UNIVERSITY OF SOUTHAMPTON

MASTERS THESIS

USPTO Patent Network: A statistical exploration of structural differences between examiner and applicant citations

Author:
Alun MEREDITH

Supervisor:
Dr. Markus BREDE

A thesis submitted in fulfillment of the requirements for the degree of Data Science MSc

in the

Agents, Interaction and Complexity
Electronics and Computer Science

September 16, 2016

UNIVERSITY OF SOUTHAMPTON

MASTERS THESIS

USPTO Patent Network: A statistical exploration of structural differences between examiner and applicant citations

Author:
Alun MEREDITH

Supervisor:
Dr. Markus BREDE

A thesis submitted in fulfilment of the requirements for the degree of Data Science MSc

in the

Agents, Interaction and Complexity
Electronics and Computer Science

September 16, 2016

UNIVERSITY OF SOUTHAMPTON

Abstract

Faculty of Physical Science and Engineering
Electronics and Computer Science

Data Science MSc

USPTO Patent Network: A statistical exploration of structural differences between examiner and applicant citations

by Alun MEREDITH

Patent citation networks describe the organisation and evolution of innovation. Here we conduct a statistical analysis of the structure of the USPTO citation network. We show that the network is currently log-normally distributed but that there has been an evolution to this state. We also separate citations based on the role of the individual citing it and show structural differences in how Examiners and Applicants make their citations and how disparate these two sub-networks are. We present evidence that the networks have evolved over time shifting to one dominated by the applicant citations.

Acknowledgements

Thanks to my supervisor Markus for all his mentor-ship over the last 3 months and everyone involved in the Data Science MSc for putting together a fantastic and inspirational course, as well as those studying with me, without whom I wouldn’t have learnt nearly as much.

Contents

List of Figures

List of Tables

List of Abbreviations

USPTO	**U**nited **S**tates **P**atent **O**ffice
XML	e**X**tensible **M**arkup **L**anguage
SGML	**S**tandard **G**eneralised **M**arkup **L**anguage
ICE	**I**nternational **C**ommon **E**lement

Chapter 1

Introduction

Patent networks are intrinsically linked to innovation and the economy. Although work has been done to extract economic meaning from these networks, economic growth through innovation is not well understood, and the value of a patent within these networks is often simplified to simply be the number of times they are cited.

Patents are a legal document claiming ownership of innovation. Through this process patent citations are legal obligations to reference 'prior art' any of which are initially missing from the application are added by the patent examiner. The objectivity of the examiner and legal basis for citations is what differentiates them from other bibliographic citation networks such as academic citations.

Patents and by proxy innovation can be viewed as a combinatorial process [31]. In this way innovation like research 'stands on the shoulders of giants'. Through a greater understanding of the formation and evolution of the network insights can be gained into the process of innovation which impacts econometrics and companies trying to identify areas to direct their research alike. Many believe that over the last ten years we are in the middle of technological information revolution [5] and understanding the extent and impact of this on innovation.

The US patent office database offers an opportunity for analysis as one of the largest patent collections openly available. Its large scale encompasses all the patents granted in the US from 1986 - 2015 in a structured although messy form, with more than 5 million patents and 110 million citations over the 30 year period. The scale of the database and large number of features allows for unique insight into the structure of patent citation networks although along with the messiness requires big data solutions to achieve this. The USPTO network is relatively enclosed network with 84.5% of its citations being internal since 2005 relative to a large scale study of academic citations which had 19-36 % internal citations [18].

The aims of this project is to use the USPTO database in order to conduct some exploratory analysis and gain some statistical insight into the structure and evolution of the network and how it may be changing over the lifespan of the network, in particular the last 10 years.

1.1 Dissertation Structure

This dissertation is organised as follows: In Chapter 2 there is a review of the literature and history of patent network research. In chapter 3 the data pipeline is explained, describing the processes by which the data has been parsed, cleaned and feature engineered before analysis could be done.

Chapter 4 describes the analysis and discussion in a narrative flow, starting with some overviews of the network and reproduction of the work of Valverde et al. before analysing the coloured network of two classes of citation based on who contributed the citation to the patent. In chapter 5 the basic results are reviewed, implications and further work discussed.

Chapter 2

Literature Review

2.1 History

The study of information networks has seen a growing interest in the past 20 years as the networked systems become more and more popular and integral to modern society through revolutions in communication with email and mobile phones, revolutions in content consumption with the internet and revolutions in system architecture with the internet of things and agent based systems.

Until recently many of these networks were treated as variations of random graphs, where edges (links) are attached to a set of vertices (nodes) sequentially through a random process [11, 10]. However many networks have been shown to display characteristics not present in random graphs so different models with different attachment mechanisms are required.

A ′Complex Network′ is a macroscopic description of systems in which emergent macroscopic behaviour is produced from the small scale interactions, for example the climate is emergent from the interaction of different weather systems [20]. These networks are loosely divided into two main classes: ′small world′ and ′scale-free′ networks. Small world networks are known for their short maximum path length between two nodes, producing characteristics like the ′Six Degrees of Separation′. These networks are highly clustered with long distance links between those clusters [27]. Scale-free networks however typically are not strongly clustered and are known for the distribution of edges across nodes (degree) following a power-law $f(x) \sim x^{-\alpha}$, this is sometimes also known by Zipf′s law or a Pareto distribution [32]. This distribution has the property that there is no characteristic scale for the network. A typical scale means that each node has about the same consistent numbers of links producing structures similar to lattices so having no typical scale produces very unstructured networks.

Scale-free networks are often characterised by a ′rich get richer′ phenomenon, known as the Matthew effect [15]. Both Price and Barabási present a ′preferential attachment′ mechanism to model this effect where new edges are attached to nodes with probability as a linear function to the number of edges of that node to produce a power-law distribution [1, 17].

2.2 Bibliographic Networks

Bibliographic networks are networks of authors and/or their works. Examples of these include collaboration networks, where links are formed from instances of author collaboration and citation networks where authors cite other works to be related. Academic citation and collaboration networks were some of the earliest identified as complex networks [1].

The statistical nature of many complex networks, including patent networks is still being debated. This forms two fundamental questions: what is the functional form of the underlying distribution and what mechanisms form this distribution, these primarily try to identify power-law distributions and preferential attachment mechanisms although more recently research has extended preferential attachment with additional mechanisms such as ageing terms and fitness models. Answering the first question is difficult as log-normal curves can be very similar to power-law curves under the correct parameters. Clauset et al. describes a method for differentiating between these distributions using bootstrapping methods to discount unlikely distributions and log-likelihood ratios to compare two functional forms [6]. Although preferential attachment produces power-laws, a power-law is not sufficient to show presence of preferential attachment. Power-laws have been shown to be produced by other mechanisms such as a multiplication of many random processes [16]. Log-normal distributions have also been shown to be produced by random multiplicative processes but also sub-linear preferential attachment [14]. Therefore preferential attachment must be identified independently of the distributions functional form.

In academic citation networks there has been recent research in the mechanics behind citations such as the propagation of errors which suggests 70 - 80% of citations are copy and pasted from a secondary source [19] or the study of redundant edges to show that the majority of references (70%) are secondary [8]. Scientific citation networks have been shown to have characteristics consistent with a preferential attachment mechanism and Weibull ageing term [3].

Multi-dimensional networks are a generalisation of networks combining networks of nodes with multiple classes of link. The study of these networks has become one of the fastest growing areas of research in network science offering generalised diagnostics which can account for differences in nature between the links as well new insights [13].

2.3 Patent Networks

The main differentiating factor between academic citations and those found in patent networks is the stricter legal scope of citations in patents. Patents have a legal requirement to cite all 'prior art' from which the patent is based. This is a more narrow definition of a citation but also more exhaustive because if a 'prior art' is not cited in the application ideally a patent examiner will add it to the document. There is a larger body of research on academic citation networks due to the inherent domain knowledge researchers have with the field and because it was among the first to be identified as scale-free [1].

In 2007 Cs'ardi et al. published the first paper examining the US patent office database from a graph theoretical perspective [9]. They did this by by applying a basic model that assumes the attractiveness of a node (rate at which new nodes attach to it) is a function of the age and number of citations of the node. Normalising for the growth of the patent network over time they found the total attractiveness of the system over time could only be replicated with a super-linear preferential attachment model. They also explore the idea that an increase in the number of citations per patent over time has been coupled with a fundamental change in the structure of the network. Finding that the level of stratification starts to increase in alignment with the higher citation rates.

Valverde et al. [24] also analyses the data-set from a graph-theoretical perspective suggesting a form for the extended power law for the in-degree distribution, this functional form converges to a power law for high orders and an exponential for low orders. They also explore the clustering and modularity of the system. The clustering coefficient being approximately inversely proportional to the number of citations, suggesting a hierarchical structure. Finally they disagree with Cs'ardi et al. finding a linear preferential attachment.

A recent study builds exposes the danger of relying on simple models studying the patent network as a whole [12]. Bernard Gress (2009) investigated the ratios between citations given and received in different technology groups. High number and diversity of citations given was treated as an indication of generality and number of citations received of productivity and originality. He then compared these measures and how they varied over time for different technology categories. He primarily concludes that these categories are fundamentally different therefore research needs to take this into account.

Byungun Yoon [30] built a patent network from weighted term similarities of patent documents rather than bibliometric citations; using standard bag of words methods and comparing these measures to analogies in citation networks. Through inspection they argue that the centrality of their network yields a more relevant approximation of impact because it is less biased by age and preferential attachment mechanisms.

As patents are a representation of innovation many studies try to measure the impact of a patent. Network based techniques often limit themselves to using the number of citations received as a measure of impact, however there is a lot of debate as to what extent this measure is valid and how these can be improved.

2.3.1 Innovation Networks

There is a body of research exploring the analogy between the evolution of innovation and biological evolution. The premise is that each invention is built from the recombination of previous inventions. The two models have their differences, for example there is a limited concept of 'death' in innovation as very old patents may still be cited by new ones and it is hard to think of bibliometric patent networks as direct lineages.

Yeoun et al. [31] explores this idea of invention as a recombination process by looking at the use of technology codes in patents as a proxy for novelty. Technology codes map the technological niche of a patent into categories and subcategories. Patents can have combinations of technology codes. They show that as the number of patents increases the number of new codes being generated falls off while new code combinations maintains a power-law, concluding new technologies has a minimal role relative to recombination. They also show that 40% of patents use existing combinations vs. new ones suggesting these are incremental improvements.

Technology code combination distributions do not age in the same way as bibliometric citations, codes appear not to age with 99% of codes being used at least once every 7 years.

They also look at the dissimilarity of the codes as a proxy for novelty. If a patent is used in a very different field from its parents it is argued that it is more likely to be a bigger leap in novelty. Categorising patents as either narrow or broad and using count as a metric, we only get a sense how novelty has changed with time and not any of the network factors which may be present here, such as the distribution of novelty could be a power law or the degree of novelty can be a measure of linkage between clusters.

The limitations of the evolutionary analogy are loosely addressed warning that citations in patents aren't directly related to lineage but about carving a legal niche and there being no good metric of fitness for patents.

Buchanan et al.[4] glosses over some of these limitations, using the number of citations a patent has as an "impact" metric, a proxy for fitness. Prior art citations also function as a proxy for combinatorial lineage. They tell the story of the most cited patents in the network over the past 30 years and show that such a distribution of citations cannot come from random natural selection and therefore must be due to adaptive selection in an evolutionary model of innovation. This argument is an evolutionary perspective on the random network vs. preferential attachment network differentiation.

They focus on showing this idea more robustly incorporating a multitude of normalisation techniques and simulating a null hypothesis random network model by sampling existing data, rather than building a clean model from scratch. Observing the familiar hallmarks of a fat-tailed distribution they conclude that these "superstars" high impact is due to adaptive features, however they do not address the role of preferential attachment here, how many of the citations received are due to 'rich getting richer' mechanics or due to the intrinsic quality of those innovations.

Their paper also investigates the dissimilarity of technology codes as a proxy for novelty making the claim that large leaps in novelty are responsible for the largest "impact" patents. However because its scope is only looking at the 20 most cited patents falls short of being able to make such a general argument about the network as a whole.

Finally Arthur et al.[2] makes the most direct contrast between the search process of a genetic algorithm and the evolution of technology. In their paper they simulate an evolving population of logical circuits starting from simple logic gates in order to meet a selection of logical needs. Analysing the evolution of the population and resulting network.

They find many of the typical evolutionary features present such as building blocks being formed as intermediary steps to complex solutions i.e. the building block hypothesis. Sub-optimal solutions slowly become extinct after better solutions emerge.

Complex features are also observed such as a loose power law distribution of edges in the network and avalanches of redundancy as new technologies replace old ones and their dependencies, the size of these redundancies follows a power law showing self-organised criticality.

The paper incorporates a standard genetic algorithm; ignoring many of the observed differences between natural selection and the evolution of innovation, such as holding a finite population therefore incorporating the "death" of patents as they are replaced, and use of random selection. Both of these were shown earlier to not be accurate in the patent network, despite this they achieve results similar to observed patent networks, further research could conclude that many of the differences observed naturally arrive from a simple model for example patent ageing could be a result from the saturation of combinatorial space around older technologies.

2.4 Conclusion

In conclusion there is ongoing research into the statistical and structural nature of the patent network. Although there are some methods designed to analytically differentiate between functional forms it appears that investigations using these methods are absent from the patent network and there are contradictory conclusions into the nature of preferential attachment in the network.

