Karlsruher Institut für Technologie

Titel der Seminar-/Abschlussarbeit

BACHELORARBEIT

von

Paul Theuer

Matrikelnummer: 2053912

Studiengang: Wirtschaftsingenieurwesen

Institut für Volkswirtschaftslehre (ECON)
Lehrstuhl für Wirtschaftspolitik

Prüfer:	Prof. Dr. Ingrid Ott
Zweitprüfer:	Prof. Dr. Kay Mitusch
Betreuender Mitarbeiter:	Name Mitarbeiter, Name Mitarbeiter 2
Bearbeitungszeit:	01.05.2019 – 01.11.2019

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Quellen und Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

- Heranführung an Thema
- Technologieräume auf unterschiedliche Weise darstellen
- Hängt von Zielrichtung ab
- Ziel: Versuch makroökonomische Fragestellungen zu veranschaulichen und zu erklären
- Klassischer Ansatz: Fokus auf patentbasierte Ansätze
- Beispiel mit Experiment (Engelsman und Raan 1994)

Eine Technologie ist ein komplexes, heterogenes Kongolmerat aus verschiedenen Feldern, sie lässt sich durch verschiedenen Aspekte charakterisieren und wird neben der Wissenschaft bzw. dem Stand der Technik durch verschiedene Faktoren beeinflusst. Dynamischen Einflussfaktoren führen dazu, dass sich die Entwicklung von Technologien, ähnlich wie in Feldern der Mathematik oder Physik, chaotisch und scheinbar unvorhersehbar verhält.(Engelsman und Raan 1994). Dabei steht und fällt aber der Erfolg von Unternehmen maßgeblich mit dem setzten auf die richtigen Technologien. Zahlreiche Studien der letzten Jahrzehnte haben gezeigt, dass die Produktivität eines Unternehmens oder einer Industrie, stark mit deren Ausgaben im Forschungs- und Entwicklungsbereich zusammenhängt.(Jaffe et al. 1993). Um zukunftsfähige Technologien in bestimmten Bereichen zu identifizieren, bietet es sich also an zunächst das Innovationsverhalten von Unternehmen in diesen Bereichen zu untersuchen.

2 Patente

- in diesem Abschnitt Funktion von Patenten erklären, da diese Basis für weitere Evaluationen darstellt
- Ein Patent gewährt dem Anmelder ein Besitz- und Verfügungsrecht für die kommerzielle Nutzung einer Erfindung
- Erfindung muss neu und nichttrivial sein (Def. aus PatG nutzen); Stand der Technik erläutern
- PatG: was keine Erfindung darstellt
- Wird ein Patent gewährt, wird ein für die Öffentlichkeit einsehbares Dokument erstellt.
- Dokument beinhaltet eine Reihe von Informationen: Erfinder, Arbeitgeber, etc., hier wichtig: Referenzen bzw. Zitationen
- Besitzen eine rechtliche Funktion: Begrenzen den Umfang der Verfügungsrechte ausgehend von dem Patent
- Sicherstellen, dass das gewährte Patent einen nichttrivialen Wissensbeitrag im Vergleich zu dessen Vorgängern erreicht.
- Im Kontext hier bedeutet: Wenn Patent Y Patent X zitiert, wird Patent Y auf dem Wissen von Patent X aufbauen.
- Kritik nach (Karki 1997)
- propensity der Patentklassen

2.0.1 Grundlegendes

In diesem Abschnitt soll die grundlegende Funktion von Patenten erklärt werden, da diese unsere Datenbasis sein werden und damit die Grundlage für alle kommenden Evaluationen darstellen. Ein Patent gewährt dem Anmelder ein Besitz- und Verfügungsrecht für die kommerzielle Nutzung einer Erfindung.

Klassifizierung

- ipc und cpc

Datenbasis

- panda dataframes
- plots

3 Jaffe

Im ersten Teil dieses Kapitels werden die Papiere (Jaffe 1986) und (Jaffe 1989) zusammengefasst. Im zweiten Teil soll ein Technologieraum, nach der Definition des Autors, für die Firma Honda und ihre Konkurrenz nachgebildet und analysiert werden.

3.1 Literatur

3.1.1 Grundlagen

Mit seinem Papier (Jaffe 1986) bietet Adam B. Jaffe einen wichtigen Grundbaustein für die Quantifizierung von Forschungs- und Entwicklungsarbeiten von Unternehmen. Offizielle innovatorische Kennzahlen von Unternehmen, wie z.B. Forschungskosten sind eindimensional und wenig aussagekräftig. Um die technologischen Position von Firmen zu charakterisieren bedient sich (Jaffe 1986) eines patentbasierten Ansatz. Jedes Unternehmen forscht in verschiedenen Sektoren. Alle relevanten Wirtschaftszweige werden mithilfe des Patentklassifikationssystems von 328 Patentklassen[1] in insgesamt 21 Clustern zusammengefasst. Es werden 260.000 Patente von 1700 Firmen über einen Zeitraum von zehn Jahren (1969 - 1979) betrachtet. Jaffe (1986) kann zeigen, dass die Produktivität von Firmen in F&E positiv von der innovatorischen Arbeit ihrer „technologischen Nachbarn“ beeinflusst wird und Firmen ihr Innovationsverhalten ändern wenn sie eine Gelegenheit dazu bekommen. (Jaffe 1989) baut inhaltlich auf (Jaffe 1986), hier werden zunächst zehn Firmen aus unterschiedlichen Sektoren in einem Technologieraum abgebildet. Dabei setzt der Autor seinen Fokus sowohl auf die technologischen Positionen der Unternehmen und die mathematischen Hintergründe des Technologieraums, als auch auf das Bilden der (industriellen) Cluster ausgehend von den Patentklassen. In Jaffe (1986), werden primär die statistischen Grundlagen für die Untersuchung von Wissensübertragungseffekte („Spillovers“), bereitgestellt.

[1] dabei handelt es sich um das Klassifikationssystem des National Bureau of Economic Research (NBER)

3.1.2 Zielsetzung

Private Firmen investieren Ressourcen in Forschung und Entwicklung um wirtschaftlich nützliches Wissen zu „generieren„. Der Erfolg dieser Investitionen variiert stark, teilweise durch nicht messbare Zufallsfaktoren und teilweise durch systematische Effekte, ausgelöst durch das wirtschaftliche und technologische Umfeld einer Firma. Es sind diese Effekte („Spillovers“) die in (Jaffe 1986) untersucht werden. Die primäre makroökonomische Zielsetzung in (Jaffe 1986) und (Jaffe 1989) ist das konkretisieren von Übertragungseffekten im Innovationskontext und deren Auswirkungen auf die Wirtschaftlichkeit und das Verhalten von Unternehmen. Spillovers bezeichnen Wissensübetragungen von einer Firma auf eine oder mehrere andere Firmen. Diese Übertragungseffekte sind im Gegensatz zu unternehmerischen Allianzen unbeabsichtigt. Darüber hinaus beschreibt der Autor einen weiteren Effekt, den der „technological opportunity“. Die „technological opportunity“ wird als exogene, technologisch bedingte, Schwankung in der Produktivität in Forschungs- und Entwicklungsarbeit definiert. (Jaffe 1986). Es soll untersucht werden, ob eine Veränderung des Innovationsverhaltens einer Firma, nicht nur markt- bzw. nachfragebasiert ist, sondern zusätzlich von Änderungen in der technologische Landschaft abhängig ist.

3.1.3 Bestimmung des Technologieraums

Um „Spillovers “ zu untersuchen müssen zunächst die technologischen Nachbarn von den zu betrachteten Firmen bestimmt werden. Im Allgemeinen soll also eine Darstellung gefunden werden, die den (technologische) Abstand von mehreren Unternehmen zueinander abbildet. Das Resultat dieser Darstellung wird als Technologieraum verstanden. In (Jaffe 1986) wird angenommen, dass die Patente von Firmen, die in verschiedenen Patentklassen verteilt sind, die Innovationsinteressen der jeweiligen Firma reflektiert. Hält also beispielsweise die Firma Honda viele Patente in dem Bereich „Hybridfahrzeuge“, so wird Honda auch zu einem großen Teil nach Technologien in dieser Richtung forschen.

Sei f_{ik} der Anteil von Firma i's Patenten in Patentklasse k. Der Vektor $f_i = (f_{i1} \dots f_{iK})$ positioniert die Firma in einem $K-$dimensionalen Technologieraum (Jaffe 1989). Es kann davon ausgegangen werden, dass zwei Firmen technologisch verwandt sind, wenn diese sich in dem Technologieraum nahe stehen. Um Ähnlichkeiten in Innovationsverhalten darzustellen, müssen die resultierenden Patentvektoren der Firmen in ein sinnvolles Verhältnis gesetzt werden. Dafür ist die Auswahl einer passenden Distanzmetrik ausschlaggebend. Eine Möglichkeit bestünde darin, zu schauen, in welche Richtung die Vektoren zeigen. In erster Linie soll das Verhältnis der Firmen zueinander bewertet werden, nicht wie viele Patente die Firmen jeweils insgesamt besitzen. Wir betrachten

also nicht die Länge, sondern die Winkel der Vektoren zueinander. Eine passendes Maß dafür bietet die „Cosine distance“, oder auch Kosinus-Ähnlichkeit.

$$P_{ij} = \frac{\sum_{k=1}^{K} f_{ik} f_{jk}}{\sqrt{\sum_{k=1}^{K} (f_{ik})^2} \sqrt{\sum_{k=1}^{K} (f_{jk})^2}} \tag{3.1}$$

Gewissermaßen misst P_{ij} den Grad der Überschneidung der Vektoren f_i und f_j. Je ähnlicher sich die Patente der Firmen i und j - im Bezug auf ihre Einteilung in die Patentklassen sind - desto größer wird der Zähler sein. Im Nenner wird das Ergebnis normalisiert. Sind i und j identisch, ist $P_{ij} = 1$. Wie auch der Korrelationskoeffizienten ist diese Metrik symmetrisch, es gilt also $P_{ij} = P_{ji}$.

In (Jaffe 1989) werden mit diesem Ansatz zunächst 10 verschiedene Unternehmen über 328 Patentklassen verglichen. Das Ergebnis bestätigt die Intuition. Alle Unternehmen in der Computerindustrie stehen sich nahe. Pharmaunternehmen sind weit entfernt von Unternehmen in den Bereichen Büro- und Schreibwaren, aber relativ nahe zu Firmen in Medizin- und Zahntechnik.

3.1.4 Clustering von Patentklassen

Weiter soll gezeigt werden, wie die Patentklassen in Cluster zusammengefasst werden, deren Gruppierung sich am Markt beziehungsweise der Industrie orientieren. Zweck einer solchen Einteilung ist es die Interaktion zwischen Marktzweigen und Firmen einzufangen. Welche Folgen hat eine Veränderung eines Marktzweiges (Cluster), auf das strategische Verhalten einer Firma und vice-versa?

Das „Clustering“ in Jaffe (1989) ist ein iterativer Prozess, der sich an den „$K-$means“ Algorithmus anlehnt. Dabei entspricht jeder Punkt im Raum des „$K-$means“ der Verteilung der Patente einer Firma über alle Patentklassen. Die geometrische Mitte eines Clusters entspricht der durchschnittlichen Verteilung über alle Patentklassen, aller Firmen in unserem Industriecluster.

Für jede Patentklasse k, wird die Anzahl der Firmen (C_k) bestimmt, die Patente in dieser Kategorie halten. Jetzt wird „ad-hoc “, die Anzahl der N-höchsten C_k bestimmt. Eine Firma wird dem Cluster zugeordnet, in dem sie am meisten Patente besitzt. Daraus ergeben sich die initialen Cluster[2]. Für jedes Cluster wird die durchschnittliche Verteilung der Patente über alle Patentklassen

[2]Bei $K-$means erfolgt die Initialisierung meistens zufällig

berechnet (Analog: die Mittelpunkte der Cluster aus $K-$means). Dann wird die Verteilung der Patente über den Patentklassen für jede Firma einzeln berechnet (Analog: die einzelnen Punkte aus $K-$means). Passt die Verteilung eines Clusters, in der sich eine Firma nicht befindet besser[3] auf die Verteilung der Firma selbst, so wechselt diese Firma in das Cluster. Dieser Prozess endet, wenn keine Firma mehr das Cluster wechselt. Zuletzt werden die Cluster je nach Zuordnung benannt, so ergeben sich beispielsweise die Cluster Chemie und Kohlenstoff, Nahrung, Medizin und Automobile.

Das „Clustering“ wird für zwei Datensätze aus den Jahren 1972 und 1973 durchgeführt. Interessant sind diejenigen Firmen, die ihre Cluster zwischen den zwei Zeitperioden wechseln, oder anders: wir betrachten die Zu- und Abflüsse der Firmen innerhalb der verschiedenen Cluster. So kann beobachtet werden, dass „generische“ Technologien, wie z.B. Beschichtungen, Fluid-Handling und Signalgebung, ihre „Mitglieder“ aus vielen verschiedenen Industrien ziehen. Außerdem werden verschiedene interdisziplinäre Beziehungen zwischen Technologien festgestellt. Textile und Papier teilen Interessen mit den Industrien aus Beschichtungen und die Pharmaindustrie hängt stark mit Industrien für Medizintechnik und Apparaturen allgemein zusammen. Besonders auffallend sind die Ergebnisse wenn die Firmen einzeln betrachtet werden. So landen beispielsweise die drei großen Automobilhersteller alle in dem Cluster Triebwerke, die Automobilzulieferer hingegen, befinden sich in dem Automobilcluster. Xerox, ein Unternehmen für Bürobedarf, zentralisiert sich im Cluster für Elektrochemie.

3.1.5 Übertragungseffekte und Technologische Opportunität

Mit dem definierten Technologieraum als Ausgangspunkt können nun die „Spillovers“ untersucht werden. Demnach werden Firmen, die vielen anderen innovativen Firmen nahe stehen von deren Position profitieren können. Die potentiellen „Spillover“ über dem Technologieraum sind nicht gleich verteilt, sie sind an den Punkten am höchsten, an denen sich die meisten Firmen befinden und somit am meisten potentielles Wissen an andere Firmen übertragen können. Alle potentiellen „Spillovers“ werden als „Spilloverpool“ definiert. Der „Spilloverpool“ S_i wird für jede Firma bestimmt und ergibt sich aus der gewichtet Summe der Innovationsarbeit aller anderen Firmen.

$$S_i = \sum_{j \neq i} P_{ij} R_j \tag{3.2}$$

[3] auf die Distanzmetrik zwischen Verteilungen wird nicht weiter eingegangen

Den Effekt den eigene Innovationen auf eine Firma hat ist in der realen Welt natürlich nicht direkt beobachtbar. In Jaffe (1989), werden dafür vier Indikatoren betrachtet: Patente, Bruttoeinnahmen, Kapitalertrag und Marktwert. Die Korrelation zwischen diesen Indikatoren und der innovatorischen Aktivität einer Firma sollte intuitiv klar sein. Ausgaben im Forschungs- und Entwicklungsbereich sind oft riskant, dennoch sollte die Findung neuer Technologien oder neuem Wissen früher oder später Einnahmen generieren und den Marktwert der Firma steigern. Mit Hilfe von Regressionstechniken wird versucht diese Indikatoren als Funktionen von Forschungsausgaben, Übertragungseffekten und weiteren Variablen zu approximieren. Beispielsweise ergibt sich die Patentfunktion aus einer modifizierte Cobb-Douglas Technologie (Jaffe 1986). k_i wird als das „neu generierte Wissen“ einer Firma definiert.

$$p_i = \beta_1 r_i + \beta_2 r_i s_i + \gamma_i s_i + \sum_{c=1}^{21} (\delta_{1c} - \alpha_c) D_{ic} + \epsilon_{1i} \tag{3.3}$$

k_i wird als das „neu generierte Wissen“ einer Firma definiert. r_i sind die Forschungsausgaben der Firma i und s_i der Spilloverpool aus 3.2. Die D_{ic}'s sind Dummyvariablen aus den oben angesprochenen Clustern, dabei wird angegeben ob eine Firma Patente in dem jeweiligen Cluster besitzt oder nicht. Ist die technologische Opportunität relevant so werden die δ_{1c}'s, die Gewichtungen der Cluster, unterschiedlich ausfallen. So könnte man beispielsweise annehmen, dass das Cluster der Automobilindsturie eine höhere Gewichtung haben sollten, als das der Kühlindustrie. ϵ_{1i} sind zufällige Störterme der einzelnen Firmen. Die Gleichung 3.3 impliziert, dass die Forschungs- und Entwicklungsarbeit andere Firmen, den Wissensoutput einer Firma direkt erhöhen können. Funktionen für Bruttoeinnahmen, Kapitalertrag und Marktwert werden analog definiert. Für weitere statistische Details siehe Jaffe (1986).

Nach Approximation der Koeffizienten kann beispielsweise berechnet werden, dass eine permanente Erhöhung der Forschungsausgaben um 10%, eine durchschnittlich Patentsteigerung um 8.8% zufolge hat. Weiter steigen die Bruttoeinnahmen um 0.3%, der Kapitalertrag um 1.8% und der Marktwert um 3.6%. Zusätzlich kann gezeigt werden, dass Firmen sich in die Cluster bewegen, die überdurchschnittlich hohe Einnahmen und Marktwerte aufweisen (Jaffe 1989).

Der in 3.2 definierte Spilloverpool spielt für die Erklärung aller vier Indikatoren eine wichtige Rolle. Zusätzlich gibt es eine Beziehung zwischen der firmeneigenen Forschungsarbeit und der Menge an erreichbaren Wissen für diese Firma. Die Produktivität der eigenen Innovationsarbeit hängt also von der Innovationsarbeit anderer, im Technologieraum naher, Firmen ab, welche wiederum von der eigenen Innovationsarbeit abhängt. Dieser Effekt führt dazu, dass der Spilloverpool für den Allgemeinfall nicht genau betrachtet werden kann. Die Firmen müssen zunächst

nach ihrer eigenen innovativen Aktivität differenziert werden.

