CAPPA – Career Paths of Patent Attorney

Funded by the EPO Academic Research Programme 2017

Final Report on Progress 30th of June 2019

Jun.-Prof. Dr. Lutz Maicher

Kazimir Menzel

lutz.maicher@uni-jena.de

kazimir.menzel@uni-jena.de

Technology Transfer Research Group
Institute for Informatics
Friedrich-Schiller-Schiller University Jena
Ernst-Abbe-Platz 2
07743 Jena
Germany

Table of Content

1 Introduction

The goal of the project "CAPPA – Career Paths of Patent Attorneys" is to provide a comprehensive, yet anonymised dataset that contains the career paths of (all) patent attorneys that acted as representatives before the European Patent Office (EPO) since 2008. This dataset is intended to assist researchers, who are interested in patent attorneys and their career paths for their own research. The CAPPA dataset has been published has as open data by end of June 2019.

The main raw data used to compile the dataset are EPO patent applications, which is regularly published by the EPO as open data. This data is, however, not completely clean and thus imposes a serious barrier to direct research. The CAPPA dataset is intended to shrink this barrier and make this data more accessible to researchers working on patent attorneys. To reach this goal, the work plan of "CAPPA" has been split into six work packages to be done over 18 months. The first three, "WP-01 Setting the Stage," "WP-02 CAPPA," and "WP-03 Gap Identification" have already been (partially) presented in the report from 14th of June 2018.

This final report summarises the entirety of the project and presents some results in the following structure. Section 2 gives an overview of the preliminary research that informed the remainder of the project. In section 3, we describe the organisational challenges faced during the project period. The technical challenges and underpinnings of the project, in particular the data gathering and harmonisation are summarised in section 4. The main output, the CAPPA dataset, is described in detail in section 5, while section 6 contains some of the results and visualisations that correspond to work packages "WP-05 Analysis" and "WP-06 Career Patterns." The Appendix contains additional material and details the output of the project to the specific work packages in the project plan. The CAPPA dataset[1] has been published on zenodo.org and is accompanied by a visualisation tutorial[2] that has been published separately on zenodo.org as well.

[1] http://doi.org/10.5281/zenodo.3265385

[2] http://doi.org/10.5281/zenodo.3267424

2 Introductory Work

2.1 Literature Review

While the literature on patent attorney career paths has been scarce at best, the literature on IP service providers and law firms is richer. Süzeroglu-Melchiors [1] discusses and investigates with considerable data the conditions under which what kind of technology company is outsourcing patent-related services to which kind of service provider. In [2], Janicke and Ren discuss partly the influence of the career path of an attorney on success in patent infringement trials. Macdonald and Lefang [3] addresses centrally the role of the patent attorney in the innovation process. Somaya, Williamson and Zhang [4] investigate the role and possible paths of patent attorneys in technology companies. While focusing mostly on the influence of patent attorneys on the inventive productivity of the company, they also discuss where and how these firms can find these employees.

The most important challenge has been the scarcity of literature on the topics that are central to the initial phase CAPPA itself. Very little recent research has been found that is specifically focused on the career paths of patent attorneys.

2.2 Conceptualising “Career Path”

The notion of a career path is not trivial. In general, it might encompass a lot of changes in position and responsibilities within as well across companies. In order to ground the working definitions sufficiently well in reality, we conducted structured interviews with practising patent attorneys that were randomly selected from the pool of all patent attorneys registered with the EPO.

The interviews, which took between 30 and 90 minutes, followed the structure that is given in Appendix 8.1. The attorneys were asked about the development of their career paths as well as motivations for their initial career choice to become a patent attorney and for changing employers. The interviews were conducted by telephone, the questions having been sent to the attorneys prior, under the promise of highest confidentiality to the interviewees to allow them to answer the questions as freely as possible.

Based on these preliminary interviews, it became apparent that we had to stick to a rather high-level notion of a career path. Within our work we define a career path as the sequence of employment relationships that a patent attorney undergoes throughout the observation period, including self-employment.

As the EPO register only contains patent attorneys that have passed the European Qualifying Examination (EQE), we do not track the career paths prior, albeit the qualitative interviews have indicated that this might be an interesting field of research. The interviews also indicated that it might be sensible to include the development of the responsibilities within one law firm or technology company in the definition, but it turned out to be not feasible to obtain the necessary data at a useful scale.

3 Organisational Challenge

The project was faced with one unexpected major organisational challenge. In the original working plan, we expected to outsource the following tasks to Fraunhofer IMW, the originator and maintainer of the IP Industry Base:

- WP 2 EPAC (PM 3 – PM 6)
    - Extending the data collection range in the IP Industry Base from 5 years to 10 years, including necessary data harmonisation activities (due PM 4, outsourced to FhG IMW)
    - Integration of the continuous generation of the CAPPA into the IP Industry Base (due PM 6, outsourced to FhG IMW)
- WP 3 Gap Identification (PM 5 – PM 10)
    - Applying the Continuous Data Quality Assessment facility to the EPAC based on the Data Quality Requirements identified in the previous data quality checks (outsourced to Fraunhofer IMW, due PM 10)
- WP 4 Gap-Filled EPAC (PM 10 – PM 15)
    - Provision of a tool that allows patent attorneys to personally verify their career profiles in the EPAC and claim corrections (through the IPIB interface) (outsourced to Fraunhofer IMW, due PM 15)
    - Augmentation of the facility for the generation of the EPAC (see WP2) by means for the continuous generation of the gap-filled EPAC (due PM 15, outsourced to FhG IMW)

The original budget for this subcontracting was 20k EUR. Due to a strategic change and a consequential significant lack of personnel resources with necessary technological skills, the Fraunhofer IMW was not able to realise the originally planned subcontracting. Fortunately, the Fraunhofer IMW was very cooperative in this situation and gave us the necessary access to the data and the software code, to conduct the necessary implementations ourselves.

Thus, we did not spend the intended 20k EUR budget for the subcontracting. On the other side, we had to reallocate a working load of approx. 6 personnel months (with 60% FTE)[3] within the working plan. In the Appendix 9.2 you can see, which tasks we reduced or cancelled from the original working plan, in order to fulfil with these reduced resources the overall goal of the project: publishing the CAPPA dataset as open data and in highest possible quality.

4 Technical Challenges

4.1 Data Sources

The data sources of CAPPA consist of two separate openly and freely accessible data sources. For data from now on, the main data source is the EPO registry of patent attorneys, tracking their affiliation to employers or self-owned law firms on a weekly basis. As this registry does not cover past affiliations, it is complemented by the information that is extracted from the patent applications that are accessible through the APIs provided by the EPO (by using the Ruby gem *epo-ops* [5] which was published as open source as well).

In the latter data source, information on the patents includes the patent law firms involved in the application process as well as the attorneys that have been involved with the patent (by a rate of roughly 70% the person is given). This information is continuously imported by the IP Industry Base (IPIB, https://s.fhg.de/IPIB) tool chain. The IPIB extracts and semi-automatically harmonises among various other information the EPO application number, the EPO application date, the associated representatives and the employers of these representatives, if available.

It should be noted that the data acquired through this procedure lacks consistency to a considerable degree. It is therefore necessary to harmonise the employing companies into *Groupings*, which aim to track companies through mergers, acquisitions, name changes, parent-child relationships within corporations or even misspellings. Patent attorneys themselves require also a considerable degree of harmonisation, as names are sometimes recorded or extracted differently as well as the possibility of two attorneys sharing the same name complicates the matter. The harmonisation has been conducted through a process that combines automated elements with human curation.

[3] Kazimir Menzel had a contract for a 60% FTE. Each personnel moths of this contract costs approx. 3,2 TEUR. In consequence, a budget of 20 TEUR is an equivalent to approx. 6 personnel months with a 60% FTE contract.

The *Input Dataset* is computed based on the EPO patent applications that have been filed since 2008-01-01 until 2018-12-31. For each application, the dataset uses the *Patent Application ID*, the *Year-Month* from the EPO application date, the *Grouping ID* and the *Attorney ID*. Each entry in the dataset corresponds to unique actual configurations of these four variables. The pre-processed dataset tracks 8,710 patent attorneys, who have been affiliated with 2,930 employers (*Groupings*) on more than 1,070,000 patent applications.

4.2 Data Processing

4.2.1 Baseline Data

The *Input Dataset* is subsequently used to detect the affiliations between patent attorneys and employers over time. This is done by assigning for each triple of (*attorney*, *employer*, *patent application*) a *year-month date*, estimated through a time series model from the *priority date*, the *EPO Publication A1 date* and the *effective date*. The structure is depicted in Table 1 below.

At this stage, we encounter the first problem, that patent application records often contain all involved attorneys and all involved law firms or patent arms of technology companies respectively without any indication of relationships between the two. It is therefore necessary, to disentangle this information through a second step, where we compute for each pair (*attorney*, *year-month*) a set of potential *employers*. Enhanced by other metrics, such as the *number of observed patent applications* for each *attorney-employer* pair and the *duration* of the affiliation of both, the resulting data set is then used as the baseline model to further computation.

attorney_id	employer_id	first_observation	last_observation	duration	n_pat
1	356	2008-07	2009-10	16	13
1	2031	2009-03	2014-02	60	27
...	...	...	...	...	...

Table 1: Extract from the Baseline Data Frame

As has already been indicated, it is necessary to solve three different problems, when improving the baseline data. First, it is necessary to find a model that correctly identifies valid relationships between *employers* and *attorneys*, second, it is necessary to embed these models into a time sequence, in order to maintain the validity of relationships over time, and third to find ways to correctly interpolate missing time periods, which is a result of solving the former two problems.

4.2.2 Gap Identification

Finding a remedy to the first issue involved considerable work in identifying and fine-tuning useful models. It has been solved by constructing a series of graphs that link attorneys and employers through weighted edges, whose weights are given by the number of patent applications within a given time period. This results in three different graphs for each time period, an *attorney-employer* graph, an *attorney-attorney* graph and an *employer-employer* graph. The time periods are rolling windows of six months and rolling windows of one year, in order to capture as much context as possible for any given month during the observation period.

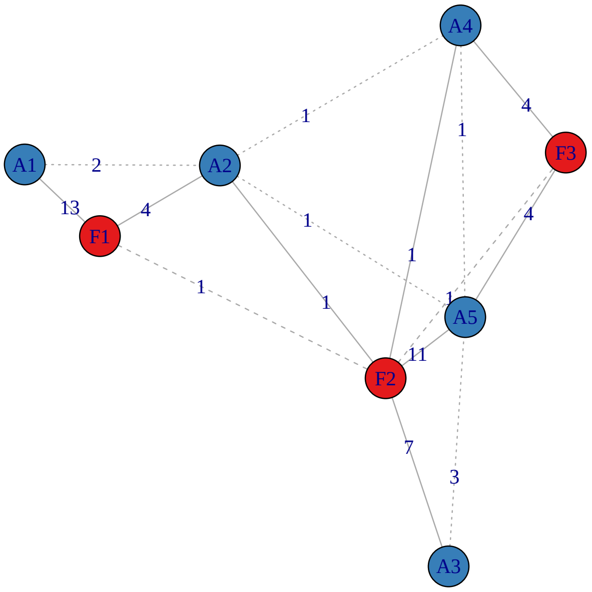


Figure 1: Combined example of the resulting model, where solid lines represent the attorney-employer *graph, dotted lines the* attorney-attorney *graph and dashed lines the* employer-employer *graph*

For each period, the weights of the *attorney-employer* graph are used to compute the probability that an attorney has been employed by a given employer and the weights of the other two graphs used to distinguish between incidental and valid relationships. Incidental relationships come up, if e.g. attorney A working for law firm X and attorney B working for a technology company Y are both involved in the application for patents 30671 and 48078 during a given time period. In this case, we detect four relationships, one between attorney A and law firm X, one between attorney A and technology company Y and the same for attorney B. Of these four relationships, only two, (A, X) and (B, Y) are valid. To determine the validity of a relationship, we analyse first the graph containing the links between employers and then the graph containing the links between attorneys. Depending on the frequency of both pairs showing up together measured against how they show up in different constellation, we estimate a pair specific parameter that makes the distinction.

To find the correct parameters, a set of 520 randomly selected patent attorneys was chosen as a test set. For the attorneys in the test set, we collected valid employment data from LinkedIn, which went through the same harmonisation process as the data from the EPO patent applications and in the same form as the data model of the baseline model. The parameters of the graph models were then trained against the test set.

The second problem was solved by using two separate resolutions, one year and six months, and then repeating the process described above for each window and each resolution. This way, it was possible to detect valid overlapping relationships and resolved further ambiguous relationships.

As a result of both strategies combined, only 63 cases of ambiguous employment relationships remained, which were sorted out by manual verification. Cases that could not be resolved, were dropped from the dataset.

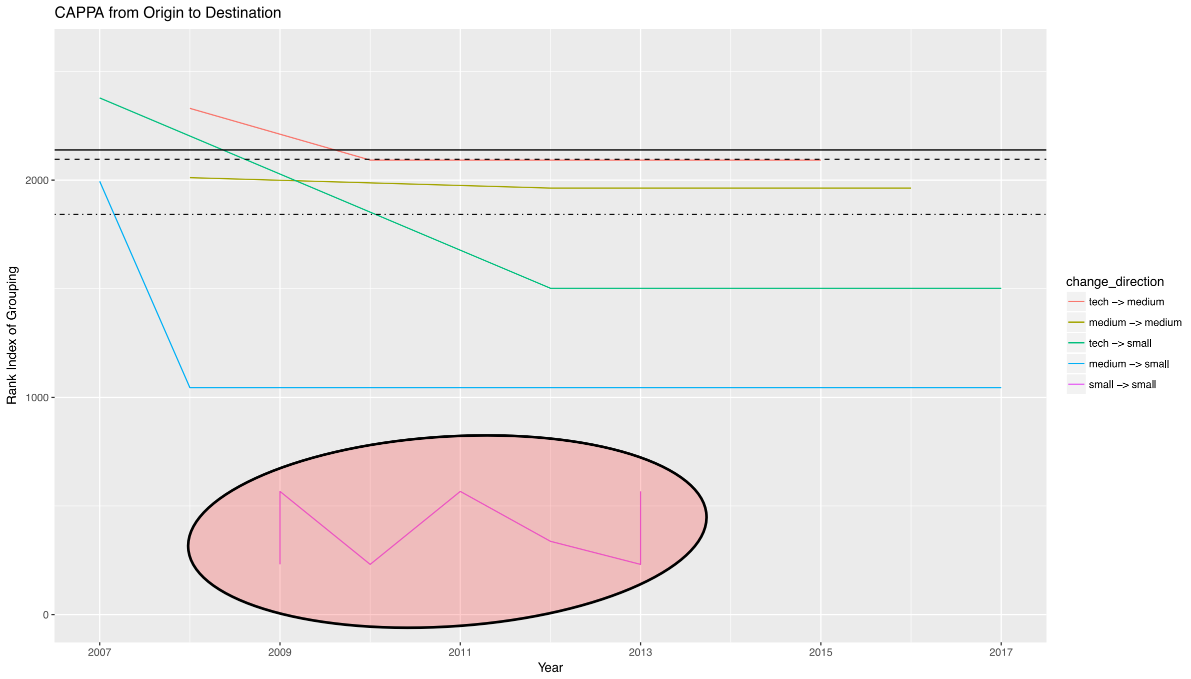


Figure 2: Graphic showing selected career paths from the baseline model, with a degenerate pattern in purple.