Finally there are significant structural changes between technology groups and evidence of structural changes within the network over time.

Chapter 3

Data Pipeline

This chapter describes the data pipeline, from downloading to analysis as shown in figure 3.1. The data is hosted in bulk on the USPTO website [23], after parsing the html of the website to extract the download urls and download the data it is parsed into two structured tables, one for information about each patent and another for each citation within each patent, i.e. a collection of edges along with some additional information to ease cross referencing strain later. After parsing the data is cleaned and tested to ensure a reasonable level of quality and to get a qualitative understanding of the accuracy and limitations of the data-set. From this point some analysis can be done on the data such as degree distribution or number of patents granted each year, for more in depth analysis feature engineering is done either in memory using R or through map-reduce and other database frameworks, for this reason the data is migrated from csv files to a database. Finally analysis is done on the feature engineered data, primarily this is done through using database frameworks to summarize or sample the data before exporting to R for the main analysis and graphical representation.

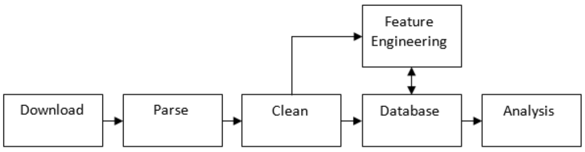


FIGURE 3.1: Flowchart of the data pipeline

3.1 Raw Data

The USPTO offers one of the largest openly available patent networks with images of their patents and trademark documents available since 1796; as text through optical character recognition since 1920 and more recently from 1976 structured full text data (excluding images and diagrams) is available. A subset of the full text data is the bibliographic data; all of the front page information such as date granted, the title of the patent and the citation information.

This data-set from January 1st 1976 to December 31st 2015 is 110Gb in size in the form of a structured documents of varying formats: From 1976 to 2001 a proprietary text format is used, from 2002 onward the USPTO follows the Patent Grant International Common Element (ICE) document type definition (DTD). While ICE provides international consistency it revised frequently, initially using sgml (ICE 1.5 and 1.6) in 2005 switching to XML for versions 4.0 - 4.4 [29]. One patent from each of the three major formats (text, sgml and xml) are included in appendix B, C and D respectively.

Up until 1996 each year contains one file containing all the patent information in that year in a single document. After this point there is a separate file for each week an associated file listing all the patent numbers present in that file and a summary file containing extra information such as notes about that week and the number of total patents granted of each type. The sgml and xml files have each patent as a single document appended into a single file.

3.2 Parsing

The above structure of the data presents a number of challenges to the parsing process. Most notably the changing formats and schema. The difficulty is parsing these in a consistent way, to minimise any systematic differences between the formats.

A single parsing function for all formats was developed; using the same functional structure to process each format (Appendix A). The core logical loop for parsing each line of data, shown in figure 3.2a, extracts the tag and contents of each line, where the tag is the expression identifying what the contents is, e.g. <date>20011004</date> indicates the contents between the brackets is a date. The function then maps the tag to an action and takes that action. This minimizes differences between formats as long as a semantically equivalent mapping can be found for each.

The parsing function builds up two tables, one for the patent level data and one for the citations. Initializing each of these as empty vectors, adding patent information and citation information to those vectors and flushing them by appending them to a csv file at the end of each citation or patent section in the document 3.2b. Some tags are not unique, such as date referring to the date a patent was granted the date a patent application was made or the date a citation patent was granted. To resolve this issue the function identifies the region of a patent document it is currently in and therefore how to interpret uniquely a non-unique tag. It also takes the opportunity to record the degree of each patent with a simple counter.

The second philosophy of this method of parsing makes it easy to adjust the variables being recorded by adding the tags to the mapping vectors (and name), this flexibility handles changes in schema over time and parsing new variables.

The variables parsed for each patent were: Patent number, degree, date granted, main and further classification codes. The variables extracted for each citation were: Parent patent number, citation patent number, citation patent granted date, country of the citation and 'cited by' which distinguishes between two classes of individual producing the citation. In this way the citations are in a tall format, in accordance with Hadley Wickham's principles of 'tidy data' [28].

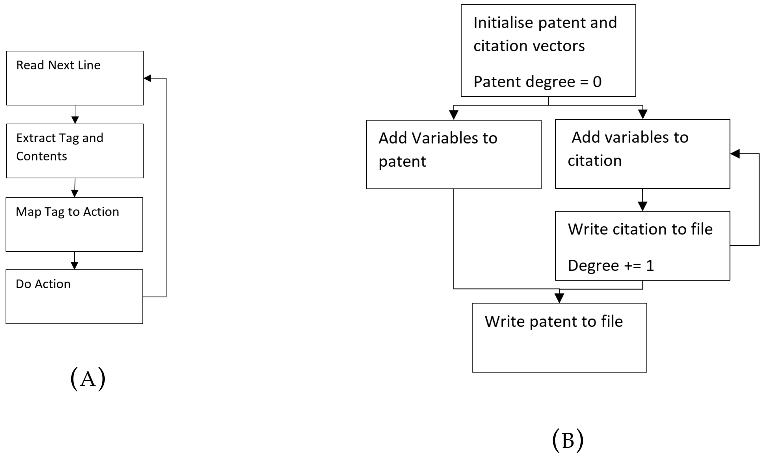


FIGURE 3.2: Flowchart showing the structure of the document parsing function. (A) describes the logical loop while (B) describes the structure in which a whole patent is parsed into a patent Vector and a citation table

3.2.1 Cleaning

After parsing the data it is cleaned to remove some sources of systematic errors. Dealing with big data, considerations must be made when cleaning the data with respect to the impact compared to the difficulty performing that cleaning. The data is tested against some reference points to assess its consistency, both internally between different formats and schema and externally against other sources.

The categorical factors such as 'citedBy' and 'Country' are cleaned by generating a list of possible values that factor can take and excluding values outside those levels. The dates are in a numeric format "YYYYMMDD" (Where Y,M,D refers to digits representing Year, Month and Day respectively). This is parsed into a date object represented by the number of days since the 1st of January 1970. This allows the dates to be more directly compared as they can be subtracted to get time differences. This process also isolates any values which do not conform to the expected format and returns NA. Using 2001 as an example only 1 date failed to parse this way. Dates on the patents must also in the range of 1976 and 2015. Citation dates can be lower as patents may cite other patents which are not in our data-set.

Technology codes and patent numbers follow specific schema, US patent numbers are numeric codes which may have a preceding letter (D | RE | PP | H | T) indicating their class, e.g. 'D' represents design patents. While it can be ensured that all US patents use this schema there are foreign patents with many different schema, it is impractical to clean for all these foreign patent schema. In practice not cleaning these foreign patent numbers is unlikely to have an impact on the analysis as the primary purpose for cleaning these patent numbers is to improve matches between citations and patents, which is only necessary for internal citations.

The patent numbers under citations had several encoding inconsistencies with those under the Patent: punctuation, white-space, additional filler '0' characters and additional digits at the end of the number when parsed from the text format were all cleaned. In 2001 94.5% of the US citations after

1976 were matched with patents parsed.

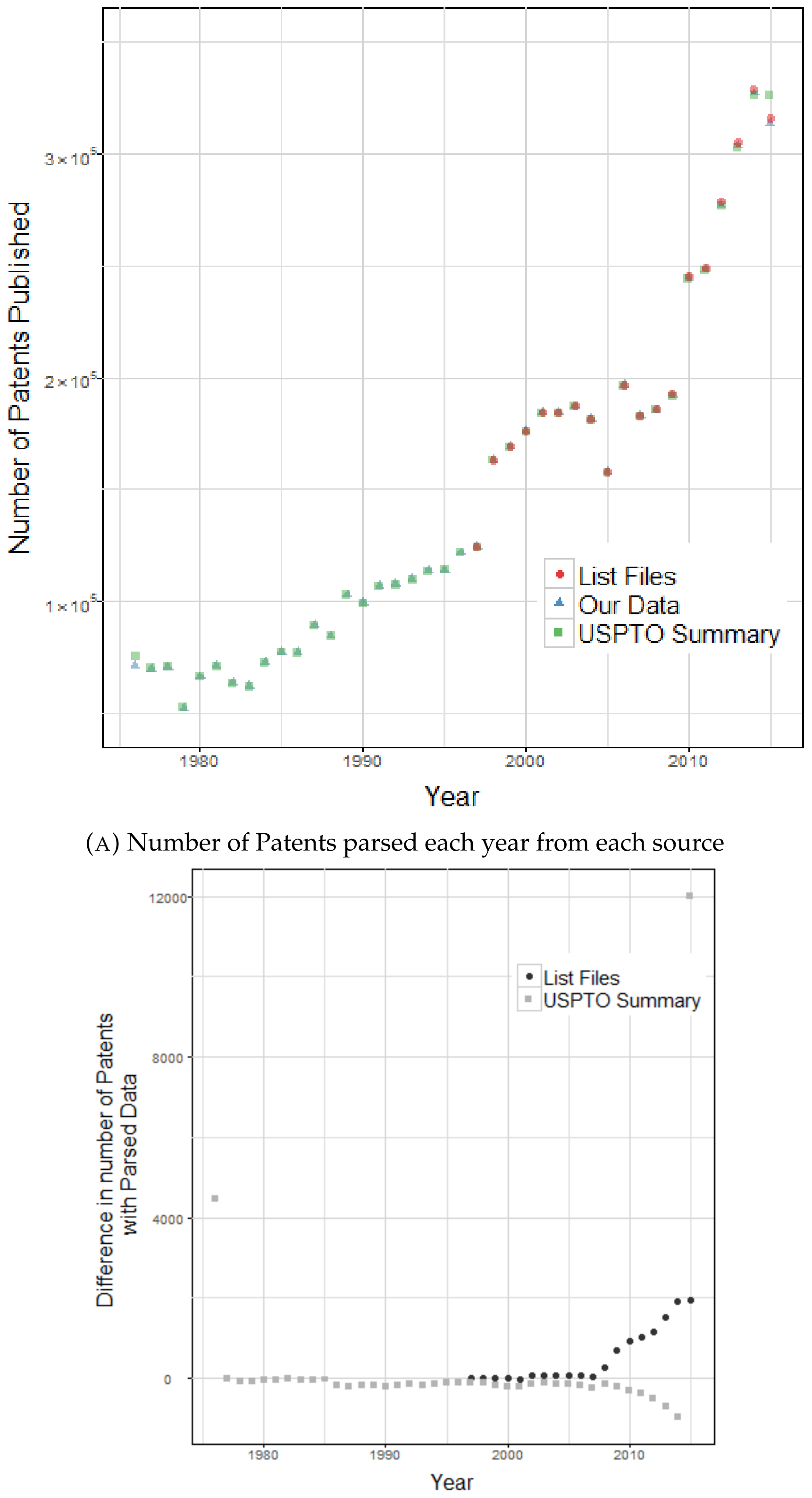
There are a variety of sources of NA values in the dataset. In the parsing phase missing data in the raw dataset is returned as NA and when reading the outputted csv each value must be the expected type (e.g. Order must be an integer). The cleaning phase also produces NA values when a factor value doesn't fit into allowed levels. For example in 2001 4 citations were completely NA, 155 numeric dates are NA and 897 dates when parsed into date format are NA. That year there are a total of 2,609,494 citations so NA values are negligible in quantity.

The parsed and cleaned data can be compared to those published by the USPTO [21], as well as the number of patents in the '.lst' files which mirror the raw files documenting the patent numbers present. If accurate, differences between these files and parsed patents represent the patents that the parsing function has failed to process. '.lst' files are only available from 1997 onward (figure 3.3). There are two large differences between patent counts at either end of the data-set; in 1976 and 2016. The difference in 2016 is due to two weeks of data having access denied during the download stage but all the weeks are present in 1976 so the difference here is unknown. Due to the relative size of these errors these years have been omitted from patent count based analysis. Before 2008 the number of patents parsed was equivalent to the USPTO summary statistics, since then however the parsed data has significantly more patents than present in the summary statistics. Manually looking up some of these additional patents reveals them to be complete non-duplicated patents so it is not clear why they may be missing. There have always been a small number of patents in the list files not parsed, however until approximately 2010 these numbers were fewer than 100 but have grown to around 1% of the data, this aligns with the summary statistics divergence suggesting they also experienced these problems to a more severe degree. The difference here doesn't line up with the introduction of XML (2005) but may be due to some of the ICE schema changes or errors in the structure of the document (trying to use an XML parsing library on these years yielded a lot of missing data due to failure to structural errors in the XML).

The second reference point for parsing the data is internal consistency. In 2001 both text and sgml data files are present. This gives the unique opportunity to parse both and compare the results. We found that 95 / 184078 patents are present in the text parsing but absent in the sgml, but none are present in the sgml and absent in the text format. Manually referencing the patent numbers with the USPTO database reveals they are valid patents missed absent from the sgml document rather than missed from the parsing method. When these patents are taken into account all citations are present in both formats. Overall these differences confirm that although there are some differences in the raw data, the parsing function is treating differing data formats equally.