In Jaffe (1986) wird dazu in drei Fällen unterschieden. Zunächst werden Firmen mit niedriger, dann mit durchschnittlicher und zuletzt in überdurchschnittlicher innovativen Aktivität betrachtet. Jetzt kann untersucht werden welchen Einfluss der Spilloverpool auf die Unternehmen hat. Die emittierten Patente und die Bruttoeinnahmen steigen mit den Übertragungseffekte für alle Firmen. Je größer der Spilloverpool, desto stärker ist der eigene technische Fortschritt. Ertrag und Marktwert sinken für wenig aktiven Firmen. Steigt die Aktivität der Firmen auf den Durchschnitt, so werden Ertrag und Marktwert positiv. Der Marktwert überdurchschnittlich aktiver Firmen steigt am stärksten

Nach diesen Ergebnissen wird der technologische Erfolg eines Nachbarn, es einer Firma also erleichtern selber technologische Fortschritte zu machen. Gleichzeitig wird dem wenig aktiven Unternehmen ein Markterfolg erschwert. Intuitiv ist diese Schlussfolgerung einleuchtend, denn technologische Nachbarn sind oft Konkurrenten. Ein schmaler technologischer Fortschritt wird für kurzfristige Mehreinnahmen sorgen, ein langfristiger Erfolg am Markt kann damit aber nicht garantiert werden. Auf der anderen Seite überwiegen die Vorteile der Übertragungseffekte für höchst innovative Firmen. Aufgrund des komplementären Effekts zwischen eigener Innovationsstärke und der, der technologischen Nachbarn, werden sich stark innovative Unternehmen nachhaltig am Markt behaupten können.

Auch die Gewichtungen δ_{1c} der Cluster sind für alle Performance Indikatoren (Beispiel: Patentindikator siehe 3.3) von großer Bedeutung. Hält man firmenspezifischen Attribute und Poolvariablen konstant, finden sich systematische Unterschiede in den Ergebnissen für verschieden Cluster. Zusätzlich sind die Cluster, die in den betrachteten Zeiträumen Firmen dazugewinnen, diese, die durchschnittlich höhere Profite und Marktwerte aufweisen. Das Modell deutet zumindest stichpunktartig auf die Nutzung innovativer Gelegenheiten hin, ob es sich dabei allerdings um einen „market-pull“ oder „technology push“ oder möglicherweise um eine Kombination beider Effekte handelt, lässt der Autor weitestgehend offen.

3.2 Der Technologieraum für die Firma Honda und ihre Konkurrenz nach Jaffe (1986)

3.2.1 Firmenauswahl

Als Datengrundlage nutzen wir die in Kapitel 4 angesprochenen Patente aus der CPC-Klasse $Y02T\_10$. Der Schwerpunkt unserer Analyse wird also stets vor dem Hintergrund sein, Firmen bzw. Patente im Bereich von innovativen, klimaneutrale Technologien in der Automobilbranche zu untersuchen. Um einen Technologieraum nach Jaffe (1986) für die Firma Honda zu erstellen, bestimmen wir zunächst alle Firmen, die wir mit Honda in ein Verhältnis setzen wollen. In Jaffe (1989), werden „willkürlich“ zehn Firmen ausgewählt, die sich in ihrer Position am Markt teilweise stark unterscheiden. Wir wollen versuchen die Position der Firma Honda und ihrer Konkurrenten möglichst realistisch darzustellen, dafür suchen wir uns zunächst die aktivsten Unternehmen in diesem Sektor. Aktivität definieren wir dabei als die Anzahl der Patente, die eine Firma in der Klasse $Y02T\_10$ besitzt. Bei der Anzahl der Firmen halten wir uns an Jaffe (1989), wir wollen einen aussagekräftigen aber gleichzeitig übersichtlichen Technologieraum. In Benner und Waldfogel (2008) wird gezeigt, dass eine geringe Datenmenge zu sehr unpräzisen, teilweise widersprüchlichen, Ergebnissen führen kann. Mit dieser Auswahl kann also außerdem garantiert werden, dass unser „patent bias“, möglichst gering gehalten wird.

In 3.1, werden die zehn aktivsten Firmen mit ihren jeweiligen Patentzahlen abgebildet. Wir haben insgesamt sechs japanische Unternehmen (Toyota, Nissan, Honda, Denso, Mazda und Hitachi), zwei Unternehmen mit amerikanischem Ursprung (Ford und GM), ein deutsches Unternehmen (Bosch) und Hyundai aus Südkorea. Bosch und Denso sind Automobilzulieferer, die restlichen Unternehmen sind Automobilhersteller im klassischen Sinne. Mit über 30000 angemeldeten Patenten, fällt der japanische Automobilhersteller Toyota klar aus dem Muster. Die neun anderen Firmen befinden sich alle in einem Intervall von (4000, 10000) Patenten. Honda liegt mit knapp über 7000 angemeldeten Patenten auf dem vierten Platz hinter Toyota, Nissan und Bosch.

Patente, die von verschiedenen Unternehmen angemeldet wurden, dennoch demselben Konzern angehören dürfen nicht vergessen werden. So haben wir beispielsweise die Unternehmen „Toyota Motor Corporation“, „Toyota Industries Corporation“ und „Toyota Central Research & Development Lab“ zusammengefasst.

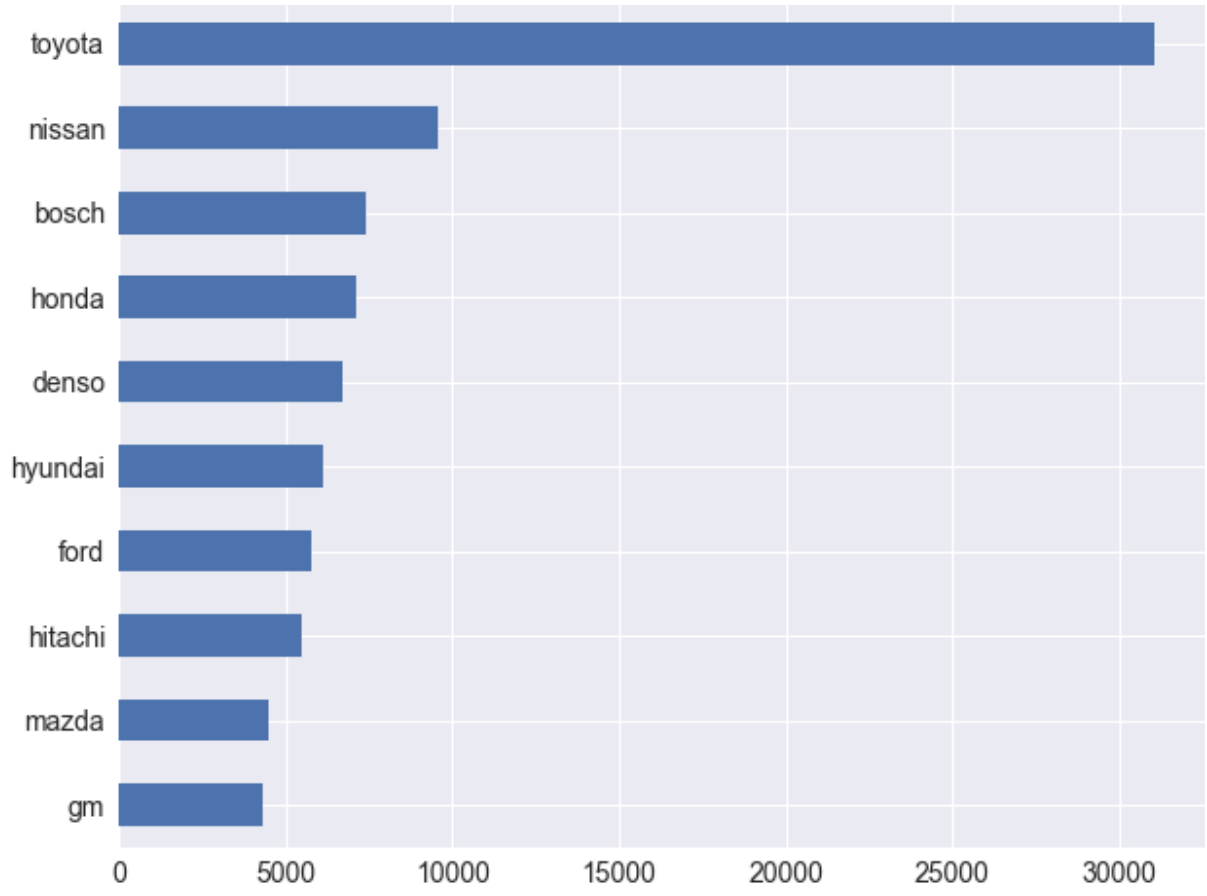

Abbildung 3.1: Patentzahl Top 10 Unternehmen

3.2.2 Bildung der Patentvektoren und der Distanzmatrix

Wir wollen die in 3.1.3 vorgestellten Vektoren f_{ik} für unsere Firmen und unsere Patentklassen bilden. Wir betrachten zunächst die ersten vier Stellen der CPC-Klassen und fassen alle Patente nach ihrer Gruppe zusammen.

Für die Vektorlänge wählen wir im ersten Schritt ad-hoc $k = 15$. In 3.2 sehen wir die Verteilung aller $Y02T\_10$ Patente auf die fünfzehn häufigsten CPC-Klassen. Bei der Gruppe F02D handelt es sich um Patente im Bereich „Controlling Combustion Engines“. Diese Klasse beinhaltet, für die Automobilindustrie eher klassische Pantente wie „Treibstoffkontrollsysteme “ und „Motorkontrollsysteme “. Interessant ist beispielsweise die Klasse B60L. Dabei handelt es sich um Patente im Bereich der Elektroantriebe. So geht es hier um Themen wie „Electric propulsion with power supply from forces of nature, e.g. sun or wind“ und „Methods, circuits, or devices for controlling the traction-motor speed of electrically-propelled vehicles“. Fraglich ist ob wir unsere Patente wie in Jaffe (1989) clustern sollten. Würde man alle Patente in der Automobilindustrie betrachten, wäre eine Einteilung in einzelne Bereiche sicher sinvoll. Da wir aber explizit Patente einer Gruppe betrachten ($Y02T\_10$), ist ein Sinnzusammenhang inhärent gegeben. Ein Clustering ist für unseren Daten also überflüssig.

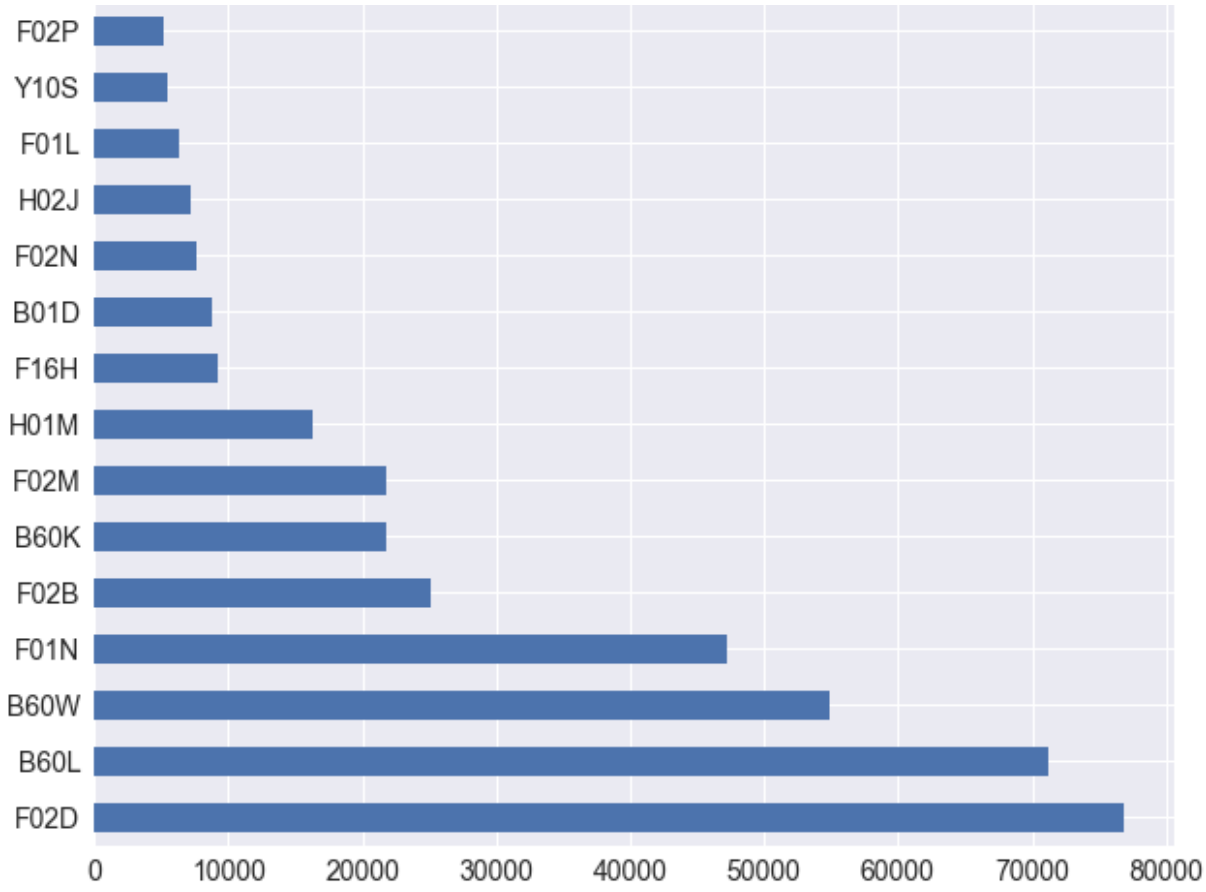

Abbildung 3.2: Anzahl der $Y02T\_10$ Patente nach Gruppierung

Für die Einteilung der Patentvektoren folgen wir dem Vorgehen von Jaffe (1989) und ordnen jeder Firma einen Vektor mit den in 3.2 abgebildeten CPC-Klassen zu. Wir gruppieren unsere Daten also nach Firmen und zählen die jeweils angemeldeten Patente pro Patentklasse. Wir erhalten also beispielsweise $f_{honda\ F02P} = 371$ und $f_{hondaF02D} = 4710$. So ergibt sich für die Firma Honda folgender Vektor:

$$f_{honda} = \{371, 773, 628, 496, 478, 520, 627, 902, 1690, 2660, 2003, 2349, 3522, 7218, 4710\}$$

Wir bilden diesen Vektor für jede Firma und wenden anschließend die Metrik aus 3.1 an. Es ergibt sich folgende Distanzmatrix:

	toyota	nissan	bosch	honda	denso	hyundai	ford	mazda	hitachi	gm
toyota	1									
nissan	0.963	1								
bosch	0.913	0.901	1							
honda	0.982	0.968	0.891	1						
denso	0.910	0.926	0.972	0.898	1					
hyundai	0.950	0.908	0.885	0.953	0.879	1				
ford	0.962	0.957	0.940	0.948	0.952	0.951	1			
mazda	0.656	0.816	0.656	0.686	0.700	0.605	0.741	1		
hitachi	0.891	0.913	0.844	0.923	0.869	0.826	0.827	0.626	1	
gm	0.974	0.902	0.883	0.948	0.855	0.936	0.929	0.555	0.814	1

Wir lesen: $P_{honda\ toyota} = P_{toyota\ honda} = 0.982$. Wie bereits erwähnt wird der Zähler bei der Kosinus-Ähnlichkeit im Nenner normalisiert. Die Distanzen befinden sich in einem Wertebereich von 0 bis 1, wobei die Firmen bei 1 identisch sind und bei 0 keine Patente in derselben CPC-Klasse halten. Betrachtet man die einzelnen Werte der Matrix, fällt auf, dass sich alle Firmen, bis auf Mazda, sehr nahe stehen. Intuitiv ist dieses Ergebnis sinnvoll. Anders als in Jaffe (1989) vergleichen wir ausschließlich Unternehmen aus einem Industriezweig. Zusätzlich sind die von uns gewählten CPC-Klassen relativ ungenau. Wir betrachten lediglich die CPC-Klassen die uns durch die Gesamtpatentanzahl vorgegeben wird, es ist also möglich, dass eine Firma Patente in einer CPC-Klasse hält, diese aber nicht in unserem Vektor vorkommt. Durch diesen Effekt „verlieren“ wir potentielle Patentklassen, welche unsere Firmen weiter unterscheiden könnten. Um das Problem zu umgehen wählen wir eine andere Vektorzusammensetzung: Sei M die Menge der von uns betrachteten CPC-Klassen. Wähle die größten 30 CPC-Klassen c_{ik} ($i = 1...30$) für jede unserer zehn Firmen k. Ist c_{ik} nicht in M. Füge c_{ik} M hinzu.

Um für bessere Interpretierbarkeit zu sorgen, wollen wir im nächsten Abschnitt die resultierende Distanzmatrix mittels multidimensionaler Skalierung in einer zweidimensionalen Ebene darstellen.

