 | 1 | - | - | Y_F24D | 4 | 4 | 7 | 3 | - |
| B_G21C | 4 | 1 | 1 | - | - | Y_F24F | 4 | 4 | 7 | - | - |
| B_G21F | 4 | 1 | 1 | - | - | Y_F24H | 4 | 4 | 7 | - | - |
| Y_A47L | 4 | 8 | 7 | - | - | Y_F25B | 4 | 4 | 7 | 3 | - |
| Y_B08B | 4 | 1 | - | - | - | Y_F25D | 4 | 4 | 7 | - | - |
| Y_B21B | 4 | - | - | - | - | Y_F28D | 4 | 4 | 7 | 3 | - |
| Y_B21D | 4 | 4 | - | - | - | Y_F28F | 4 | 4 | 7 | 3 | - |
| Y_B22C | 4 | - | - | - | - | Y_G05D | 4 | 4 | 7 | 3 | - |
| Y_B22D | 4 | 1 | - | - | - | Y_G21C | 4 | 1 | 1 | - | - |
| Y_B22F | 4 | 1 | 1 | 6 | - | Y_G21D | 4 | 1 | - | - | - |
| Y_B23K | 4 | 1 | 1 | 6 | - | Y_G21F | 4 | 1 | - | - | - |
| Y_B23P | 4 | 4 | 6 | - | - | B_A23N | 5 | - | - | - | - |
| Y_B24C | 4 | - | - | - | - | B_A47K | 5 | 3 | 1 | - | - |
| Y_B33Y | 4 | - | - | - | - | B_A62B | 5 | - | - | - | - |
| Y_B60H | 4 | 4 | 7 | - | - | B_A62D | 5 | 3 | 4 | - | - |
| Y_B60N | 4 | - | - | - | - | B_B01D | 5 | 3 | 4 | 5 | 3 |
| Y_B64C | 4 | 4 | 6 | 3 | - | B_B01F | 5 | 3 | 4 | 5 | - |
| Y_B64D | 4 | 4 | 6 | 3 | - | B_B01J | 5 | 3 | 4 | 5 | 3 |
| Y_B64F | 4 | - | - | - | - | B_B02C | 5 | 1 | 1 | - | - |
| Y_C21D | 4 | 1 | 1 | - | - | B_B03B | 5 | 1 | - | - | - |
| Y_C22C | 4 | 1 | 1 | 6 | - | B_B03C | 5 | 3 | 4 | - | - |
| Y_C22F | 4 | - | - | - | - | B_B03D | 5 | - | - | - | - |
| Y_D06F | 4 | 8 | 7 | 3 | - | B_B04B | 5 | - | - | - | - |
| Y_F01D | 4 | 4 | 6 | 6 | - | B_B04C | 5 | 3 | - | - | - |
| Y_F01K | 4 | 4 | 6 | 6 | - | B_B07B | 5 | 1 | 1 | - | - |
| Y_F02C | 4 | 4 | 6 | 6 | - | B_B09B | 5 | 3 | 4 | - | - |

Note: a dash (-) indicates that a node is not in the network.