3.3 Database processing

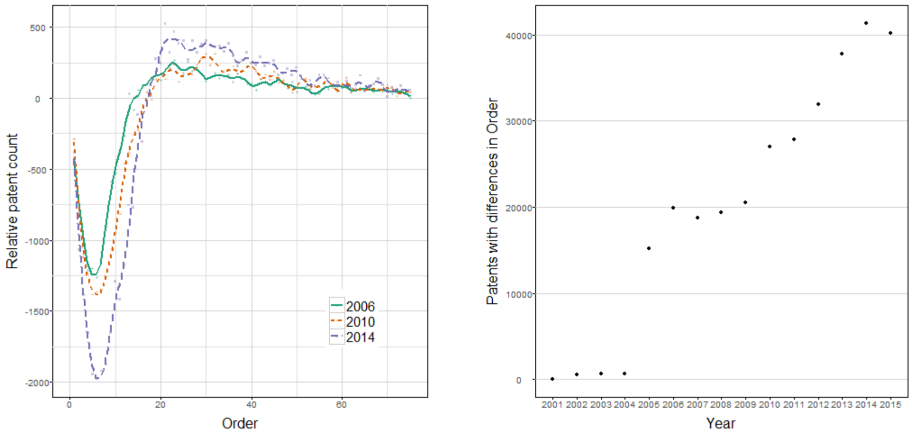
Once the data has been parsed and cleaned some analysis can be done on it, but most analysis requires feature engineering. Because the data does not fit in memory this can be done in batch but often the data must be



(A) Number of Patents parsed each year from each source

(B) Difference between number of Patents parsed from different sources and parsed data

FIGURE 3.3: The number of Patents granted each year according to 3 sources: Parsed data; a USPTO report and the .lst files.



(A)

(B)

FIGURE 3.4: Number of patents with different order based on parsing method. (A) against order for the years 2006, 2010 and 2014 and (B) for each year.

reordered or regrouped, for this reason a database was used. Mongodb was chosen for this as it is designed around big data with its map-reduce and aggregation frameworks making these feature engineering operations to be done in parallel, having javascript and JSON as primary interfaces allowed for ease of use.

Using a map-reduce the order for each patent was calculated as the number of citations matching each patent. There are some differences between these orders and the ones produced with the parsing function (figure 3.4). We can see that these differences are only non-negligible after 2005, when the XML format is introduced. Manual inspection shows that the map-reduced orders are the ones which are incorrect but it is not obvious what the cause of the error is. The map-reduce over represents small orders and under-represents large orders, with the same total number of citations parsed. This suggests that the map-reduce may be interpreting a single patent as multiple different patents. This error only becomes significant in the most recent years but represents 1% of the data in 2015, so caution is advised for these years.

In summary the parsing and cleaning of the data have focused heavily on maximizing the consistency of the data parsed from the design of the parsing function to the cleaning steps. We have analysed these results and find that it out-performs USPTO summary statistics and non-negligible errors appear to be limited to 1976 and the last few years of the data-set. It is important to know where these errors are and be cautious when analysing them but we can't dismiss these sections of the data-set.

Chapter 4

Analysis

4.1 Data Summary

After parsing and cleaning the data is stored in two tables, a patent table and a citation table as shown in tables 4.1 and 4.2. Definitions of variables are described in section 3.2.1.

The dataset covers the time period from 1976 to 2015, over this period 5,803,302 patents have been processed along with 114,585,378 citations. Since 2001 where the data dictating who cited each citation was initially released 25.7% of citations have been the Examiner and the rest being other sources. Since 2005 where accurate country codes for foreign patents have been released 85.5% of citations have been internal, i.e. to other US patents.

TABLE 4.1: First 6 entries in parsed and cleaned patent table, 1976

	Patent	Date	Order	Date2
1	RE28671	19760106	8	2196
2	RE28672	19760106	7	2196
3	RE28673	19760106	3	2196
4	RE28674	19760106	5	2196
5	RE28675	19760106	13	2196
6	3930271	19760106	2	2196

TABLE 4.2: First 6 entries in parsed and cleaned citation table

	Patent	Citation	Date	CitedBy	Country	Date2
1	D607176	534632	18950200	cited by examiner	US	-27362
2	D607176	D28957	18980600	cited by other	US	-26146
3	D607176	D45899	19140600	cited by other	US	-20303
4	D607176	D59909	19211200	cited by examiner	US	-17563
5	D607176	D96223	19350700	cited by examiner	US	-12603
6	D607176	D96224	19350700	cited by examiner	US	-12603

4.2 Network Summary

In this section some general summaries of the network are computed. Some of these summaries have been previously carried out by other authors allowing validation of this/their analysis.

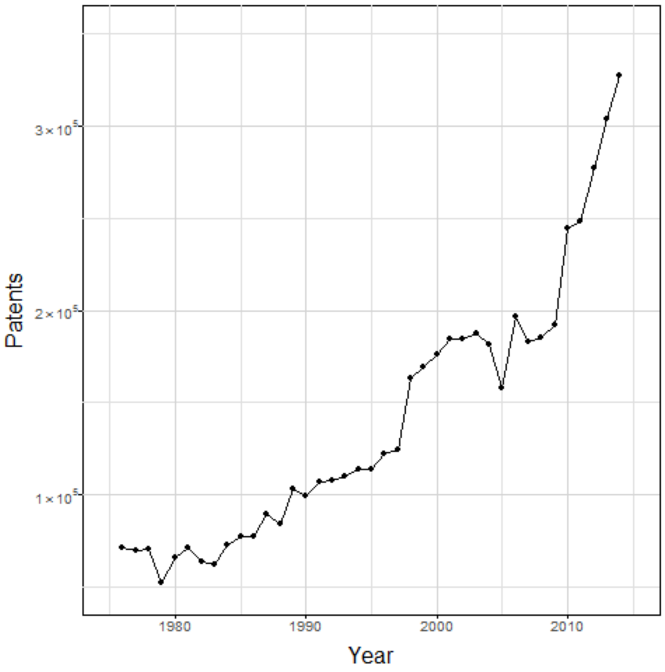


FIGURE 4.1: Number of patents granted each year from 1976 to 2014

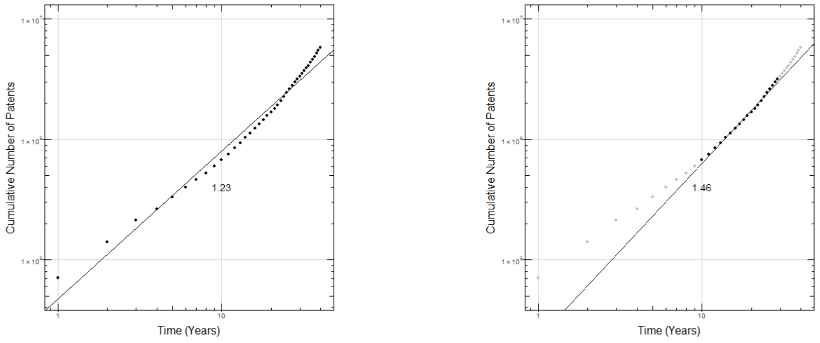
4.2.1 Patents per Year

Each patent was grouped by the year in which it was published and frequency computed. The year 2015 was omitted due to the presence of two missing months of data due to a download error on the part of the USPTO as discussed in the previous section.

Figure 4.1 shows the number of patents granted each year on linear scales. There are two main features: a large jump between 1997 and 1998 and a transition to very rapid growth after 2009. Note that these events don't align themselves with changes in the data format which occur from 2001-2002 and from 2004-2005. This transition may be caused by socio-economic or political effects as it occurs directly after the 2008 recession, regardless the continued growth after 2009 indicates a change in the evolution of the network. Valverde et al. [24] conducts a similar analysis, over the years of 1986 to 2005, which has the same general shape but some minor differences; most visibly in the size of the gap between 1997 and 1998 which is bigger in this analysis than the former. In the previous chapter patent counts per year were compared to other sources, USPTO summary statistics and the 'lst' files, consistency of these points to some possible difference in methodology or parsing errors in the case of Valverde's paper.

Valverde et al. claims that this metric of patents granted per year follows a power law distribution. To demonstrate this they plot the cumulative number of patents on log-log scales and make a linear fit. To replicate this analysis patent counts per year the cumulative number of patents in the network is computed and plotted on logarithmic scales. A generalised linear model is fitted using least squares regression and a log-log transform.

Figure 4.2b shows the cumulative number of patents granted over time starting in 1976 and ending in 2014. Time is given as number of years from the origin. Figure (A) shows a power-law model fitted to the entire dataset



(A)

(B) G

FIGURE 4.2: The cumulative number of patents granted from 1976 to 2015, plotted on a log-log scales. Fitted with generalised linear model of the form $y = x^{\alpha}$ using (A) all the data and (B) only the data in over the years 1986 to 2005.

and its exponent whereas figure (B) conducts this method only on the years 1986 to 2005. The cumulative distribution is often used to identify power-law behaviour because often the tail of the distribution has rare events with large amounts of statistical noise, the cumulative integrates the noise into a smooth function. If $p(x) = x^{\alpha}$ the cumulative distribution is also a power law with smaller exponent:

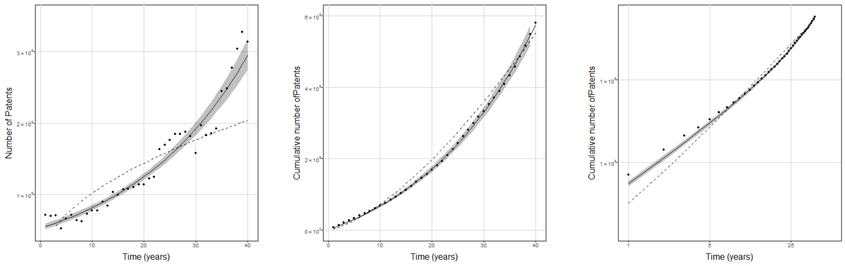
$$P(x) = \frac{1}{\alpha - 1} x^{(\alpha - 1)}$$

While not necessary in this instance due to the absence of a noisy tail, a CDF was used to closely replicate the Valverde methodology and is general best practice. Although using least squares regression on a log-log transformed cumulative distribution is not the most rigorous method for fitting power-law distributions introducing small bias to the exponent, this appears to be the methodology used in the Valverde paper (although it is not clear) additionally the exact value of the exponent is not important only the general shape of the model.

From 4.2a there is a clear divergence from the fit over the first few years. In Valverde's paper it appears that in their paper they may disregarded these first few points as outliers, this can be shown by figure 4.2b which omits these points yielding a very similar exponent (1.46 vs. 1.45). Since 2005 where their data-set ended new data-points do not fit this model well. This is more apparent in figure 4.3a which shows this same fit on the non-cumulative plot.

Using the same methodology as before an exponential fit to the data is produced, 4.3. This is shown on figure 4.3 with a 95% confidence interval along side the original power-law fit.

It is clear that a power-law fit doesn't fit the data, it is plausible that an exponential model may be more accurate (fig.4.3). However the presence of two major shifts in the distribution in 1997/8 and 2009/10 indicate that large external effects are causing changes in this data-set. These changes are significant enough that without taking them into account or aggregating over a longer period of time or smaller timescales making a statistically significant claim about the underlying distribution here would be difficult.



(A) Patents granted per year

(B) Cumulative number of patents granted

(C) Cumulative patents granted on a log-log scale

FIGURE 4.3: Power law and exponential fits for cumulative number of patents published each year

4.2.2 Degree Distribution

The degree distribution is the frequency distribution of patents with respect to the total number of citations they make. It can inform how new citations are made and copes well with the evolution of the network over time as analysing years separately they do not need to be normalised. This has been separated into different years and the years 1984,1992, 2002 and 2012 have been on a log-log scale and fitted with a generalised linear model as before to replicate the third plot in the Valverde paper 4.4.

The exponent of the power-law is increasing over time, indicating both more citations over time but more extreme values also. These values however are quite different from those found by Valverde who found an exponent of -4.0 in 2002. This may be due to a number of methodology differences, most likely that our analysis includes foreign patents. We see that the frequency increases to a peak giving the distributions a mode of 6 to 8 depending on year, after around 10 to 20 this appears to converge to a power law.

As before some best practices have been ignored in order to reproduce the Valverde et al. paper better. Valverde claims that this distribution follows an extended power law of the form $P(k) \sim (k + k_0)^{-\gamma}$, where k is the degree and k_0 and γ are constants. An extended power-law converges to a power law of $k >> k_0$ and converges to an exponential for $k << k_0$, it is therefore functionally similar to the popular model of a power-law with an exponential cut-off. They do not however state how they make their fit or conduct any tests to identify the efficacy of such a fit.

Considering the 4 different years presented in figure 4.4 and the 4 most popular discrete distributions described in table 4.4, each of these functional forms is fitted to the cumulative frequency order distribution by maximising their likelihood functions. The cumulative distribution is used as it mitigates noise caused by low frequency events. This is done while optimising the cut-off value x_{min} for that distribution. Optimising x_{min} uses a method proposed by Clauset et al. which searches over a range of values and minimises the difference in probability distributions of the data and the fit [7]. The difference metric used is the Kolmogorov-Smirnov statistic; the maximum distance between the two distributions.