3.2.3 Multidimensionale Skalierung

Multidimensionale Skalierung (MDS) ist ein Berechnungsverfahren, das dabei helfen soll verdeckte Strukturen in Daten aufzuzeigen (Kruskal 1978). Angenommen man hat eine Karte, die verschiedenen Position amerikanischer Städte zeigt. Durch einfaches messen der Abstände zwischen den Städten, lässt sich eine Tabelle mit den jeweiligen Abstandswerten ausfüllen. Die Rückrichtung dieser Aufgabe ist das Problem der MDS (Kruskal 1978). Im Prinzip geht es also darum N Objekte, geometrisch durch N Punkte zu repräsentieren. Dabei sollen die Abstände zwischen den Punkte im zweidimensionalen Raum, die Unterschiede der Objekte (hier: unsere zehn Firmen) reflektieren können. Bei der klassischen MDS werden die Distanzen zwischen den Objekten mit der euklidischen Metrik berechnet. Da wir bereits eine „eigene“ Distanzmetrik haben, fällt dieser Schritt weg. Ausgehend von unserer Matrix folgen alle weiteren Schritte der MDS dem Shepard-Kruskal Algorithmus. Das Ergebnis der multidimensionalen Skalierung ist, bis auf Rotation und Skalierung, eindeutig. Umgesetzt wird die multidimensionale Skalierung mithilfe der freien Software-Bibliothek „Scikit-learn“.

```
# Importieren der Bibliothek
from sklearn.manifold import MDS

# Setze Anzahl der CPC-Klassen pro Firma und random_state für Reproduzierbarkeit:
def mds(anzahl=30, random_sate = 42):

    # Berechnung der Distanzen, Metrik ist die Kosinus-Ähnlichkeit
    jt = jaffetable(cosine_distance, anzahl)
    d = jt.to_numpy()
    N = d.shape[0]

    # Korrektur der Indizes
    for i in range(N):
        for j in range(i+1,N):
            d[i,j] = d[j,i]

    # Initalisierung der MDS, zwei Dimensionen, Distanzmatrix ist vorgegeben
    embedding = MDS(n_components=2, dissimilarity = 'precomputed', random_state = random_state
        )

    # Durchführung der MDS
    m = embedding.fit_transform(d)

    # Erstellung des Plots
    fig = plt.figure(figsize=(8,8))
    ax = fig.add_subplot()
    plt.axis('equal')

    # Übertragung der Koordinanten in den Plot
    plt.scatter(m[:, 0], m[:, 1], color='navy')

    # Beschriftung der Punkte
    for i in range(N):
        plt.annotate(jt.index[i].lower(),(m[i,0] + 0.01, m[i,1]))
```

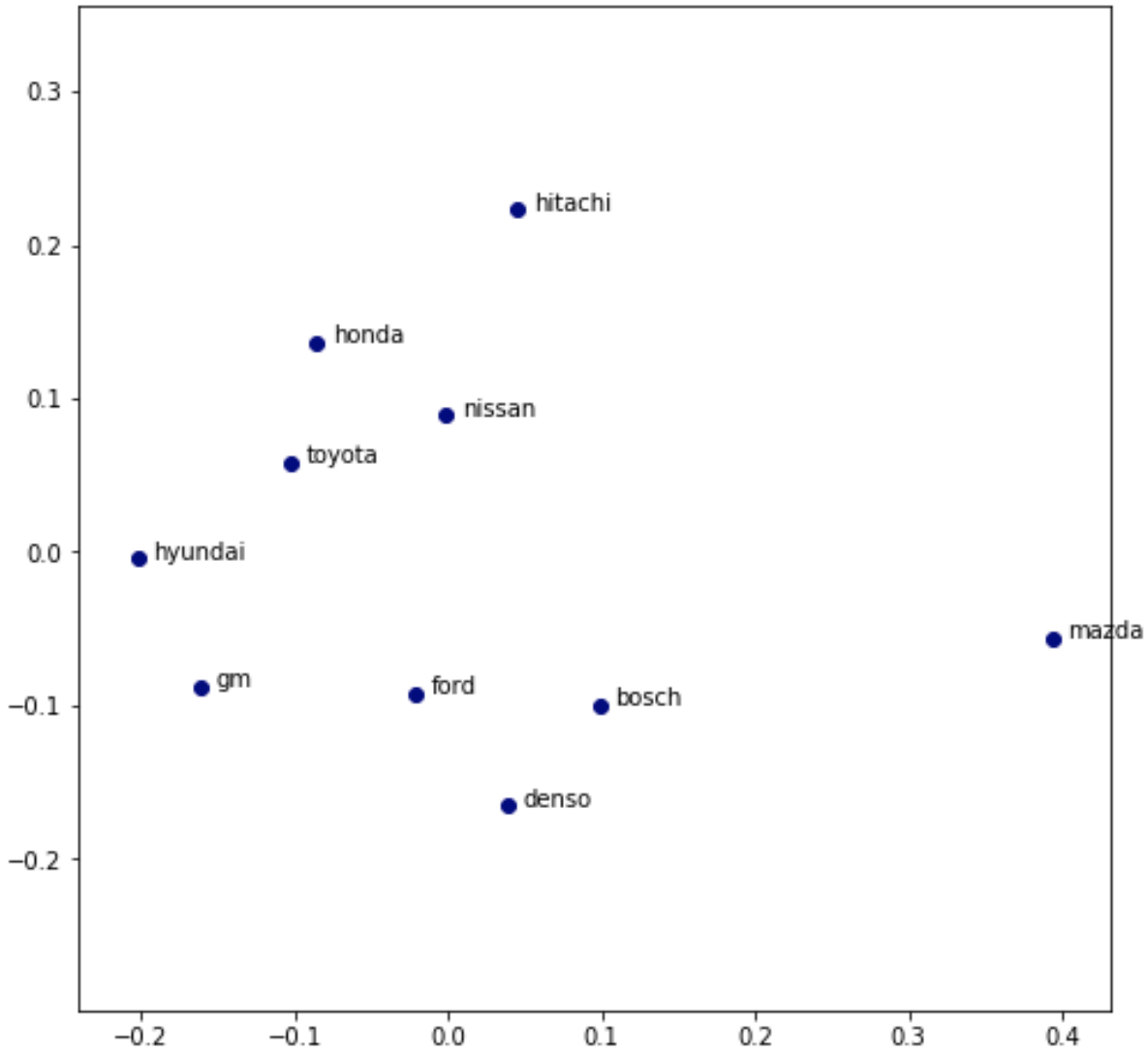

Abbildung 3.3: Graph der multidimensionalen Skalierung nach Patentvektoren

3.2.4 Interpretation des Graphen

Bei der Interpretation eines MDS-Graphen ist es immer wichtig zu beachten, dass sowohl die Achsen, als auch die absoluten Positionen der Firmen keine Rolle spielen. Es geht lediglich um das Verhältnis der Firmen zueinander. Würde man beispielsweise eine weitere Firma in die Berechnung aufnehmen, werden alle Firmen ihre Positionen ändern, je nachdem in welchem Verhältnis diese zu der neuen, und den anderen, Firmen stehen.

Wir erkennen zunächst ein kleineres Firmencluster, bestehend aus Ford, Bosch und Denso. Da Bosch und Denso Automobilzulieferer sind, könnte man davon ausgehen, dass beide Unter-

nehmen in ähnlichen technologische Sektoren aktiv sind. So stellen Automobilzulieferer oft Einzelteile wie z.B. Zündkerzen und Schrauben her. Das Cluster Ford, Bosch, Denso (und GM) unterscheidet sich primär durch starke Aktivität in der Patentklassen[4] F02D (Verbrennungsmotor) und vergleichsweise schwache Aktivitäten in der Patentklasse mit Hauptfokus auf Elektromobilität (B60L). Als zweite Gruppe beobachten wir die drei größten japanischen Automobilhersteller: Toyota, Honda und Nissan. Mit 32% aller Neuzulassungen in Japan liegt Toyota an der Spitze, gefolgt von Honda und Nissan zu je 15%.[5]. Die Graphik legt nahe, dass das Patentierverhalten der drei Firmen ähnlich sein muss. Tatsächlich sind die Unternehmen, bis auf kleinere Abweichungen, in denselben Patentklassen aktiv. Primär beobachten wir überdurchschnittliche Aktivitäten in der CPC-Klassen B60L (Elektromobilität). Am interessantesten jedoch ist die Position von Mazda. Während alle anderen Unternehmen mehr oder weniger nahe zueinander stehen, fällt Mazda gänzlich aus dem Muster. Grund dafür ist eine beachtliche Patentzahl in der CPC-Klasse F02D. Genauer handelt es dabei um die Patente der Klassen F02B53/00 und F02B2053/005. Die zugrundeliegende Technologie dieser Patente ist der Wankelmotor. "Mazda hat in Japan einen außergewöhnlichen Hybrid-Antrieb mit Wankelmotor patentieren lassen. Das System arbeitet mit ingesamt drei E-Motoren und soll laut Mazda vor allem leichter sein als aktuelle elektrische Allradantriebe."[6]. Um Aussagen über die Entwicklung der Technologien zu machen und mögliche Zielrichtung der Unternehmen festzustellen wäre es interessant zu sehen, wie sich die Position der Firmen über mehrere Zeiträume hinweg verändern.

3.2.5 Zeitliche Entwicklung

In Jaffe (1986) wird untersucht welche Technologiecluster über die Zeit an Mitglieder „gewinnen“. Wir wollen zeigen in welche Richtungen sich die Firmen in unserem MDS Plot über zwei Zeiträume hinweg bewegen. Dabei ergibt sich jedoch folgendes Problem: Die Abstände der Firmen zueinander ist immer eindeutig, allerdings können die Firmen im resultierende MDS Graph rotieren. Für dieselbe Lösung gibt es also eine Vielzahl unterschiedlicher Graphen. Das macht es für uns zunächst unmöglich zwei Graphen miteinander zu vergleichen, beziehungsweise die Bewegungsrichtung einer Firma aufzuzeigen. Mithilfe des setzen einer „random_state“-Variable, lassen sich zwar Graphen genau reproduzieren, allerdings verändern sich unseren Ausgangsdaten für den zweiten Zeitraum, damit besteht das Problem also weiterhin. Wir schlagen folgende Lösung vor: wir berechnen den Graphen für den ersten Zeitraum und merken uns die Koordinaten der Punkte. Für den zweiten Graphen berechnen wir den Abstand der resultierenden Koordinaten zu den Koordinaten des ersten Graphen für (beispielsweise) 10000 verschiedene „random_states“. Wir verwenden schließlich die Koordinaten, die den geringsten Abstand zu den Koordinaten des ersten Graphen aufweisen. Damit garantieren wir, dass sich die beiden

[4]Wir betrachten die Patentklassen in Relation zu der jeweilige Gesamtpatentanzahl
[5]Quelle: https://www.autoscout24.de/auto/japanische-automarken/ Stand: 6.10.2020
[6]Quelle: https://www.electrive.net/2020/04/20/ Stand: 06.10.2020

Graphen sinnvoll vergleichen lassen, ohne dabei falsch implizierte Bewegung der Firmen, durch Rotation des Graphen aufgrund der MDS zu erhalten.

Wir betrachten die Zeiträume der Jahre 2000 bis 2010 und 2011 bis 2018. Daraus ergeben sich die Graphen **??**

- Plot

3.2.6 Robustheit

Um die Robustheit der Ergebnisse zu garantieren wurden folgende Maßnahmen durchgeführt. Naheliegend ist es zunächst die Länge der Patentvektoren (K) zu variieren. Je nach Vektorlänge werden mehr oder weniger Patentklassen für die Berechnung berücksichtigt. Um alle relevanten Patentklassen in die Berechnung aufzunehmen, sollte K also möglichst hoch gehalten werden, dabei ist das Ergebnis bei einem Wert von $K = 30$, schon sehr robust, eine weitere Erhöhung der Vektorlänge führt zu keiner signifikanten Veränderung im Graphen. Als nächstes wurde die Granularität der betrachten Patentklassen verändert. Bis jetzt haben wir nur die ersten vier Stellen der CPC-Klassen betrachtet, davon ausgehend wurden die Patentzahlen der Firmen gruppiert. Betrachtet man beispielsweise die ersten sieben Stellen der CPC-Klassen werden die Patentvektoren „feiner“. So wird die CPC-Klasse F02D, beispielsweise zu F02D_13 und F02D_41. Auch hier finden wir keine signifikante Veränderung des Graphen. Da die Wahl der Distanzmetrik Ergebnisse stark beeinflussen kann, wurde im letzten Schritt statt der Kosinus-Ähnlichkeit die in Bar und Leiponen (2012) beschriebene „min-complement distance“.

$$P_i j = 1 - \sum_{k=1}^{K} \min\{p_{ik}, p_{jk}\} \tag{3.4}$$

Die Motivation hinter der wohldefinierten Metrik ist es irrelevante Patentklassen für die Berechnung der Distanzen zwischen den Firmen i und j nicht zu berücksichtigen. Irrelevante Patentklassen gelten als solche, wenn eine Firma in dieser Patentklasse keine Patente hält. Bar und Leiponen (2012) behauptet, dass die technologische Distanz zwischen zwei Firmen, von den Patentklassen abhängen sollte, in denen beide Firmen aktiv sind. In Gleichung 3.4 werden jeweils die Minima der Patentzahlen pro Klasse und Firma betrachtet. Ergibt sich für den Term $\min\{p_{ik}, p_{jk}\} = 0$, hat das keinen Einfluss auf die Summe. Im Kontext der Robustheitsprüfung ergab das Anwenden der „min-complement distance“, die größte Veränderung (siehe **??**). Allerdings bleiben die oben angesprochenen Gruppierungen der Firmen weiterhin bestehen.

Letztendlich lässt sich also sagen, dass der Graph der MDS und unser Technologieraum nach Jaffe (1986) in seinem Ergebnis robust ist.

3.2.7 Zusammenfassung

4 Der Technologieraum basierend auf Patentzitaten

In diesem Kapitel wird ein Technologieraum auf Basis von Patentzitaten vorgestellt. Zunächst werde ich die Paper Stuart und Podolny (1996) und Kitahara und Oikawa (2017) zusammenfassen und anschließend die darin vorgestellten Modelle implementieren. In Kapitel 3 haben wir den Technologieraum für die Firma Honda und ihre Konkurrenz nach dem Ansatz von (Jaffe 1989) dargestellt und interpretiert. Grundlage dabei war es die Firmen aufgrund von Patentklassen zu unterscheiden. Als nächstes wollen wir versuchen etwas näher an den Daten zu arbeiten um gemeinsame Wissensgrundlagen von Unternehmen für die Formulierung des Technologieraums zu nutzen.

4.1 Literatur

4.1.1 Grundlagen

Stuart und Podolny (1996) stellen neben Jaffe (1986) eine andere Möglichkeit vor, wie Unternehmen und deren Technologien in einem Technologieraum dargestellt werden können. Zunächst stellen die Autoren die „lokale Suche“ als eine zentrale Voraussetzung für die Lokalisation von Firmen in einem Technologieraum fest. Bei der lokalen Suche geht es um die Annahme, dass Firmen ihre Forschungs- und Entwicklungskapazitäten in die Bereiche investieren, in denen sie bereits erfahren und profiliert sind. Somit ist es wahrscheinlicher, dass Daimler beispielsweise im nächsten Quartal ihre Forschungsgelder in die Weiterentwicklung des autonomen Fahrens investiert, als in Tierfutter. Ein zitatbasierter Ansatz setzt voraus, dass Unternehmen auf vergangenem Wissen aufbauen. Würden Firmen ihre Innovationsrichtungen also willkürlich wählen, wäre ein solche Herangehensweise sinnlos. Im Hauptteil ihrer Arbeit stellen die Autoren eine Methodik vor, wie sich - anhand von Überschneidungen in Patentzitaten - ein Maß für die Ähnlichkeiten verschiedener Firmen im Forschungs- und Entwicklungsbereich berechnen lassen.

Kitahara und Oikawa (2017) erweitern in ihrer Publikation den traditionellen Ansatz von Stuart und Podolny (1996), um Aussagen über die Entwicklung des Innovationsverhalten von Firmen

in den Vereinigten Staaten zu treffen. Um präzisere Ergebnisse zu erzielen, berücksichtigen die Autoren dabei nicht nur „first-order overlaps“ sondern auch „second-order overlaps“. Dieses Vorgehen ermöglicht es zusätzlich zu den Patentzitaten auf der ersten Ebene, die von den Patentzitaten auf der ersten Ebene wiederum zitierten Patente - die der zweiten Ebene - in das Modell aufzunehmen.

4.1.2 Zielsetzung der Studien

Ziel der Arbeit von Stuart und Podolny (1996) ist es ihren „neu definierten“ Technologieraum durch ein eigenes Beispiel zu motivieren. Mithilfe des definierten Modells wir die Veränderungen der technologischen Positionen zehn japanischer Firmen in der Halbleiterindustrie über einen Zeitraum von 1982 bis 1992 in drei Zeitintervallen dargestellt. Die Darstellung erfolgt mittels multidimensionaler Skalierung. Änderung des strategischen und innovativen der Firmen Verhaltens sollen mittels der resultierenden Graphik begründet werden. So kann beispielsweise der innovative Erfolg einer Firma, mit ihrer Bewegung im Technologieraum, begründet werden. Weiterhin versuchen die Autoren, Gruppierungen von Firmen im Technologieraum zu interpretieren, um möglicherweise die Formierung strategischer Allianzen aufzuzeigen.

In Kitahara und Oikawa (2017) wird ein allgemeineres Bild betrachtet. Anstatt den Effekt der technologischer Distanz auf Firmenebene zu betrachten, untersuchen die Autoren die Beziehung zwischen innovativem Output und Technologieraumformationen. Es wird versucht folgende Kernfrage zu beantworten: Welche Art von Verteilung stimuliert Innovation? Konstruiert man ein Beispiel wird die Fragestellung klarer. Wir nehmen an, alle Firmen in dem Technologieraum sind auf eine technologische Position konzentriert. Wie Jaffe (1986) suggeriert, werden sich alle Unternehmen aufgrund von Spillovers unfreiwillig gegenseitig helfen, denn sind sich alle Unternehmen ähnlich ist es einer Firma schnell möglich innovative Technologien der anderen Firmen selbst umzusetzen. Auf der anderen Seite sind Übertragungseffekte aber gering, wenn sich die Firmen über den kompletten Technologieraum verteilen. Kitahara und Oikawa (2017) untersuchen die Mitte dieser beiden Extrema. Angenommen es gibt zwei, voneinander entfernte Pole, um die sich die Firmen gruppieren. Die beiden Firmengruppen nutzen jeweils unterschiedliche Technologien und wetteifern darum, welche Technologie zum Marktstandard wird.

4.1.3 Der Technologieraum nach Stuart und Podolny (1996)

Nimmt man an, dass Unternehmen ihre Forschungsrichtungen „lokal suchen“, ist die patentbasierte Definition eines Technologieraums relativ intuitiv. Haben Firmen ähnliche Wissensquellen werden sie ähnliche Technologien besitzen und in ähnlichen Technologiesektoren forschen.

Demnach werden sich die Firmen im Technologieraum nahe stehen. Die technologische Distanz ist nach Stuart und Podolny (1996) also ein Maß nach Gemeinsamkeiten in Wissensbasen.

