Metodološki zvezki, Vol. 2, No. 2, 2005, 195-208

Analysis of U.S. Patents Network: Development of Patents over Time

Nataša Kejžar[1]

Abstract

The NBER network of U.S. patents from 1963 to 1999 (Hall, Jaffe, Tratjenberg 2001, USPTO) is an example of a very large citation network (3774768 vertices and 16522438 arcs).

Using islands algorithm (Zaveršnik, Batagelj, 2004) for the Search Path Count (SPC) weights (Hummon and Doreian 1989; Batagelj 2003) the most powerful theme in the entire network was determined. From this we selected a group of companies and categories that appeared and split the entire network into subnetworks according to selected companies and technological categories. We study the general trends and features of the subnetworks over the past thirty-seven years.

We propose another approach for studying patents' network as a temporal network. Vertices from the same category in the same time slice are shrunk and then the obtained smaller networks over time are studied.

By studying development patterns of the network over time we are trying to determine the general trends in the research and development for the selected companies and categories over the past three decades.

1 Introduction

Patent datasets usually contain different information about innovation development. There is detailed information about innovation, the inventors and assignees, the year the innovations were granted etc. They include citations of previous patents and of other scientific literature. If the citation part is computerized, the dataset can be used to study the connections among patents (i.e. the connections among assignees or different inventions). There are also restrictions one must consider when using such a dataset. Not every invention is patented, and in most cases not the entire dataset is computerized.

The NBER database of U.S. patents was developed between 1975 and 1999. It includes patents granted in the United States between January 1963 and December 1999. There are 2923922 patents with text descriptions and 850846 patents represented with pictures (for a total of 3774768 patents). There are 16522438 citations between them.

Patent data were used for research already in the 1960s. Later, in 1990, Trajtenberg (Hall et al., 2001) presented the finding that citations are correlated with the importance (value) of innovation. From that time on many works were undertaken to demonstrate the

[1] University of Ljubljana, Slovenia; natasa.kejzar@fdv.uni-lj.si

usefulness of citations in many different purposes (as indicators of spillovers of knowledge (Jaffe, Trajtenberg and Henderson, 1993), for construction of new measures for individual patents (Hall, Jaffe and Henderson, 1997, Hall et al., 2001).

The idea of our work was to look at patents' data as a large network. The vertices of the network are the patents and directed links among them are the citations. For every patent its grant year and application year (year in which the inventor applied for a patent licence) is known and we used these dates to look at the network over time. The intention of this article is to regard the patents network as a temporal network. We present two approaches.

The first approach is to look at the substructure of the entire network. In order to identify an interesting subnetwork we could explore further we searched for the most important themes (Kejžar et. al, 2004) in the entire network. We used Search Path Count Method (Batagelj, 2003) to determine the weights of every patent and every citation and then used the algorithm for determining islands (Zaveršnik, Batagelj, 2004). We extracted the subnetwork of the company that assigned the largest amount of patents in the most important (most powerful) line island (the line island with the largest minimal line weight). Four more networks were extracted according to the number of patents (assigned by other companies) this company cites the most and the category they primarily belong to. We tried to determine the general trends in research and development for the selected companies over the past 37 years.

The idea of the second approach is to study the entire patent network as a temporal network. We shrunk all vertices from a single technological category (aggregated United States Patent and Trademark Office (USPTO) patent classes (Hall et al., 2001)) in the same time slice to one vertex to obtain smaller networks. We studied these smaller networks (one network describes the relationships among technological categories in one time slice) over time. In order to look more closely at a particular segment of the network subcategories or assignee numbers can be used to shrink vertices.

2 The NBER patents network

The entire network has characteristics of a citation network (Kejžar et. al, 2004). It is directed and has no directed cycles (with the exception of one loop). Minimal, maximal and average degrees of the network are shown in Table 1.