The accuracy of these fits is assessed using a bootstrapping algorithm, the distribution is sampled with replacement n times, each time the KS statistic is computed between this sampled distribution and the original

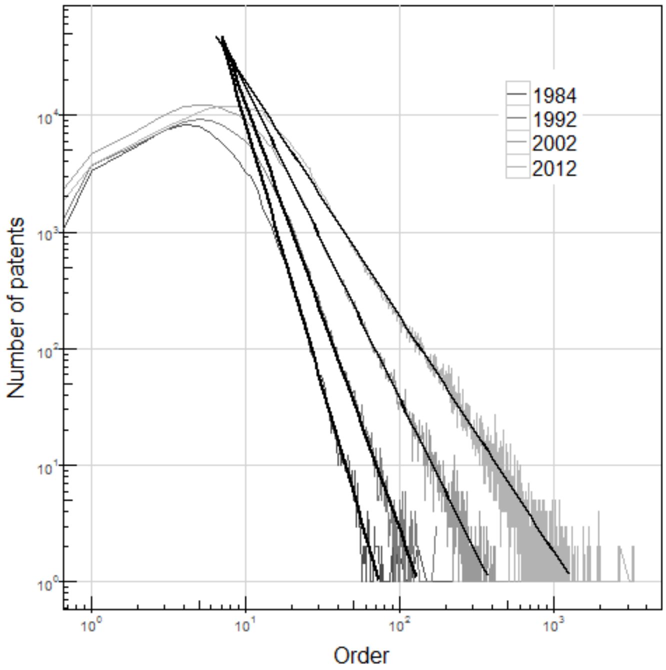


FIGURE 4.4: Degree distribution (number of citations for each patent) on a log-log scale shown for years 1984,1992,2002,2012. Fit with generalized linear model of the form $y \sim x^\alpha$ yielding exponents of minus 6.31, 6.92, 7.77 and 8.46 respectively

data. The P value is given as the proportion of sample distributions less similar to the original distribution than the fit. A rule of thumb given by Clauset et al. states to achieve an error of approximately ϵ the number of samples used should be at least $\frac{1}{4}\epsilon^{-2}$. 2500 samples were used to achieve $\epsilon \approx 0.01$ [6].

From figure 4.5 we see the increasing gradient of the power-law exponent over time, additionally the distribution appears to become less straight and has greater changes in shape at the tail, which was harder to identify using the non-cumulative plot 4.4. As the distribution becomes less straight the power-law fit becomes less accurate especially failing to capture the shape of the tail. The Poisson and exponential distribution never successfully fit the main body of the data, the Poisson distribution especially fails in this regard fitting either poorly in 1984 or to increasingly small subsections of the tail in later years. The exponential function follows a similar trend, fitting the main body of the data poorly initially before trying to fit only the tail of the distribution, matching the exponential with the power-law suggests that an exponential cut-off is plausible however the log-normal distribution also appears to be a good fit, capturing some of the behaviour of the tail in a way none of the other distributions were able to.

Note that this bootstrapping procedure is conducted on a sample of 1000 observations due to heavy performance issues with the algorithm while the distribution fits use the entire dataset, this can lead to large differences between the two as different x_{min} values are optimised. For example in 2002 and 2012 the Poisson distribution has a very large x_{min} and therefore a higher p value would be expected than when it poorly fits to the whole

TABLE 4.3: Functional form and maximum likelihood estimate (MLE) of distributions tested

Distribution	$f(x)$	MLE
Power Law	$x^{-\alpha}$	$\hat{\alpha} = 1 + n[\sum_{i=1}^{n} ln\frac{x_i}{x_{min}}]^{-1}$
Log-Normal	$\frac{1}{x}exp[-\frac{(lnx-\mu)^2}{2\sigma^2}]$	$\hat{\mu} = \frac{\sum lnx}{n}, \hat{\sigma^2} = \frac{\sum_k (lnx_k-\mu)^2}{n}$
Exponential	$\exp^{-\lambda x}$	$\hat{\lambda} = \frac{1}{\bar{x}}$
Poisson	$\frac{\lambda^x \exp^{-\lambda}}{x!}$	$\hat{\lambda} = \sum_{i=1}^{n} k_i n$

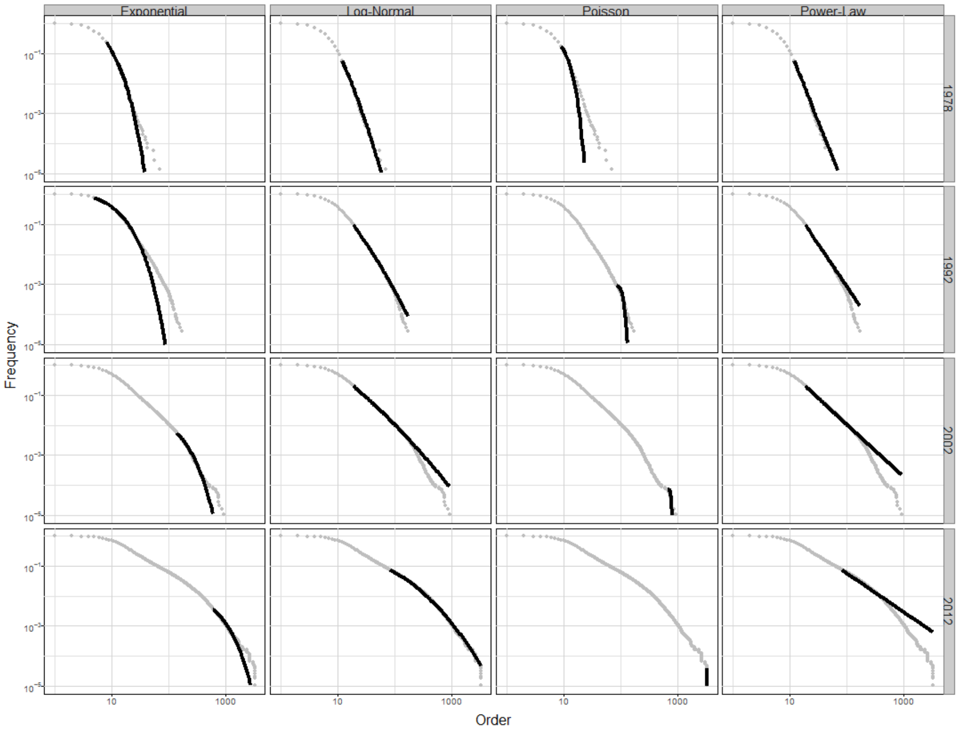


FIGURE 4.5: Cumulative density function fitted using maximum likelihood estimates for Poisson, Exponential, Log-normal and Power-law distributions

TABLE 4.4: P values fitting distributions to the degree distribution obtained through bootstrapping method on a n = 10000 sample of the data

Year	Power-Law	Log-Normal	Poisson	Exponential
1984	0.494	0.302	0.000	0.001
1992	0.820	0.934	0.000	0.005
2002	0.534	0.399	0.108	0.000
2012	0.266	0.280	0.000	0.448

TABLE 4.5: P values giving upper limit of if power-law is true (one-sided) or both power-law and log-normal distributions are equally good (two-sided) using Vuong's test

Year	One-sided	Two-sided
1984	3.87e-1	7.74e-1
1992	1.39e-4	2.78e-4
2002	1.1e-10	2.3e-10
2012	9.3e-86	1.9e-85

curve. While the small sample size relative to the size of the data may limit the insight from this test, the consistently low and often 0 values of the Poisson and Exponential distributions means they can likely be dismissed as candidates for this distribution.

While the power-law and log-normal distributions are consistently significant it is not sufficient to say that the power-law is better than the log-normal. To compare the models directly Vuong's test is used [26], this is based on the null hypothesis that both fits are equally far from the true distribution, if true the log-likelihood is expected to have a Normal distribution. This test is done using the complete data (without sampling) using the fits shown in figure 4.3.

The power law is never the more likely distribution based on this test but in 1984 it is significant to the degree that the two-sided test, indicating both distributions are equally likely has high p-value. However this likelihood of a power law being true consistently decreases over time so that in 2012 there is a 9.3×10^{-86} chance of a power-law in contrast to the log-normal distribution.

4.3 Coloured Network Analysis

One of the features parsed was who contributed the citation to the patent document. From 2001 to 2013 this is split into two categories "Cited by Examiner" and "Cited by Other". The Examiner has a duty to fulfil the legal requirements of the patent application citing any prior art which was absent from the patent previously. Cited by Other "indicates those cited in a protest, by an attorney or agent not acting in a representative capacity but on behalf of a single inventor, and by the applicant" [21]. In reality all sources of citation besides the applicant are rate, this is shown in 2014 and 2015 where the 'cited by other' field was split into two, 'cited by applicant' and 'cited by third party'. Over these two years 99.999% of these citations were 'cited by applicant' rather than the third party. For the sake of consistent analysis these have been combined again into a 'cited by other' category.

This 'citedBy' field was analysed because of the differing motivations and responsibilities of the parties involved. The Examiner with a legal and professional responsibility is likely to act differently from the applicant and many of the features that make the patent network unique are the legal responsibilities that citations have so studying this sub-network could yield more varied results than the network as a whole, especially when compared to other bibliographic networks.

4.3.1 Degree Distribution

The degree distribution of for the two sub-networks are calculated in the same way as section 4.2.2 shown in figure 4.6b and the mean degree for each of the two sub-networks and total are calculated for each year 4.6a.

The average number of citations for each patent changes very little over time for the Examiners but increases rapidly over the available data range for 'Other'. As Other dominates the Examiner the total value follows this closely this means that the structure of the Examiner network is masked. It is unclear how this relationship held over the previous years, it may be that the two sources of citation had a more harmonious relationship in the past as over the period of time from 1976 to 200 the increase in average citation over time is fairly constant and similar to the increase in time of the Examiner network.

The exponent for the degree distribution is more negative for the examiner indicating that it operates over a smaller range of values, reaching a maximum at approximates 100 rather than 1000. The Examiner distribution also appears to be a better fit in the tail of the distribution and has higher frequencies at low citation counts.

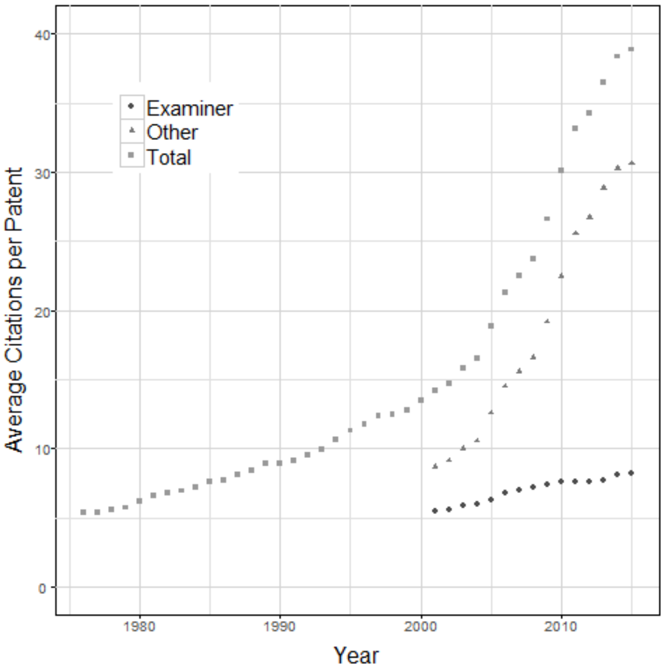
All of these features of the Examiner network relative to the Other network are not unexpected as they indicate more conservative and consistent practices which would be fairly typical of a professional. The large range of values may suggest a tendency for the applicants to occasionally over-cite their patents to an extreme.

The changes in average citation number being caused by differences only in the applicants activity could damage the idea that there is a structure of innovation itself over this time period because it suggests that the patents themselves aren't necessarily becoming more interdependent as an increase in average citations per patent otherwise suggests. However this doesn't discount the idea as the Examiner acts after the applicant so changes to the patents/innovations themselves could be accounted for by their actions alone although this is a less likely explanation.

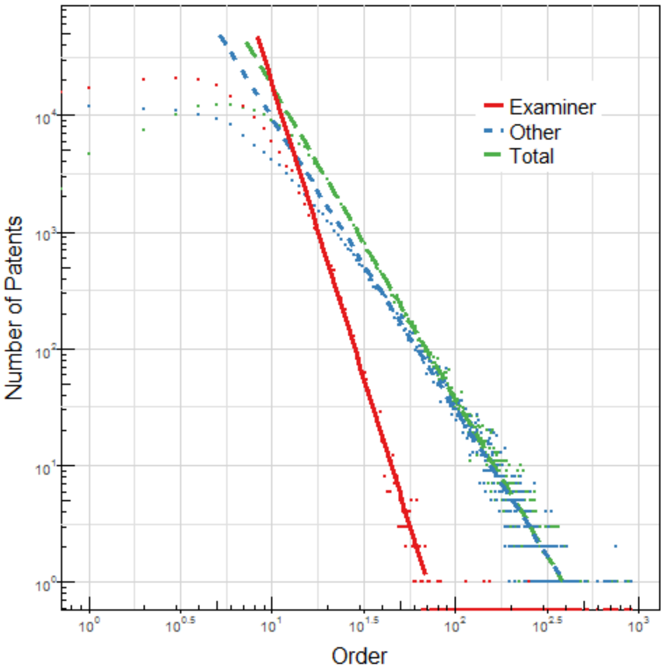
4.3.2 Correlation

In order to investigate the relationship between the Examiner citations and Other citations a number of analyses of correlation between the order of each is measured (number of Examiner citations vs. number of Other citations for each patent).