Der Koeffizient α_{ij} repräsentiert das Verhältnis, der Anzahl an Wissensquellen auf der Firma j aufbaut, die auch von Firma i genutzt werden. Werden Patentzitate als Analogien für Wissensquellen behandelt so ist der Wert α_{ij} der Anteil der zitierten Patente von Firma i, die auch Firma j zitiert. Die „competition-coefficients“ sind Werte zwischen 0 und 1, wobei die Firmen bei $\alpha_{ij} = 0$ auf keine gemeinsamen Patente zitieren. Für ein System von N Firmen, kann eine asymmetrische NxN Matrix aufgestellt werden. Die Einträge dieser „community matrix“ sind die Koeffizienten α_{ij} und α_{ji} für $(j = 1, 2, ..., N; i = 1, 2, ..., N; i \neq j)$.

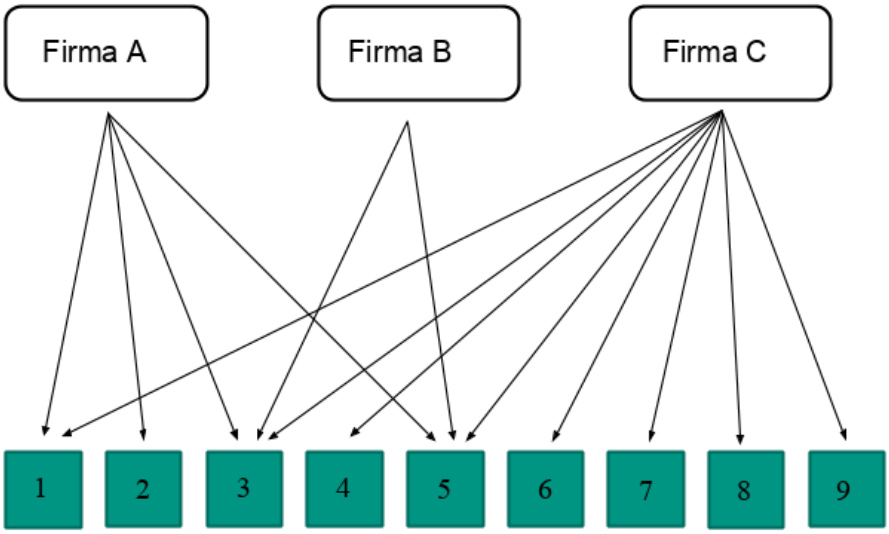

Abbildung 4.1: Beispiel: drei Firmen zitieren Patente

Die Abbildung 4.1 zeigt drei Firmen A, B und C und die Patente 1 bis 9. Der Pfeil von einer Firma zu einem Patent impliziert ein Zitat. So zitiert beispielsweise Firma A insgesamt vier Patente (1, 2, 3 und 5). Firma B zitiert lediglich zwei Patente (3 und 5). Der Koeffizient α_{AB}, ist der Anteil der zitierten Patente von Firma A, die auch von Firma B zitiert werden. Es ergibt sich $\alpha_{AB} = \frac{2}{4} = 0.5$. Da Firma B alle Patente zitiert, die auch von Firma A zitiert werden ergibt sich umgekehrt $\alpha_{BA} = 1$. Die Werte entlang der Hauptdiagonale liefern keine Informationen und werden mit 0 besetzt. Für die Firmen A, B und C ergibt sich folgende „community matrix“:

$$\begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \\ \end{array} \begin{pmatrix} 0 & 0.50 & 0.75 \\ 1 & 0 & 1 \\ 0.375 & 0.25 & 0 \end{pmatrix} \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \\ \end{array}$$

Die „community matrix“, dient als Ausgangspunkt. Firma A baut auf 100% des Wissens der Firma B und 37.5% der Firma C auf. Der Spaltenvektor einer Firma lässt sich auch als deren „technologische Nische“ verstehen. Um zu untersuchen, wie sich die Firmen im Technologieraum über die Zeit bewegen, definieren die Autoren zunächst die Distanz zwischen Firma i und j zum festen Zeitpunkt t_m. Die (euklidische) Distanz zwischen Firmen ist ein Maß davon, inwieweit sich die technologischen Nischen der Firmen i und j mit denen aller anderen Firmen k überschneiden.

$$d_{jit_m} \equiv d_{ijt_m} = \left\{ \sum_{k=1}^{n} [(\alpha_{ikt_m} - \alpha_{jkt_m})^2 + (\alpha_{kit_m} - \alpha_{kjt_m})^2] \right\}^{\frac{1}{2}}, k \neq i, j \tag{4.1}$$

Die α's sind die Wettbewerbskoeffizienten der i'ten und j'ten Dyaden zum Zeitpunkt t_m. Der erste Term $(\alpha_{kit_m} - \alpha_{kjt_m})$ beschreibt die Differenz Nischenüberschneidungen zwischen Firma i und anderen Firmen k und der Nischenüberschneidungen von Firma j und anderen Firmen k. Der zweite Term $(\alpha_{ikt_m} - \alpha_{kjt_m})$ ist die Gegenrichtung. Er gibt an inwieweit sich die technologischen Nischen der Firma k mit denen der Firmen i und j überschneiden.

Um die Distanz zwischen Firmen zu zwei verschiedenen Zeitpunkten festzuhalten, wird weiterhin eine Matrix nach folgender Metrik definiert.

$$d_{it_ljt_m} \equiv d_{jt_mit_l} = \left\{ \sum_{k=1}^{n} [(\alpha_{ikt_l} - \alpha_{jkt_m})^2 + (\alpha_{kit_l} - \alpha_{kjt_m})^2] \right\}^{\frac{1}{2}}, k \neq i, j \tag{4.2}$$

Die Idee ist Analog zu 4.1, nur werden jetzt die Differenzen der Nischenüberschneidungen über die Zeitpunkte t_l und t_m gebildet. Sind alle Firmen in jeder Zeitperiode aktiv ergeben sich für die symmetrische Distanzmatrix D, $N * T$ Zeilen und $N * T$ Spalten, wobei N die Anzahl der Firmen, und T die Anzahl der Zeitperioden ist. Im letzten Schritt wird die Matrix mittels multidimensionaler Skalierung (siehe Abschnitt 3.2.3) graphisch dargestellt.

4.1.4 Die Firma Mitsubishi im Technologieraum

Stuart und Podolny (1996) wenden die oben beschrieben Methodik an um Positionsveränderungen zehn japanischen Firmen in der Halbleiterindustrie darzustellen. Für den Technologieraum wurden die Jahre 1982, 1987 und 1992 gewählt. Da es zehn Firmen zu je drei Zeiträumen gibt, hat der resultierende Graph 30 Punkte. Stuart und Podolny (1996) teilen den Technologieraum zunächst in zwei Gruppen ein. Auf der einen Seite befinden sich Firmen, die „einfache“, auf den Mainstream-Konsum basierte Technologien produzieren. Auf der anderen Seite befinden sich Unternehmen die sich in komplexeren, industriellen Technologiesektoren aufhalten. So produzieren diese Firmen die Kerntechnologien die in einem jeden Computer stecken (beispielsweise Schaltungslogik für Prozessoren).

Das Unternehmen Mitsubishi, beschloss Mitte der 80er Jahre aufgrund von Prognosen einen strategischen Wechsel von dem Konsummarkt in den Industriemarkt vorzunehmen. Maßnahmen waren dabei unter anderem eine drastische Erhöhung der Forschungs- und Entwicklungsausgaben, um in den Markt für Speicherentwicklung vorzudringen, sowie die eigene Herstellung von Intel's Mikroprozessoren. Mithilfe ihrer Darstellung (siehe 4.2) können die Autoren zeigen, dass dieser Strategiewechsel erfolgreich war. Mitsubishi bewegt sich innerhalb des betrachteten Zeitraums im Technologieraum weiter weg von Firmen wie Sony und immer näher zu innovativen Unternehmen wie Hitachi und Toshiba.

4.1.5 Strategische Allianzen

Im Gegensatz zu Spillovers hat eine Firmenallianz symbiotische Eigenschaften. Austausch von Wissen und Technik, soll einen positiven Einfluss auf die Marktposition aller beteiligte Firmen haben. Stuart und Podolny (1996) fanden im Rahmen ihrer Recherche 35 Fälle von Wissensaustausch zwischen den zehn Firmen.

Wie oben bereits erwähnt teilen die Autoren ihren Technologieraum in zwei Bereiche ein. Dabei werden die Firmen der linken Technologieraumhälfte als „core“ bezeichnet. Diese Firmen sind besonders innovativ und können hohe Marktanteile aufweisen. Dazu gehören unter anderem die Firmen Hitachi, Toshiba, NEC, Fujitsu und Mitsubishi nach ihrer Neuorientierung. Auf der anderen Seite des Technologieraums befinden sich Firmen der „periphery“. Darunter befinden sich Firmen wie Sony und Sanyo, die sich auf den Konsumhandel konzentrieren.

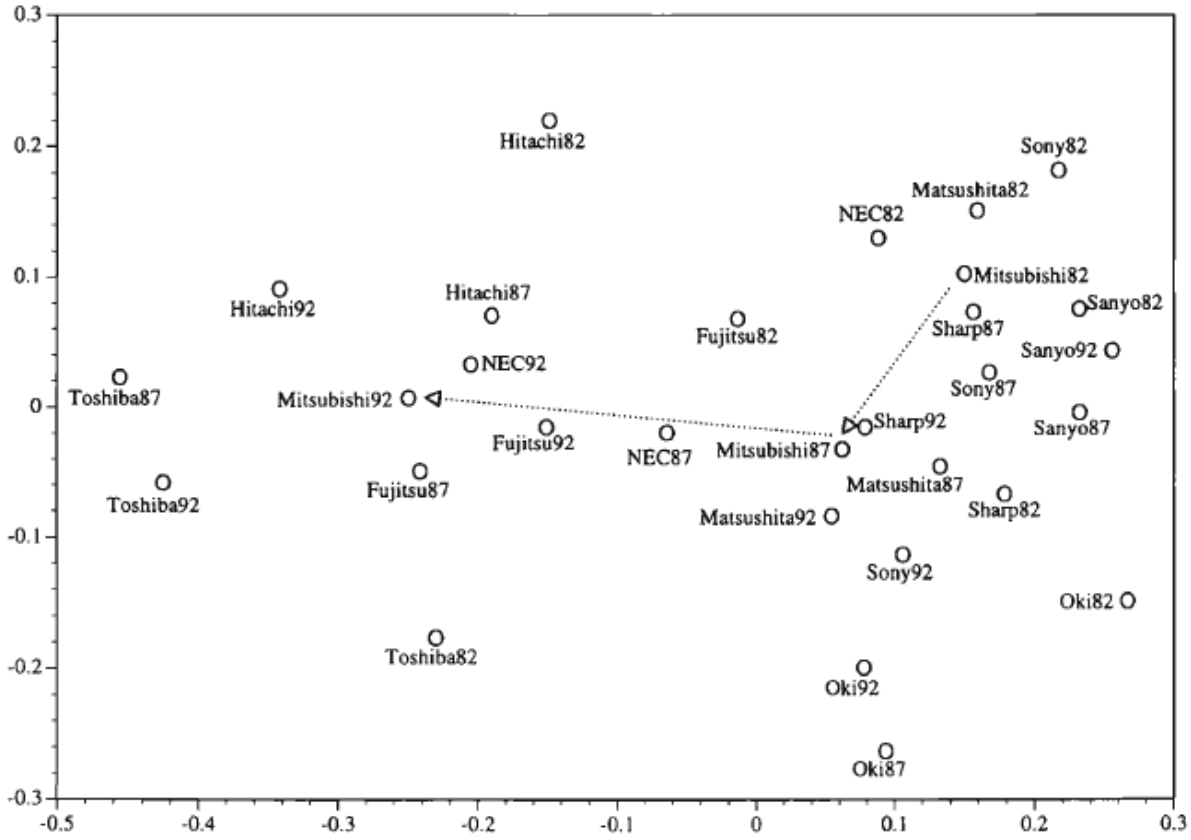

Abbildung 4.2: Technologische Positionen von Firmen in der japanischen Halbleiterindustrie (Stuart und Podolny 1996, S. 30)

Firmen im „core“-Bereich des Technologieraums sind nach Stuart und Podolny (1996) an insgesamt 31 der 35 Partnerschaften beteiligt. Strategische Allianzen finden sich also fast ausschließlich in die Richtungen „core-to-core“, beziehungsweise „core-to-periphery“. Mitsubishi, die Firma die sich innerhalb des Technologieraumes am extremsten bewegt hat, war in allen drei Jahren an den meisten Firmenpartnerschaften beteiligt. Im allgemeinen korreliert die Bewegungsdistanz einer Firma im Technologieraum positiv mit der Anzahl ihrer eingegangenen Partnerschaften (Stuart und Podolny 1996, S. 34). Gerade das Wechseln des Technologieclusters ist für Firmen nur in seltenen Fällen mit eigener Innovationskraft zu bewältigen. Nach Stuart und Podolny (1996) besteht damit ein klarer Zusammenhang zwischen der Affinität einer Firma strategischen Allianzen einzugehen und ihrer Positionierung im Technologieraum. Auch wir konnten im vorherigen Kapitel eine Annäherung der Firma Ford und der Firma Toyota feststellen. Diese Bewegung könnte mitunter durch eine, im Jahre 2011 eingegangene Partnerschaft im Hybridsektor[1] begründet sein.

[1] Quelle: Forbes Stand: 09.10.2020

4.1.6 Erweiterung des Ansatzes nach Kitahara und Oikawa (2017)

Kitahara und Oikawa (2017) erweitern den von Stuart und Podolny (1996) entwickelten, auf Zitaten basierenden Ansatz, um eine weitere Ebene. Die Intuition dahinter ist, dass Zitate von Zitaten (Zitate der „zweiten Ebene“) für die Wissensflüsse weitere, wichtige Informationen beinhalten können. In diesem Abschnitt soll das Modell anhand eines eigenen Beispiels veranschaulicht werden.

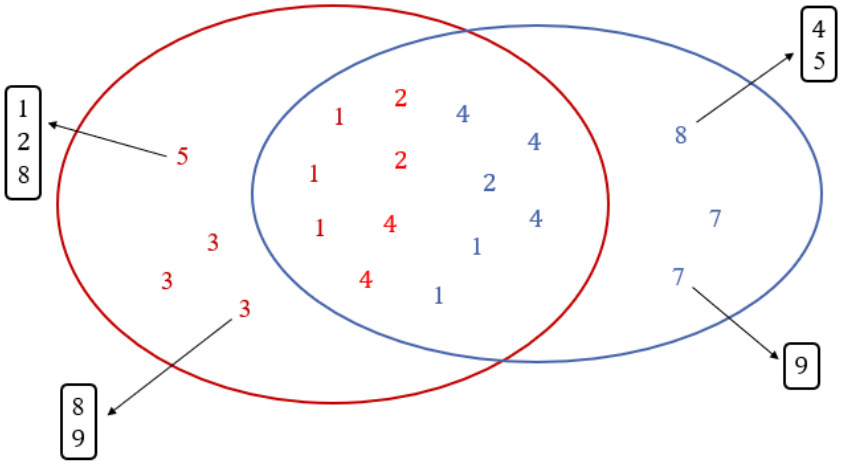

Abbildung 4.3: Patentzitate der Firma Blau (links) und Firma Rot

Die Abbildung 4.3 zeigt zwei fiktive Firmen (Rot und Blau)[2] und ihre jeweiligen Patentzitate. So zitiert Firma R Die roten Patente 1, 2, 3 und 5 unterschiedlich oft. Analog zitiert Firma B die blauen Patente 1, 2, 4, 7 und 8. Die Patentzitate der zweiten Ebene werden durch Pfeile impliziert. So zitiert beispielsweise Patent 3, wiederum die Patente 8 und 9. Kitahara und Oikawa (2017) führen vier Gleichungen ein, die alle Zitationsbeziehungen einfangen.

$$\omega_{ij}^{1} \equiv |O(P_i, P_j)| + |O(P_j, P_i)| \tag{4.3}$$

Sei $|P|$ die Anzahl der Patente in einer Liste P. Seien weiterhin $O(P_i, P_j)$, alle Patente in P_i, die sich mit den Patente in P_j überschneiden (Wiederholungen inklusive). Wir erhalten für $P_R = \{1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 5\}$ und $P_B = \{1, 1, 2, 4, 4, 4, 7, 7, 9\}$, $|P_R| = 11$ und $|P_B| = 9$. ω_{ij}^{1} zählt die Patente im Schnittbereich der beiden Firmen. Nach Gleichung 4.3 erhalten wir für unsere Firmen: $\omega_{RB}^{1} = \omega_{BR}^{1} = 13$. Im Unterschied zu Stuart und Podolny (1996), werden auch Mehrfachzitate in die Berechnung aufgenommen. Auch hier ist die Intuition klar. Zitiert

[2] Weiter: Firma R und Firma B

eine Firma dasselbe Patent mehrfach, so wird es für sie eine bedeutendere technologische Rolle spielen, als ein einfach zitiertes Patent.

Auch wenn sich Zitate nicht direkt überschneiden, könnten die Patente indirekt technologisch verknüpft sein. Seien also $\tilde{P}_{ij}$[3], die Patente aus P_i, die sich nicht mit P_j überschneiden (Patente außerhalb der Schnittmenge). Weiterhin seien $C_i(p)$, Patent p's Zitate, mit $p \in \tilde{P}_{ij}$. Auch hier berücksichtigen wir also nur die Patent(zitate), die noch nicht durch die Gleichung 4.3 abgedeckt wurden. Den Elementen aus $\tilde{P}_{ij}$ wird ein positives Gewicht k zugeordnet, wenn p_k ein Patent in $\tilde{P}_{ij}$ zitiert oder von einem Patent in $\tilde{P}_{ji}$ zitiert wird. Für die erste Komponente der „second-order overlaps“ gilt:

$$\omega_{ij}^{21} = \sum_{k=1}^{n_{ij}} \frac{|O(C_i(p_k), \tilde{P}_{ij})|}{|(C_i(p_k)|} \tag{4.4}$$

Wir erhalten für $\omega_{RB}^{21} = \frac{\{|8|\}}{1} + \frac{\{|8|\}}{2} + \frac{\{|8|\}}{2} + \frac{\{|8|\}}{2} = 1 + 0.5 + 0.5 + 0.5 = 2.5$, veranschaulicht in Abbildung 4.4. ω_{ji}^{21} wird analog zur Gleichung 4.4 definiert. Für unser Beispiel erhalten wir $\omega_{BR}^{21} = 1$.