The third problem involved the interpolation of missing relationships. A very common case is an artefact of the necessity to aggregate of larger periods to build useful graph models and then having the sliding series to correct for the resulting lack of resolution. These cases were resolved building a model that determined an attorney-specific threshold parameter when a change of employment would be considered valid within a six months window after the first change has been detected. Missing months were then interpolated between the first valid observation of an attorney and the last valid observation of an attorney, even if this results several cases of dual employment over a six months period in the final data set. The upside is, that these cases would have been lost otherwise due to ambiguity or resulted in oscillating patterns that contain no information as shown in figure 2.

5 CAPPA Data Structure

In the previous section on the computation of the data, the four central data frames of the final CAPPA dataset have been briefly mentioned. While then the focus was more on the processing aspects, the following section describes the output structures of the final CAPPA dataset in more detail. All these data structures can be queried in conjunction or separately, depending on the interest. The finalised CAPPA dataset was published as open data and open access via zenodo.org on the 28th June 2019 und the DOI 10.5281/zenodo.3265385[4]. Furthermore, a set of visualisations has been published with commentary and example Code in a Notebook on ZENODO with the DOI 10.5281/zenodo.3267424[5].

5.1 Attorney Data Frame

5.1.1 Description

The *Attorney Data Frame* contains unique and cumulative data on the observed patent attorneys. The data frame contains the *Attorney ID* for (anonymised) identification. *Number of Employments* holds the observed employments. The variable *Number of Patent Applications* contains the count of patent applications with which the attorney has been affiliated over the observation period. *Months Active* contains the time difference between the first and the last observation. The structure of this data frame is depicted in table 2.

[4] http://doi.org/10.5281/zenodo.3265385
[5] http://doi.org/10.5281/zenodo.3267424

attorney_id	n_emp	n_pat	first_observation	last_observation	months_active
1	1	5	2009-03	2009-11	5
2	2	166	2008-04	2018-11	67
...	...	...	...	...	...

Table 2: Extract from the Attorney Data Frame

5.1.2 Representation

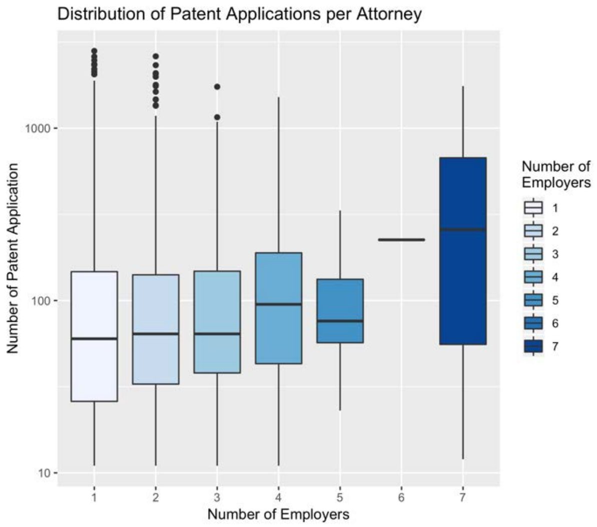


Figure 3: Patent applications per attorney grouped by the number of employers.

The *Attorney Data Frame* can be used in two different representations. The most important direct interaction with the data structure is its list form. This representation is best suited to investigate the attorneys directly. Figure 3 illustrates one such use case. Note that the y-axis is scaled in log$_2$-scale. A unit difference along the y-axis corresponds to a multiplication by the factor 2 if moving upwards.

A different take on the data is the attribute representation. In this form, the data structure is intended to facilitate filtering and visualising the *Career Path* and *Change Data Frames*. This is particularly helpful, when targeting network-related graph aspects such as subgraphs, trees or paths.

5.2 Employer Data Frame

5.2.1 Description

The second data structure is the *Employer Data Frame*. Like the *Attorney Data Frame,* it contains the employer specific individual data. Note, that employer refers to the concept of *groupings* discussed above. The directly computed variables include the *Employer ID*, *Number of Patent Applications* and the *Number of Attorneys*, which are both aggregated over the full observation period. The structure of this data frame is depicted in figure 3.

The variable *Mean Retainment* holds the average duration that the attorneys stayed with the employer, *Median Absolute Deviation of Retainment* corresponds to the median absolute deviation of the attorneys' stay. The variable *Median Patent*

Applications per Attorney is calculated over the total period the attorneys' stay with the employer.

employer_id	rank	type	n_att	mean_ret	mad_ret	n_pat	med_ppa
1	878	small	1	125	0	22	22
2	769	small	1	100	0	15	15
...	...	...	...	...	...	...	...

Table 3: Extract from the Employer Data Frame

Type assigns the employers to one of four categories, *technology companies*, which refers to in-house patent divisions of (usually larger) companies, *large, medium* and *small* firms. The first group is computed during the harmonisation process, which separates between IP service providers that were only active for companies within their own grouping from those that were active for more than one grouping. The last three groups are based on a *Zeta*-distribution over the *Number of Attorneys*, which corresponds roughly to slicing the interval between the logarithm of the smallest *Grouping* and the logarithm of the largest *Grouping* into three even intervals and then assign the types according to the interval into which each individual firm falls. *Rank* is a mixed ordering that orders the firms according to *Type*, with technology companies being the highest and small the lowest rank. Within these groups, the firms are ordered by *Number of Attorneys*, ties are broken through *Number of Patent Applications*. Remaining ties are subsequently broken by *Patent Applications per Attorney*.

5.2.2 Representation

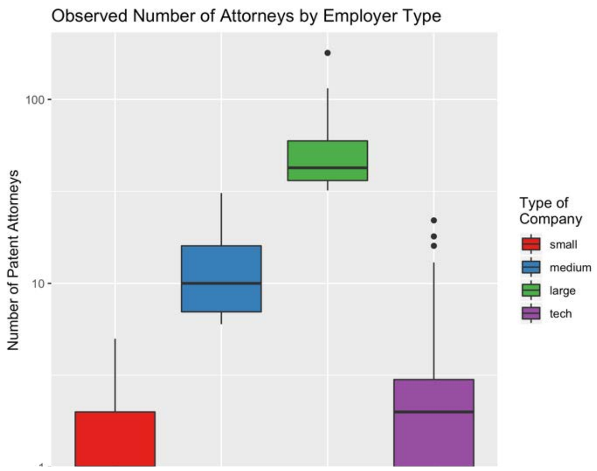


Figure 4: The types of employers by the number of attorneys

The *Employer Data Frame* is mostly amenable to three representations. In rectangular form, it is intended to investigate the attributes of the firms in aggregation. Figure 4 shows one use case, where the type is plotted against the distribution of total filed patent applications over the observation period. As in the section before, the y-axis unit distances correspond to multiplication by 2.

The second representation of this data structure is as a generator for scales or subordinated scales for ordinal variables. This is used in the subsequent plots of career paths and changes. It allows to impose a meaningful ordering on the employers so that it is possible to infer visual information from plots. The scale representation in the *Employer Data Frame* is incorporated in the *Rank* variable, which has proven helpful in highlighting and exploring the structure of the service provider market.

A third representation is as attribute to the vertices of the *Change Data Frame* when used in graph representation. This is currently used only in the computation of the variable *Change Category* in the *Change Data Frame*.

5.2.3 Data Issues

Depending on the organisation of the patent attorney profession across countries, it is difficult to identify “shadow groupings.” One case in point is the *Association of Intellectual Property Law Firms in Sweden* (SEPAF), another are IP-related, organised groups between universities or other research institutes. If these are known, then they are excluded, albeit this often involves considerable human curation.

5.3 Career Path Data Frame

5.3.1 Description

The third data structure is the *Career Path Data Frame*. It contains the variables *Attorney ID* pulled from the *Attorney Data Frame* as identifier of the path, the variable *Employer ID*, which is pulled from the *Employer Data Frame* and serves as identifier of the stations of employment in this path. In addition, it contains the variables *First Observation* and *Last Observation* and the *Number of Patent Applications* with the employer over the given period. The structure of this data frame is depicted in table 4.

attorney_id	employer_id	first_observation	last_observation	n_pat
1	1820	2008-10	2009-11	3
2	574	2008-01	2012-07	7
2	2853	2012-01	2018-07	22
...	...	...	...	...

Table 4: Extract from the Path Data Frame

5.3.2 Representation

The paths of all graphs lend themselves to an interpretation of a time series with monthly granularity that tracks the *Changes* in time and shows for each attorney a separate path through the space of *Employers* in time. That way, it can be used to explore patterns of path behaviour. Figure 5 shows the career paths of selected patent attorneys. Each line corresponds to an attorney. The separating solid horizontal line separates technology companies from patent law firms and the dotted lines separate the small, medium and large patent law firms in ascending order upwards.

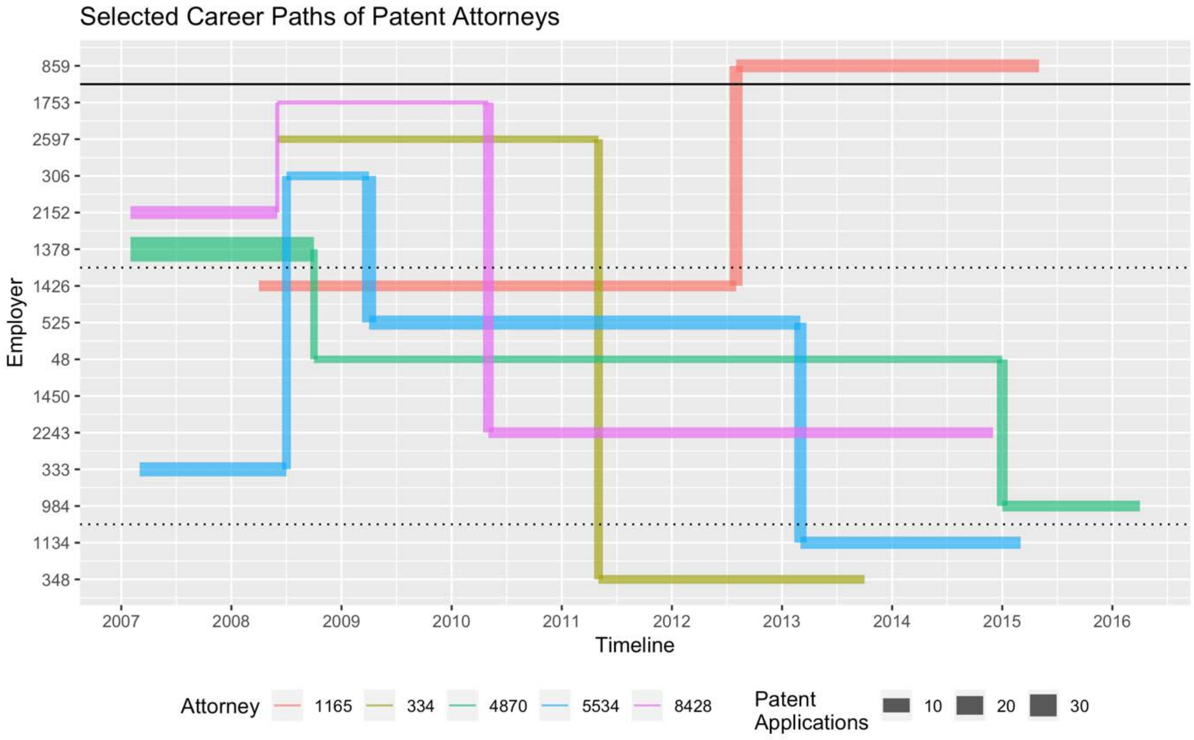


Figure 5: Selected career paths of patent attorneys. The solid line separates large law firms from technology companies, *the dotted lines separate* small, medium *and* large *law firms in ascending order. The thickness of the path segment reflects the number of patent applications for the respective attorney during this employment station.*

5.4 Change Data Frame

5.4.1 Description

The last data structure is the *Change Data Frame*. It holds the changes that have been extracted by the series of graph models. Each change is associated to an *Attorney ID*, the *Time of Change* it occurred, the *Original Employer* and the *Target Employer*. *Change Category* labels the changes according to *Type* from the *Firm*

Data Frame. Each change is put in one of 16 categories of the form *“employer (source firm) —> employer (target firm)”*. The *Change Data Frame* currently contains 4707 changes. The structure of this data frame is depicted in table 5.

attorney_id	time_of_chg	origin	origin_type	target	target_type	change_type
7343	2011-05	1576	Tech	458	small	tech -> small
4482	2008-11	1695	Medium	559	medium	medium -> medium
...	...	...	...	...	...	...

Table 5: Extract from the Change Data Frame

5.4.2 Representation

The fundamental model of the *Change Data Frame* is that of a sequence of directed multi-edged graphs, where each edge represents an *attorney changing* and the vertices are given by the firms. However, we found that this kind of visualisation is very difficult to interpret and not very insightful to work with.

The second representation is a sequence of *change matrices*, where each point corresponds to an *attorney changing* and *source* and *target* represent the corresponding pairs of coordinates in the space of firms. This way we visualise changes in clusters of changes and investigate change free zones or areas where the changes are more or rather less frequent than to be expected. This last interpretation has been proven to be particularly helpful, when exploring changes in conjunction with the other two data structures as well as when verifying the plausibility of patterns.

Figure 6 compares employment changes between two periods over two and a half years. The top right corner, marked by the dashed lines contains employment changes between technology companies, the rectangular regions directly below contain changes from technology companies to law firms and the rectangular region on the top directly to the left the changes from law firms to technology companies. The position of the companies along the x- and the y-axes are derived from the scale representation of the *Employer Data Frame*. Lower positions imply smaller companies in general. It is also possible to read the change type from the colour map.