A scatter-plot of all patents from 2001 to 2015 was made using a log-log scale, comparing the number of citations made by Examiners to those made by Others, because these orders are discrete a uniform random noise of ± 0.5, this expands the data to uniformly occupy the area which rounds to its discrete value. This allows the density to be seen more clearly. The scatter-plot is made with a sample of 100,000 observations and a linear fit is added using least square regression to represent the correlation between the two.



(A) Average degree by citedBy for the years 2001 to 2015 and average total degree from 1976 to 2015



(B) Degree distribution by citedBy on a log-log scale, fitted with generalised linear model of the form $y \sim x^{-\alpha}$, here total refers only to the years 2001 to 2015

FIGURE 4.6

To more clearly represent the density of the scatter-plot a contour density graph is built on a log-log scale using two-dimensional kernel density estimation. This evaluates a bivariate normal kernel over the log-log transformed data to produce a density estimate. This is done without the random uniform noise because the discrete density is well represented already. The bandwidth of the kernel for a data vector x is the normal reference bandwidth given by equation 4.1 [25, page 130]:

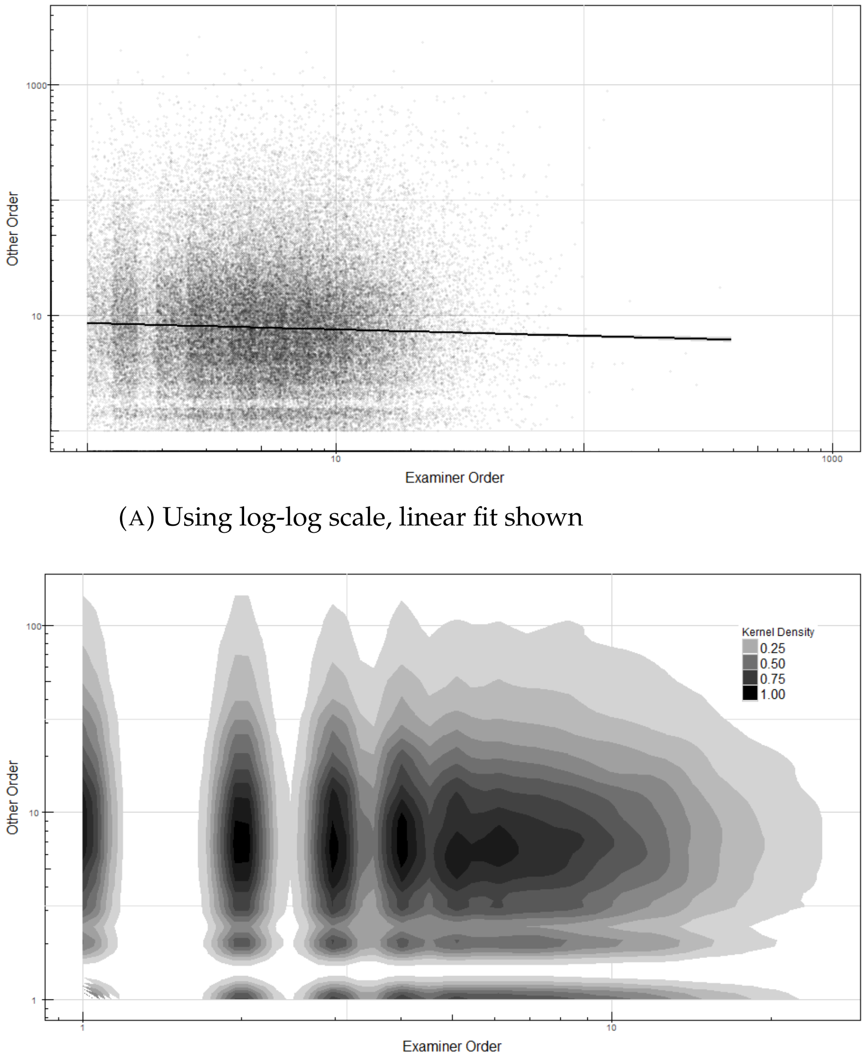
$$\begin{aligned} a &= \frac{Q_3(x) - Q_1(x)}{1.34} \\ b &= \sqrt{\sigma_x} \\ bandwidth &= 4 \times 1.06 \times min(a, b) \times N_x^{-1/5} \end{aligned} \tag{4.1}$$

The mean and median of one class is computed for logarithmically binned intervals of the other class. I.e. for the set of observations with Examiner degree in the first interval the mean and standard deviation of the Other degree is calculated (figure 4.8b) as well as median and interquartile range (figure 4.8c). The median is computed because heavy tailed distributions can be biassed using the mean and standard deviation only represents error (when transformed to standard error) under the assumption of a normal distribution.

Finally correlation metrics are computed with p-values using the fBasics R package [22]. Pearson correlation is computed non-parametrically using a bootstrapping procedure. As Pearson correlation is a measure of linear correlation between variables a log-log transformed version was also computed. Kendall's Tau and Spearman's rho are rank based measures so work on non-parametric data natively, this means that the rho and the tau are not measures of a linear relationship like the Pearson product-moment correlation.

In figures 4.7a and 4.7b it is clear that any correlation between the two orders are very small in nature, this is supported by all the correlation metrics computed (table 4.6). This is fairly unexpected, under a model that each patent has some underlying true number of citations and applicants have some fairly constant level of completeness for this true number that the correlation would be significant and positive as the greater true citation number the greater the number of citations from both examiner and applicant in this model. However under a model where the true number of citations is fairly constant but the distribution of completeness from the applicant varies a negative correlation would be expected because as the applicant finds most the relevant citations there are fewer left for the examiner to complete the patent. This model however assumes that both the examiner and the applicant are getting citations from a limited set of correct citations, if either is making a significant portion of their citations from outside this set then correlations would decrease. This is likely what is occurring either some of the parties are making citations from outside the set of correct citations and/or the set is not well defined.

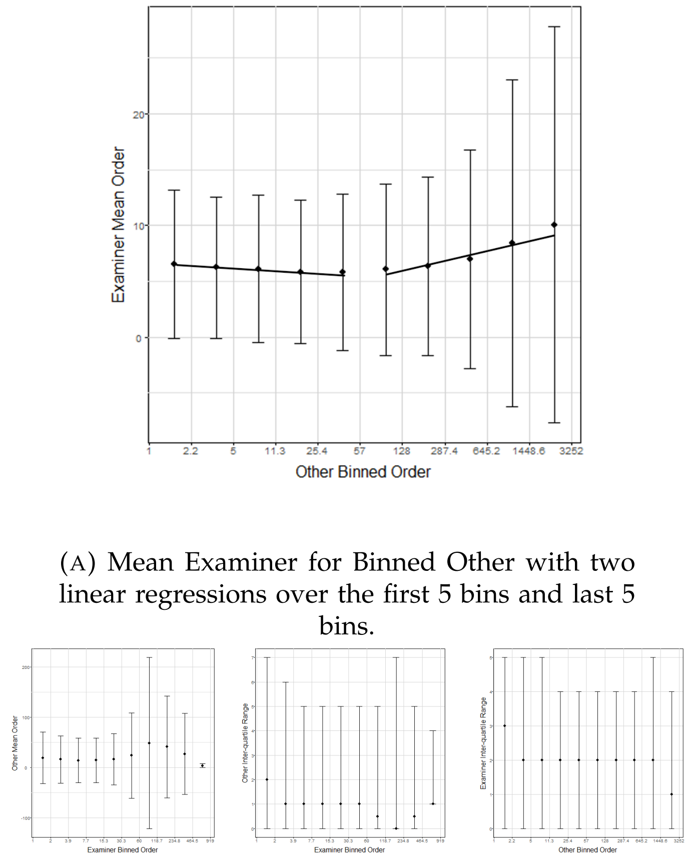
The data may support these models of correlation however, figure 4.8a shows a slight negative correlation for the bulk of the data but a positive correlation for the large order values and each of the correlation metrics are slightly negative. Figure 4.8a shows that extreme values have slightly



(A) Using log-log scale, linear fit shown

(B) Contour density graph using Kernel Density Estimator

FIGURE 4.7: Scatterplot of ’Examiner’ and ’Other’ orders for each patent using an n = 100,000 sample.



(A) Mean Examiner for Binned Other with two linear regressions over the first 5 bins and last 5 bins.

(B) Mean Other for Binned Examiner

(C) Median Other for Binned Examiner

(D) Median examiner for binned Other

FIGURE 4.8: Comparison of 'Other' and 'Examiner' subgraphs degree correlation by logarithmically binning one and calculating metrics about the other (mean/median)

positive correlation, while the error is large enough that this effect isn't statistically significant and the discrete nature of the median makes it difficult to identify trends in the other plots, the Kernal density plot does appear to corroborate this slight positive correlation for extreme orders. The Mean of the other for binned Examiner shows a similar effect but the positive correlation peaks before decreasing again, there are very low sample sizes for these high values of Examiner.

4.3.3 Fitting Distributions

As in section 4.2.2 four different functional forms are fitted to the cumulative degree distribution using maximum likelihood estimators while optimising x_{min} to minimise the Kolmogorov-Smirnov statistic as a measure of goodness-of-fit. Here the year 2002 is used as a year present in the Valverde analysis which has Examiner and Other networks analysable. A bootstrapping method using 2500 samples to compute a P value to an approximate accuracy of 0.1 is used using the same methods as previously. This method is conducted on a sample of 10000 observations due to the performance limitations of the method. Due to this sample being a relatively small subset of

TABLE 4.6: Correlation metrics and their P values. Pearson transformed refers to Pearson correlation conducted on log-log transformed data

Statistic	Value	One-sided (Less)	Two-sided
Kendall's Tau	-0.0889	2.2e-16	2.2e-16
Pearson Correlation	-0.008	5.5e-3	1.1e-2
Pearson transformed	-0.0378	2.2e-16	2.2e-16
Spearman's rho	-0.1243	2.2e-16	2.2e-16

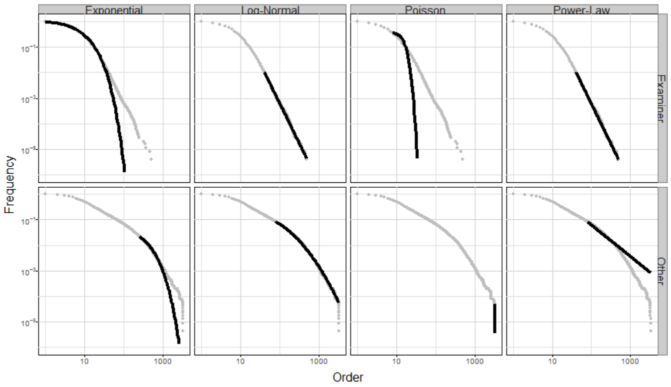


FIGURE 4.9: Fitted 2002 degree distributions for "Other" and "Examiner" sub-graphs using Maximum Likelihood for discrete distributions: Poisson, Exponential, Log-normal, Power-law.

the data results from this are taken with caution. A log-likelihood ratio test is also used to compare distributions, this uses all the available data rather than a sample but first refits the distributions so that both fits share the same x_{min}, this is done by setting the x_{min} as the greater of the two previous fits, as it is more conservative, before solving for the other parameter(s) by maximising the likelihood function.

As in section 4.2.2 a P value of less than 0.1 is taken as a rule of thumb to be discounted. Here all the P values are significantly lower than before and none of the Poisson or Exponential distributions get a significant result. While the Other network shows very similar P values for Power-Law and Log-Normal the Examiner network has very low values for the Log-Normal, lower than the Exponential and lower than the threshold for

TABLE 4.7: P values fitting distributions to the degree distribution obtained through bootstrapping method on a n = 10000 sample of the data

CitedBy	Power-Law	Log-Normal	Poisson	Exponential
Examiner	0.177	0.024	0.000	0.090
Other	0.343	0.327	0.037	0.000

TABLE 4.8: P values for likelihood data is power law distributed in comparison to log-normal using likelihood ratio test

CitedBy	Log Likelihood Ratio	P: One sided	P: Two Sided
Examiner	0.352	0.637	0.725
Other	-14.78	9.75e-50	1.95e-49

significance. When using the log-likelihood ratio test to compare distributions however the Examiner gives a high 2 sided P value. The two sided test indicates the probability that the data comes from both distributions i.e. that they are both good fits to the data and can't be analytically distinguished. This is in conflict with the bootstrapping results although as previously stated the small sample size makes these results fairly unreliable. The Other class however completely has vanishingly small p values for a power law suggesting that this can be discounted and a log-normal is the better fit.

When comparing these results to the time varying ones from section 4.2.2 the Examiner sub-network shows similar results to the 1984 analysis whereas the Other sub-network shows similar results to the more recent (2002/2012) analysis. This helps to support the idea that the relative dominance of the applicant citations is a structural change in the network.

4.3.4 Patent growth over time

In order to try to identify further the changes in structure of the network over time, we look at how the parameters of power-law and log-normal fits have changed over time. Each year is fitted with a power-law distribution, maximising the likelihood while searching over a range of x_{min} values as before. After the year 2001 this is also computed separately for the Examiner and Other sub-network respectively. This process is repeated for log-normal distribution. While the power-law only has one parameter, α, the log-normal has two, μ and σ.