## Table A1 (continued)

| | | Length | | | | | | Length | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 | | All | 2-4 | 5-8 | 9-14 | 15-28 |
| B_B09C | 5 | 3 | 4 | - | - | Y_A47K | 5 | 3 | - | - | - |
| B_B30B | 5 | 1 | 8 | - | - | Y_A62D | 5 | 3 | 4 | - | - |
| B_B65F | 5 | 1 | 8 | - | - | Y_B01D | 5 | 3 | 4 | 5 | 3 |
| B_C01B | 5 | 3 | 4 | 5 | 1 | Y_B01F | 5 | 3 | 4 | - | - |
| B_C01C | 5 | - | - | - | - | Y_B01J | 5 | 3 | 4 | 5 | 3 |
| B_C01D | 5 | - | - | - | - | Y_B02C | 5 | 1 | 1 | - | - |
| B_C01F | 5 | 3 | 4 | - | - | Y_B03B | 5 | 1 | 1 | - | - |
| B_C02F | 5 | 3 | 4 | 5 | - | Y_B03C | 5 | 3 | - | - | - |
| B_C05F | 5 | - | - | - | - | Y_B03D | 5 | - | - | - | - |
| B_C05G | 5 | 3 | - | - | - | Y_B07B | 5 | 1 | 1 | - | - |
| B_C07B | 5 | 3 | 4 | 5 | - | Y_B09B | 5 | 3 | 4 | - | - |
| B_C07C | 5 | 3 | 4 | 5 | 2 | Y_B09C | 5 | - | - | - | - |
| B_C10B | 5 | 3 | - | - | - | Y_B30B | 5 | 1 | - | - | - |
| B_C10G | 5 | 3 | 4 | 5 | - | Y_B65F | 5 | 1 | 8 | - | - |
| B_C10J | 5 | 3 | 4 | - | - | Y_C01B | 5 | 3 | 4 | 5 | - |
| B_C10K | 5 | - | - | - | - | Y_C01C | 5 | 3 | 4 | - | - |
| B_C10L | 5 | 3 | 4 | 5 | - | Y_C01F | 5 | 3 | - | - | - |
| B_C10M | 5 | 3 | 4 | 5 | 1 | Y_C01G | 5 | 3 | 4 | - | - |
| B_C10N | 5 | - | 4 | 5 | - | Y_C02F | 5 | 3 | 4 | 5 | - |
| B_C21B | 5 | 1 | 4 | - | - | Y_C05F | 5 | 3 | 4 | - | - |
| B_C21C | 5 | 1 | 4 | - | - | Y_C05G | 5 | 3 | - | - | - |
| B_C22B | 5 | 3 | 4 | - | - | Y_C07B | 5 | 3 | 4 | 5 | - |
| B_C25C | 5 | 3 | 4 | - | - | Y_C07C | 5 | 3 | 4 | 5 | - |
| B_D06B | 5 | - | - | - | - | Y_C10B | 5 | 3 | 4 | - | - |
| B_D21B | 5 | - | - | - | - | Y_C10G | 5 | 3 | 4 | 5 | - |
| B_D21C | 5 | 1 | 1 | - | - | Y_C10J | 5 | 3 | 4 | - | - |
| B_D21D | 5 | - | - | - | - | Y_C10K | 5 | 3 | 4 | - | - |
| B_D21F | 5 | 1 | 1 | - | - | Y_C10L | 5 | 3 | 4 | 5 | - |
| B_E03C | 5 | 3 | 7 | - | - | Y_C10M | 5 | - | - | - | - |
| B_E03D | 5 | 3 | - | - | - | Y_C11C | 5 | 3 | - | - | - |
| B_E03F | 5 | 3 | 1 | - | - | Y_C12M | 5 | 3 | 4 | 5 | - |
| B_F01N | 5 | 7 | 4 | 5 | 3 | Y_C21B | 5 | 1 | 4 | - | - |
| B_F23G | 5 | 3 | 4 | - | - | Y_C21C | 5 | 1 | 4 | - | - |
| B_F23J | 5 | 3 | 4 | - | - | Y_C22B | 5 | 3 | 4 | 1 | - |
| B_F25J | 5 | 3 | 4 | 5 | - | Y_C25C | 5 | 3 | - | - | - |
| B_F26B | 5 | 3 | 4 | - | - | Y_D21B | 5 | 1 | 1 | - | - |
| B_F27B | 5 | 1 | 4 | - | - | Y_D21C | 5 | 1 | 1 | - | - |
| B_F27D | 5 | 1 | 4 | - | - | Y_D21F | 5 | - | - | - | - |
| B_G01T | 5 | - | 4 | 5 | 3 | Y_E03B | 5 | 3 | - | - | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A1 (continued)**