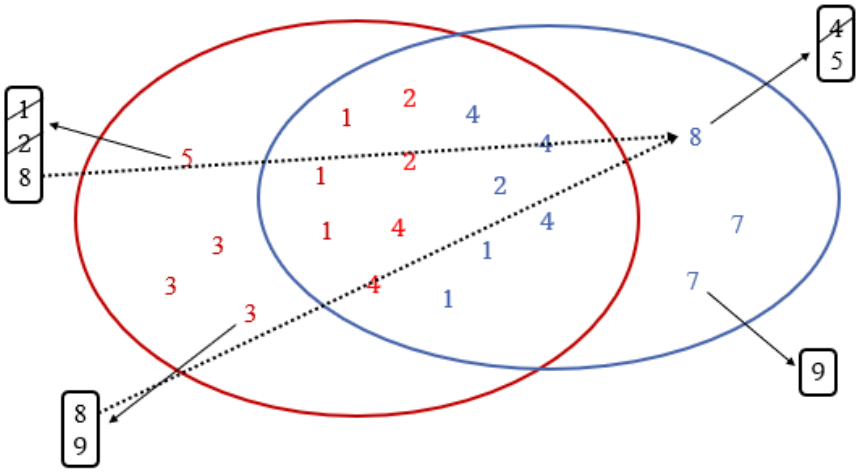

Abbildung 4.4: Patentzitate der zweiten Ebene, ω_{RB}^{21}

Mit der zweiten Komponente der „second-order overlaps“, werden alle „übrigen“ Patente verrechnet. Seien $\tilde{P}'_{ij}$ also die Patente in $\tilde{P}_{ij}$, die nicht mit den Patente in $\tilde{P}_{ji}$ überlappen und $\tilde{n}'_{ij}$ die

[3]Man beachte: $\tilde{P}_{ij}$ kann auch Wiederholungen beinhalten

Anzahl der Patente in $\tilde{P}'_{ij}$. Es gilt:

$$\omega_{ij}^{22} = \sum_{k=1}^{n'_{ij}} \frac{|O(C_i(p_k), C_j(\tilde{P}_{ji}))|}{|C_i(p_k)|} \tag{4.5}$$

ω_{ji}^{22} ist wieder analog definiert. Da die Menge $\tilde{P}'_{RB}$ leer ist erhalten wir für $\omega_{RB}^{22} = 0$. Für die Gegenrichtung (Abbildung 4.5) erhalten wir $\omega_{BR}^{22} = \frac{|O(C_B(7),\{9\})|}{1} + \frac{|O(C_B(7),\{9\})|}{1} = \frac{\{|9|\}}{1} + \frac{\{|9|\}}{1} = 2$.

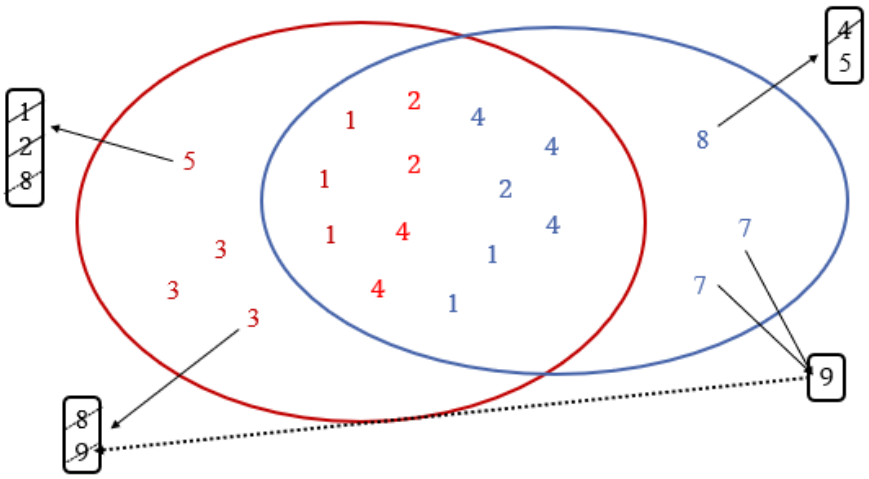

Abbildung 4.5: Patentzitate der zweiten Ebene, ω_{BR}^{22}

Als nächstes wird die gewichtete Summe aller ω gebildet, wobei $\eta \in (0, 1)$.

$$\omega_{ij} = \frac{\omega_{ij}^{1} + \eta(\omega_{ij}^{21} + \omega_{ji}^{21}) + \eta^2(\omega_{ij}^{22} + \omega_{ji}^{22})}{|P_i| + |P_j|} \tag{4.6}$$

η lässt sich als Diskontierungsfaktor für die technologische Relevanz der„second-order Overlaps“ interpretieren. Je niedriger η gewählt wird, desto weniger haben die Patentzitate der zweiten Ebene einen Einfluss auf das Endergebnis. Weiterhin wird ω_{ij} im Nenner normalisiert. Es gilt also: $\omega_{ij} \in [0, 1]$. Für $\eta = 0.5$ erhalten wir: $\omega_{BR} = \omega_{RB} = \frac{13+0.5(2.5+1)+0.5^2(0+2)}{11+9} = 0.7625$.

Im letzten Schritt überführen Kitahara und Oikawa (2017) die berechneten Indizes in eine symmetrische Distanzmatrix.

$$d_{ij} = -\log(\omega_{ij}) \tag{4.7}$$

Die resultierende Matrix kann anschließend erneut mithilfe der multidimensionalen Skalierung auf eine zweidimensionale Ebene projiziert werden.

4.1.7 Technologische Polarisationen

Die Polarisation eines Technologieraums ist eine Konfiguration bei der sich ein Großteil der Unternehmen an zwei, voneinander entfernten, Punkten im Technologieraum gruppieren. Wie bereits in Abschnitt 4.1.2 erwähnt untersuchen Kitahara und Oikawa (2017) Technologieraumformationen um durch Polarisation induzierte Innovationsaktivitäten zu begründen. Mit dem oben beschriebenen Ausgangsmodell wird die Verteilung von US-amerikanische Firmen in der Halbleiterindustrie in Zeiträumen von jeweils fünf Jahren dargestellt.

Wo Jaffe (1986) und Stuart und Podolny (1996) Technologien auf Firmenebene untersuchen, wollen Kitahara und Oikawa (2017) ein möglichst umfangreiches Bild schaffen. So betrachten die Autoren alle Firmen des Industriesektors. Um Polarisationen innerhalb der Verteilungen möglichst gut zu veranschaulichen, führen die Autoren außerdem einer Kerndichtenschätzung durch.[4]. Dabei wird die Gruppen- bzw. Clusterbildung von Unternehmen durch eine „Heatmap“ sichtbar gemacht.

In der ersten Reihe der Abbildung 4.6 sehen wir das Ergebnis der MDS, jedes Unternehmen wird durch einen Punkt im Raum repräsentiert. Die zweite Reihe zeigt die Konturen der Kerndichtenschätzung. Dabei suggerieren helle Punkte eine große Firmendichte.

Im weiteren Verlauf ihrer Arbeit motivieren Kitahara und Oikawa (2017) weitere Indikatoren. So wird beispielsweise die Gesamtdurchschnittsdistanz aller Firmen im Technologieraum zueinander über mehrere Jahre berechnet. Außerdem wird ein Polarisationsindex mittels zweidimensionaler KDE aufgestellt. Im Ergebnis können die Autoren zunächst durchschnittlich steigende technologische Distanzen und zunehmende Polarisationen US-amerikanischer Firmen feststellen. Zusätzlich wird eine Beziehung zwischen Polarisationen und Anzahl an Patentzitaten dargestellt. So können die Autoren zeigen, dass zunehmende Polarisationen die durchschnittliche

[4] Verfahren: multivariate kernel desinity estimation (KDE)

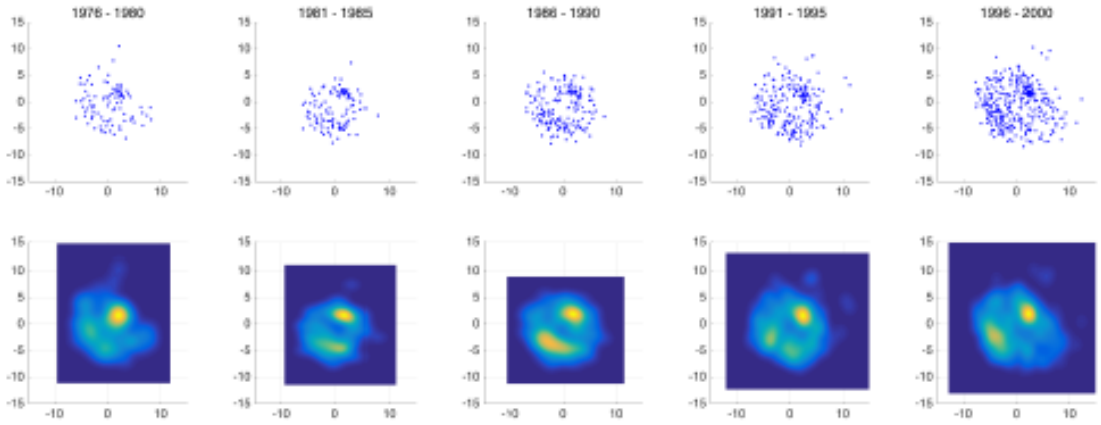

Abbildung 4.6: Bewegung US-amerikanischer Firmen über die Zeit (Kitahara und Oikawa 2017, S. 12)

Anzahl an Patentzitaten steigert und somit Innovation induziert. Dieser Effekt lässt sich allerdings nur für Patentdaten vor dem Jahre 1990 beobachten. Die Autoren führen diese Tatsache auf US Patentreformen in den 80er Jahren zurück.

4.2 Der Technologieraum für die Firma Honda und ihre Konkurrenz nach Stuart und Podolny (1996) und Kitahara und Oikawa (2017)

4.2.1 Umsetzung des Modells nach Stuart und Podolny (1996) und Intepretation der Ergebnisse

Wir wollen das in Abschnitt 4.1.3 vorgestellten Modell eines Technologieraum, basierend auf Patentzitaten implementieren und das resultierende Bild anschließend mit dem aus 3.2.3 vergleichen.

Wir erweitern unsere Datenbasis (alle $Y02T\_10$-Patente) um ihre jeweiligen Zitationen. Im Durchschnitt zitiert eine Firma ca. 29000 Patente. Die Anzahl der Zitationen pro Firma wird in Abbildung 4.7 für die ersten 15 Unternehmen dargestellt.

Da wir die Technologieräume miteinander vergleichen wollen, wählen wir zunächst dieselben zehn Firmen wie in Kapitel 3. Die Wettbewerbskoeffizienten (α_{ij}) berechnen wir gemäß Definition (Abschnitt 4.1.3). Die gemeinsame Wissensbasis zweier Firmen erhalten wir durch Bildung der Patentschnittmenge.

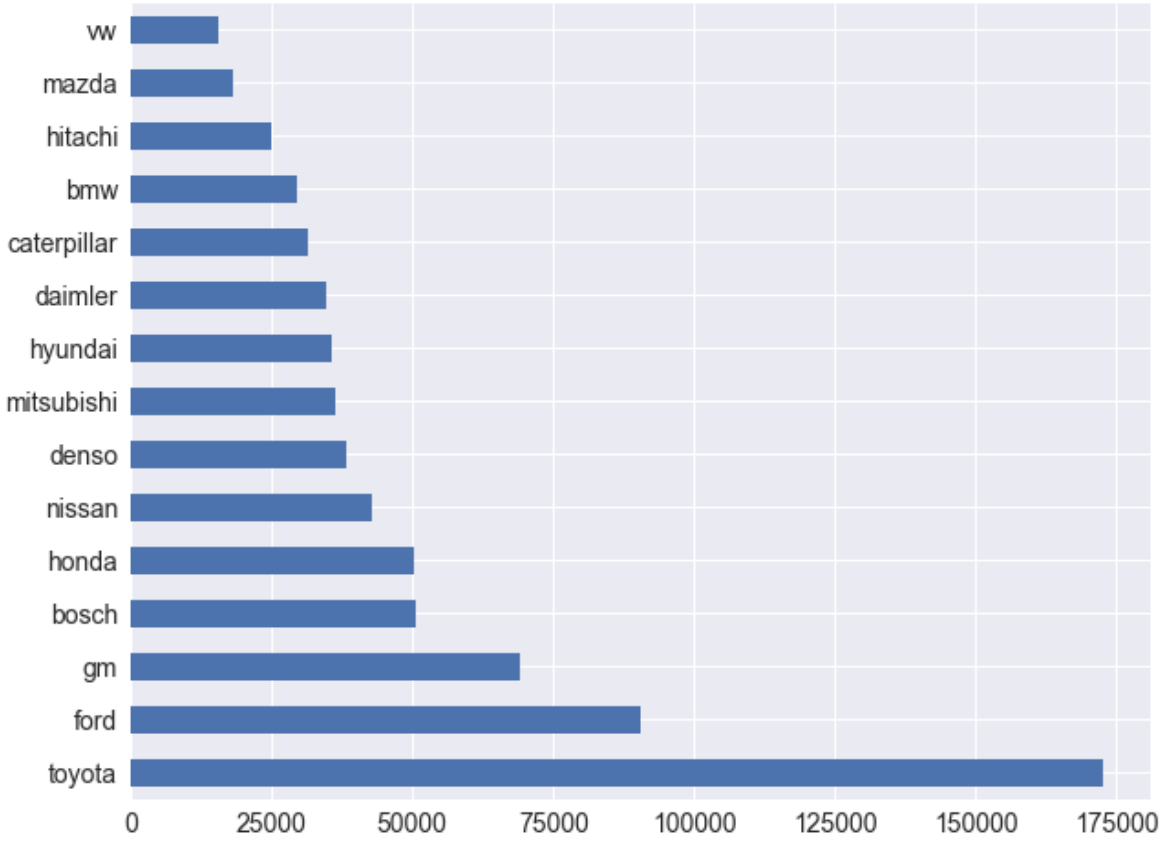

Abbildung 4.7: Anzahl Zitationen pro Firma

```python
def getAlpha(i,j):

    # Hole Firma i und Firma j aus Liste
    firmaI = firmenlist[i]
    firmaJ = firmenlist[j]

    # Alle Patente, die von Firma i zitiert werden
    dfI = df4.loc[df4.firma == firmaI]
    patI = set(dfI.cited_family_id.unique())

    # Alle Patente, die von Firma j zitiert werden
    dfJ = df4.loc[df4.firma == firmaJ]
    patJ = set(dfJ.cited_family_id.unique())

    # Patente, die von Firma i und j zitiert werden
    patSchnitt = patI & patJ

    # Anteil der Patente, die Firma i zitiert, die auch von Firma j zitiert werden.
    aij = len(patSchnitt) / len(patI)
    return aij
```

Um ein möglichst Umfangreiches Bild zu schaffen betrachten wir Patentdaten für den Gesamtzeitraum (1970 - 2019). Für unsere zehn Firmen erhalten wir eine 10x10 „community matrix“.

Anschließend berechnen wir die Distanzen der Firmen zueinander nach der oben definierten Metrik 4.1. Im letzten Schritt wenden wir wie die multidimensionale Skalierung auf unsere Distanzmatrix an.

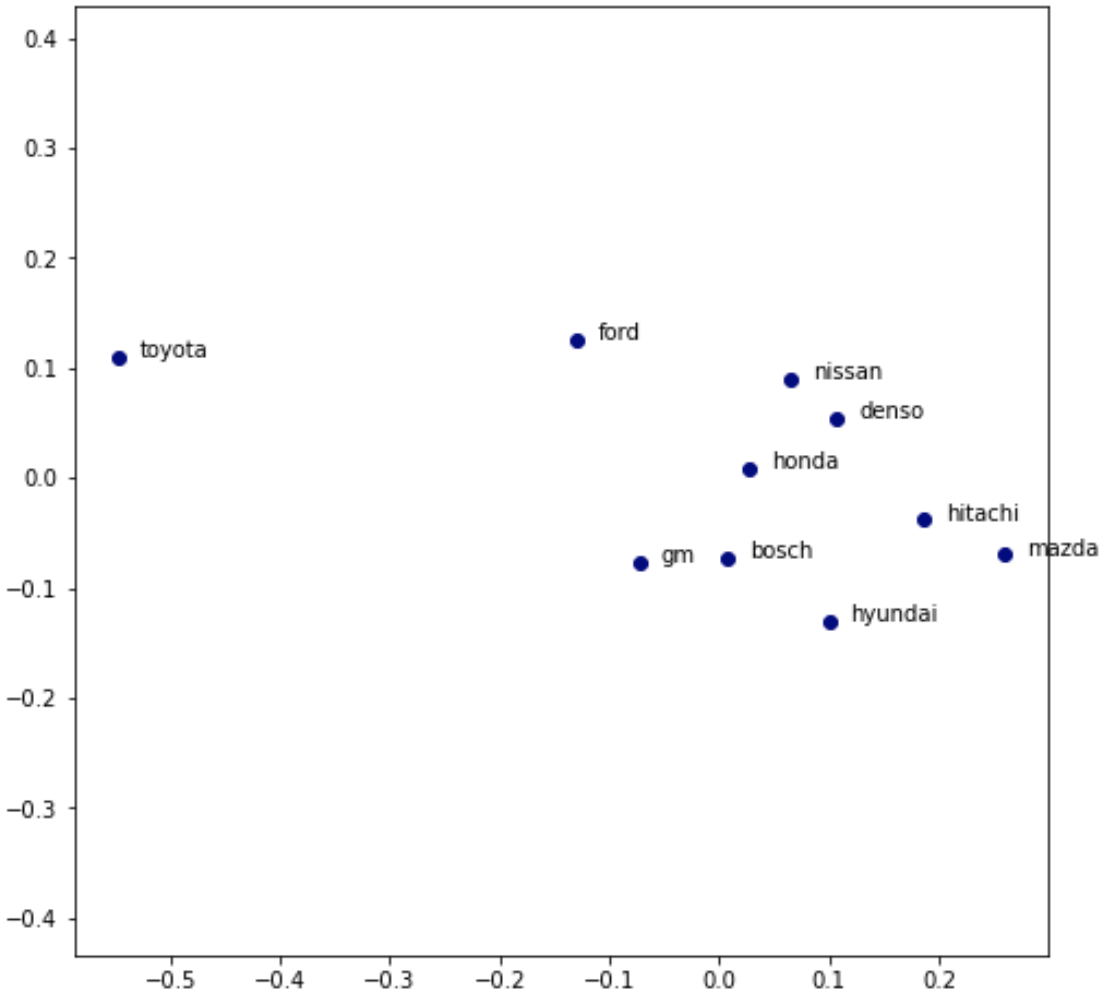

Abbildung 4.8: Graph der multidimensionalen Skalierung nach Patentzitaten

Der Graph in Abbildung 4.8 zeigt den Technologieraum nach Stuart und Podolny (1996) für unsere zehn Firmen. Auf den ersten Blick erkennen wir große Unterschiede zu unserem Graphen in Kapitel 3. So sehen wir auf der einen Seite des Raumes die Firma Toyota. Auf der anderen Seite befinden sich alle restlichen neun Unternehmen eng gruppiert. Trotz der offensichtlichen Unterschiede erkennen wir kleinere Gemeinsamkeiten zu unserem Technologieraum nach **jaffe1989technological**. Um diese Gemeinsamkeiten besser aufzuzeigen, betrachten wir in Abbildung 4.9 den Technologieraum ohne Toyota. Wir können diese Änderung im MDS-Graphen relativ bedenkenlos durchführen, da sich alle anderen Firmen zu Toyota ähnlich verhalten.