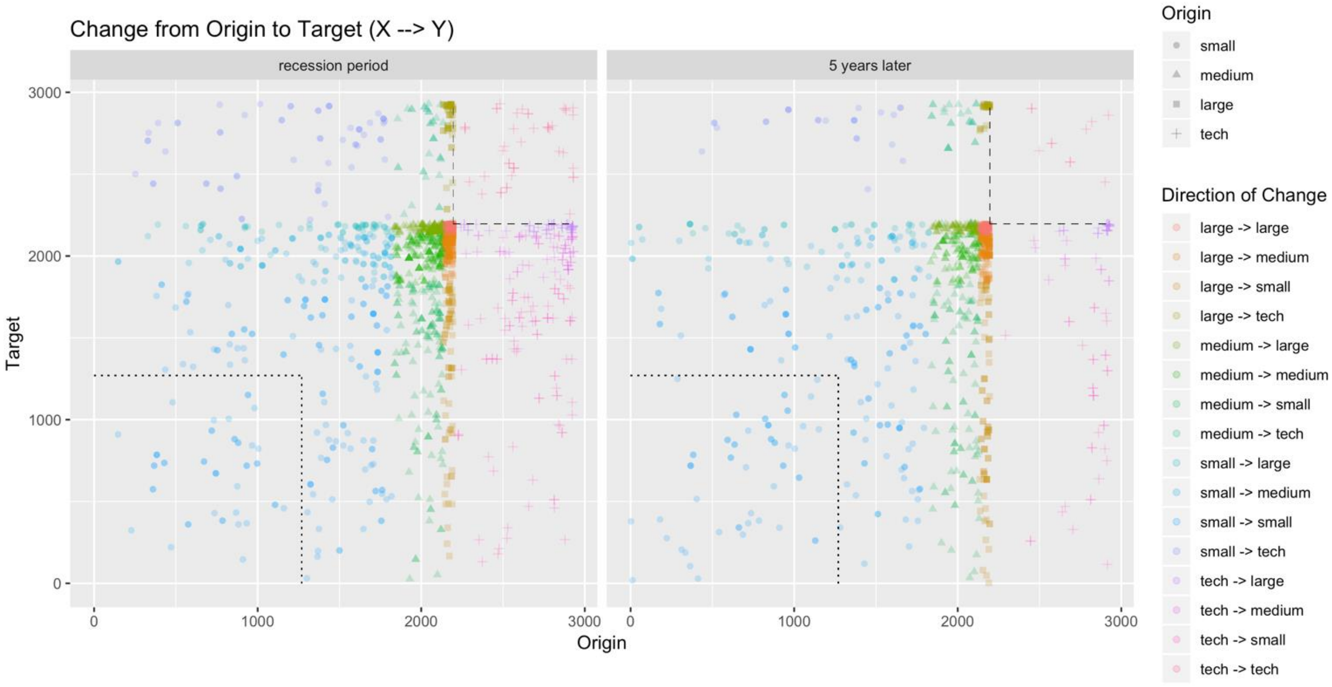


Figure 6: Comparison of employment changes by patent attorneys from July 2008 to December 2011 on the left side (recession period) and the with the period from July 2013 to December 2016 (5 years later).

6 Descriptive Statistics and Limitations

6.1 Overview

The dataset as a whole comprises four sub-datasets, each containing different aspects. Table 6 summarises the datasets in terms of size. As can be read off, CAPPA contains data 8,710 patent attorneys, 2,930 employers of patent attorneys (including patent law firms or single patent attorney offices), 10,786 career stations and 4,707 employer changes over the course of 11 years or 132 months.

	Attorney Data Frame	Employer Data Frame	Career Path Data Frame	Change Data Frame
Observations	8,710	2,930	10,786	4,707
Features	6	8	5	7
Entries	52,260	23,440	53,930	32,949
Time Range	2008-01 – 2018-12		2008-01 – 2018-12	2008-01 – 2018-12

Table 6: Summary of the CAPPA dataset

6.2 Limitations

Due to the scope of the data retrieval, the dataset suffers several important limitations. Anonymity is required by the entry into force of the new GDPR. It does not allow to make explicit, for which firms an attorney worked at a given time, while at the same time keeping the dataset open and widely accessible, as it would make the individual attorneys easy to identify. The authors recognise that this might cause difficulties for researchers trying to combine CAPPA with other datasets.

The focus on patent applications constrains the dataset to only this part of a patent attorney's work. This implies in turn that some essential variables to reconstruct a more holistic depiction of individual career paths. A particularly interesting second data source would have been data on oppositions. Leaving out these data means that career stations for attorneys with very few patent applications are not always exact or complete. The authors decided to retain them in the dataset, though, to provide as complete data as possible.

Not a prevalent but sufficiently common phenomenon is parallel employment of patent attorneys. While the reasons for this are manifold as has been determined in interviews with selected attorneys. This makes it often difficult to decide whether any attorney is just caught in a transition phase or working in parallel. In these cases, two change dates have been recorded and in the accompanying data paper, we recommended to treat the first date as authoritative.

EPO patent applications being the most import data source, in turned out to be impossible to retrieve and extract information on an attorney or an affiliated employer for a sizable subset of these data. The lack of these information led to the patent applications being excluded from the dataset computation process. This lack amounts to approximately 2100 registered patent attorneys that are included in the current version of CAPPA. The same lack of information retrieval led to reduction of patents that could be assigned to the included patent attorneys and employers., which means that not all patent applications from the covered period could be included in the computation of the dataset.

A general feature of the source data is conducive to a very common and fundamental in retrieval. Albeit the EPO patent applications as they are accessible contain some structure, this structure is by no means complete. In particular, many data elements contain unstructured strings and are not entirely consistent across applications. This results in inconsistent naming for patent attorneys, patent law firms and technology companies among others. This exacerbated by the fact that names of individuals and even more important firms and companies tend to change over time, in latter case very frequently. In case of companies and law firms, this is further compounded by frequent mergers and acquisitions, while old names are retained. Together with the unstructured way in which the data is recorded, this complicates the process to identify the main entities of the here concerned research correctly. Consequently, some of the assignments are not necessarily correct for every point in time.

7 Findings

7.1 General Change Patterns

Preliminary exploratory analyses indicated several interesting tendencies that are summarised in Figure 7. The overall tendency of employment changes occurs between similar categories of law firms and that changes to small law firms dominate otherwise. Changes between *technology companies* are an exception to the first pattern. This is to be expected as patent attorneys are deeply involved in the process of managing innovation. It is therefore expected that technology companies will try to retain these employees or at least try to avoid losing them to direct competitors.

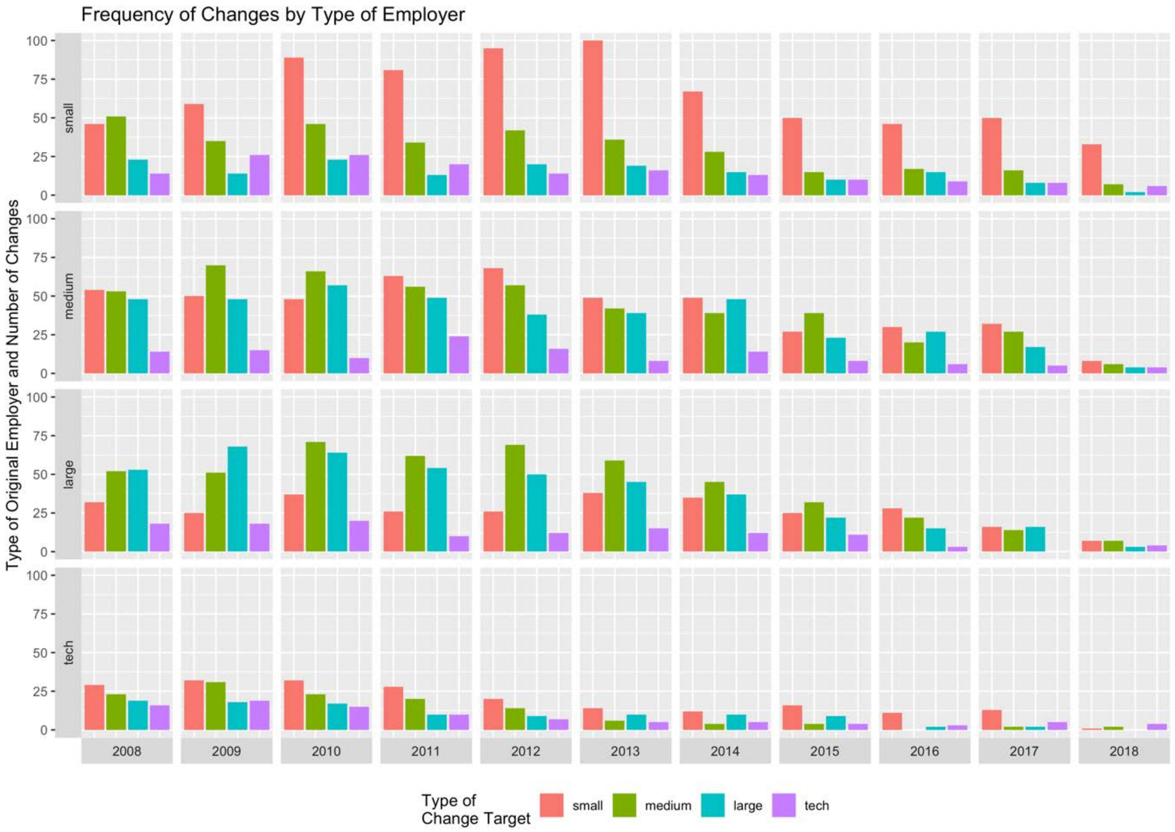


Figure 7: Frequency of changes from 2008 to 2018 separated between original employers in the rows and target employers by colour.

the other changes, the figure below indicates that a slight wave effect through the years and the classes from *large* to *small* might be found. Qualitative research conducted through the interviews suggests that this is often a consequence of mergers. The restructuring among the client base of patent law firms in the aftermath of the financial crisis appears to have induced a small wave of mergers among patent law firms as well. It appears that this process begun among larger patent law firms and was then reciprocated among the smaller law firm with some delay.

7.2 Specific Patterns

Inspecting the change matrices in Figure 8, we see some of our suspicions confirmed. In general, it can be noted that most changes occur between medium and large firms and from larger to smaller firms. To see this, note the relatively populated area in the upper right quarter of the yearly matrices.

This plot and Figure 6 as well highlight an interesting phenomenon. The bottom left square encompasses an area, where we would have not expected any changes as these are law firms for which we have only one patent attorney observed. We have investigated all of these cases manually and it appears that these either involve law firms with a certain specialisation, where one attorney is responsible for filing the patent application even though other patent attorneys also work for the law firm or the interesting case that a single patent attorney is working for two separate law firms. A typical case appears to be a patent attorney working for a law firm that employs also regular lawyers, while also taking cases on the side through her private law firm.

The third aspect is that apparently, many of the changes from service provider firms to tech companies occur from smaller and medium companies. This can be inferred from the upper left stripe on the matrices. This might be due to the relatively close or individual services a provider gives to the tech companies, which might result in being taken in.

7.3 Path Patterns

In order to visualise path pattern, we can generally take two perspectives. The upper part of Figure 9 groups each attorney path from the perspective of the type of her last employer during the observation period, while the lower part of Figure 9 groups the attorney paths from the perspective of the type of their first employer during the observation period.

We notice two patterns. The first pattern appears to be a trickle-down effect. Career paths tend to “descend” from larger to smaller employers. The second pattern is that changes between different types of employers become much less frequent after 2012. This is not only an effect of the reduced amount of changes after 2012 that is also visible in Figures 7 and 8, but a genuine trend in itself. In general, we notice that there is a tendency of patent attorneys to become partners in smaller law firms, be it boutique or single attorney law firms.

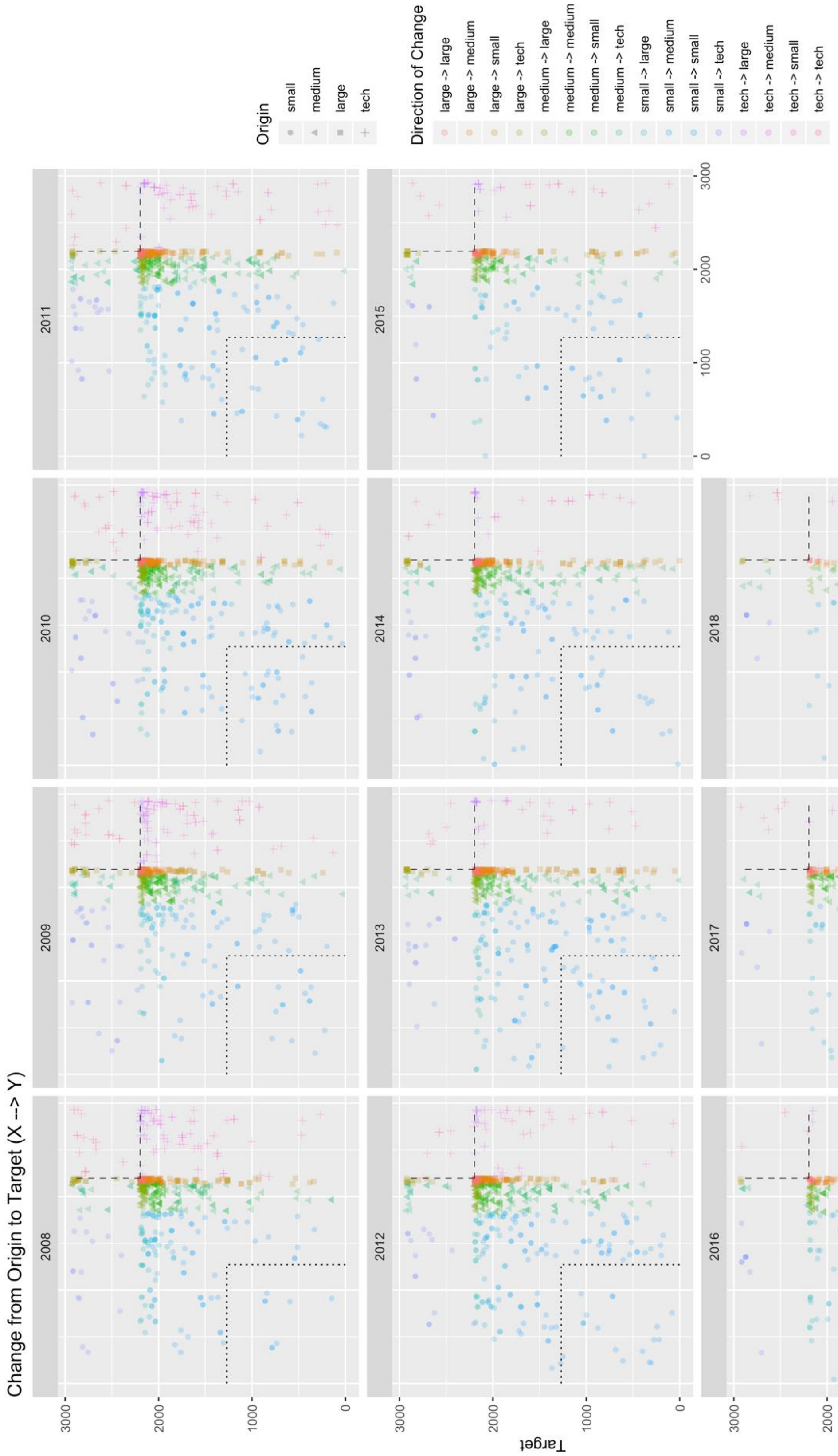


Figure 8: Change matrices on yearly resolution.