Table 1: Degrees of vertices.

	min	max	avg
input	0	779	4.377
output	0	770	4.377
all	1	793	8.754

Figure 1 shows three graphs of vertex degree distributions (input, output and all respectively). The graphs are in logarithmic scale. A one is added to all indegrees and outdegrees in order to satisfy logarithm definability. The degree distributions have a power law tail (Albert, Barabási, 2002), which indicates that the probability that a randomly selected

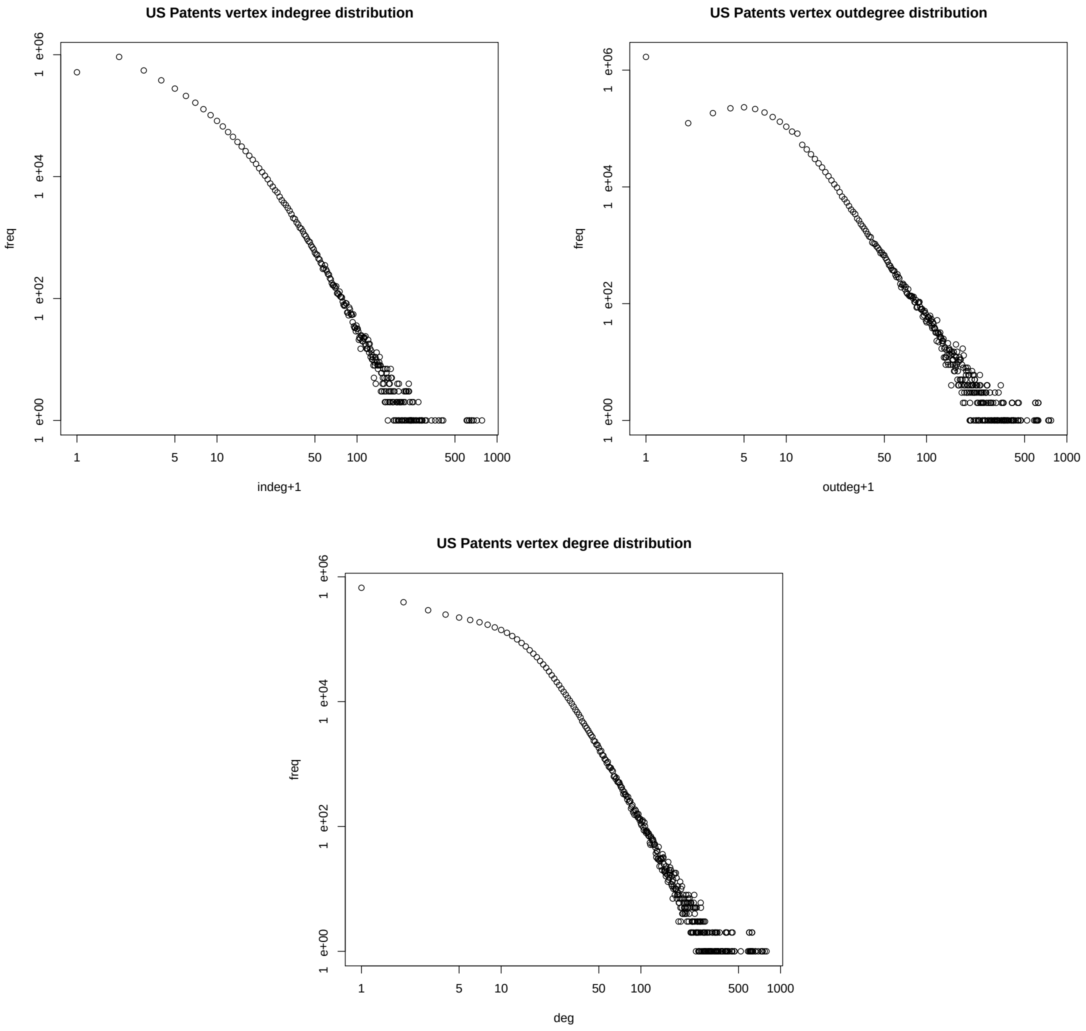

Figure 1: Vertex indegree, outdegree and overall degree distributions.

vertex has k lines is distributed according to $k^{-\gamma}$ where γ is a constant. These kinds of networks are called *scale-free* networks (Albert, Barabási, 2002, Batagelj, Ferligoj, 2003).

The network degree distribution follows a power-law for degrees larger than 20 (degree where linearity in the graphs begins). It is necessary to emphasize here that the observed network is truncated in time. Only patents granted from 1963 to 1999 are included in the network and there is no citation data for the period from 1963 to 1975. All the patents granted during that time are therefore represented as if they did not cite.

Patents of a different grant age have a varying number of received citations due to time truncation. For example, a patent granted in the year 1975 has had 24 years of receiving citations whereas the newest patent is not old enough to be cited already. Therefore it is to be expected that the variability of the received citations is smaller for recently granted patents and higher for patents granted a long time ago.