Die Gesamtkonstellation fällt etwas anders aus, die Gruppierungen bleiben aber fast gleich. Wir sehen sowohl die Firmengruppierung der Automobilzulieferer Denso und Bosch, als auch die Gruppierung der Konkurrenten Honda und Nissan. Die Firma Mazda befindet sich wieder am

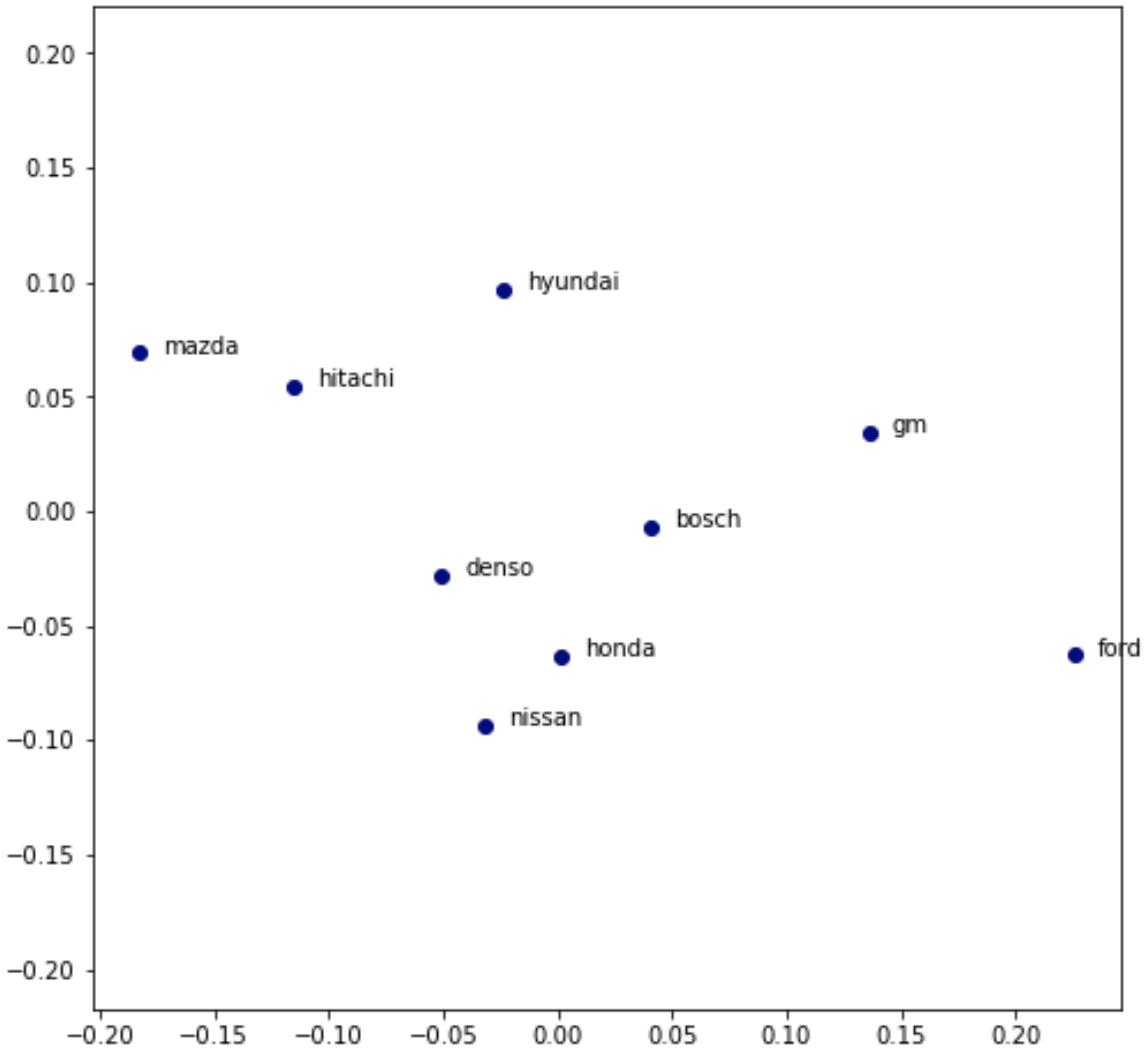

Abbildung 4.9: Graph der multidimensionalen Skalierung nach Patentzitaten ohne die Firma Toyota

Rand des Technologieraums.

Wir wollen versuchen die Position von Toyota zu begründen. Wie in Kapitel 3 bereits gezeigt wurde nimmt die Firma Toyota eine Sonderstellung ein. Die Firma besitzt in der von uns betrachteten Patentklasse mit Abstand die meisten Patente. Eine große Patentzahl impliziert indirekt eine hohe absolute Zahl an Patentzitationen. Abbildung 4.7 bestätigt diese Annahme. Diese Tatsache wiederum führt dazu, dass Toyota eine sehr breite technologische Nische einnimmt. Da die restlichen Unternehmen nur einen geringen Teil dieser technologischen Nische besetzen werden die Wettbewerbskoeffizienten (α_{ij}) dieser Unternehmen im Vergleich zu Toyota relativ klein ausfallen. Zusätzlich ist die von Stuart und Podolny (1996) definierte Metrik, eine euklidische Distanzmetrik. Die Vektorlängen der neun Unternehmen sind sehr klein, die Länge des Vektors der Firma Toyota jedoch relativ groß. Im Ergebnis erhalten wir schließlich die ausfallende Position von Toyota im Technologieraum.

Um diese Annahme zu untermauern nehmen wir zusätzlich eine Firma in die Berechnung auf, bei der wir davon ausgehen können, dass diese nur einen sehr geringen Teil der technologischen Nischen aller anderen Firmen einnehmen wird. Mit knapp 13000 Patentzitaten in der $Y02T\_10$ Klasse ist die Firma LG Electronics nicht weit hinter der Firma Mazda (18000). Allerdings handelt es sich bei LG keineswegs um ein Unternehmen in der Automobilbranche. Somit sollte die Firma nur einen geringen Teil der technologischen Nischen unsere zehn Firmen einnehmen können. Wir gehen also davon aus, dass sich LG am anderen Ende des Technologieraums befinden wird. Der Graph der Abbildung 4.10 bestätigt diese Annahme.

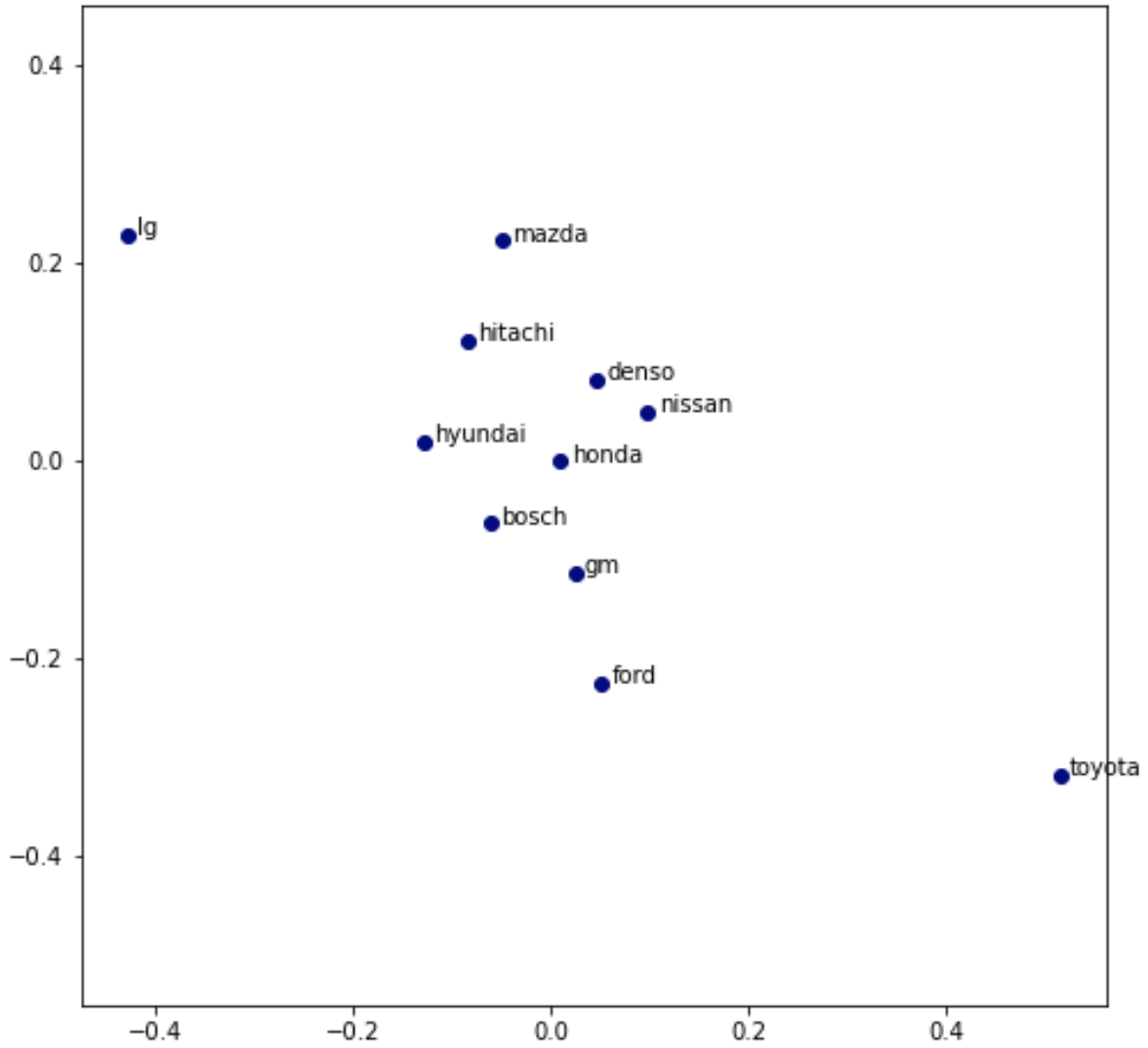

Abbildung 4.10: Graph der multidimensionalen Skalierung nach Patentzitaten mit der Firma LG

4.2.2 Umsetzung des Modells nach (Kitahara und Oikawa 2017) und Intepretation der Ergebnisse

Wir wollen sehen wie sich unser Technologieraum verändert wenn wir Patentzitate auf der zweiten Ebene zulassen. In diesem Kontext erwies sich die algorithmische Umsetzung für die Berechnung der ω-Werte als Herausforderung (Code siehe Anhang A.1). Um Laufzeitprobleme zu vermeiden haben wir sowohl die Hin- als auch die Rückrichtung aller second-order Zitate in je einem „dictionary“ gespeichert (Zugriff in $O(1)$). Wir führen die Berechnung in zwei Schritten durch. Im ersten Schritt lokalisieren wir die Patente anhand der Zugehörigkeit in den verschiedenen ω-Klassen und speichern zusätzlich deren Gewichte. Die Gewichtungen hängen davon ab in welcher ω-Klasse sie sich befinden. Im zweiten Schritt berechnen wir die verschiedenen ω-Werte anhand ihrer Gewichtungen und die Distanz der Firmen nach den Formeln 4.6 und 4.7.

Abbildung 4.11 zeigt den Graph der multidimensionalen Skalierung nach Patentzitaten inklusive Patentzitaten der zweiten Ebene für einen Diskontierungsfaktor von 0.6. Wir beobachten eine Veränderung des Technologieraums im Vergleich zu Abschnitt 4.2.1. Das Zulassen der second-order Zitate „glättet“ das Ergebnis in dem Sinne, als dass es jetzt zu deutlich mehr Überschneidungen in Patentzitaten kommt. Je höher wir η wählen desto größer ist der Einfluss der second-order Zitate. Das Verhältnis der Firmen zueinander ändert sich dabei wenig, allerdings rücken die Firmen näher zusammen (Die Achsen des MDS-Graphen werden kleiner).

Für diesen Ansatz sind die Ergebnisse relativ konsistent zu den Ergebnissen in Kapitel 3. Wir beobachten Mazda weit entfernt von anderen Firmen am Rand des Technologieraums. Die Firmen Bosch und Denso stehen sich wiederum nahe. Weiterhin finden wir Toyta, Nissan und Honda im selben Quadranten des Technologieraums. Auch die Positionen der Firmen GM, Ford und Hyundai verhalten sich ähnlich zu deren Positionen in Kapitel 3.

In (Kitahara und Oikawa 2017) wird die gesamte Technologieraumformation für einen Wirtschaftszweig betrachtet. Wir wollen versuchen eine größere Anzahl an Firmen zu betrachten um möglicherweise eine Polarisation im Innovationssektor der Automobilbranche festzustellen. Um die Übersichtlichkeit des Graphen zu gewährleisten wählen wir die Top 31[5] Unternehmen (nach Patentzahl) und schauen wie sie sich in unserem Technologieraum anordnen (Abbildung 4.12).

Wir erkennen einen Pol um den sich die etablierten OEM's verteilen. An der Peripherie des Pols befinden sich weniger etablierte Automarken und Unternehmen aus der Elektronik- und Halbleiterbranche. Intuitiv ergibt dieses Bild Sinn. Auch wenn nicht-Automobilunternehmen

[5] Top 30 Unternehmen und die Firma Porsche (Platz 31)

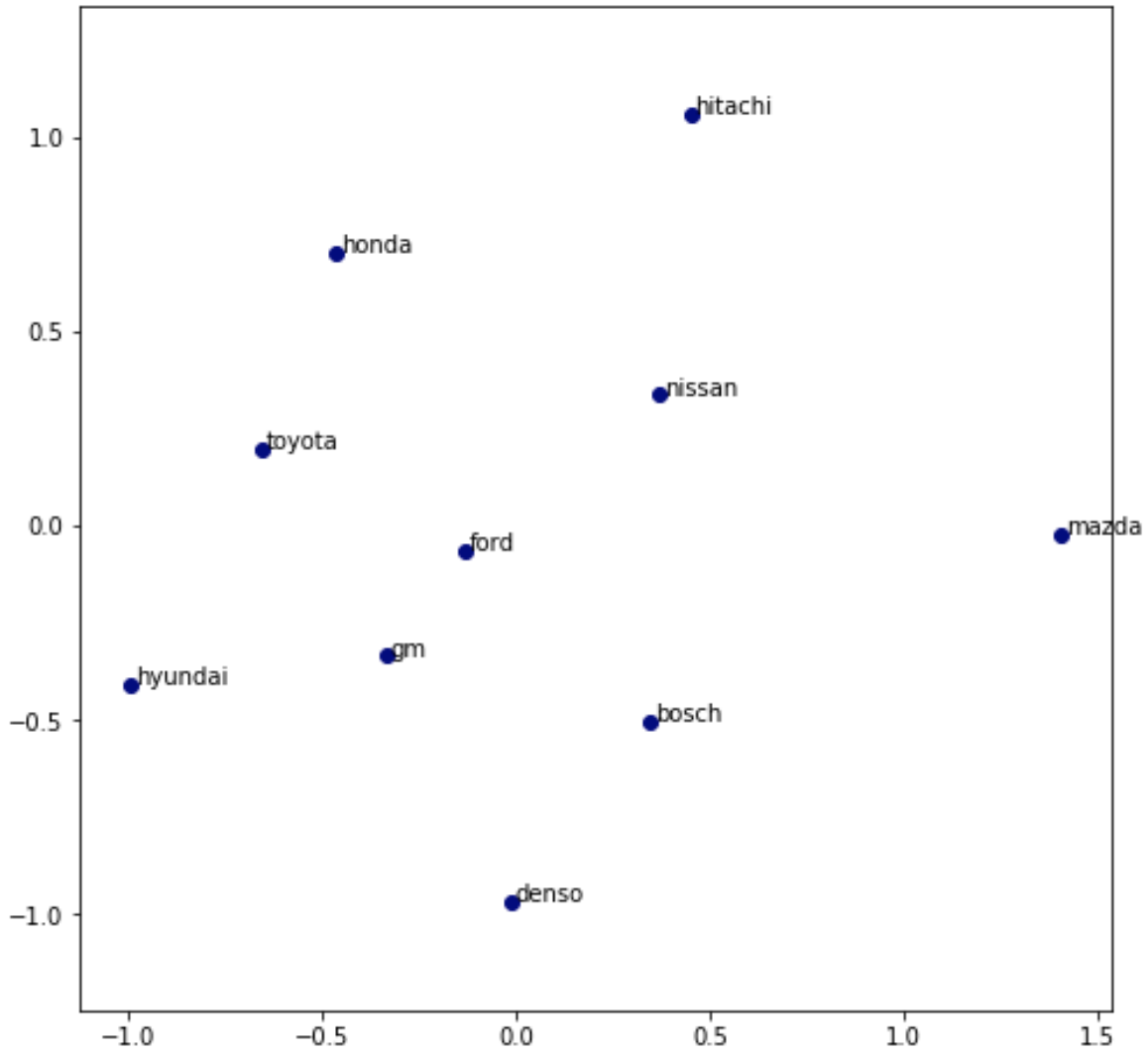

Abbildung 4.11: Graph der multidimensionalen Skalierung nach Patentzitaten inklusive second-order Zitate $\eta = 0.6$

einige Patente der $Y02T\_10$-Kategorie halten, werden sich ihre Wissensgrundlagen weniger mit den Unternehmen der Automobilindustrie decken. Zudem vernetzten sich die Unternehmen der Automobilindustrie gerade bei der Suche nach nicht-fossilen Fortbewegungsmitteln in den letzten Jahren kontinuierlich.