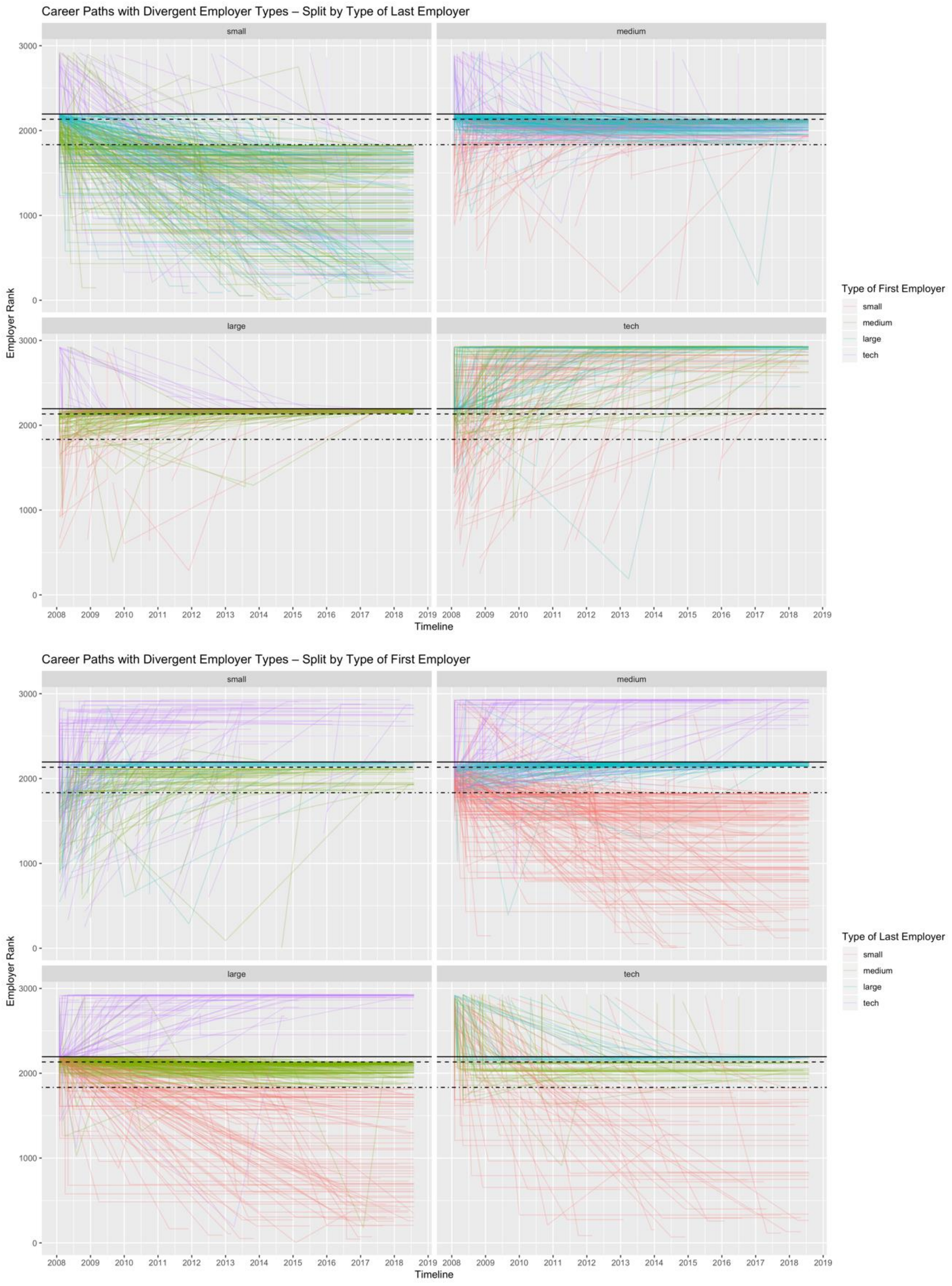


Figure 9: Path plots showing the career paths of all attorneys whose employers' type at the beginning of the observation period differed from type of their last employer. Each line path refers to one attorney, in the upper part, each subplot refers to the last employer’s type, the colour to the first employer’s type, in the lower part the relation is inversed.

8 References

[1] Süzeroglu-Melchiors (2017): “The supply side of IP management: Understanding firms choices’ regarding IP intermediaries,” World Patent Information (50), 55-63.

[2] Janicke and Ren (2006): “Who wins patent infringement cases, AIPLA Quarterly Journal (34), 1-43.

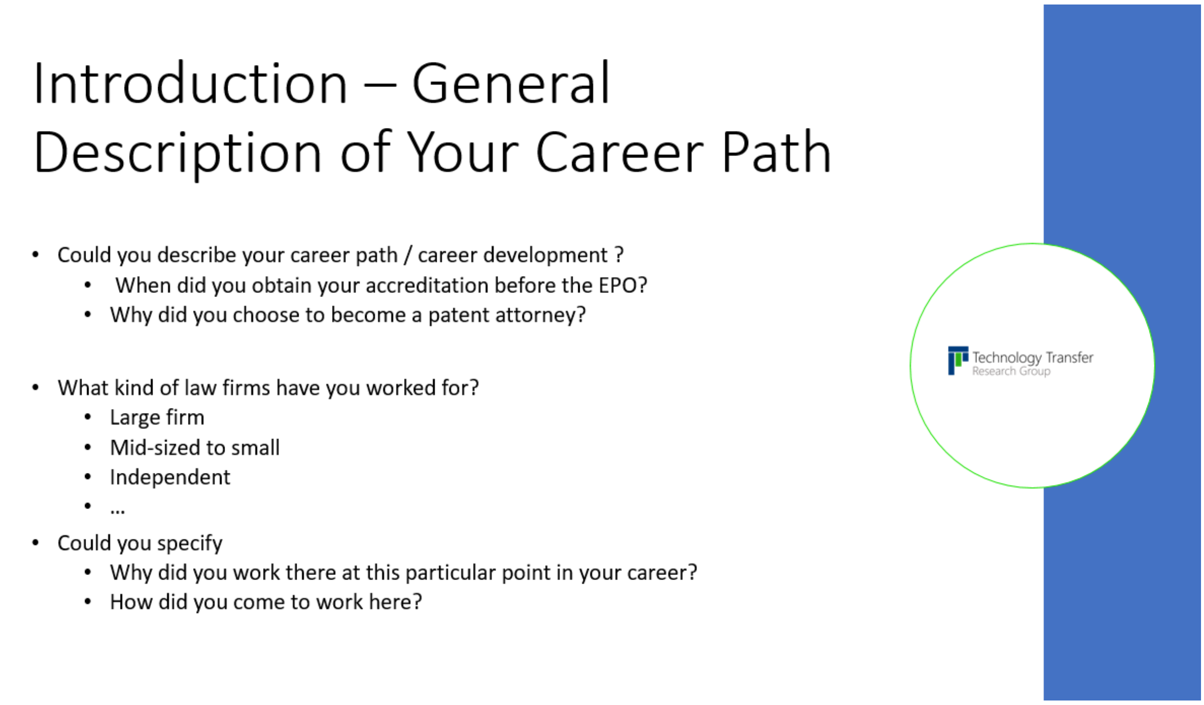
[3] Macdonald and Lefang (1998): “Measuring innovation – The patent attorney as an indicator of innovation,” Computer and Law Security Report (14), 8-12.

[4] Somaya, Williamson and Zhang (2007): “Combining patent law expertise with R&D for patenting performance,” Organization Science (18), 922-937.

[5] Kießling, Terbach & Prilop (2017): “epo-ops,” https://www.rubydoc.info/gems/epo-ops/0.3.1.

9 Appendix

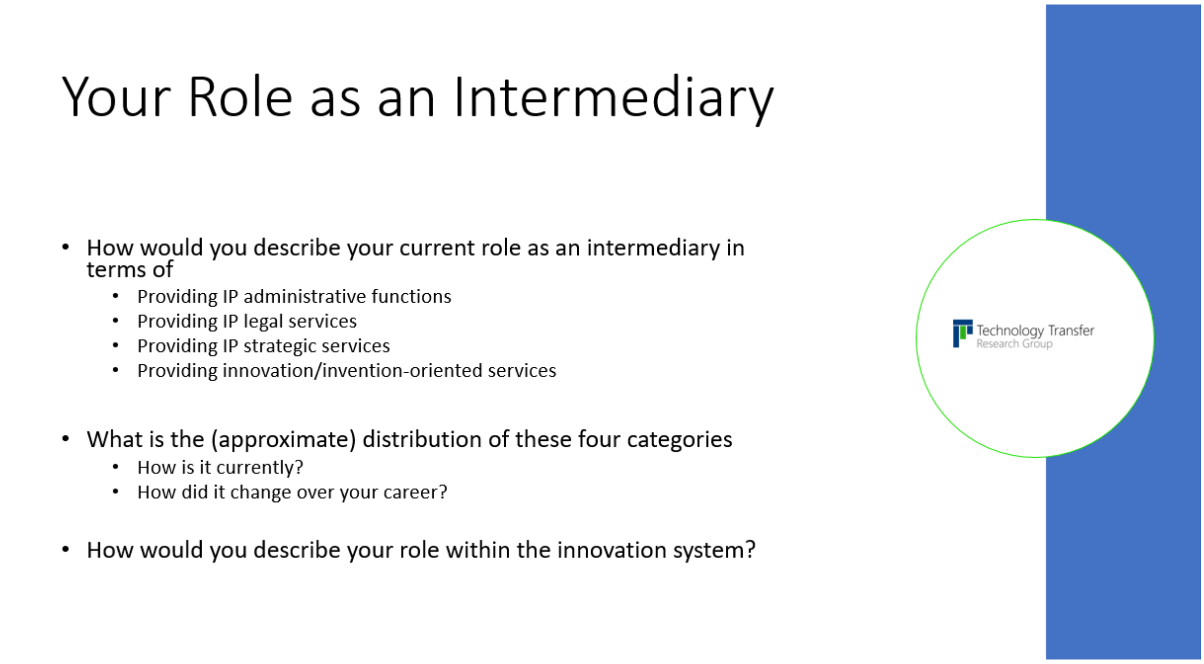
9.1 Guideline for the Structured Interviews



Introduction – General Description of Your Career Path

- Could you describe your career path / career development ?
  - When did you obtain your accreditation before the EPO?
  - Why did you choose to become a patent attorney?
- What kind of law firms have you worked for?
  - Large firm
  - Mid-sized to small
  - Independent
  - …
- Could you specify
  - Why did you work there at this particular point in your career?
  - How did you come to work here?

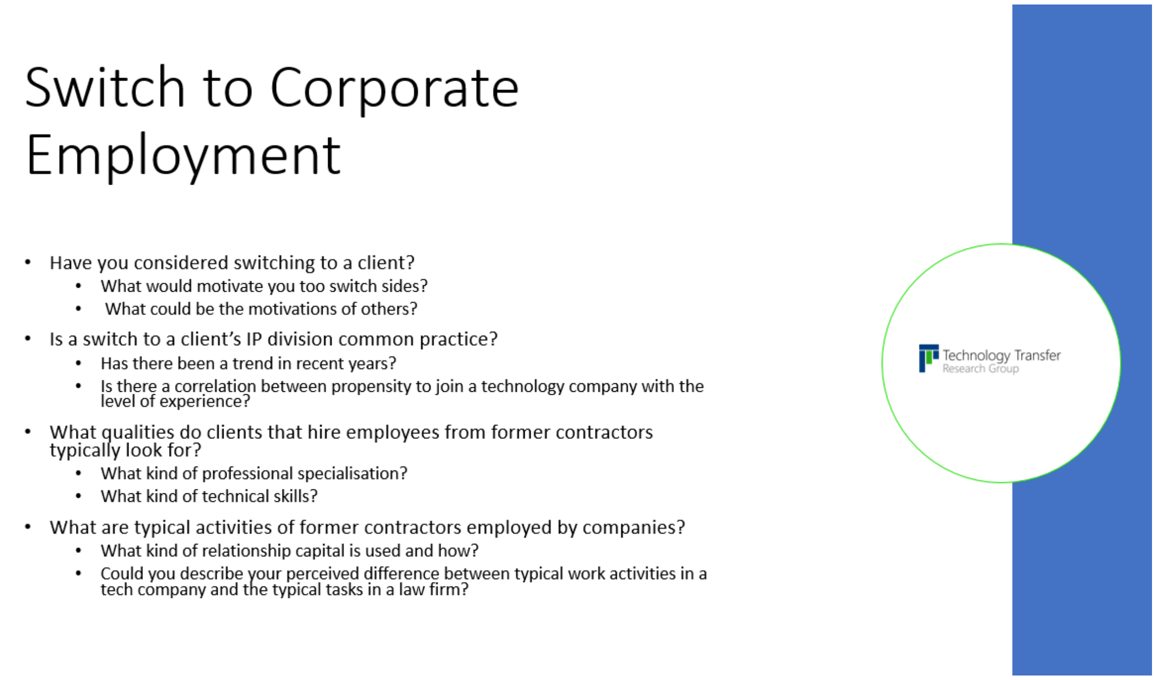
Technology Transfer
Research Group



Your Role as an Intermediary

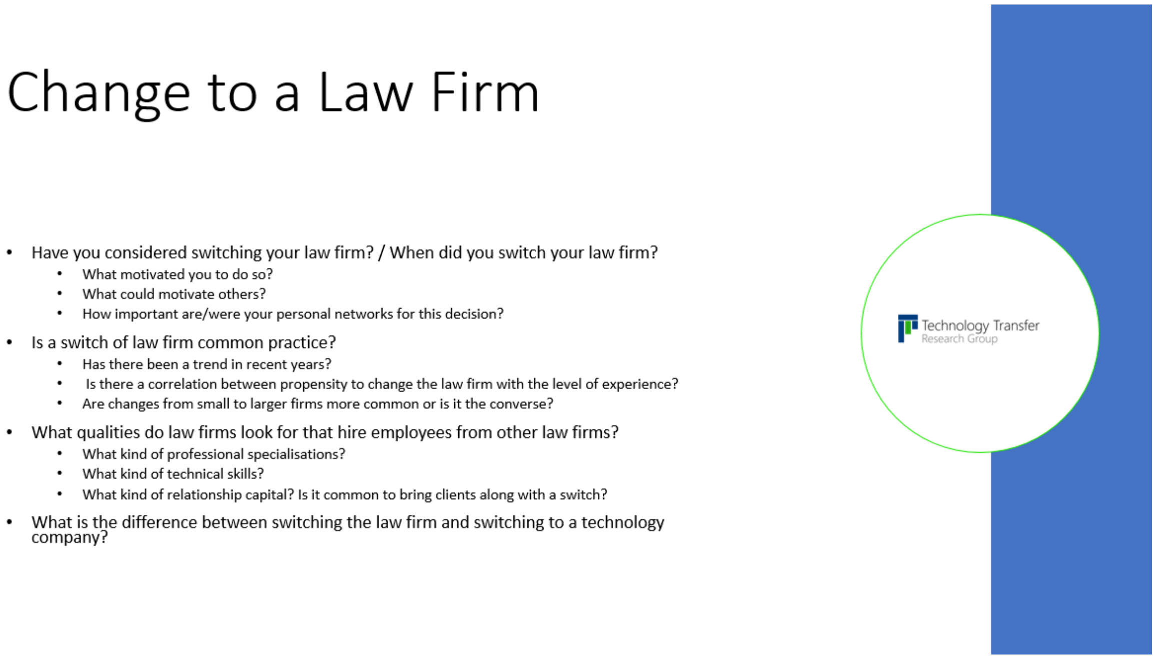
- How would you describe your current role as an intermediary in terms of
  - Providing IP administrative functions
  - Providing IP legal services
  - Providing IP strategic services
  - Providing innovation/invention-oriented services
- What is the (approximate) distribution of these four categories
  - How is it currently?
  - How did it change over your career?
- How would you describe your role within the innovation system?