The differences in citations are also caused by divergent practices of the US Patent Office during the 37 years that are highly correlated with computerization of the data. The number of granted patents per year in the recent years is higher than in previous years. This makes for an increase in the total number of citations made, which could mean that patents get more citations than they did before. The total numbers of citations received for different grant years over a fixed time interval are therefore not easily comparable.

These are just some of the issues one has to consider when analyzing the entire network.

Distribution of the size of weak components was also computed. Table 2 shows the not fully detailed distribution. There exist several small weak components and one huge one.

Table 2: Rough weak components distribution.

size	1 & 2	3 & 4	5 & 6	7 & 8	9 & 10	11 & 12	13 & 14	15 & 16	19	3764117
number	2830	583	276	72	35	12	6	2	1	1

3 Subnetworks: Extraction and analysis

In this section we describe the procedure we used for the identification and extraction of interesting subnetworks as the first approach of studying NBER U.S. patents network as a temporal network. We analyzed the obtained subnetworks over time.

There are two time variables (application year and grant year) in the NBER U.S. patents dataset. Because the actual timing of the patented inventions is closer to the *application year* than to the grant year (Hall et al., 2001), we used application year as the variable which determines the age of a patent. However, there is a drawback to our decision. An application year variable exists only for patents granted since 1967, but because citations exist only for patents granted since 1975 this truncation has an even larger effect on the data than application year.

Hall, Jaffe and Trajtenberg (Hall et al., 2001) showed that the invention is granted within an average of 2 years, with standard deviation of about 1 year after the assignee applied for it. Because of all the described truncations of the data our analysis was done for the time period from 1974 to 1996.

3.1 Extraction

We decided to divide the entire network into subnetworks. To obtain interesting subnetworks we searched for the most important themes (Kejžar et. al, 2004) in the entire network. Search Path Count Method (SPC) (Batagelj, 2003) was used to determine weights of citations. SPC method counts all possible paths that run through arcs in a directed acyclic network. The weight of an arc equals the relative number of paths that run through it. Weights of the citations can be interpreted as indicative of citations' *relative importance in the network.*

We determined line islands (Zaveršnik and Batagelj, 2004) of size $[2, 1000]$ on the new (weighted) network. Line islands are connected subnetworks (groups of vertices with lines) that (locally) dominate according to the weights of lines. We were interested in the island with largest minimal line weight since this island is the most powerful. It turned out it is also one of the two largest islands (islands of 1000 vertices). It consists mainly of technological category 53 (Motors, Engines & Parts). Technological categories for the data were aggregated by Hall, Jaffe, and Trajtenberg (Hall et al., 2001) from USPTO patent classes. These were further aggregated into 6 main categories: Chemical, Computers and Communications, Drugs and Medical, Electrical and Electronics, Mechanical, and Others. The distribution of particular technological categories can be seen in Table 3.

Table 3: Percentage of categories represented in the largest island.

category	13	46	55	21	69	19	43	45	22	53
percent	0.1	0.1	0.1	0.2	0.2	0.5	4.0	4.1	7.7	*78.2*

Figure 2 shows the island. Older patents are presented at the bottom of the figure and the newest patents are at the top. We examined also the titles of the patents. They were obtained from the website of USPTO (USPTO). We automatized the search using the statistical package R (R package) and its package XML. The most important patents in the island are displayed as darker boxes with their titles and application years. As already found (Kejžar et. al, 2004), patents in this island deal with one main topic — *fuel evaporation in internal combustion engines*.

We were interested in the assignees for patents from the most powerful island. Assignee identifications were obtained from the file on the USPTO website (USPTO). There were 48 patents that lacked this information. Figure 3 presents the assignee distribution according to the number of patents they have in the most powerful island. The three largest contributors of patents can be seen in Table 4.

Table 4: Three largest contributors of patents from the largest island.

# patents	assignee id	assignee
136	404330	Nissan Motor Company, Limited
126	582975	Toyota Jidosha K.K.
98	256400	Honda Giken Kogyo Kabushiki Kaisha (Honda Motor Co., Ltd.)

Because we were interested in studying of the most powerful part of the network over time, we extracted the subnetwork of the company that assigned the largest amount of patents to the most important island. This is the Nissan company. The subnetwork we extracted consists of all the patents assigned by Nissan and all the patents cited by Nissan. Table 5 shows five companies whose patents are most frequently cited by Nissan patents.