4.2.3 Zusammenfassung

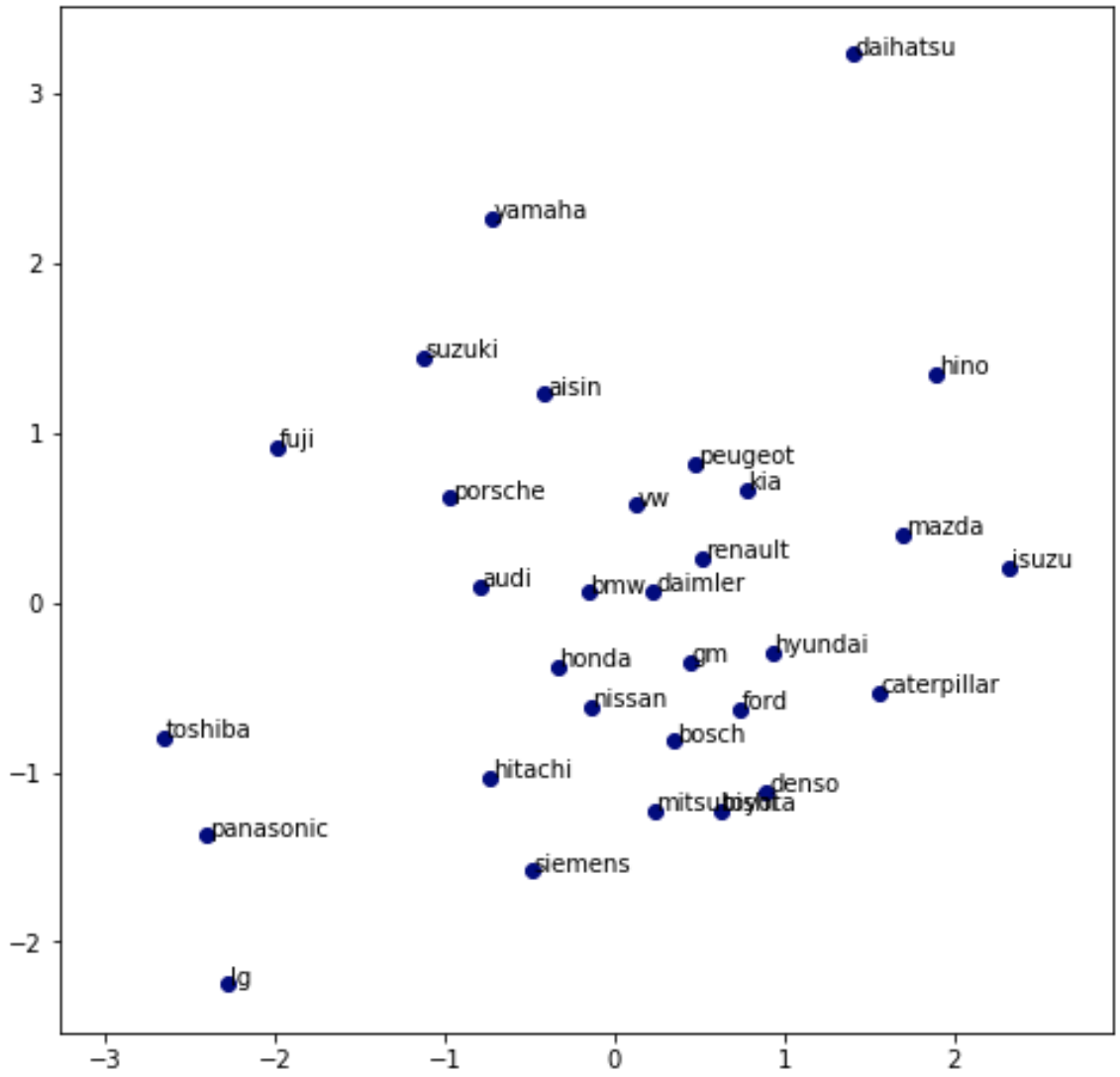

Abbildung 4.12: Graph der multidimensionalen Skalierung nach Patentzitaten inklusive second-order Zitate für 31 Firmen

5 Der Technologieraum auf Basis von Patentauszügen

In Kapitel 3 wurde ein Technologieraum auf der Basis von Patentklassen vorgestellt. Im nächstens Schritt haben wir einen engeren Datenbezug vorausgesetzt und Überschneidungen in Patentzitaten genutzt um einen weiteren Technologieraum für die Firma Honda und ihre Konkurrenz zu formulieren. Als letztes wollen wir noch eine Datenebene tiefer gehen. Dabei sollen die Auszüge und die Titel der Patente als Grundlage für den Technologieraum dienen. Mithilfe von „Topic Modeling“, wollen wir den einzelnen Patente Themen zuordnen und anschließend nach unseren Firmen gruppieren. Im ersten Teil des Kapitels werde ich die Funktionsweise des Topic Modelings zusammenfassen und das Themenmodell im Anschluss auf unsere Patenttexte anwenden.

5.1 Latent Dirichlet allocation

5.1.1 Das Problem

Die Methode des Topic Modeling bietet die Möglichkeit Textsammlungen thematisch zu explorieren. Ein Thema stellt dabei eine Gruppe gewichteter Wörter dar. Der Sinnzusammenhang der Wörter soll dabei im Idealfall auf ein bestimmtes Thema rückschließen lassen. Angenommen wir wollen also fünf Zeitungsartikel nach Inhalt klassifizieren. Für den Menschen erfolgt diese Einteilung während des Leseprozesses meist intuitiv. So könnte man annehmen, dass die Artikel die Themen Sport, Politik und Wissenschaft behandeln. Für den Menschen erfolgt die Themeneinteilung während des Leseprozesses meist intuitiv. Angenommen wir haben keine fünf Artikel, sondern eine Millionen Dokumente. Die algorithmische Umsetzung der Themenzuweisung auf die Dokumente ist das Problem der „latent Dirichlet allocation“.

5.1.2 Grundlagen

Die „latent Dirichlet allocation“ (LDA) ist ein iteratives stochastisches Model für eine Sammlung diskreter Daten (z.B. eines Textkorpus) und wird in (Blei et al. 2003) als das moderne Themenmodell vorgestellt. LDA basiert auf einem dreistufigen Bayes'schen Model. Dabei wird jedes Dokument einer Kollektion aus einer endliche Mischung verschiedener Themen modelliert. Jedes Thema entspricht wiederum einer endlichen Mischung verschiedener Wörter (Blei et al. 2003).

Mit der „unsupervised machine learning“ Technik lässt sich eine beliebige Anzahl an Dokumenten anhand von Themen sortieren, ohne deren wahre Verteilung vorher zu kennen. So eignet sich das Modell primär für die Klassifizierung großer Textdaten. Abstrakt lässt sich die LDA als eine Maschine, die Dokumente produziert, beschreiben. Mit einer sehr geringen Wahrscheinlichkeit wird also beispielsweise der Text der Bibel oder der, der Unabhängigkeitserklärung produziert. Diese „Maschine“ hat verschiedene Einstellungen. Der Algorithmus sucht iterativ die Einstellungen der Maschine, die das zugrundeliegende Dokument (Textinput) am wahrscheinlichsten produziert.

5.1.3 Definition

Behandeln wir die LDA weiterhin analog einer Maschine, so ergibt sich für den Entwurf dieser Maschine folgendes Bild.

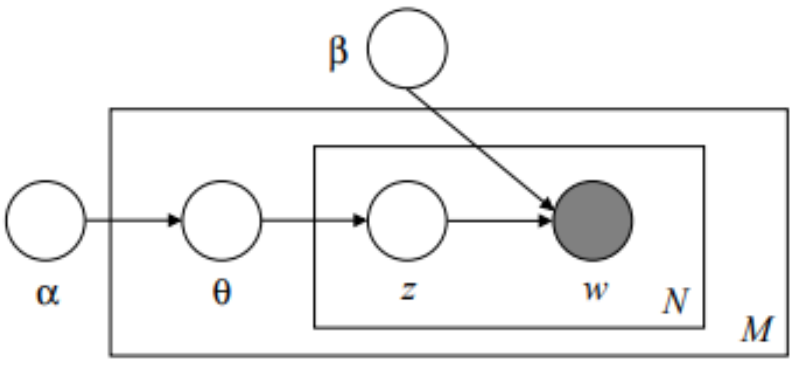

Abbildung 5.1: Graphisches Modell der LDA (Blei et al. 2003, S. 997)

Dabei sind α und β Parameter von Dirichletverteilungen, Θ ist eine Multinomialverteilung. Aus der Multinomialverteilung entstehen die Themen z und die Wörter w. Aus den N Wörtern

erhalten wir den aus M Dokumente bestehenden Textkorpus. Gegeben den Parametern α und β ergibt sich folgende Wahrscheinlichkeit für die Maschine ein Dokument zu produzieren.

$$P(W, Z, \Theta, \phi; \alpha, \beta) = \prod_{j=1}^{M} P(\Theta; \alpha) \prod_{i=1}^{K} P(\phi; \beta) \prod_{t=1}^{N} P(Z_{jt}|\Theta_j) P(W_{jt}|\phi_{z_{jt}}) \tag{5.1}$$

Die ersten zwei Terme sind Dirichletverteilungen, der dritte und der vierte Term sind wiederum Multinomialverteilungen. Beide Verteilungen behandeln jeweils Themen und Wörter des Dokumentes. In den nächsten zwei Abschnitten werde ich genauer auf die beiden Verteilungen eingehen. Dabei werde ich die LDA weiterhin analog einer Maschine beschreiben.[1]

5.1.4 Die Einstellungen der Maschine

Wir betrachten zunächst den ersten Term: $\prod_{j=1}^{M} P(\Theta; \alpha)$. Grundsätzlich lässt sich die Dirichletverteilung als eine geometrischen Dichtefunktion verstehen. Gegeben drei Themen: Sport, Politik und Wissenschaft. Zusätzlich existiert ein Dreieck, wobei ein Thema jeweils einer Ecke des Dreiecks zugeordnet wird. Die Dokumente/Artikel sind Punkte in diesem Dreieck. Diese Punkte unterliegen, abhängig von ihrer Position, ebenfalls einer Verteilung. So könnte man für den linken Punkt in Abbildung 5.2 von einer Verteilung aus 40% Sport, 40% Wissenschaft und 10% Politik ausgehen. Das andere Dokument könnte einer Verteilung von 90% Politik, 5% Wissenschaft und 5% Sport entsprechen. Für drei Themen erhalten wir ein Dreieck. Für N-Themen positionieren sich die Punkte der Verteilung im Raum eines N-Dimensionalen simplex. Analog gibt der Term: $\prod_{i=1}^{K} P(\phi; \beta)$, die Verteilung der Themen an. Dabei entsprechen die Ecken des Simplex den Wörtern der Themen. So erhalten wir einmal eine Assoziation zwischen Dokumenten und deren Themen und zusätzliche eine Beziehung zwischen Themen und Wörtern. Beide Verteilungen zusammen sind die „Einstellungen“ der LDA.

5.1.5 Die Zahnräder der Maschine

Die Multinomialverteilungen sind im Kontext der LDA, die Verteilungen, die auf Basis der Dirichletverteilungen, Dokumente „produzieren“. So werden für den Term $P(Z_{jt}|\Theta_j)$, nach der zugrundeliegenden Verteilung der Dokumente in den Themen $P(\Theta; \alpha)$, für jedes der Dokumente eine Anzahl von Themen gewählt. Für das Beispiel in Abbildung 5.2 erhält man für das linke Dokument, eine Anzahl an Themen entsprechend der Wahrscheinlichkeitsverteilung aus 40% Sport, 40% Wissenschaft und 10% Politik. Jedes Thema repräsentiert eine Verteilung von

[1] Idee: Latent Dirichlet Allocation Stand: 12.10.2020

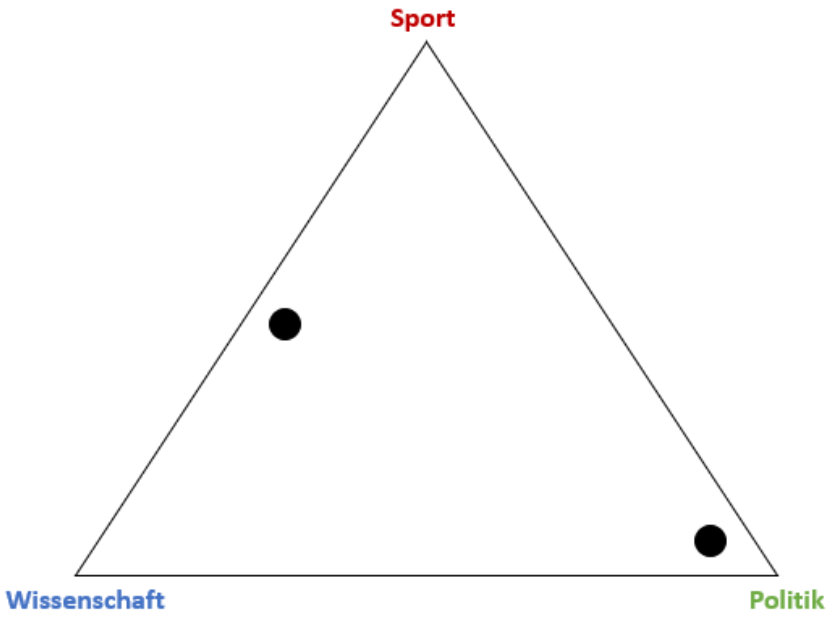

Abbildung 5.2: Verteilung zweier Artikel auf Themen in einer Dirichletverteilung

Wörtern. Für jedes der gewählten Themen aus $P(Z_{jt}|\Theta_j)$, werden für den Term $P(W_{jt}|\phi_{z_{jt}})$ wiederum Wörter nach der zugrundeliegenden Wörter-Themen Wahrscheinlichkeitsverteilung $P(\phi;\beta)$ gewählt (Abbildung 5.3). Die resultierenden Wortgruppen der jeweiligen Themen ist letztlich das Dokument der Maschine.

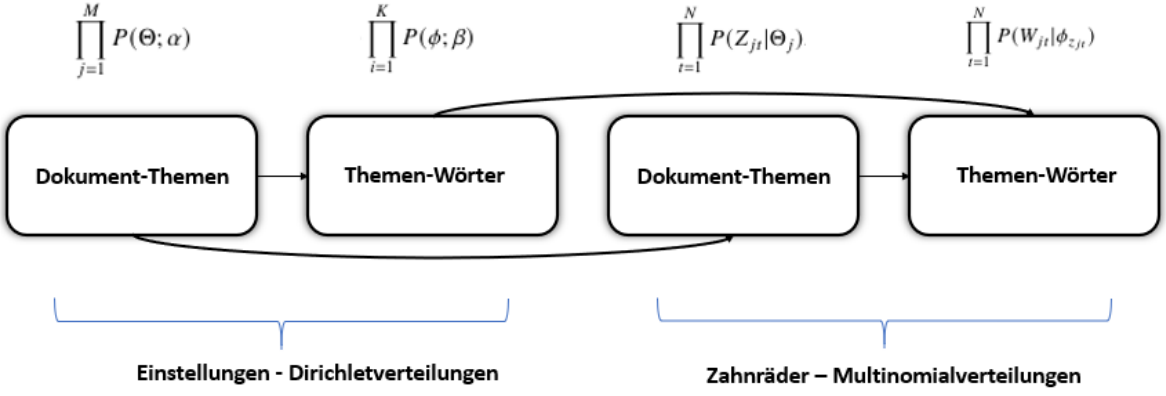

Abbildung 5.3: Zusammenhang der Verteilungen

Der Term 5.1, berechnet mit welcher Wahrscheinlichkeit, dieses Dokument dem Dokument der Eingabe entspricht. In der Realität wird diese Wahrscheinlichkeit sehr gering sein. Entsprechen jedoch die Einstellungen, beziehungsweise die zugrundeliegenden Dokument-Themen und

Themen-Wörter Verteilungen: $\prod_{j=1}^{M} P(\Theta;\alpha) \prod_{i=1}^{K} P(\phi;\beta)$, nicht den Verteilungen der Eingabe, wird die resultierende Wahrscheinlichkeit noch geringer ausfallen.

Der Algorithmus der LDA maximiert also - durch Veränderung der Einstellungen - iterativ die Wahrscheinlichkeit, mit der das resultierende Dokument dem Text der Eingabe entspricht.

5.2 Topic Modeling der Patentauszüge

5.2.1 Datenvorverarbeitung

Zunächst holen wir uns die Texte aller englischen $Y02T\_10$-Patente aus der Datenbank. Da die einzelnen Titel der Patente weitere Informationen enthalten fügen wir sie den Textauszügen hinzu. Bevor die Patenttexte verarbeitet werden können müssen die Daten bereinigt werden. Im ersten Schritt der Datenvorverarbeitung entfernen wir alle „stopwords“ aus den Texten. Bei den Stopwörtern handelt es sich um Wörter, die für das Ergebnis des Themenmodells unwichtig sind, da sie keine Informationen beinhalten. Dabei geht es beispielsweise um Konjunktionen wie „and“ oder den Artikel „the“. Patentspezifische Wörter, wie „Problem“ und „Solution“, die in fast jedem Textauszug vorkommen, fügen wir der Liste manuell hinzu. Insgesamt enthält die Liste der „stopwords“ 185 Einträge (Liste siehe Anhang A.2).