Figure 4.10 shows the power-law parameters as they change over time. Note the Other distribution has its exponent changing over a wide range in a noisy manner. The power-law is a poor fit for this distribution and so over different years the power-law is fitting to different parts of the curve by varying x_{min} by large amounts, this is demonstrated in figure 4.11. When x_{min} is relatively small the parameter shadows that of the total as you would expect. As seen earlier the Examiner has smaller exponents however for this data over this time period it is hard to identify any diverging trends between the two if any exist.

The parameters for the log-normal, are more insightful 4.12a. The total has a tight cluster of years with a few outliers, the Other occupies only this cluster while the Examiner has points both in the cluster and the outlier region. Zooming in on the cluster there is a similar story where the total occupying a cluster but also some space away from that tight cluster. The cluster are the fits from the last 10 years and are clearly separable from before this point. The other occupies a parameter space very close to

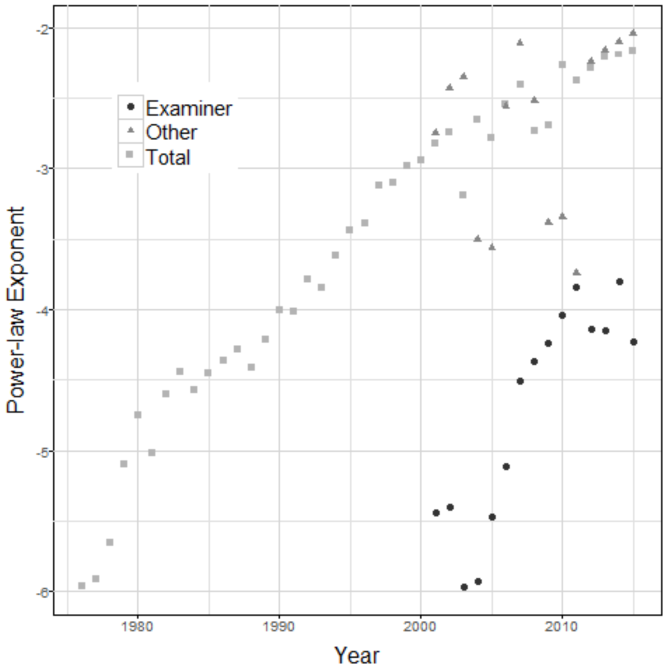
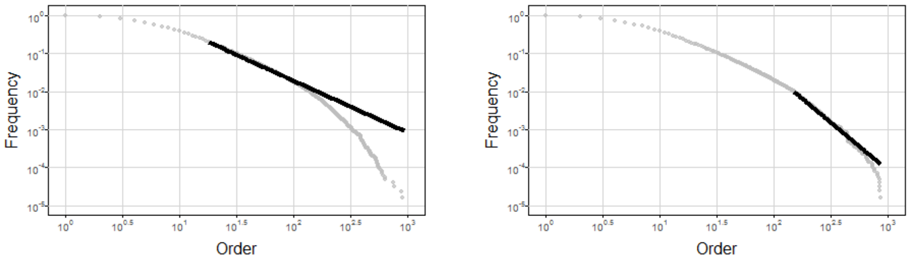


FIGURE 4.10: Variation in the exponent of the Power-Law fit for degree distribution from 1976:2015, where available "Examiner" and "Other" sub-graphs are also computed

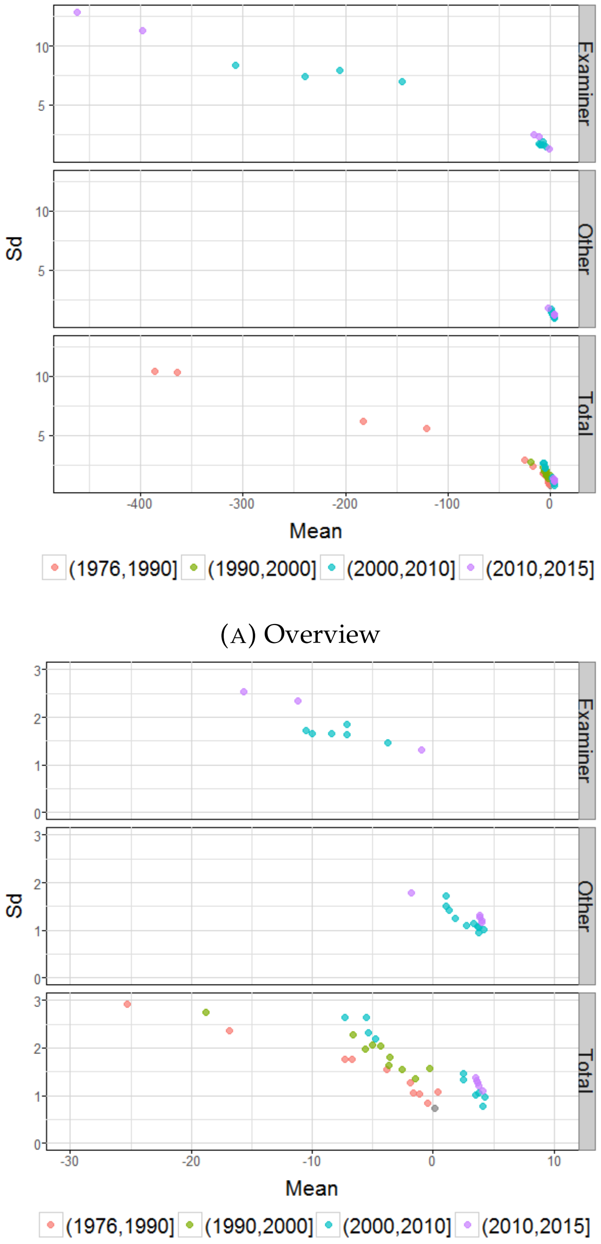


(A) Power-Law fit for "Other" sub-graph in 2003

(B) Power-Law fit for "Other" sub-graph in 2004

FIGURE 4.11

this cluster where the Examiner occupies a parameter space which is representative of the Total before the last 10 years. This further supports the hypothesis that the network was once much more strongly influenced by the examiner sub-network as the structure of the network now when fitted with the log-normal distribution is similar to the whole network fitted with this distribution 10-20 years ago.



(A) Overview

(B) Zoomed in on cluster

FIGURE 4.12: Variation in the mean and standard deviation of log-normal fits each year 1976 to 2015 for cited by "Examiner", cited by "Other" and combined. Colour representing the decade

Chapter 5

Conclusion

5.1 Review

While we have not tested for distributions such exponential cut-off to power law distributions or extended power laws, we have analytically shown the log-normal is the most likely fit to the overall degree-distribution of the current network. The chance of a power-law fit is sufficiently unlikely that it can be discounted as a hypothesis through bootstrapping and log-likelihood comparisons.

The network has evolved to this point at the beginning of the data-set comparison techniques couldn't distinguish between power-law and log-normal distributions but were in favour of power-law while over time the distribution is became less similar to a power-law and more similar to a log-normal.

We coloured the network into two classes from 2001 to 2015 depending on the role of the individual making the citation, the Examiner network and the Other network. We found that these are fairly disparate networks with small negative correlations between the two and different functional forms (although small positive correlations at extreme order values). The Other citations dominate in number so their structure is similar to that found in the aggregate over these years but the Examiner network seems to emulate the aggregate network in the years before these classes could be distinguished (1976 to 2000), with log-normal parameters fits similar to these times and Power-law/log-normal P values calculated yielding similar results, i.e. that neither power-law and log-normal can be discounted but the power-law is more likely (P = 0.637). Finally the average number of citations changing with time showing significant growth in the Other network further drowning out the Examiner network but supporting a hypothesis that in the past the Other network smaller relative to the Examiner and that this change in how applicants are interacting with patents is the cause of a lot of the changes in the patent network over the last decade.

5.2 Further Research

As mentioned in the introduction showing a log-normal or power-law form is not sufficient to identify preferential attachment or other such mechanisms. Considering the contradicting research in the USPTO data-set research identifying the presence and nature of any preferential attachment mechanisms is important future research.

One of the main research questions of this project was to identify the structural changes to the patent network over time and relate that to claims of an information revolution or economic identifiers. While we have shown there are major structural changes and that they are linked to the change in interaction of the applicant of a patent. This hasn't been linked to external factors. Further research could seek to achieve this in a number of ways, correlating changes to the patent system with econometric information answering questions like how is innovation affected by recession. More importantly would be through studying the how the technology sectors vary, identified large structural differences between technology sectors, relating this research with temporal changes can give a much better idea of how innovation shifts between different technological areas [12].

The nature of the different classes of citation form different types of relationship between the same nodes. This can be represented as an unweighted multi-layer network. Representing the network in this manner can allow for techniques which study the whole network in a way which doesn't ignore the different nature of relationship between an Examiner citation and a Applicant citation [13].

Finally a more directly practical direction for further research can focus on the prediction of success, and evaluation of value to the network of different patents. While research has been done on these topics adding the additional features and information from the Examiner network may improve the accuracy of these models. Additionally the extra features may prove useful in other ways such as clustering algorithms and anomaly detection in order to identify patent trolls.

Appendix A

Document Parsing Function

```
parse <- function(input_path, type, output_path_patent, output_path_citation) {
    # Libraries
    require(stringr)
    require(readr)

    # Save environment to reference
    Env <- environment()

    # Create output paths if they don't exist
    dirs <- str_replace_all(c(output_path_patent, output_path_citation), "/.+.csv$", "")
    sapply(dirs, function(dir) {
        if (!dir.exists(dir)) dir.create(dir, recursive = T)
    })

    # Actions ########################################################

    # Initialise output vecotrs for patent / citations upon reaching a new patent
    initialise_result <- function() {
        # If files don't exist write their column names
        if (!file.exists(output_path_patent)) {
            write(patent_colnames, output_path_patent, sep = ",", append = F,
                  ncolumns = length(patent_colnames))
        }
        if (!file.exists(output_path_citation)) {
            write(citation_colnames, output_path_citation, sep = ",", append = F,
                  ncolumns = length(citation_colnames))
        }
        patent <<- vector("character", length = length(patent_colnames))
        citations <<- NULL
        citation_current <<- vector("character", length = length(citation_colnames) - 1)
    }

    # Write patent / citation information to file upon finding the end of the current patent
    flush_patent <- function() {
        if (patent[1] != "") {
            # Add degree to dataframe
            nrow_citations <- nrow(citations)
            if (is.null(nrow_citations)) nrow_citations <- 0
            patent[which(patent_colnames == "Order")] <- nrow_citations
            write(patent, output_path_patent, sep = ",",
                  append = T, ncolumns = length(patent_colnames))
            write_csv(as.data.frame(citations), output_path_citation, append = T)
        }
    }

    # Add current citation to citations matrix after finding the end of the current citaiton
    flush_citation <- function() {
        if (sum(citation_current != rep("", length(citation_colnames) - 1)) > 0) {
            citations <<- rbind(citations, c(patent[1], citation_current))
            citation_current <<- vector("character", length = length(citation_colnames) - 1)
        }
    }

    # Upon finding relevant information add to patent/citation output
    add_patent_information <- function(tag = line_tag, State = state,
                                       vars = tag_add_patent_information) {
        if (State != "patent") return()
        ii <- match(tag, vars)
        patent[ii] <<- contents()
    }

    add_citation_information <- function(tag = line_tag, State = state,
                                         vars = tag_add_citation_information) {
        if (State != "citation") return()
        ii <- match(tag, vars)
        citation_current[ii] <<- contents()
    }

    # State change: upon reaching different sections change state
    state_change_patent <- function() {
        state <<- "patent"
    }
    state_change_citation <- function() {
        state <<- "citation"
    }
    state_change_none <- function() {
        state <<- "none"
    }

    # Function to get contents (so not evaluated every line)
    contents <- function(line = Env$line, type = Env$type) {
        ifelse(type == "text",
               str_trim(substring(line, 5)),
               str_replace_all(line, "<.*?>", ""))
    }

    # If sgml check for citation and country information hidden inside the tag
    sgml_exceptions <- function(line) {
        if (grepl("(<CITED-BY-OTHER/>)|(<CITED-BY-OTHER>)", line)) {
            citation_current[which(citation_colnames == "Cited␣by") - 1] <<- "cited␣by␣other"
        } else if (grepl("(<CITED-BY-EXAMINER/>)|(<CITED-BY-EXAMINER>)", line)){
            citation_current[which(citation_colnames == "Cited␣by") - 1] <<- "cited␣by␣examiner"
        }
        if (grepl("<PARTY-US>", line)) {
            citation_current[which(citation_colnames == "Country") - 1] <<- "US"
        }
        # Remove some extra tags which create duplicate matches
        line <- str_replace(line, "(</DOC>)|(<CITED-BY-OTHER/>)|(<CITED-BY-EXAMINER/>)|
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣(<PARTY-US>)|(<CITED-BY-OTHER>)|(<CITED-BY-EXAMINER>)", "")
        return(line)
    }

    ## initialisation ##################################################
    state <- "patent"
    # List the tags associated with each action function
    switch(type,
           text = {
               tag_initialise_result <- "PATN"
               tag_flush_patent <- "PATN"
               tag_flush_citation <- c("UREF", "FREF", "OREF", "DRWD", "PATN")
               tag_add_patent_information <- c("WKU␣", "ISD␣")
               tag_add_citation_information <- c( "PNO␣", "ISD␣")
               tag_state_change_citation <- c("UREF", "FREF")
               tag_state_change_patent <- "PATN"
               tag_state_change_none <- c("INVT","CLAS")
           },
           xml = {
               tag_initialise_result <- '<?xml␣version="1.0"␣encoding="UTF-8"?>'
               tag_flush_patent <- '</us-patent-grant>'
               tag_flush_citation <- c("</us-citation>", "</citation>")
               tag_add_patent_information <- c("<doc-number></doc-number>", "<date></date>",
                                               "<main-classification></main-classification>",
                                               "<further-classification></further-classification>")
               tag_add_citation_information <- c("<doc-number></doc-number>", "<date></date>",
                                                 "<category></category>", "<country></country>")
               tag_state_change_citation <- c("<us-citation>", "<citation>")
               tag_state_change_patent <- c("<publication-reference>", "<classification-national>")
               tag_state_change_none <- c("</publication-reference>", "</us-ciation>",
                                          "</citation>", "</classification-national>")
               citation_colnames <- c("Patent", "Citation", "Date", "CitedBy", "Country")
               patent_colnames <- c("Patent", "Date", "MainClassification", "FurtherClassification", "Order")
           },
           sgml = {
               tag_initialise_result <- '<SDOBI>'
               tag_flush_patent <- "</SDOBI>"
               tag_flush_citation <- "</B561>"
               tag_add_patent_information <- c("<B110><DNUM><PDAT></PDAT></DNUM></B110>", # patent number
                                               "<B140><DATE><PDAT></PDAT></DATE></B140>", # Date
                                               "<B521><PDAT></PDAT></B521>", # Main classification ,
                                               "<B522><PDAT></PDAT></B522>") # Further classification
               tag_add_citation_information <- c("<DOC><DNUM><PDAT></PDAT></DNUM>", # Patent number
                                                 "<DATE><PDAT></PDAT></DATE>", # Patent Date
                                                 "<CTRY><PDAT></PDAT></CTRY>") # Country
               tag_state_change_citation <- "<B561>"
               tag_state_change_patent <- '<SDOBI>'
               tag_state_change_none <- NULL
               citation_colnames <- c("Patent", "Citation", "Date", "CitedBy", "Country")
               patent_colnames <- c("Patent", "Date", "MainClassification", "FurtherClassification", "Order")
           })

    # Initialise total taglist and response index
    tag_list <- list(tag_flush_citation, tag_flush_patent, tag_state_change_patent,
                     tag_state_change_citation, tag_state_change_none, tag_initialise_result,
                     tag_add_citation_information, tag_add_patent_information)
    tag_all <- unlist(tag_list)
    tag_index <- rep(1:length(tag_list), sapply(tag_list, length))
    # Vectorise action functions in same order as tags which call them
    action_functions <- list(flush_citation, flush_patent,  state_change_patent,
                             state_change_citation, state_change_none, initialise_result,
                             add_citation_information, add_patent_information)

    # Read data
    text <- read_lines(input_path)

    ## LOOP ############################################################
    pb <- txtProgressBar(min = 0, max = length(text), initial = 0, style = 3)
    initialise_result()
    for (i in seq_along(text)) {
        line <- text[i]
        if (i %% 1000 == 0) setTxtProgressBar(pb, i)

        # Turn current line into a tag (acts differently for text / sgml)
        line_tag <- ifelse(type == "text",
                           substring(line, 1, 4),
                           str_replace_all(line, ">.*?<", "><")
```

```
                          )

        if (type == "sgml") line_tag <- sgml_exceptions(line_tag)

        # Check if it matches any in the list
        matches <- tag_index[tag_all %in% line_tag]

        # If match occurs invoke the appropriate action function
        if (length(matches) > 0) {
            lapply(matches, function(x) action_functions[[x]]())
        }
    }
    if (type == "text") flush_patent() #Because text doesn't have close tags
    close(pb)
}
```