Technology Transfer
Research Group



Switch to Corporate Employment

- Have you considered switching to a client?
  - What would motivate you too switch sides?
  - What could be the motivations of others?
- Is a switch to a client’s IP division common practice?
  - Has there been a trend in recent years?
  - Is there a correlation between propensity to join a technology company with the level of experience?
- What qualities do clients that hire employees from former contractors typically look for?
  - What kind of professional specialisation?
  - What kind of technical skills?
- What are typical activities of former contractors employed by companies?
  - What kind of relationship capital is used and how?
  - Could you describe your perceived difference between typical work activities in a tech company and the typical tasks in a law firm?



Change to a Law Firm

- Have you considered switching your law firm? / When did you switch your law firm?
  - What motivated you to do so?
  - What could motivate others?
  - How important are/were your personal networks for this decision?
- Is a switch of law firm common practice?
  - Has there been a trend in recent years?
  - Is there a correlation between propensity to change the law firm with the level of experience?
  - Are changes from small to larger firms more common or is it the converse?
- What qualities do law firms look for that hire employees from other law firms?
  - What kind of professional specialisations?
  - What kind of technical skills?
  - What kind of relationship capital? Is it common to bring clients along with a switch?
- What is the difference between switching the law firm and switching to a technology company?

9.2 Mapping of Planned Work Packages to Outputs

Work Package	Description (Originally planned work)	Reported Work (Output)
WP-1	systematic compilation of the state of the discussion on the role of patent attorneys for innovation systems	The results of the literature review have shortly been discussed in section 2.1.
	structured opened interviews with 50 patent attorneys	After intensive acquisition work, 42 practising patent attorneys were interviewed. The insights were used to guide the conception of career path and examine the intermediate results of the data compilation process as discussed in in sections 2.2 and 4.2.2.
WP-2	definition of a data model for the uniform representation of career paths based on existing dataset in the IPIB and the analysis results of WP-1	The final data model of the CAPA dataset was presented and discussed in sections 4.1 and 5.
	prototypical generation of the CAPPA based on the developed data model and the existing dataset in the IPIB harmonisation activities	The prototypical generation of the CAPPA dataset has been discussed and presented in the Report from 15/06/2018 and is briefly covered in section 4.1.
WP-3	checking LinkedIn Profiles (600) against the career path in the CAPPA	520 LinkedIn profiles of practising patent attorneys were collected and processed to serve as a test set to train the relationship detection algorithm.
	requesting 250 individuals to actively check the data in the CAPPA and request to augment and correct the data in a structured way	Due to the current sensibility on data protection (as a positive result of the GDPR) we avoided to directly contact patent attorneys in order to not harm the project success. Hence, we used the "Profile needs update?" feature which has been integrated during the project into the

		IP Industry Base. Feedback being received through this facility is directly added to harmonisation processes of the IPIB.
	Applying CDQA to CAPPA	In order to save resources (as discussed in section 3) we have to decided not to add this additional level of data quality assurance. However, we believe that the techniques we are currently using to calculate the final CAPPA dataset do assure a high level of data quality.
WP-4	provision a tool that allows patent attorneys to personally verify their career profiles in the CAPPA data set	The IPIB now provides the above mentioned "Profile needs update" facility. This tool is in frequent use. However, as discussed above, we do not actively ask patent attorneys to verify their profile.
	investigate gap filling techniques	The techniques that have been used to provide clean and gap-filled data have been described in section 4.2.
	augmentation of the facility for the generation of CAPPA releasing the GAP filled CAPPA	The gap-filled CAPPA dataset has been published as open data and in open access on the 28th of June 2019 on ZENODO with the DOI 10.5281/zenodo.3265385 and is intended to be updated every six months.
WP-5	visualisation of the CAPPA	Some of the visualisations have been shown in section 5. A further set of visualisations has been published with commentary and example Code in a Notebook on ZENODO with the DOI 10.5281/zenodo.3267424.
	Calculation and interpretation of network properties of the CAPPA	A short overview of the findings is discussed in section 6 of the report.
WP-6	correlating CAPPA with success indications	In order to save resources (as discussed in section 3) we were forced to cancel this task from our working plan. However, we

		hope that our data set will support other researcher to move forward in that field of interest.
	comparing results between CAPPA and gap-filled CAPPA	The results of the CAPPA have been discussed in section 4.2. Finally, we have decided to not distinguish between both types of CAPPA dataset. The published CAPPA dataset is what we called the gap-filled CAPPA dataset in the work programme.
	validation of results through feedback interviews	In order to save resources (as discussed in section 3) we were forced to cancel this task from our working plan.
	presentation of the results at the EPO workshops	A presentation of previous results has been conducted at the workshop in June 2018. A presentation of the final results will be given at the concluding ARP workshop in December 2019

9.3 Additional Ideas from the Intermediate Workshop

Request	Description (Originally requested work)	Reported Work (Output)
1	analysis of market geography and mobility of patent attorneys across national borders in Europe	In order to save resources (as described in section 3) we have conducted preliminary explorations, without a special focus in geography and mobility.
2	matching data of patent attorneys in the data set with the EQE list	In order to save resources (as described in section 3) we were not able to include this task.
3	link between specialisation profile of patent attorneys and their mobility	Useful specialisation profiles would have required more data points for most of the attorneys, so it was not possible within the constraints (described in section 3) of the project to investigate this further.

9.4 Data Paper Included with Dataset

CAPPA – A Dataset on the Career Paths of EPO Patent Attorneys

DOI: 10.5281/zenodo.3265385
Online available: https://zenodo.org/record/3265385

Authors: Kazimir Menzel[1], Lutz Maicher[1,2]

Affiliations:
[1] Technology Transfer Research Group, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

[2] Fraunhofer-Center for International Management and Knowledge Economy, Neumarkt 9-19, 04109 Leipzig, Germany

Contact email: lutz.maicher@uni-jena.de

Abstract

The CAPPA dataset contains anonymised information on the career paths of patent attorneys that are registered with the European Patent Office (EPO) tracking their employment from 2008 to 2018. It consists of four different sub-datasets (Attorney Data Frame, Employer Data Frame, Change Data Frame, and Career Path Data Frame) that are composed for specific information needs. The dataset has been computed, cleaned and made consistent based on publicly available patent records that are available through the EPO.

Value of the data

- The CAPPA dataset contains anonymised information on the career paths of patent attorneys registered with the European Patent Office, derived from all patent applications at the European Patent Office, published since 2008.
- The CAPPA dataset is to be updated every six months, released as new versions.
- The dataset and all its versions are provided as open data under the Creative Commons Attribution 4.0 International license.

Data

Patents are an important source of open data, heavily used in innovation economics for research and, through patent information systems, by practitioners. While innovation activities are extensively investigated using patent data, its usage for research about the involved patent attorneys is limited. A major reason for this is that patent data requires significant harmonisation [1], which poses a high barrier to research. In order to lower this barrier somewhat, we created the CAPPA dataset.

Based on a stringent harmonisation of the patent attorney and patent firm entries retrieved from the EP patent applications (against identifiers of the IP Industry Base[6]), the dataset contains the anonymised affiliations of patent attorneys with patent law firms and technology companies on a monthly basis since January 2008. This information is derived from all available patent applications before the European Patent Office (EPO) over the last 12 years that contain sufficient information.

The dataset was computed based on 922,959 patent applications that contained sufficient information. From these, career paths for 8,710 patent attorneys could be recorded, which were affiliated to a total of 2930 employers. The data used covers a period of 12 years, from 01/01/2008 to 31/12/2018. In order to ensure anonymity of the individual patent attorneys, the attorneys and their employers have been anonymised and are accessible only through their unique identifier within the dataset.

Structure of the dataset

The dataset consists of four data frames. The first, the *Attorney Data Frame*, has the following structure:

attorney_id	n_emp	n_pat	first_observation	last_observation	months_active
1	1	5	2009-03	2009-11	5
2	2	166	2008-04	2018-11	67
...	...	...	...	...	...

- attorney_ID. A unique and anonymised identifier for the patent attorney that allows to track the individual patent attorney.

[6] http://s.fhg.de/IPIB

- n_emp. Number of employers with which the attorney was affiliated over the observation period.
- n_pat. Number of patents, which have been assigned to the patent attorney over the observation period.
- first_observation. The first month given in “yyyy-mm” format, in which the attorney has been observed on a patent.
- last_observation. The last month given in “yyyy-mm” format, which the attorney has been observed on a patent.
- months_active. The duration of an attorney’s career span according to patent applications given in months.

The second data frame, the *Employer Data Frame*, has the structure:

employer_id	rank	type	n_att	mean_ret	mad_ret	n_pat	med_ppa
1	878	small	1	125	0	22	22
2	769	small	1	100	0	15	15
...	...	...	...	...	...	...	...

- employer_id. A unique and anonymised identifier for the registered employers of patent attorneys that allows to track the individual employer. Note that also self-employed attorneys are tracked as employers.
- rank. A rank over all employers that is synthetically derived from the cumulative number of employed attorneys, a period-based measure and the activity (number of patents). It is intended to propose one way of natural scaling when working with the dataset and serves also to distinguish between types of patent attorney employers. The rank is in ascending order, i.e. the higher number implies the higher rank.
- type. Typology of employers of patent attorneys, distinguishing between technology companies (tech) and patent law firms and within the latter between large-, medium- and small-sized companies (*large*, *medium*, *small*).
- n_att. The number of patent attorneys that have been observed as affiliated with the employer over the entire observation period.
- mean_ret. The mean duration of retaining a patent attorney given in months.
- mad_ret. The median absolute deviation of retaining a patent attorney given in months.
- n_pat. The number of patents that have been assigned to the employer over the observation period.
- med_ppa. The median of patents assigned to employed attorneys.

The third data frame, the *Career Path Data Frame*, contains the actual career paths, consisting of the stations of employment for each attorney. In total, 10,786 such stations have been recorded and are available through the following structure:

attorney_id	employer_id	first_observation	last_observation	n_pat
1	1820	2008-10	2009-11	3
2	574	2008-01	2018-07	7
2	2853	2008-01	2018-07	22
...	...	...	...	...

- attorney_id. A unique and anonymised identifier for the patent attorney that allows to track the individual patent attorney. The attorney_id is the same as in the *Attorney Data Frame*.
- employer_id. A unique and anonymised identifier for the registered employers of patent attorneys that allows to track the individual employer. Note that also self-employed attorneys are tracked as employers. The ID is the same as in the *Employer Data Frame*.
- first_observation. The month in which an attorney has been first observed as affiliated with an employer, given in “yyyy-mm” format.
- last_observation. The month in which an attorney has been observed last as affiliated with an employer, given in “yyyy-mm” format.
- n_pat. The number of patents that have been assigned to the attorney and the employer during his employment period.

The fourth data frame, the *Change Data Frame*, contains the changes of employment, when one attorney has actually changed from one employer to another, including work for two employers in parallel. In total, 4707 such changes have been recorded. The *Change Data Frame* has the structure:

attorney_id	time_of_chg	origin	origin_type	target	target_type	change_type
7343	2011-05	1576	tech	458	small	tech -> small
4482	2008-11	1695	medium	559	medium	medium -> medium
...	...	...	...	...	...	...

- attorney_id. A unique and anonymised identifier for the patent attorney that allows to track the individual patent attorney. The attorney_id is the same as in the *Attorney Data Frame*.
- time_of_chg. The month, in which a change of employment has been observed given in the “yyyy-mm” format. The date is set to the first month, during which an attorney has been observed with his new employer.
- origin. The employer_id of the employer with which the attorney was associated before the change. The employer_id is the same as in the *Employer Data Frame*. Note that a new entrance is not recorded as a change, i.e. only changes from one employer to another are recorded. New entrances can be readily seen in the *Career Path Data Frame*.
- origin_type. The type of the original employer as given in the *Employer Data Frame*.
- target. The employer_id of the employer with which the attorney is associated after the change. The employer_id is the same as in the *Employer Data Frame*. Note that it is not recorded as a change, when an attorney is no longer observed, i.e. only changes from one employer to another are recorded.
- target_type. The type of the target employer as given in the *Employer Data Frame*.
- change_type. The type of change as a combination of origin and target

Limitations

Due to the scope of the data retrieval, the dataset suffers several limitations that should be taken into account when working with it.

- Anonymity. Due to the entry into force of the new GDPR, it is not possible to make explicit, for which firms an attorney worked at a given time as this would allow it easily to identify the attorneys. The authors are aware of the fact that this might cause some difficulties, when seeking to combine the information with other data.
- Focus on patent applications. As the dataset does not take into account other aspects of a patent attorney's work, the career paths are solely derived from patent application data. This means in turn that some of the quantitative variables of patent attorneys as well as their employers reflect only a fraction of the actual activity of the attorney. This also means that the career stations of attorneys with very few patent applications are not always exact nor complete. The authors decided to keep them in the dataset in order to deliver as complete a dataset as possible.
- Parallel employment. A not prevalent but also not uncommon phenomenon is that patent attorneys are found in parallel employment. While the reasons for this are manifold as has been determined during interviews with selected attorneys, it makes it often difficult to decide whether an attorney is found in a transition phase between two employments or working in parallel. In these cases, there are two changes recorded, one when the attorney is first observed with her new employer and when the attorney is last observed with her old employer. In cases of doubt, it is recommended to take the first date as authoritative.
- Insufficient or ambiguous patent records. As most of the information has been retrieved from EPO patent applications, it was for a sizeable subset not possible to either retrieve information on an attorney or her affiliated employer. The lack of such information led to an exclusion from the dataset so that approximately 2100 registered patent attorneys are not covered in this dataset. This also led to a reduction of patents that could be assigned to the attorneys and their employers, which means that not all patents from the covered period could be included in the quantitative measures in the data set.

Acknowledgements

The creation of the CAPPA dataset has been funded by the Academic Research Programme of the European Patent Office under the grant agreement number 2017/8405.

References

[1] Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B. H., & Harhoff, D. (2010). Harmonizing and combining large datasets-An application to firm-level patent and accounting data (No. w15851). National Bureau of Economic Research.

9.5 Visualisation Tutorial for CAPPA

Visualising the Career Paths of Patent Attorneys

Kazimir Menzel

DOI: 10.5281/zenodo.3267424
Online available: https://zenodo.org/record/3267424

This notebook introduces some ways to visualise the CAPPA dataset [1] and is intended to serve as short guide to working with the dataset. It uses R with the tidyverse library [2] for its ease of interaction with data and in creating plots.

The dataset contains data on employment stations of patent attorneys registered with the European Patent Office that have been extracted from patent application data. It consists of four data frames, each containing different data that can be examined and visualised according to the research interests of the researcher.