| | | Length | | | |
|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 |
| Y_E03F | 5 | 3 | - | - | - |
| Y_F01N | 5 | 7 | 4 | 5 | 3 |
| Y_F23G | 5 | 3 | 4 | - | - |
| Y_F23J | 5 | 3 | 4 | - | - |
| Y_F25J | 5 | 3 | 4 | - | - |
| Y_F26B | 5 | 3 | 4 | - | - |
| Y_F27B | 5 | 1 | 4 | - | - |
| Y_F27D | 5 | 1 | 4 | - | - |
| B_A01H | 6 | 5 | 5 | 4 | - |
| B_A01J | 6 | - | - | - | - |
| B_A01K | 6 | 3 | 5 | 4 | - |
| B_A01N | 6 | 5 | 5 | 4 | 2 |
| B_A01P | 6 | 5 | 5 | 4 | - |
| B_A21D | 6 | - | 5 | - | - |
| B_A22C | 6 | - | - | - | - |
| B_A23B | 6 | 5 | 5 | - | - |
| B_A23C | 6 | 5 | 5 | 4 | - |
| B_A23D | 6 | - | 5 | 4 | - |
| B_A23F | 6 | - | - | - | - |
| B_A23G | 6 | 5 | 5 | 4 | - |
| B_A23J | 6 | - | 5 | - | - |
| B_A23K | 6 | 5 | 5 | 4 | - |
| B_A23L | 6 | 5 | 5 | 4 | 2 |
| B_A45D | 6 | - | - | - | - |
| B_A61B | 6 | 5 | 5 | 4 | 2 |
| B_A61C | 6 | 5 | 5 | 4 | - |
| B_A61F | 6 | 5 | 5 | 4 | 1 |
| B_A61H | 6 | - | - | - | - |
| B_A61J | 6 | 1 | 5 | - | - |
| B_A61K | 6 | 5 | 5 | 4 | 2 |
| B_A61L | 6 | 5 | 5 | 4 | 1 |
| B_A61M | 6 | 5 | 5 | 4 | - |
| B_A61N | 6 | 5 | 5 | 4 | - |
| B_A61P | 6 | 5 | 5 | 4 | 2 |
| B_A61Q | 6 | 5 | 5 | 4 | 2 |
| B_A63B | 6 | 5 | - | - | - |
| B_B01L | 6 | 5 | 5 | 4 | - |
| B_B07C | 6 | - | 2 | - | - |
| B_B27K | 6 | - | - | - | - |
| B_C07D | 6 | 5 | 5 | 4 | 2 |
| B_C07F | 6 | 5 | 5 | 4 | 1 |
| B_C07H | 6 | 5 | 5 | 4 | 2 |
| B_C07J | 6 | 5 | 5 | - | - |
| B_C07K | 6 | 5 | 5 | 4 | 2 |
| B_C08B | 6 | 5 | 5 | 4 | - |
| B_C11B | 6 | 3 | 5 | - | - |
| B_C11C | 6 | - | 5 | - | - |
| B_C11D | 6 | 5 | 5 | 4 | 2 |
| B_C12C | 6 | - | - | - | - |
| B_C12M | 6 | 3 | 5 | 4 | - |
| B_C12N | 6 | 5 | 5 | 4 | 2 |
| B_C12P | 6 | 5 | 5 | 4 | 2 |
| B_C12Q | 6 | 5 | 5 | 4 | 2 |
| B_C12R | 6 | 5 | 5 | 4 | 2 |
| B_C13K | 6 | - | - | - | - |
| B_C40B | 6 | - | - | - | - |
| B_G01N | 6 | 5 | 5 | 4 | 2 |
| Y_A01H | 6 | 5 | - | - | - |
| Y_A01K | 6 | 3 | - | - | - |
| Y_A01N | 6 | 5 | 5 | 4 | - |
| Y_A01P | 6 | - | - | - | - |
| Y_A23B | 6 | 5 | 5 | - | - |
| Y_A23K | 6 | 5 | 5 | - | - |
| Y_A23L | 6 | 5 | 5 | - | - |
| Y_A61B | 6 | 5 | 5 | - | - |
| Y_A61F | 6 | - | - | - | - |
| Y_A61K | 6 | 5 | 5 | 4 | 2 |
| Y_A61L | 6 | 5 | 5 | - | - |
| Y_A61M | 6 | 5 | - | - | - |
| Y_A61P | 6 | 5 | 5 | 4 | 2 |
| Y_B07C | 6 | - | - | - | - |
| Y_C07D | 6 | 5 | 5 | 4 | 2 |
| Y_C07F | 6 | 5 | 5 | 4 | - |
| Y_C07H | 6 | 5 | 5 | 4 | - |
| Y_C07J | 6 | 5 | - | - | - |
| Y_C07K | 6 | 5 | 5 | 4 | 2 |
| Y_C08B | 6 | - | - | - | - |
| Y_C11B | 6 | 3 | - | - | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A1 (continued)**

| | Length | | | | |
|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 |
| Y_C11D | 6 | - | - | - | - |
| Y_C12N | 6 | 5 | 5 | 4 | 2 |
| Y_C12P | 6 | 5 | 5 | 4 | - |
| Y_C12Q | 6 | 5 | 5 | 4 | - |
| Y_C12R | 6 | 5 | - | - | - |
| Y_C23F | 6 | - | - | - | - |
| Y_G01N | 6 | 5 | 5 | 4 | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A2. Cluster membership of brown/green & countries, by trajectory length**