Appendix B

Example Raw Data Patent: Text format

```
PATN
WKU  RE0286710
SRC  5
APN  500649\&
APT  2
PBL  E
ART  315
APD  19740826
TTL  Hydrophone damper assembly
ISD  19760106
NCL  18
ECL  13
EXA  Basinger; Sherman D.
EXP  Blix; Trygve M.
NDR  2
NFG  10
INVT
NAM  Widenhofer; James W.
CTY  Jackson
STA  MI
ASSG
NAM  Sparton Corporation
CTY  Jackson
STA  MI
COD  02
REIS
COD  50
APN  151269
APD  19710609
PNO  03701175
ISD  19721031
CLAS
OCL    9  8R
XCL  340  2
XCL  340  3T
XCL  340  7R
XCL  340  8R
EDF  2
ICL  B63B 2152
ICL  B63B 5102
FSC    9
FSS  8 R
FSC  340
FSS  2;3 T;8 S;8 R;7
FSC  114
FSS  206 R
UREF
PNO  2790186
ISD  19570400
NAM  Carapellotti
OCL    9  8R
UREF
PNO  3329015
ISD  19670700
NAM  Bakeke et al.
OCL  340  2
UREF
PNO  3377615
ISD  19680400
NAM  Lutes
OCL  340  2
UREF
PNO  3543228
ISD  19701100
NAM  Farmer
OCL  340  2
UREF
PNO  3543228
ISD  19701100
NAM  Farmer et al.
OCL    9  8R
UREF
PNO  3711821
```

```
ISD  19730100
NAM  Dale et al.
OCL    9  8R
UREF
PNO  3720909
ISD  19730300
NAM  Sikora
OCL  340  2
UREF
PNO  3803540
ISD  19740400
NAM  Mar et al.
OCL  340  2
LREP
FRM  Beaman \& Beaman
ABST
PAL  A damper for use in submerged hydrophone suspension systems including an
       elongated mass cylinder defined by a tube of flexible synthetic plastic
       film utilizing a check valve located at each end permitting water to enter
       the tube and preventing egress. Additionally, each tube end is provided
       with a disk transversely disposed to the tube length and of a diameter
       substantially greater than that of the tube to provide drag and
       hydrodynamic mass damping. The tube and disk are of a configuration to
       eliminate vortex shedding and the entire damper assembly is capable of
       being folded and packed within a concise configuration prior to
       deployment.
```

Appendix C

Example Raw Data Patent: Sgml format

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE PATDOC SYSTEM "ST32-US-Grant-025xml.dtd" [
<!ENTITY USD0468513-20030114-D00000.TIF SYSTEM "USD0468513-20030114-D00000.TIF" NDATA TIF>
<!ENTITY USD0468513-20030114-D00001.TIF SYSTEM "USD0468513-20030114-D00001.TIF" NDATA TIF>
<!ENTITY USD0468513-20030114-D00002.TIF SYSTEM "USD0468513-20030114-D00002.TIF" NDATA TIF>
<!ENTITY USD0468513-20030114-D00003.TIF SYSTEM "USD0468513-20030114-D00003.TIF" NDATA TIF>
]>
<PATDOC DTD="2.5" STATUS="Build 20020918">
<SDOBI>
<B100>
<B110><DNUM><PDAT>D0468513</PDAT></DNUM></B110>
<B130><PDAT>S1</PDAT></B130>
<B140><DATE><PDAT>20030114</PDAT></DATE></B140>
<B190><PDAT>US</PDAT></B190>
</B100>
<B200>
<B210><DNUM><PDAT>29152580</PDAT></DNUM></B210>
<B211US><PDAT>29</PDAT></B211US>
<B220><DATE><PDAT>20011227</PDAT></DATE></B220>
</B200>
<B300>
<B310><DNUM><PDAT>1999-2430 CA</PDAT></DNUM></B310>
<B320><DATE><PDAT>19991006</PDAT></DATE></B320>
<B330><CTRY><PDAT>CA</PDAT></CTRY></B330>
</B300>
<B400>
<B472>
<B474><PDAT>14</PDAT></B474>
</B472>
</B400>
<B500>
<B510>
<B511><PDAT>0101</PDAT></B511>
<B516><PDAT>7</PDAT></B516>
</B510>
<B520>
<B521><PDAT>D 1104</PDAT></B521>
</B520>
<B540><STEXT><PDAT>Lollipop</PDAT></STEXT></B540>
<B560>
<B561>
<PCIT>
<DOC><DNUM><PDAT>1257779</PDAT></DNUM>
<DATE><PDAT>19180200</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY-US>
<NAM><SNM><STEXT><PDAT>Anderson</PDAT></STEXT></SNM></NAM>
</PARTY-US>
<PNC><PDAT>273146</PDAT></PNC></PCIT><CITED-BY-EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>D58042</PDAT></DNUM>
<DATE><PDAT>19210500</PDAT></DATE>
<KIND><PDAT>S</PDAT></KIND>
</DOC>
<PARTY-US>
<NAM><SNM><STEXT><PDAT>Paine</PDAT></STEXT></SNM></NAM>
</PARTY-US>
<PNC><PDAT>D 1106</PDAT></PNC></PCIT><CITED-BY-EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>1539015</PDAT></DNUM>
<DATE><PDAT>19250500</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY-US>
<NAM><SNM><STEXT><PDAT>Michell</PDAT></STEXT></SNM></NAM>
</PARTY-US>
<PNC><PDAT>273146</PDAT></PNC></PCIT><CITED-BY-EXAMINER/>
</B561>
```

```
<B561>
<PCIT>
<DOC><DNUM><PDAT>1786606</PDAT></DNUM>
<DATE><PDAT>19301200</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY–US>
<NAM><SNM><STEXT><PDAT>Gordon</PDAT></STEXT></SNM></NAM>
</PARTY–US>
<PNC><PDAT>426  91  X</PDAT></PNC></PCIT><CITED–BY–EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>2036706</PDAT></DNUM>
<DATE><PDAT>19360400</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY–US>
<NAM><SNM><STEXT><PDAT>Law</PDAT></STEXT></SNM></NAM>
</PARTY–US>
<PNC><PDAT>426  85</PDAT></PNC></PCIT><CITED–BY–EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>2191352</PDAT></DNUM>
<DATE><PDAT>19400200</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY–US>
<NAM><SNM><STEXT><PDAT>Oprean</PDAT></STEXT></SNM></NAM>
</PARTY–US>
<PNC><PDAT>426101</PDAT></PNC></PCIT><CITED–BY–EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>2589823</PDAT></DNUM>
<DATE><PDAT>19520300</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY–US>
<NAM><SNM><STEXT><PDAT>Krens</PDAT></STEXT></SNM></NAM>
</PARTY–US>
<PNC><PDAT>426421</PDAT></PNC></PCIT><CITED–BY–EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>3062662</PDAT></DNUM>
<DATE><PDAT>19621100</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY–US>
<NAM><SNM><STEXT><PDAT>McDonald</PDAT></STEXT></SNM></NAM>
</PARTY–US>
<PNC><PDAT>426  91  X</PDAT></PNC></PCIT><CITED–BY–EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>3400932</PDAT></DNUM>
<DATE><PDAT>19680900</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY–US>
<NAM><SNM><STEXT><PDAT>Conrad</PDAT></STEXT></SNM></NAM>
</PARTY–US>
<PNC><PDAT>273146</PDAT></PNC></PCIT><CITED–BY–EXAMINER/>
</B561>
<B561>
<PCIT>
<DOC><DNUM><PDAT>3459296</PDAT></DNUM>
<DATE><PDAT>19690800</PDAT></DATE>
<KIND><PDAT>A</PDAT></KIND>
</DOC>
<PARTY–US>
<NAM><SNM><STEXT><PDAT>Berg</PDAT></STEXT></SNM></NAM>
</PARTY–US>
<PNC><PDAT>426134  X</PDAT></PNC></PCIT><CITED–BY–EXAMINER/>
</B561>
</B560>
<B570>
<B577><PDAT>1</PDAT></B577>
<B578US><PDAT>1</PDAT></B578US>
</B570>
<B580>
<B583US><PDAT>D  1100−129</PDAT></B583US>
<B582><PDAT>D  1199</PDAT></B582>
<B582><PDAT>426  85</PDAT></B582>
<B582><PDAT>426  90</PDAT></B582>
<B582><PDAT>426  91</PDAT></B582>
<B583US><PDAT>426100−104</PDAT></B583US>
<B583US><PDAT>426132−134</PDAT></B583US>
<B582><PDAT>426249</PDAT></B582>
<B582><PDAT>426801</PDAT></B582>
<B582><PDAT>426421</PDAT></B582>
<B582><PDAT>426660</PDAT></B582>
<B582><PDAT>273146</PDAT></B582>
<B582><PDAT>273260</PDAT></B582>
<B582><PDAT>  21372</PDAT></B582>
<B582><PDAT>  21373</PDAT></B582>
```