According to Table 5 and the category Nissan patents mostly belong to (category 53), we extracted four more subnetworks (Toyota, Honda, General Motors and Ford). The important characteristics of the five subnetworks can be seen in Table 6.

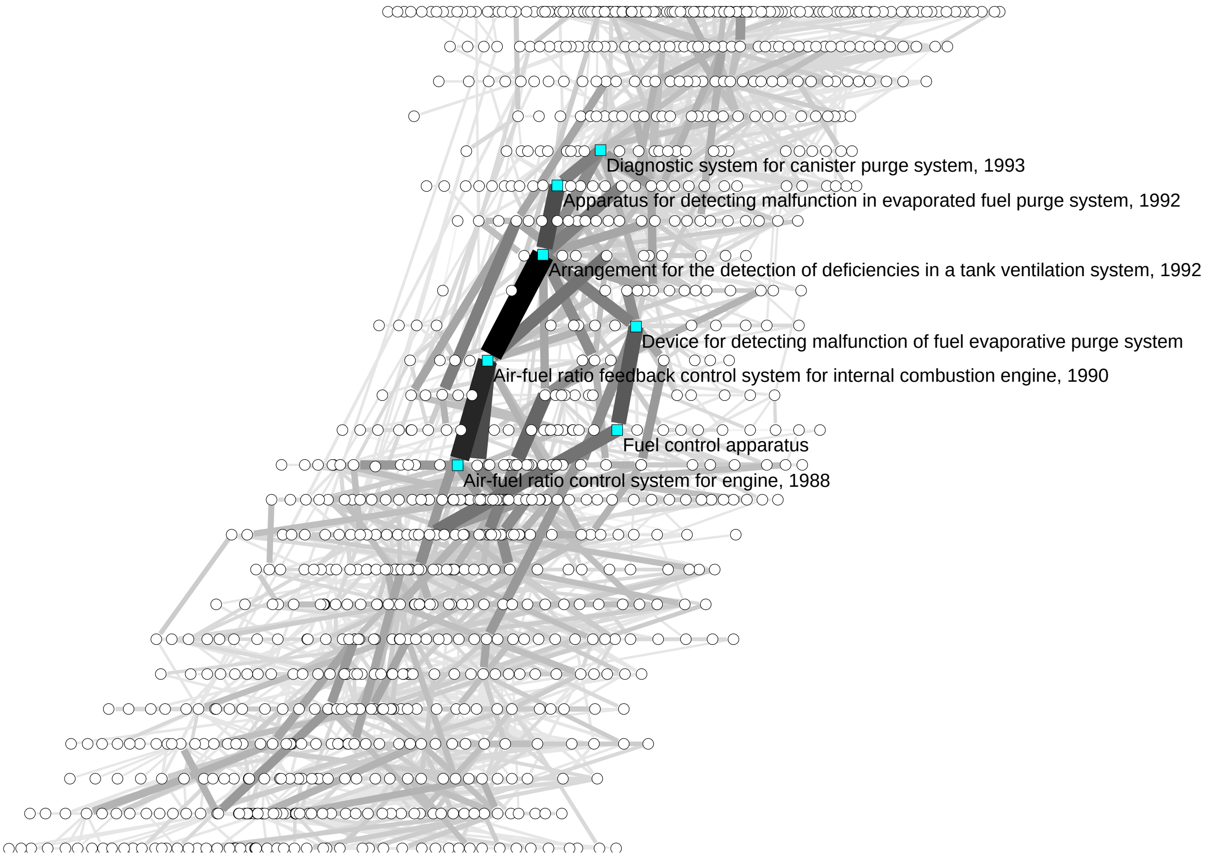
Diagnostic system for canister purge system, 1993
Apparatus for detecting malfunction in evaporated fuel purge system, 1992
Arrangement for the detection of deficiencies in a tank ventilation system, 1992
Device for detecting malfunction of fuel evaporative purge system
Air-fuel ratio feedback control system for internal combustion engine, 1990
Fuel control apparatus
Air-fuel ratio control system for engine, 1988

Figure 2: Largest island, square root of line weights multiplied by half a million is used.