| | Length | | | | | | Length | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 | | All | 2-4 | 5-8 | 9-14 | 15-28 |
| B_?? | 1 | 1 | 1 | 1 | 1 | Y_PT | 1 | 1 | - | - | - |
| B_AT | 1 | 1 | 1 | 1 | 1 | Y_SG | 1 | 1 | 2 | - | - |
| B_BR | 1 | 1 | 2 | - | - | Y_SI | 1 | - | - | - | - |
| B_CH | 1 | 1 | 1 | 1 | 1 | Y_TR | 1 | 1 | - | - | - |
| B_CZ | 1 | 1 | 1 | - | - | Y_ZA | 1 | 2 | 1 | - | - |
| B_DE | 1 | 1 | 1 | 1 | 1 | B_AR | 2 | - | - | - | - |
| B_DK | 1 | 1 | 1 | 1 | 1 | B_BE | 2 | 1 | 1 | 1 | 1 |
| B_ES | 1 | 1 | 1 | 1 | 1 | B_CA | 2 | 2 | 2 | 1 | 1 |
| B_FR | 1 | 1 | 1 | 1 | 1 | B_CN | 2 | 3 | 2 | 1 | 2 |
| B_IT | 1 | 1 | 1 | 1 | 1 | B_GB | 2 | 2 | 2 | 1 | 1 |
| B_LI | 1 | - | - | - | - | B_GR | 2 | 2 | 1 | - | - |
| B_MY | 1 | - | - | - | - | B_HK | 2 | - | - | - | - |
| B_NL | 1 | 1 | 1 | 1 | 1 | B_HR | 2 | - | - | - | - |
| B_PL | 1 | 1 | 1 | 1 | - | B_HU | 2 | 3 | 1 | 1 | - |
| B_PT | 1 | 1 | 2 | - | - | B_IL | 2 | 2 | 2 | 1 | - |
| B_SI | 1 | - | 2 | - | - | B_IN | 2 | 2 | 2 | 1 | 1 |
| B_SK | 1 | - | - | - | - | B_JP | 2 | 3 | 2 | 1 | 1 |
| B_TR | 1 | 1 | 1 | 1 | - | B_KR | 2 | 3 | 2 | 1 | 2 |
| Y_?? | 1 | 1 | 1 | 2 | 2 | B_LU | 2 | 1 | 1 | 1 | - |
| Y_AT | 1 | 1 | 1 | 2 | - | B_MC | 2 | - | - | - | - |
| Y_BE | 1 | 1 | 1 | 2 | - | B_MX | 2 | 2 | - | - | - |
| Y_BR | 1 | 1 | 1 | - | - | B_RO | 2 | - | - | - | - |
| Y_CH | 1 | 1 | 1 | 2 | 2 | B_RU | 2 | 2 | 2 | 1 | - |
| Y_CZ | 1 | 1 | 1 | - | - | B_SA | 2 | 2 | 2 | - | - |
| Y_DE | 1 | 1 | 1 | 2 | 2 | B_SG | 2 | 2 | 2 | 1 | - |
| Y_DK | 1 | 1 | 1 | 2 | - | B_SU | 2 | 1 | - | - | - |
| Y_ES | 1 | 1 | 1 | 2 | - | B_TW | 2 | 3 | 2 | 1 | - |
| Y_FR | 1 | 1 | 1 | 2 | 2 | B_UA | 2 | - | - | - | - |
| Y_GB | 1 | 2 | 2 | 2 | 2 | B_US | 2 | 2 | 2 | 1 | 1 |
| Y_GR | 1 | 1 | - | - | - | Y_CN | 2 | 3 | 2 | 2 | 2 |
| Y_HU | 1 | 1 | - | - | - | Y_HK | 2 | - | - | - | - |
| Y_IE | 1 | 2 | 1 | - | - | Y_IL | 2 | 2 | 1 | 1 | - |
| Y_IT | 1 | 1 | 1 | 2 | 1 | Y_IN | 2 | 2 | 2 | 1 | - |
| Y_LU | 1 | 1 | 1 | - | - | Y_JP | 2 | 3 | 2 | 2 | 2 |
| Y_MX | 1 | 2 | - | - | - | Y_KR | 2 | 3 | 2 | 1 | 2 |
| Y_NL | 1 | 1 | 1 | 2 | - | Y_MY | 2 | - | - | - | - |
| Y_NZ | 1 | 2 | 1 | - | - | Y_RU | 2 | 2 | 2 | 2 | 2 |
| Y_OTHER_Cntry | 1 | 1 | 1 | 1 | - | Y_SA | 2 | 2 | - | - | - |
| Y_PL | 1 | 1 | 1 | - | - | Y_TW | 2 | 3 | 2 | 2 | - |

Note: a dash (-) indicates that a node is not in the network.

**Table A2 (Continued)**

| | Length | | | | |
|---|---|---|---|---|---|
| | All | 2-4 | 5-8 | 9-14 | 15-28 |
| Y_US | 2 | 2 | 2 | 2 | 2 |
| B_AU | 3 | 2 | 2 | 1 | - |
| B_FI | 3 | 4 | 3 | 1 | 1 |
| B_IE | 3 | 2 | 2 | 1 | - |
| B_NO | 3 | 4 | 3 | 1 | - |
| B_NZ | 3 | 2 | 2 | 1 | - |
| B_OTHER_Cntry | 3 | 1 | 2 | 1 | - |
| B_SE | 3 | 4 | 3 | 1 | 2 |
| B_ZA | 3 | 2 | 2 | 1 | - |
| Y_AU | 3 | 2 | 1 | 1 | - |
| Y_CA | 3 | 2 | 2 | 1 | - |
| Y_CL | 3 | - | - | - | - |
| Y_FI | 3 | 4 | 3 | 1 | - |
| Y_NO | 3 | 4 | 3 | 2 | - |
| Y_SE | 3 | 4 | 3 | 2 | - |

Note: a dash (-) indicates that a node is not in the network.