```
<B582><PDAT> 21347</PDAT></B582>
<B582><PDAT> 21499</PDAT></B582>
</B580>
<B590><B595><PDAT>3</PDAT></B595><B596><PDAT>9</PDAT></B596><B597US/>
</B590>
</B500>
<B600>
<B630><B632><PARENT—US><CDOC><DOC><DNUM><PDAT>29/152580</PDAT></DNUM></DOC></CDOC><PDOC><DOC><DNUM><PDAT>29/121438</PDAT></DNUM><D
</B600>
<B700>
<B720>
<B721>
<PARTY—US>
<NAM><FNM><PDAT>Lisa Jane</PDAT></FNM><SNM><STEXT><PDAT>Wolfe</PDAT></STEXT></SNM></NAM>
<ADR>
<STR><PDAT>8700 No. 52nd St.</PDAT></STR>
<CITY><PDAT>Paradise Valley</PDAT></CITY>
<STATE><PDAT>AZ</PDAT></STATE>
<PCODE><PDAT>85253</PDAT></PCODE>
</ADR>
</PARTY—US>
</B721>
<B721>
<PARTY—US>
<NAM><FNM><PDAT>Jane Margaret</PDAT></FNM><SNM><STEXT><PDAT>Bachynski</PDAT></STEXT></SNM></NAM>
<ADR>
<STR><PDAT>&num;6−55 Whitemarl Dr.</PDAT></STR>
<CITY><PDAT>Rockcliffe Park Ontario</PDAT></CITY>
<CTRY><PDAT>CA</PDAT></CTRY>
<PCODE><PDAT>K1L 8J9 </PDAT></PCODE>
</ADR>
</PARTY—US>
</B721>
</B720>
<B740>
<B741>
<PARTY—US>
<NAM><ONM><STEXT><PDAT>Fulbright &amp; Jaworski, LLP</PDAT></STEXT></ONM></NAM>
</PARTY—US>
</B741>
</B740>
<B745>
<B746>
<PARTY—US>
<NAM><FNM><PDAT>Alan P.</PDAT></FNM><SNM><STEXT><PDAT>Douglas</PDAT></STEXT></SNM></NAM>
</PARTY—US>
</B746>
<B747>
<PARTY—US>
<NAM><FNM><PDAT>Linda</PDAT></FNM><SNM><STEXT><PDAT>Brooks</PDAT></STEXT></SNM></NAM>
</PARTY—US>
</B747>
<B748US><PDAT>2911</PDAT></B748US>
</B745>
</B700>
</SDOBI>
<SDODE>
</SDODE>
<SDOCL>
<CL>
<CLM ID="CLM−00001">
<PARA ID="P−00010" LVL="7"><PTEXT><PDAT>The ornamental design for a lollipop, as shown and described.</PDAT></PTEXT></PARA>
</CLM>
</CL>
</SDOCL>
</PATDOC>
```

Appendix D

Example Raw Data Patent: XML format

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE us-patent-grant SYSTEM "us-patent-grant-v42-2006-08-23.dtd" [ ]>
<us-patent-grant lang="EN" dtd-version="v4.2 2006-08-23" file="US07646781-20100112.XML" status="PRODUCTION" id="us-patent-grant" co
<us-bibliographic-data-grant>
<publication-reference>
<document-id>
<country>US</country>
<doc-number>07646781</doc-number>
<kind>B2</kind>
<date>20100112</date>
</document-id>
</publication-reference>
<application-reference appl-type="utility">
<document-id>
<country>US</country>
<doc-number>11753703</doc-number>
<date>20070525</date>
</document-id>
</application-reference>
<us-application-series-code>11</us-application-series-code>
<us-term-of-grant>
<us-term-extension>402</us-term-extension>
</us-term-of-grant>
<classifications-ipcr>
<classification-ipcr>
<ipc-version-indicator><date>20060101</date></ipc-version-indicator>
<classification-level>A</classification-level>
<section>H</section>
<class>04</class>
<subclass>L</subclass>
<main-group>12</main-group>
<subgroup>56</subgroup>
<symbol-position>F</symbol-position>
<classification-value>I</classification-value>
<action-date><date>20100112</date></action-date>
<generating-office><country>US</country></generating-office>
<classification-status>B</classification-status>
<classification-data-source>H</classification-data-source>
</classification-ipcr>
</classifications-ipcr>
<classification-national>
<country>US</country>
<main-classification>370412</main-classification>
<further-classification>370235</further-classification>
</classification-national>
<invention-title id="d0e53">Methods, systems, and computer program products for selectively discarding packets</invention-title>
<references-cited>
<citation>
<patcit num="00001">
<document-id>
<country>US</country>
<doc-number>2006/0072576</doc-number>
<kind>A1</kind>
<name>Miao et al.</name>
<date>20060400</date>
</document-id>
</patcit>
<category>cited by examiner</category>
<classification-national><country>US</country><main-classification>370394</main-classification></classification-national>
</citation>
<citation>
<patcit num="00002">
<document-id>
<country>US</country>
<doc-number>2006/0291384</doc-number>
<kind>A1</kind>
<name>Harris et al.</name>
<date>20061200</date>
</document-id>
</patcit>
<category>cited by examiner</category>
<classification-national><country>US</country><main-classification>370229</main-classification></classification-national>
</citation>
<citation>
```

```
<patcit num="00003">
<document-id>
<country>US</country>
<doc-number>2007/0201365</doc-number>
<kind>A1</kind>
<name>Skoog et al.</name>
<date>20070800</date>
</document-id>
</patcit>
<category>cited by examiner</category>
<classification-national><country>US</country><main-classification>3702301</main-classification></classification-national>
</citation>
<citation>
<patcit num="00004">
<document-id>
<country>US</country>
<doc-number>2009/0129313</doc-number>
<kind>A1</kind>
<name>Tamura et al.</name>
<date>20090500</date>
</document-id>
</patcit>
<category>cited by examiner</category>
<classification-national><country>US</country><main-classification>370328</main-classification></classification-national>
</citation>
<citation>
<nplcit num="00005">
<othercit>Web page published by &#x201c;vonage.nmhoy.net/qos.html&#x201d;; http;//web.archive.org/web/20060314042029/http://vonage.nmhoy
</nplcit>
<category>cited by other</category>
</citation>
</references-cited>
<number-of-claims>23</number-of-claims>
<us-exemplary-claim>1</us-exemplary-claim>
<us-field-of-classification-search>
<classification-national>
<country>US</country>
<main-classification>370234</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370229</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370230</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370232</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370235</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>3702351</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370352</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370412</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370428</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370429</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>37039521</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>3703954</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>37039541</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>37039542</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>37039552</main-classification>
</classification-national>
<classification-national>
<country>US</country>
<main-classification>370470</main-classification>
```

```
</classification-national>
</us-field-of-classification-search>
<figures>
<number-of-drawing-sheets>5</number-of-drawing-sheets>
<number-of-figures>5</number-of-figures>
</figures>
<us-related-documents>
<related-publication>
<document-id>
<country>US</country>
<doc-number>20080291935</doc-number>
<kind>A1</kind>
<date>20081127</date>
</document-id>
</related-publication>
</us-related-documents>
<parties>
<applicants>
<applicant sequence="001" app-type="applicant-inventor" designation="us-only">
<addressbook>
<last-name>Campion</last-name>
<first-name>Nicholas F.</first-name>
<address>
<city>Rochester</city>
<state>MN</state>
<country>US</country>
</address>
</addressbook>
<nationality>
<country>omitted</country>
</nationality>
<residence>
<country>US</country>
</residence>
</applicant>
<applicant sequence="002" app-type="applicant-inventor" designation="us-only">
<addressbook>
<last-name>Cramer</last-name>
<first-name>Keith D.</first-name>
<address>
<city>Pine Island</city>
<state>MN</state>
<country>US</country>
</address>
</addressbook>
<nationality>
<country>omitted</country>
</nationality>
<residence>
<country>US</country>
</residence>
</applicant>
<applicant sequence="003" app-type="applicant-inventor" designation="us-only">
<addressbook>
<last-name>Morrison</last-name>
<first-name>Donald A.</first-name>
<address>
<city>Rochester</city>
<state>MN</state>
<country>US</country>
</address>
</addressbook>
<nationality>
<country>omitted</country>
</nationality>
<residence>
<country>US</country>
</residence>
</applicant>
<applicant sequence="004" app-type="applicant-inventor" designation="us-only">
<addressbook>
<last-name>Strauss</last-name>
<first-name>Daniel J.</first-name>
<address>
<city>Rochester</city>
<state>MN</state>
<country>US</country>
</address>
</addressbook>
<nationality>
<country>omitted</country>
</nationality>
<residence>
<country>US</country>
</residence>
</applicant>
</applicants>
<agents>
<agent sequence="01" rep-type="attorney">
<addressbook>
<orgname>Cantor Colburn, LLP</orgname>
<address>
<country>unknown</country>
</address>
</addressbook>
</agent>
</agents>
</parties>
```

```
<assignees>
<assignee>
<addressbook>
<orgname>International Business Machines Corporation</orgname>
<role>02</role>
<address>
<city>Armonk</city>
<state>NY</state>
<country>US</country>
</address>
</addressbook>
</assignee>
</assignees>
<examiners>
<primary-examiner>
<last-name>Nguyen</last-name>
<first-name>Brian D</first-name>
<department>2416</department>
</primary-examiner>
</examiners>
</us-bibliographic-data-grant>
<abstract id="abstract">
<p id="p-0001" num="0000">A method, system, and computer program product are provided for selectively discarding packets in a network de
</abstract>
</us-patent-grant>
```

Bibliography

[1] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Reviews of modern physics* 74.1 (2002), p. 47.

[2] W Brian Arthur and Wolfgang Polak. “The Evolution ofTechnology within a Simple Computer Model”. In: *Complexity and the Economy* (2014).

[3] Katy Börner, Jeegar T Maru, and Robert L Goldstone. “The simultaneous evolution of author and paper networks”. In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), pp. 5266–5273.

[4] Andrew Buchanan, Norman H Packard, and Mark A Bedau. “Measuring the evolution of the drivers of technological innovation in the patent record”. In: *Artificial life* 17.2 (2011), pp. 109–122.

[5] Manuel Castells. *The rise of the network society: The information age: Economy, society, and culture*. Vol. 1. John Wiley & Sons, 2011.

[6] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data”. In: *SIAM review* 51.4 (2009), pp. 661–703.

[7] Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch. “On the frequency of severe terrorist events”. In: *Journal of Conflict Resolution* 51.1 (2007), pp. 58–87.

[8] James R Clough et al. “Transitive reduction of citation networks”. In: *Journal of Complex Networks* 3.2 (2015), pp. 189–203.

[9] Gábor Csárdi et al. “Modeling innovation by a kinetic description of the patent citation system”. In: *Physica A: Statistical Mechanics and its Applications* 374.2 (2007), pp. 783–793.

[10] Paul Erd6s and A Rényi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hungar. Acad. Sci* 5 (1960), pp. 17–61.

[11] P ERDdS and A R&WI. “On random graphs I”. In: *Publ. Math. Debrecen* 6 (1959), pp. 290–297.

[12] Bernard Gress. “Properties of the USPTO patent citation network: 1963–2002”. In: *World Patent Information* 32.1 (2010), pp. 3–21.

[13] Mikko Kivelä et al. “Multilayer networks”. In: *Journal of Complex Networks* (2014). DOI: 10.1093/comnet/cnu016. eprint: http://comnet.oxfordjournals.org/content/early/2014/07/14/comnet.cnu016.full.pdf+html. URL: http://comnet.oxfordjournals.org/content/early/2014/07/14/comnet.cnu016.abstract.

[14] P. L. Krapivsky, S. Redner, and F. Leyvraz. “Connectivity of Growing Random Networks”. In: *Phys. Rev. Lett.* 85 (21 Nov. 2000), pp. 4629–4632. DOI: 10.1103/PhysRevLett.85.4629. URL: http://link.aps.org/doi/10.1103/PhysRevLett.85.4629.

[15] Robert K Merton et al. “The Matthew effect in science”. In: *Science* 159.3810 (1968), pp. 56–63.

[16] Mark EJ Newman. “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary physics* 46.5 (2005), pp. 323–351.

[17] Derek de Solla Price. “A general theory of bibliometric and other cumulative advantage processes”. In: *Journal of the American society for Information science* 27.5 (1976), pp. 292–306.

[18] Sidney Redner. “Citation statistics from more than a century of physical review”. In: *arXiv preprint physics/0407137* (2004).

[19] Mikhail V Simkin and Vwani P Roychowdhury. “Stochastic modeling of citation slips”. In: *Scientometrics* 62.3 (2005), pp. 367–384.

[20] Karsten Steinhaeuser, Nitesh V Chawla, and Auroop R Ganguly. “Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science”. In: *Statistical Analysis and Data Mining* 4.5 (2011), pp. 497–511.

[21] Patent Technology Monitoring Team. *U.S. Patent Statistics Chart Calendar Years 1963 - 2015*. 2016. URL: http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm.

[22] Rmetrics Core Team et al. *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3011.87. 2014. URL: https://CRAN.R-project.org/package=fBasics.

[23] By Topic. *USPTO bulk data portal*. URL: https://bulkdata.uspto.gov/.

[24] Sergi Valverde et al. “Topology and evolution of technology innovation networks”. In: *Physical Review E* 76.5 (2007), p. 056118.

[25] William N Venables and Brian D Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[26] Quang H Vuong. “Likelihood ratio tests for model selection and non-nested hypotheses”. In: *Econometrica: Journal of the Econometric Society* (1989), pp. 307–333.

[27] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), pp. 440–442.

[28] Hadley Wickham et al. “Tidy data”. In: ().

[29] *XML Resources*. URL: https://www.uspto.gov/learning-and-resources/xml-resources.

[30] Byungun Yoon and Yongtae Park. “A text-mining-based patent network: Analytical tool for high-technology trend”. In: *The Journal of High Technology Management Research* 15.1 (2004), pp. 37–50.

[31] Hyejin Youn et al. “Invention as a combinatorial process: evidence from US patents”. In: *Journal of The Royal Society Interface* 12.106 (2015), p. 20150272.

[32] George Kingsley Zipf. “The psycho-biology of language.” In: (1935).