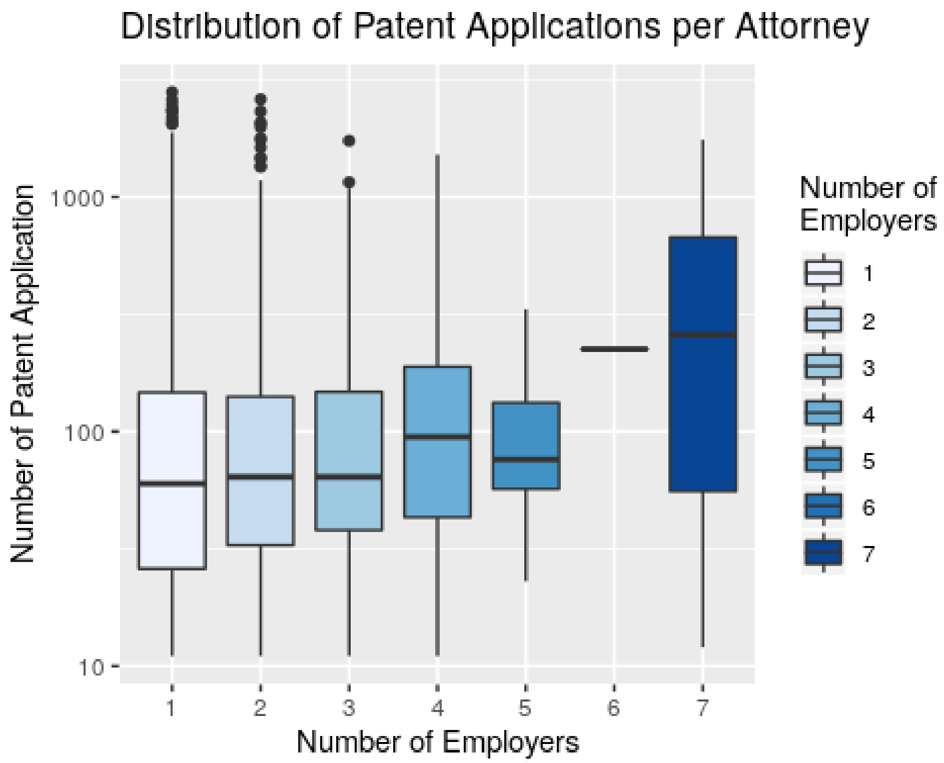
1 The Attorney Data Frame

The first data frame that we will look at is the `attorney data frame`, which contains general data on the careers of attorneys and provides an easy way to assess the group of patent attorneys as a whole.

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────── tidyverse 1.
2.1 ──

## ✓ ggplot2 3.2.1     ✓ purrr   0.3.2
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   0.8.3     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0

## ── Conflicts ────────────────────────────────────── tidyverse_conflicts
() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()

file_path = '~/Documents/CAPPA/final_data_set'

read_csv(
  file = str_c(file_path, 'attorney_data_frame.csv', sep = '/'),
  col_types = 'ciicci') %>%
  separate(                                  # These lines are a simple way
    col = 'first_observation',               # to deal with the difficulties
    into = c('first_year', 'first_month'), # of monthly resolution in R.
    sep = '-') %>%
  transform(
    first_year = as.integer(first_year),
    first_month = as.integer(first_month)) %>%
  separate(
    col = 'last_observation',
    into = c('last_year', 'last_month'),
    sep = '-') %>%
  transform(
    last_year = as.integer(last_year),
    last_month = as.integer(last_month)) -> attorneys
```

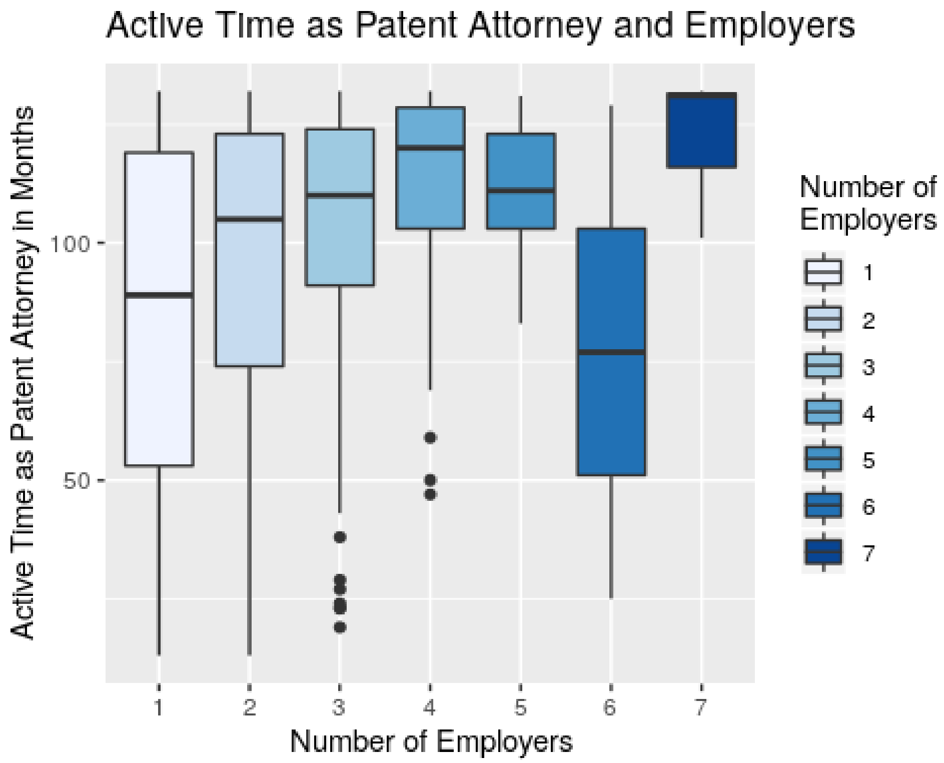
Having imported the `attorney data frame`, it is now time to inspect the most essential properties. Let us begin with graphic that summarises the number of patent applications of attorneys given the number of employers. In order to remove some outliers, we only look at attorneys that have filed between 10 and 3000 patent applications.

```
attorneys %>%
  filter(n_pat > 10 & n_pat < 3000) %>%
  transform(n_emp = as_factor(n_emp)) %>%
  ggplot() +
  ggtitle('Distribution of Patent Applications per Attorney') +
  geom_boxplot(
    mapping = aes(
      x = n_emp,
      y = n_pat,
      fill = n_emp)) +
  scale_x_discrete(name = 'Number of Employers') +
  scale_y_log10(name = 'Number of Patent Application') +
  scale_fill_brewer(
      name = 'Number of\nEmployers')
```



The distribution of the amount of patent applications per attorney is fairly stable across the number of employers, with a slight increase in the number of employers. We can now inspect another variable of interest to gain some intuition about the relationship of employment stations and the duration of activity as a patent attorney, considering only attorneys that were active for at least one year.

```
attorneys %>%
  transform(n_emp = as_factor(n_emp)) %>%
  filter(months_active > 12) %>%
  ggplot() +
  ggtitle('Active Time as Patent Attorney and Employers') +
  geom_boxplot(
    mapping = aes(
      x = n_emp,
      y = months_active,
      fill = n_emp)) +
  scale_x_discrete(
    name = 'Number of Employers') +
  scale_y_continuous(name = 'Active Time as Patent Attorney in Months') +
  scale_fill_brewer(
    name = 'Number of\nEmployers')
```

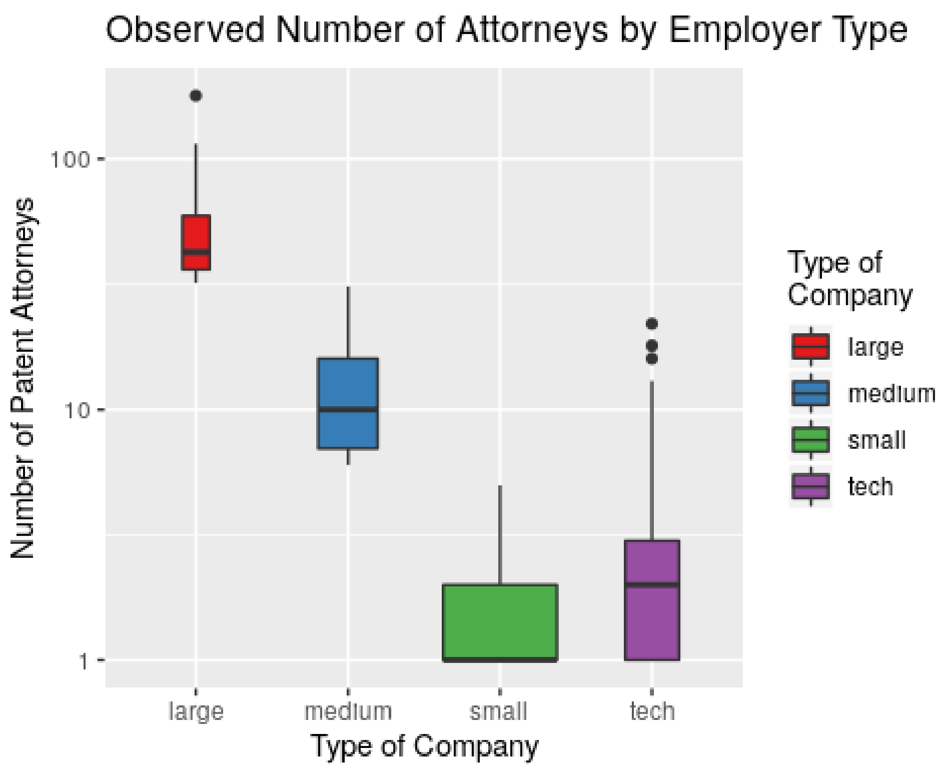


Here we can clearly observe an increase in median when moving from one to four employers. Seven employers seems to be in line with the trend as well. Interestingly, this connection seems to break apart among the attorneys that worked for five and six employers.

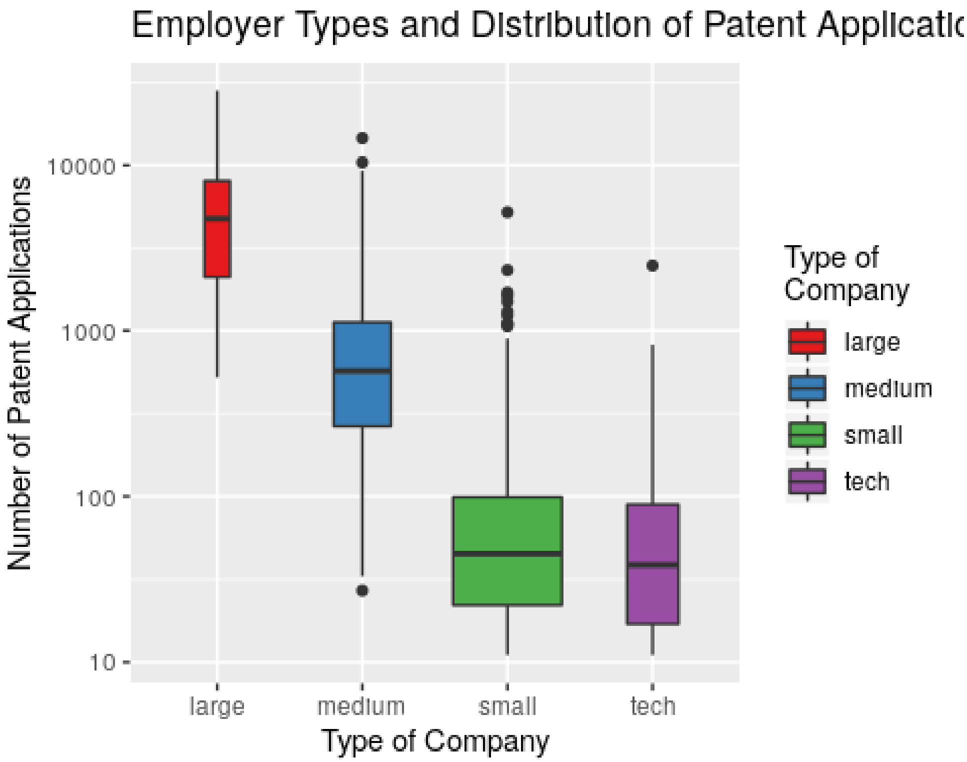
2 The Employer Data Frame

As with the `attorney data frame`, we can use the `employer data frame` to inspect some general features of the dataset. The following two boxplots show the distributions of the number of patent applications and the number of patent attorneys by the type of the employer. The employer types are assigned based on whether the employer is the patent arm of a technology company (`tech`) or wether it is a `large`, a `medium`-sized, or a `small` law firm. The prevalence of the respective type of employer in the dataset has been mapped to the width of the boxes.

```r
read_csv(
  file = str_c(file_path, 'employer_data_frame.csv', sep = '/'),
  col_types = 'ciciddid'
) -> employers

employers %>%
  filter(n_pat > 10) %>%
  ggplot() +
  ggtitle('Observed Number of Attorneys by Employer Type') +
  geom_boxplot(
    mapping = aes(
      x = type,
      y = n_att,
      fill = type),
    varwidth = TRUE) +
  scale_y_log10(name = 'Number of Patent Attorneys') +
  scale_x_discrete(name = 'Type of Company') +
  scale_fill_brewer(
    name = 'Type of\nCompany',
    palette = 'Set1')
```



```
employers %>%
  filter(n_pat > 10) %>%
  ggplot() +
  ggtitle('Employer Types and Distribution of Patent Applications') +
  geom_boxplot(
    mapping = aes(
      x = type,
      y = n_pat,
      fill  = type),
    varwidth = TRUE) +
  scale_y_log10(name = 'Number of Patent Applications') +
  scale_x_discrete(name = 'Type of Company') +
  scale_fill_brewer(
    name = 'Type of\nCompany',
    palette = 'Set1')
```



We can observe that the number of patent applications and the number of attorneys, perhaps unsurprisingly, follow a similar pattern. We can also see that *medium*-sized and *small* law firm are by far the most common employers in the dataset.

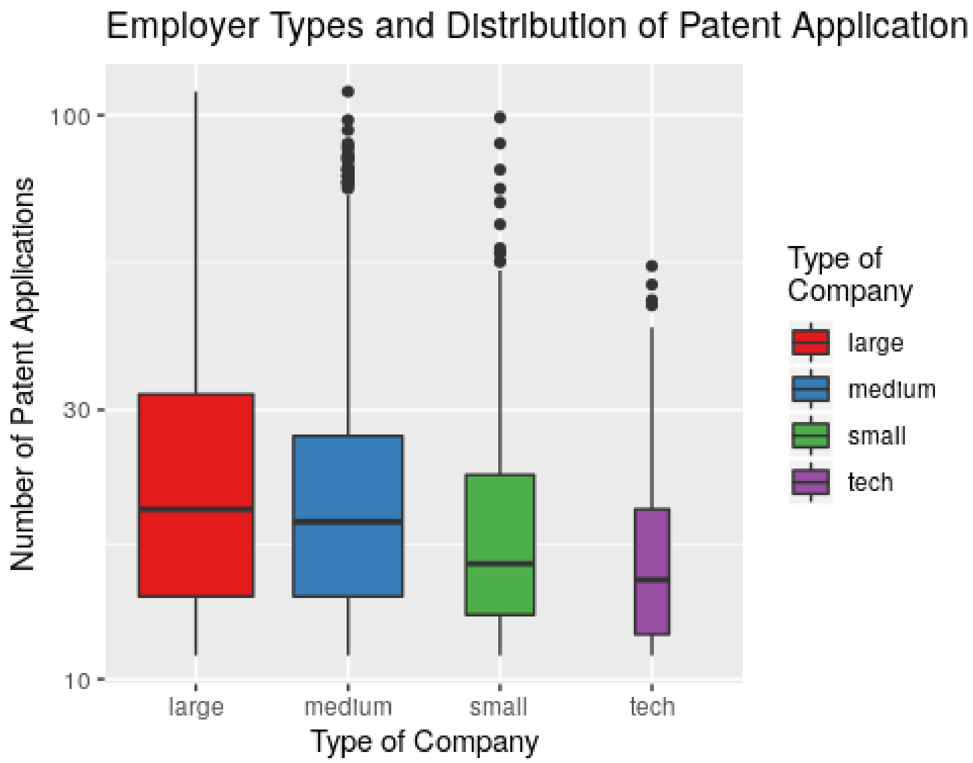
3 The Career Path Data Frame

Let us now inspect the central data frame of the CAPPA dataset. The *career path data frame* contains the information for all the career paths of the patent attorneys.