Im zweiten Schritt werden die Daten lemmatisiert. Wörter des selben Wortstamms werden dabei auf ihre Grundform zurückgeführt. So wird beispielsweise aus dem Wort „measuring“, das Wort „measure“, oder aus dem Wort „detected“, „detect“. Ohne Berücksichtigung der Orthographie werden dabei zusätzlich sämtliche Wörter in Kleinschreibung überführt. Sinn der Lemmatisierung ist es, inhaltsgleiche Wortgruppen für den Algorithmus „sichtbar“ zu machen. Die Lemmatisierung und die Entfernung der Stopwörter wurde mit Hilfe der Open-Source Softwarebibliothek „spaCy“ durchgeführt (Anhang **??**)

Im letzten Schritt der Datenvorverarbeitung, filtern wir die Extremwörter. Wörter, die im gesamten Textkorpus zu selten beziehungsweise zu häufig vorkommen, können das Ergebnis des Themenmodells negativ beeinflussen. So wollen wir beispielsweise Rechtschreibfehler ausschließen. Wie in vielen Teilen des Modellierungsprozesses, gibt es für die Parameter des Extremwortfilters keine allgemeingültigen Werte. Die Wahl der Parameter hängt letztendlich immer von den zugrundeliegenden Textdaten ab. In unserem Fall hat sich eine Wahl von: nicht weniger als fünf und nicht mehr als 50% als sinnvoll ergeben. So werden Wörter die weniger als fünf mal vorkommen nicht berücksichtigt. Weiter werden auch Wörter gefiltert, die in mehr als 50% der Patentauszüge vorkommen. Für die 205713 betrachteten Patente erhalten wir im

Ergebnis 12427 verschiedene Wörter in unserem Wörterbuch.

Wir betrachten den Textauszug mit Titel eines zufälliges Patents vor und nach Bearbeitung:

A battery capacity measuring device in accordance with the present invention has a fully-charged state detector (80e), a detected current integrator (80a), a divider (80b), and a corrector (80c) incorporated in a microcomputer (80). The fully-charged state detector detects that a battery is fully charged. The detected current integrator integrates current values that are detected by a current sensor during a period from the instant the battery is fully charged to the instant it is fully charged next. The divider divides the integrated value of detected current values by the length of the period. The corrector corrects a detected current using the quotient provided by the divider as an offset. Furthermore, a remaining battery capacity calculating system comprises a voltage detecting unit (50), a current detecting unit (40), an index calculating unit, a control unit, and a calculating unit. The voltage detecting unit detects the voltage at the terminals of a battery. The current detecting unit detects a current flowing through the battery. The index calculating unit calculates the index of polarization in the battery according to the detected current. The control unit controls the output voltage of an alternator so that the index of polarization will remain within a predetermined range which permits limitation of the effect of polarization on the charged state of the battery. When the index of polarization remains within the predetermined range, the calculating unit calculates the remaining capacity of the battery according to the terminal voltage of the battery, that is, the open-circuit voltage of the battery. APPARATUS FOR BATTERY CAPACITY MEASUREMENT AND FOR REMAINING CAPACITY CALCULATION.

Tabelle 5.1: Ein Patent vor der Datenvorverabeitung

'battery', 'capacity', 'measure', 'device', 'accordance', 'present', 'invention', 'fully', 'charge', 'state', 'detect', 'current', 'integrator', 'divider', 'corrector', 'incorporate', 'fully', 'charge', 'state', 'detector', 'detect', 'battery', 'fully', 'charge', 'detect', 'current', 'integrator', 'integrate', 'current', 'value', 'detect', 'current', 'sensor', 'period', 'instant', 'battery', 'fully', 'charge', 'instant', 'fully', 'charge', 'next', 'divider', 'divide', 'integrate', 'value', 'detect', 'current', 'value', 'length', 'period', 'corrector', 'correct', 'detect', 'current', 'use', 'quotient', 'provide', 'divider', 'offset', 'remain', 'battery', 'capacity', 'calculate', 'system', 'comprise', 'voltage', 'detecting', 'unit', 'current', 'detect', 'unit', 'index', 'calculate', 'unit', 'unit', 'calculate', 'unit', 'voltage', 'detecting', 'unit', 'detect', 'voltage', 'terminal', 'battery', 'current', 'detect', 'unit', 'detect', 'current', 'flow', 'battery', 'index', 'calculate', 'unit', 'calculate', 'index', 'polarization', 'battery', 'accord', 'detect', 'current', 'control', 'unit', 'control', 'index', 'polarization', 'remain', 'predetermine', 'range', 'permit', 'limitation', 'effect', 'polarization', 'charge', 'index', 'polarization', 'remain', 'predetermine', 'range', 'calculate', 'unit', 'calculate', 'remain', 'capacity', 'battery', 'accord', 'terminal', 'voltage', 'battery', 'open', 'circuit', 'voltage', 'battery', 'measurement', 'remain', 'capacity', 'calculation'

Tabelle 5.2: Das Patent nach Datenvorverabeitung

5.2.2 Anzahl der Themen

Wie bereits erwähnt ist die Wahl sinnvoller Parameter im Verlauf des Modellierungsprozesses essenziell für die spätere Interpretierbarkeit der Ergebnisse. Dabei stellte für uns die Wahl der Themenanzahl eine besondere Herausforderung dar. Eine große Anzahl an Themen wird dazu führen, dass die Granularität der Themen zunimmt. Allerdings kann dabei auch der Sinnzusammenhang der einzelnen Wortgruppen an Präzision verlieren. Einzelne Themen lassen sich, aufgrund repetitiver Wortsammlungen nicht mehr sinnvoll voneinander unterscheiden. Auf der anderen Seite führt die Wahl einer niedrigen Themenzahl möglicherweise zu Informationsverlusten.

Die richtige Wahl der Themenanzahl kann durch die Berechnung von Kohärenzwerten unterstützt werden. Dabei sollte man allerdings beachten, dass es sich dabei um kein absolutes Maße handelt. Ob das gewählte Themenmodell aussagekräftig ist, bleibt letztendlich der Erfahrung des Bearbeiters überlassen. Allgemein wird die Kohärenz eines Themas wie folgt definiert.

$$Coherence = \sum_{i<j} score(w_i, w_j) \tag{5.2}$$

Nach Gleichung 5.3 ist die Kohärenz die Summe der paarweisen „word scores“ für die Wörter $w_1, ...w_n$. Im Prinzip ist die Kohärenz ein Maß, für die Qualität des berechneten Themas, oder genauer: wie gut die einzelnen Wörter des Themas zueinander passen. Die Kohärenz des gesamten Themenmodells ergibt sich nach der durchschnittlichen Kohärenz aller Themen. Für die „scores“ gibt es verschiedene Metriken (UMASS, CV, CLI). Beispielsweise wird die CLI-Metrik wird nach Newman et al. (2010) wie folgt definiert.

$$score_{CLI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i), p(w_j)} \tag{5.3}$$

$p(w_i)$ ist die Wahrscheinlichkeit, für das Wort w_i in einem zufälligen Dokument gefunden zu werden. Analog gibt $p(w_i, w_j)$ die Wahrscheinlichkeit an, beide Wörter w_i, w_j in demselben, zufälligen Dokument zu finden. Die Wahrscheinlichkeiten werden aufgrund eines externen Datensatzes, basierend auf Wikipedia Einträgen, berechnet (Newman et al. 2010) .

Für die Berechnung der Kohärenzen unserer Themenmodelle verwenden wir den etwas komplexeren „cv-score“ basierend auf „normalized pointwise mutual information (NPMI)“ (**syed2017full**).

Dabei ergeben sich Kohärenzwerte zwischen 0 und 1. Wobei die beiden betrachteten Wörter eines Themas bei einem Wert von 1 identisch sind. In der Regel werden Werte zwischen 0.5 und 0.7 als „gut“ angesehen.

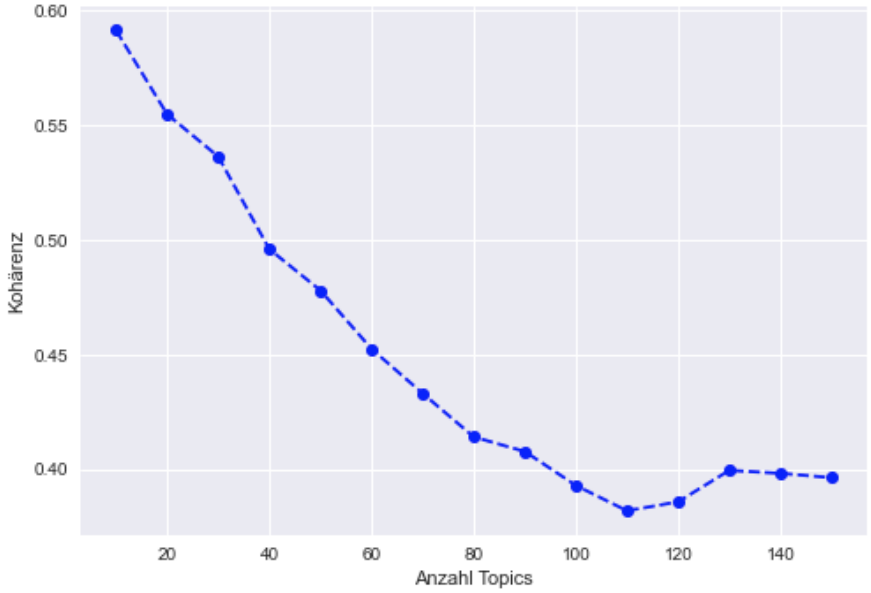

Abbildung 5.4: Kohärenzwerte pro Themenzahl

Der Graph in Abbildung 5.4 stellt die Kohärenzen in Relation zu den jeweiligen Anzahlen an Themen (Punkte im Graph). Mit steigender Themenzahl lässt sich ein klarer Abwärtstrend für die Kohärenzwerte erkennen, wobei sich das Maximum bei einer Anzahl von $K = 10$ Themen befindet. Wie oben aber bereits erwähnt dient die Kohärenz lediglich der Orientierung. Für die weitere Evaluationen haben wir uns zunächst auf drei Themenmodelle fokussiert ($K = 10$, $K = 30$ und $K = 80$). Sinn dieser Auswahl war ein Modell für jeweils eine Themengröße genauer zu betrachten. Durch die Anpassung einiger Parameter[2] gelang es uns die Kohärenzwerte der drei Themenmodelle etwas zu erhöhen (siehe Abbildung **??**).

5.2.3 Das Themenmodell mit $K = 30$ Themen

In diesem Abschnitt soll das Topic Model für eine Anzahl von 30 Themen durch die Zuordnung einiger, zufällig gewählter Patente, auf die Themen vorgestellt werden.

[2]Namentlich eine Erhöhung der „chunksize“ und der „passes“

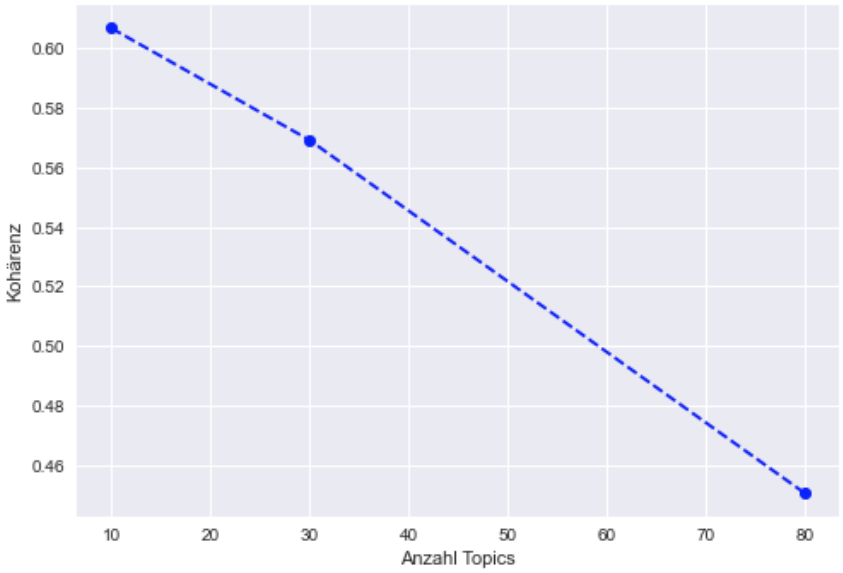

Abbildung 5.5: Kohärenzwerte pro Themenzahl, 10, 30 und 80

A Appendix

A.1 Berechnung der Distanzen nach Kitahara und Oikawa (2017)

```
1
2  # Berechnet die Distanz zweier Firmen nach Kitahara und Oikawa (2017)
3  def omega(z0,z1,eta=0.5):
4
5      '''
6      z0, z1 - Zitate von Firma_0 und Firma_1 als dictionary,
7      eta - technologischer Diskontierungsfaktor,
8      zitate2 - Zitate der zweiten Ebene als dictionary - gibt an welches Patent zitiert,
9      zitate2R - Rückrichtung der Zitate auf der zweiten Ebene als dictionary - gibt an von
           welchem Patente zitiert wird
10     '''
11
12     '''
13     In diesem Teil des Codes wird nur der Omegatyp und die Gewichtung der Patente bestimmt,
           Mehrfachzitate werden bei der Berechnung der Omegawerte
14     im zweiten Teil des Codes berücksichtigt
15     '''
16
17     # omega^1, Schnittmenge, erste Zitierebene
18     ov = set(z0) & set(z1)
19
20     # Übrige Patente (~P_ij und ~P_ji)
21     A0 = set(z0) - ov
22     A1 = set(z1) - ov
23
24     # Gewichtung der omega^1
25     g = dict()
26     for p in ov:
27         g[p] = (1,0)
28
29     # omega^21 und omega^22 in Richtung 01 (ij), Gewichtung
30     for p in A0:
31         gewicht = 0
32         typ = 0
33         if p in zitate2:
34             a = zitate2[p] - ov # Einschränkung der Menge auf noch nicht berechnete Zitate (~P
                   '_ij,)
35             schnitt = a & A1
36             if len(schnitt) > 0:    # omega^21-typ
```

```
                gewicht = len(schnitt)/len(a)
                typ = 1
            else:                   # evtl omega^22-typ
                for q in a:
                    if q in zitate2R:
                        qv = zitate2R[q]
                        schnitt = qv & A1
                        gewicht += len(schnitt)/len(a)
                        typ = 2

            if gewicht > 0:
                g[p] = (gewicht,typ)

    # omega^21 und omega^22 in Richtung 10 (ji), Gewichtung
    for p in A1:
        gewicht = 0
        typ = 0
        if p in zitate2:
            a = zitate2[p] - ov
            schnitt = a & A0
            if len(schnitt) > 0:    # omega^21-typ
                gewicht = len(schnitt)/len(a)
                typ = 3
            else:                   # evtl omega^22-typ
                for q in a:
                    if q in zitate2R:
                        qv = zitate2R[q]
                        schnitt = qv & A0
                        gewicht += len(schnitt)/len(a)
                        typ = 4

            if gewicht > 0:
                g[p] = (gewicht,typ)


    '''
    In zweiten Teil des Codes berechnen wir die verschiedenen Omegawerte anhand ihrer
        Gewichtung
    '''


    om01_1 = sum([g[k][0]*v for (k,v) in z0.items() if k in g and g[k][1] == 0])
    om01_1 += sum([g[k][0]*v for (k,v) in z1.items() if k in g and g[k][1] == 0])
    om01_21 = sum([g[k][0]*v for (k,v) in z0.items() if k in g and g[k][1] == 1])
    om01_22 = sum([g[k][0]*v for (k,v) in z0.items() if k in g and g[k][1] == 2])
    om10_21 = sum([g[k][0]*v for (k,v) in z1.items() if k in g and g[k][1] == 3])
    om10_22 = sum([g[k][0]*v for (k,v) in z1.items() if k in g and g[k][1] == 4])

    # Berechnung des finalen Omegawertes
    om = om01_1 + eta * (om01_21 + om10_21) + eta * eta * (om01_22 + om10_22)
    nenner = sum(x for x in z0.values()) + sum(x for x in z1.values())

    # Berechnung der Distanz d_ij
    return  -np.log(om/nenner)
```

A.2 Stopwords

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', ït's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', ßhould've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', ären't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', ïsn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', ßhan't", 'shouldn', ßhouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", 'wherein', 'furthermore', 'solution', 'problem', 'solve', 'method']

A.3 spaCy

```
def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']):
    """https://spacy.io/api/annotation"""
    texts_out = []
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append([token.lemma_ for token in doc if token.pos_ in allowed_postags])
    return texts_out

def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc)) if word not in stop_words] for doc
        in texts]
```

Literaturverzeichnis

Bar, T.; A. Leiponen (2012): A measure of technological distance. In: *Economics Letters* 116(3), S. 457–459.

Benner, M.; J. Waldfogel (2008): Close to you? Bias and precision in patent-based measures of technological proximity. In: *Research Policy* 37(9), S. 1556–1567.

Blei, D. M.; A. Y. Ng; M. I. Jordan (2003): Latent dirichlet allocation. In: *Journal of machine Learning research* 3(Jan), S. 993–1022.

Engelsman, E. C.; A. F. van Raan (1994): A patent-based cartography of technology. In: *Research policy* 23(1), S. 1–26.

Jaffe, A. B. (1986): *Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value*. Techn. Ber. national bureau of economic research.

Jaffe, A. B. (1989): Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. In: *Research Policy* 18(2), S. 87–97.

Jaffe, A. B.; M. Trajtenberg; R. Henderson (1993): Geographic localization of knowledge spillovers as evidenced by patent citations. In: *the Quarterly journal of Economics* 108(3), S. 577–598.

Karki, M. (1997): Patent citation analysis: A policy analysis tool. In: *World Patent Information* 19(4), S. 269–272.

Kitahara, M.; K. Oikawa (2017): Technology Polarization. In: *Tokyo Center for Economic Research (TCER) Paper*(E113).

Kruskal, J. B. (1978): *Multidimensional scaling*. 11. Sage.

Newman, D.; J. H. Lau; K. Grieser; T. Baldwin (2010): Automatic evaluation of topic coherence. In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, S. 100–108.

Stuart, T. E.; J. M. Podolny (1996): Local search and the evolution of technological capabilities. In: *Strategic management journal* 17(S1), S. 21–38.