```
read_csv(
  file = str_c(file_path, 'career_path_data_frame.csv', sep = '/'),
  col_types = 'cccci'
  ) %>%
  separate(                                 # These lines are a simple way
    col = 'first_observation',              # to deal with the difficulties
    into = c('first_year', 'first_month'), # of monthly resolution in R.
    sep = '-') %>%
  transform(
    first_year = as.integer(first_year),
    first_month = as.integer(first_month)) %>%
  separate(
    col = 'last_observation',
    into = c('last_year', 'last_month'),
    sep = '-') %>%
  transform(
    last_year = as.integer(last_year),
    last_month = as.integer(last_month)) -> career_paths
```

After importing the `career path data frame`, we can now use it together with the `employer data frame`, to repeat the visualisation from above to find out, whether the type of an employer has any bearing on the number of patent applications filed by a patent attorney. The following plot diverges from the last one only in that it counts the number of patent applications *per attorney* and *per employment period*.

```r
career_paths %>%
  left_join(
    y = employers %>%
          select('employer_id', 'type'),
    by = 'employer_id') %>%
  filter(n_pat > 10) %>%
  ggplot() +
  ggtitle('Employer Types and Distribution of Patent Applications') +
  geom_boxplot(
    mapping = aes(
      x = type,
      y = n_pat,
      fill = type),
    varwidth = TRUE) +
  scale_y_log10(name = 'Number of Patent Applications') +
  scale_x_discrete(name = 'Type of Company') +
  scale_fill_brewer(
    name = 'Type of\nCompany',
    palette = 'Set1')
```

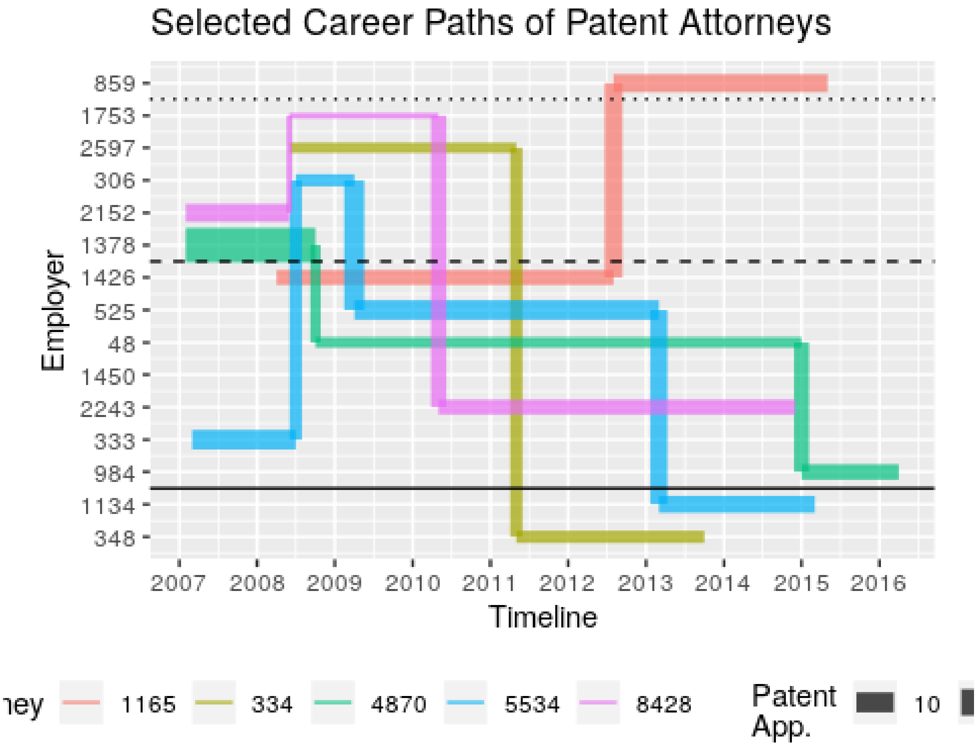


We can now see that the number of patent applications is strongly related to the employer type. It seems that larger employers provide more opportunities for patent attorneys to file patent applications than smaller employers. It also worth noting that attorneys in technology companies do not file as many patent applications as those working in patent law firms.

In this plot, the width of the boxes reflects the relative number of patent attorney stations. We can see that most single stations in the dataset are with `medium` and `Large` patent law firms.

We can now turn to inspect specific career paths. The career paths shown in the following plot have have been selected for illustrative purposes.

```
career_paths %>%
  filter(
      attorney_id == '8428' |
      attorney_id == '1165' |
      attorney_id == '4870' |
      attorney_id == '5534'|
      attorney_id == '334') %>%
  mutate(
    begin = (first_year - 2008) * 12 + first_month,
    end = (last_year - 2008) * 12 + last_month)  %>%
  select(attorney_id, employer_id, begin, end, n_pat)  %>%
  rename('timeline' = begin) %>%
  group_by(attorney_id) %>%
  do(
    add_row(.,
      attorney_id = last(.$attorney_id),
      employer_id = last(.$employer_id),
      timeline = last(.$end),
      end = NA,
      n_pat = last(.$n_pat))) %>%
  select(attorney_id, employer_id, timeline, n_pat) %>%
  left_join(
    y = employers %>% select(employer_id, rank, type),
    by  = 'employer_id') %>%
  ungroup() %>%
  mutate(
    employer_order = as.integer(as_factor(rank))) %>%
  arrange(attorney_id, timeline) %>%
  ggplot() +
    ggtitle('Selected Career Paths of Patent Attorneys') +
    geom_step(
      mapping = aes(
        x = timeline,
        y = employer_order,
        colour = attorney_id,
        size = n_pat),
      alpha = .7) +
  geom_hline(yintercept = 14.5, lty = 'dotted') +
  geom_hline(yintercept = 9.5, lty = 'dashed') +
  geom_hline(yintercept = 2.5) +
  scale_x_continuous(
    name = 'Timeline',
    breaks = seq(0, 132, by = 12),
    labels = 2007:2018) +
  scale_y_continuous(
    name = 'Employer',
    breaks = 1:15,
    labels = c('348', '1134', '984', '333', '2243', '1450', '48', '525', '
1426', '1378', '2152', '306', '2597', '1753', '859')) +
  scale_color_discrete(name = 'Attorney') +
  scale_size_continuous(name = 'Patent\nApp.') +
  theme(legend.position = 'bottom') +
  coord_fixed(ratio = 5)
```



The plot traces the career paths of five patent attorneys and highlights the difference of in the number of their patent applications by the size of the segments. The ordering of the `employers` follows their rank variable. The different `types` of `employers` can be inferred from the black lines separating the types. The solid line separates `small` from `medium`-sized employers, the dashed line `medium`-sized from `large` employers and the dotted line separates `large` employers from `tech`nology companies.

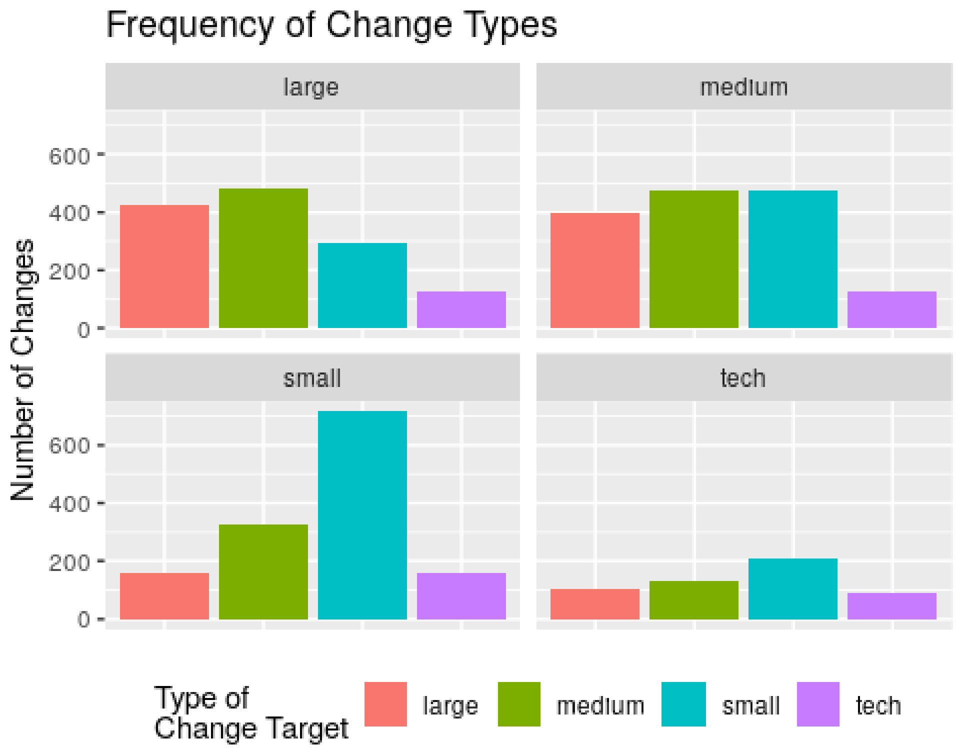
4 Change Data Frame

The last data frame to be presented in the *`change data frame`*. It contains data for all changes and can be used to investigate the nature of changes more precisely than just using the *`career path data frame`*.

```r
read_csv(
  file = str_c(file_path, 'change_data_frame.csv', sep = '/'),
  col_types = 'ccccccc'
  ) %>%
  separate(                                # These lines are a simple way
    col = 'time_of_chg',                   # to deal with the difficulties
    into = c('year', 'month'),             # of monthly resolution in R.
    sep = '-') %>%
  transform(
    year = as.integer(year),
    month = as.integer(month)) -> changes
```

We will begin with inspecting the frequency of the various change types.

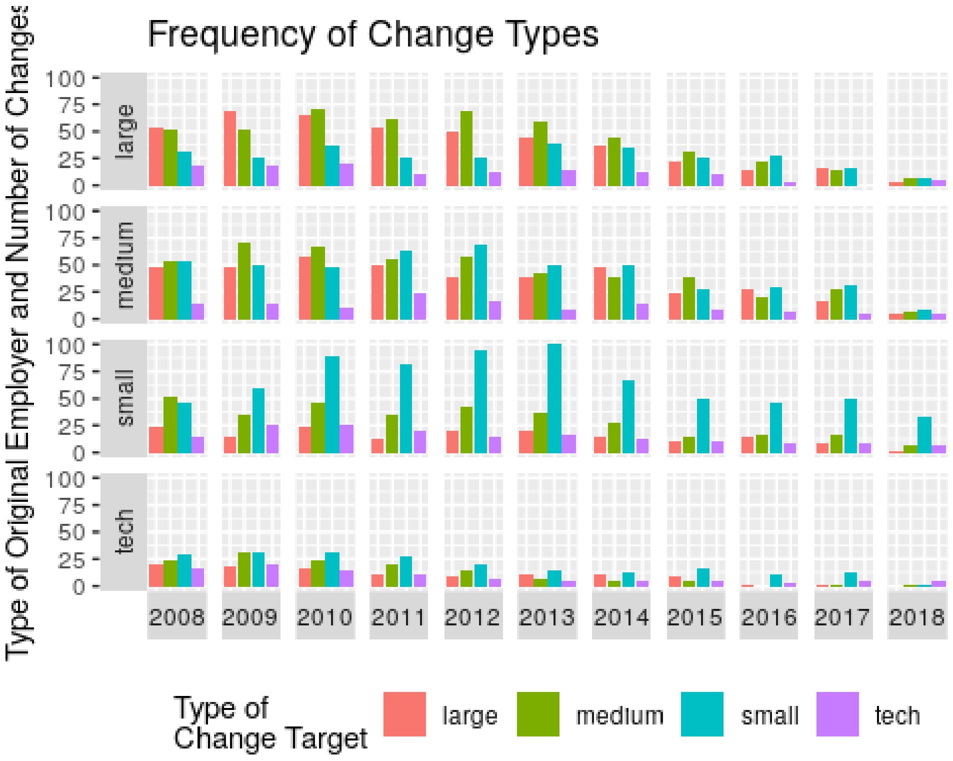
```r
changes %>%
  ggplot() +
  ggtitle('Frequency of Change Types') +
  geom_bar(
    mapping = aes(
      x = target_type,
      fill = target_type)) +
  facet_wrap(~ origin_type, ncol = 2) +
  scale_y_continuous(name = 'Number of Changes') +
  scale_fill_discrete(name = 'Type of\nChange Target') +
  theme(
    legend.position = 'bottom',
    axis.title.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.text.x = element_blank())
```



In this plot, each colour corresponds to the type of the target company of a change, each subplot to the type of the original employer. We can see here that in aggregate the most typical change has been to smaller law firms. The exception to this are attorneys that work for large law firms. The latter tend to change overwhelmingly to either medium-sized or large law firms. Interestingly, for any original employer type, we find that technology companies form the rarest change target.

We can further inspect how these change patterns vary over time:

```
changes %>%
  ggplot() +
  ggtitle('Frequency of Change Types') +
  geom_bar(
    mapping = aes(
      x = target_type,
      fill = target_type)) +
  facet_grid(
    origin_type ~ year,
    switch = 'both') +
  scale_y_continuous(name = 'Type of Original Employer and Number of Chang
es') +
  scale_fill_discrete(name = 'Type of\nChange Target') +
  theme(
    legend.position = 'bottom',
    axis.title.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.text.x = element_blank())
```



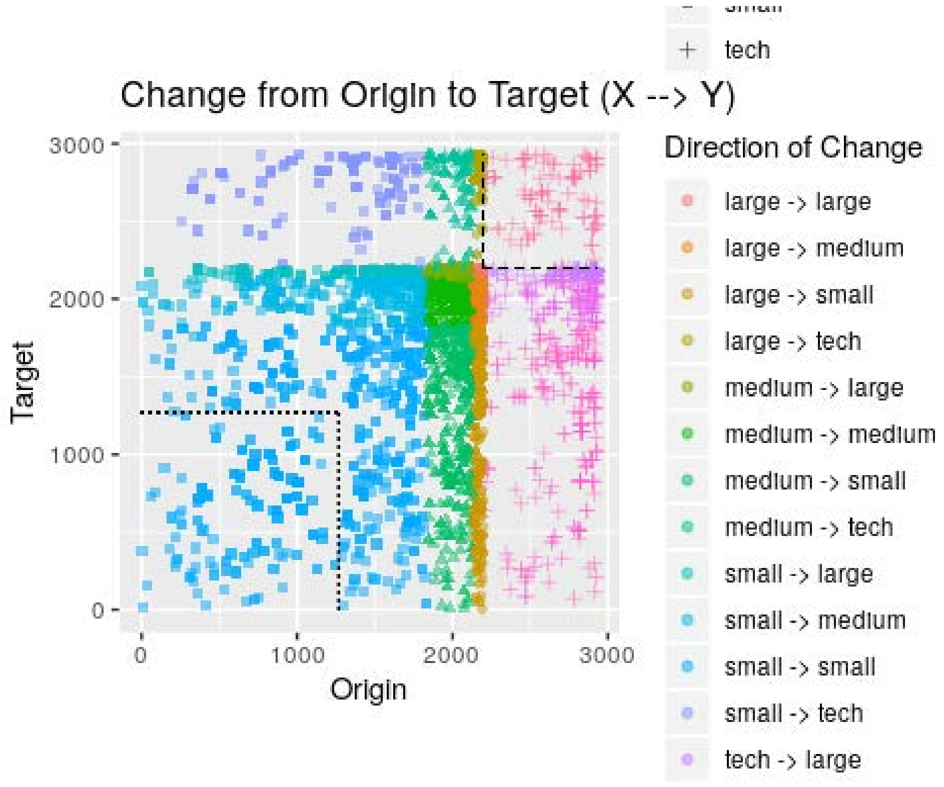
Here we can observe that the change pattern has changed markedly over time. We find a particularly strong wave of changes from *`small`* to *`small`* law firms between 2010 and 2014 and a generally strong wave of changes between 2009 and 2013. Anecdotal evidence from qualitative interviews suggests that these changes might be related to the aftereffects of the recession from 2008 to 2009, which led to changes in the relationship of client companies and patent law firms.

A second way to visualise changes on a higher perspective is through a change matrix. We will first define two regions of special interest, before we render the change matrix. At first, we would like to highlight changes between single attorney law firms. It might come at a surprise but during the preparation of the dataset, we found that this is far more common than one would expect.

```r
employers %>%
  filter(n_att < 2 & type != 'tech') %>%
  pull(rank) %>%
  max() -> const_single
```

Secondly, we mark the region where changes between technology companies might appear. Depending on the research interest, one could also define arbitrary regions in this fashion.

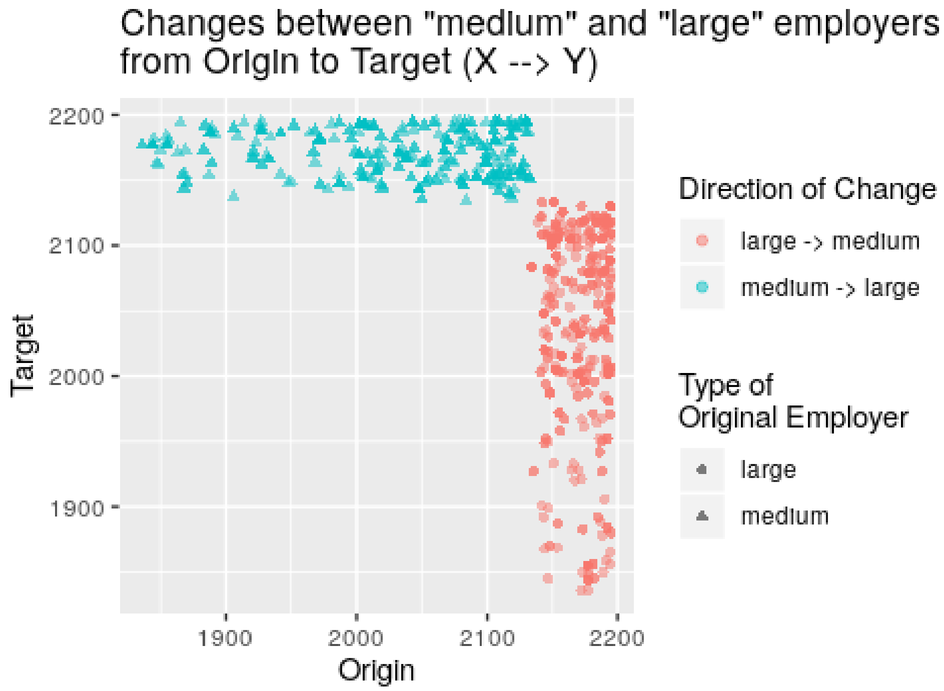
```r
employers %>%
  filter(type == 'tech') %>%
  pull(rank) %>%
  min() -> const_tech

changes %>%
  left_join(
    y = employers %>%
      select(employer_id, rank) %>%
   rename('origin' = employer_id),
  by = 'origin') %>%
 rename('origin_rank' = rank) %>%
 left_join(
  y = employers %>%
   select(employer_id, rank) %>%
   rename('target' = employer_id),
  by = 'target') %>%
 rename('target_rank' = rank) %>%
 ggplot() +
  ggtitle('Change from Origin to Target (X --> Y)') +
  scale_x_continuous(name = 'Origin') +
  scale_y_continuous(name = 'Target') +
  scale_shape_discrete(name = 'Origin') +
  scale_colour_discrete(name = 'Direction of Change') +
  geom_point(
   mapping = aes(
         y = target_rank,
         x = origin_rank,
         colour = change_type,
         shape = origin_type),
   alpha = 1/2) +
  geom_segment(
   mapping = aes(
    x = 0, y = const_single, xend = const_single, yend = const_singl
e),
   lty = 3, size = .25, alpha = 1/2) +
  geom_segment(
   mapping = aes(
    x = const_single, y = 0, xend = const_single, yend = const_singl
e),
   lty = 3, size = .25, alpha = 1/2) +
  geom_segment(
   mapping = aes(
    x = const_tech, y = const_tech, xend = max(employers$rank), yend =
const_tech),
   lty = 2, size = .1, alpha = .8) +
  geom_segment(
   mapping = aes(
    x = const_tech, y = const_tech, xend = const_tech, yend = max(empl
oyers$rank)),
   lty = 2, size = .1, alpha = .8) +
  coord_fixed(ratio = 1)
```



This matrix representation allows us to form an intuition about the nature of changes and spot areas of interest in which we can zoom in further. The top right corner and the bottom left corner been already highlighted. The top right corner contains changes between technology companies and the bottom left corner contains changes between law firms for which only one patent attorney has been observed.

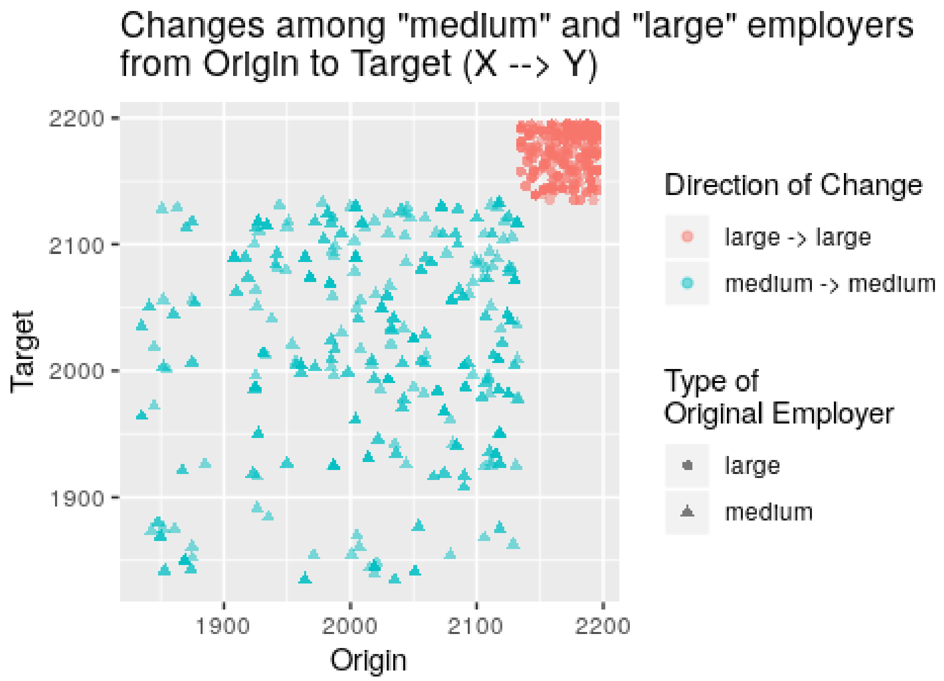
Adding the *filter* option to the above plot, we can now further inspect the changes that are of interest to us. One way would be to inspect only the changes between particular groups.

```
changes %>%
  left_join(
    y = employers %>%
      select(employer_id, rank) %>%
      rename('origin' = employer_id),
    by = 'origin') %>%
  rename('origin_rank' = rank) %>%
  left_join(
    y = employers %>%
      select(employer_id, rank) %>%
      rename('target' = employer_id),
    by = 'target') %>%
  rename('target_rank' = rank) %>%
  filter((change_type == 'medium -> large' | change_type == 'large -> medi
um')) %>%
  ggplot() +
    ggtitle('Changes between "medium" and "large" employers\nfrom Origin t
o Target (X --> Y)') +
    scale_x_continuous(name = 'Origin') +
    scale_y_continuous(name = 'Target') +
    scale_shape_discrete(name = 'Type of\nOriginal Employer') +
    scale_colour_discrete(name = 'Direction of Change') +
    geom_point(
      mapping = aes(
                  y = target_rank,
                  x = origin_rank,
                  colour = change_type,
                  shape = origin_type),
      alpha = 1/2) +
    coord_fixed(ratio = 1)
```

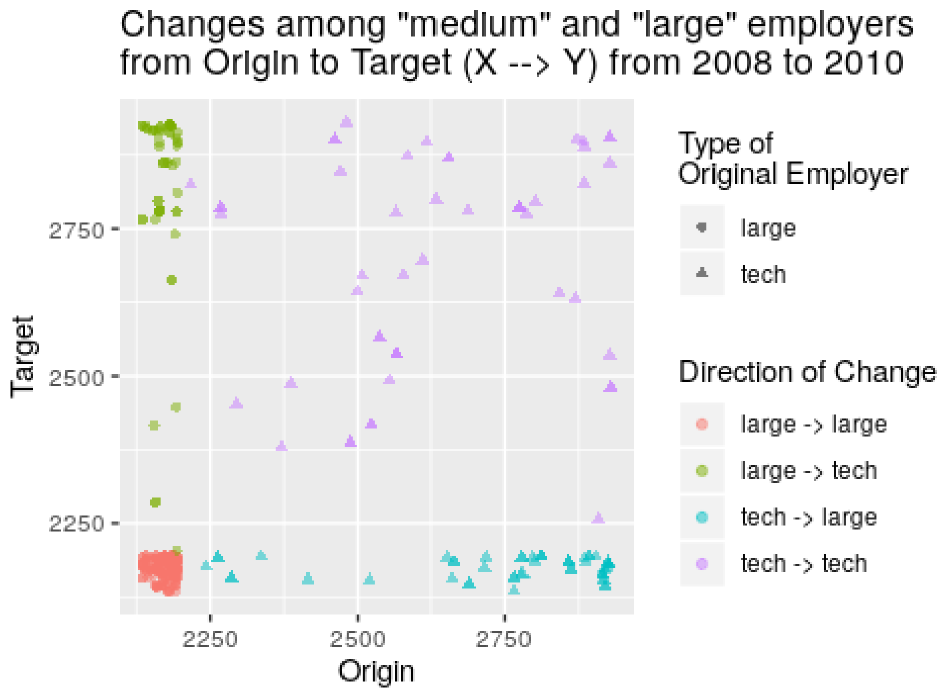


Notice that higher numbers along the `origin` and `target` axes imply that the law firm is larger in terms of attorneys and number of patent applications filed. For more, we refer to the data paper that accompanies the dataset.

```
changes %>%
  left_join(
    y = employers %>%
      select(employer_id, rank) %>%
      rename('origin' = employer_id),
    by = 'origin') %>%
  rename('origin_rank' = rank) %>%
  left_join(
    y = employers %>%
      select(employer_id, rank) %>%
      rename('target' = employer_id),
    by = 'target') %>%
  rename('target_rank' = rank) %>%
  filter((origin_type == target_type) & (origin_type == 'medium' | origin_
type == 'large')) %>%
  ggplot() +
    ggtitle('Changes among "medium" and "large" employers\nfrom Origin to
Target (X --> Y)') +
    scale_x_continuous(name = 'Origin') +
    scale_y_continuous(name = 'Target') +
    scale_shape_discrete(name = 'Type of\nOriginal Employer') +
    scale_colour_discrete(name = 'Direction of Change') +
    geom_point(
      mapping = aes(
                  y = target_rank,
                  x = origin_rank,
                  colour = change_type,
                  shape = origin_type),
      alpha = 1/2) +
    coord_fixed(ratio = 1)
```



```
changes %>%
  left_join(
    y = employers %>%
      select(employer_id, rank) %>%
      rename('origin' = employer_id),
    by = 'origin') %>%
  rename('origin_rank' = rank) %>%
  left_join(
    y = employers %>%
      select(employer_id, rank) %>%
      rename('target' = employer_id),
    by = 'target') %>%
  rename('target_rank' = rank) %>%
  filter(
    (origin_type == 'tech' | origin_type == 'large') &
    (target_type == 'tech' | target_type == 'large') &
    (2008 <= year & year < 2011)) %>%
  ggplot() +
    ggtitle('Changes among "medium" and "large" employers\nfrom Origin to 
Target (X --> Y) from 2008 to 2010') +
    scale_x_continuous(name = 'Origin') +
    scale_y_continuous(name = 'Target') +
    scale_shape_discrete(name = 'Type of\nOriginal Employer') +
    scale_colour_discrete(name = 'Direction of Change') +
    geom_point(
      mapping = aes(
                  y = target_rank,
                  x = origin_rank,
                  colour = change_type,
                  shape = origin_type),
      alpha = 1/2) +
    coord_fixed(ratio = 1)
```



We notice that many changes between *tech*nology companies and *Large* law firms involve technology companies that employ rather more patent attorneys and are found on the higher spectrum, when considering the number of patent applications. This indicates that technology companies that provide more employment opportunities for patent attorneys also attract more patent attorneys from lager law firms than those who do not.

This concludes our small introduction into the visualisation of the CAPPA dataset. We have shown how to interact with the dataset through the *tidyverse* library for the R programming language and how to visualise the most important aspects of the dataset. We hope that the dataset serves the research into patent attorneys and their career paths.

5 References

[1] *CAPPA - Career Paths of Patent Attorneys*, DOI 10.5281/zenodo.3265385, https://zenodo.org/record/3265385#.XR3fzi2Q3mE.

[2] *`tidyverse`*, https://www.tidyverse.org.