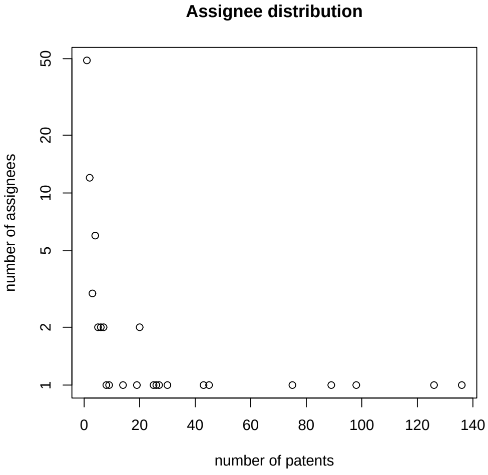

Figure 3: Assignee distribution for patents from the largest island.

Table 5: Five companies most frequently cited by Nissan.

# cited patents	company
981	**Toyota** Jidosha K.K.
943	**General Motors** Company
585	**Honda** Giken (Honda Motor Co., Ltd.)
519	Robert Bosch GMBH
448	**Ford** Motor Company

Table 6: Basic features of extracted subnetworks.

subnetwork	# all patents	# patents from the company
Nissan	27035	5893
Toyota	24730	5281
Honda	26369	5456
General Motors	47447	9141
Ford	34353	4757

3.2 Analysis

Figure 4 presents outdegree distribution only for Nissan patents in the Nissan subnetwork. For all other patents in the Nissan subnetwork the outdegree is clearly 0. The same pattern also appears in the other 4 subnetworks. It tells us that most of the patents cite fewer than 20 other patents, although there are patents that cite a very large number of other patents.

The distribution of weak components also represents an interesting feature of the networks. The rough distribution of weak components for the Nissan company can be seen in Table 7.

Table 7: Weak components distribution for Nissan.

size	2 & 3	4 & 5	6 & 7	8 & 9	10 & 11	12 & 13	17 & 19	23 & 25	33 & 37	45	*2220*
number	272	43	19	7	4	3	4	2	2	1	1

There is one huge weak component of 2220 vertices as well as some extremely small components. There is not much difference in the other four subnetworks. The sizes of the largest 3 weak components for the other networks can be seen in Table 8. The smallest difference between the largest and second largest component can be seen in the Ford network.

Due to large weak components it is interesting to look at *self citations* (Hall et al., 2001) (i.e. citations made to the same assignee) in the networks. Because not all patents contain information about assignee (unassigned patents make comprise about 18.4% of the data (Hall et al., 2001)), the lower and upper bound of self citations are computed. The equation to compute lower bound is

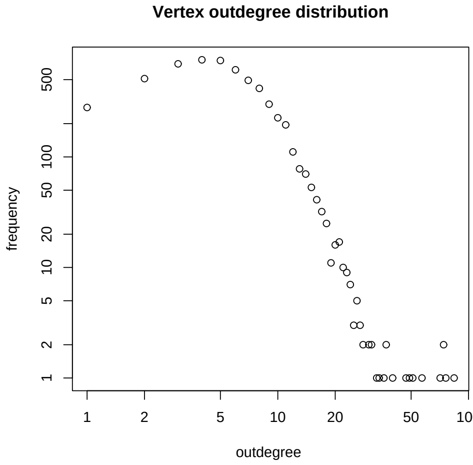

Figure 4: Logarithmic outdegree distribution for Nissan patents.

Table 8: Sizes of the largest three weak components for subnetworks of Toyota, Honda, GM and Ford.

company	3rd largest component	2nd largest component	largest component
General Motors	33	49	3713
Toyota	26	37	1835
Honda	36	37	2252
Ford	61	109	961

$$\text{lower bound [\%]} = \frac{\text{\# citations made to patents of same assignee}}{\text{\# of all citations made}} \times 100\ .$$

For upper bound the denominator changes to the number of citations where patents *have* an assignee code.

Mean percentage of self citations for the entire network and the subnetworks follows in Table 9. As seen already from the difference in the largest weak components Ford has the lowest self citation percentage (even less than for the network as a whole) whereas the other companies have a significantly larger self citation percentage than the entire network.

We were interested in how the percentage of self citations is changing over time. The results can be seen in Figures 5 and 6. It is noticeable that *Japanese companies* have some rather distinct local peaks. These correspond to strong research and development in particular fields during a particular period of time. Self citations have an overall tendency to increase. On the other hand *American companies* stay roughly within the same percentage throughout the three decades with the exception of the interval around 1975 for Ford.

Table 9: Mean percentage of self citations for specified networks.

	patents (Hall et al., 2001)	Nissan	Toyota	Honda	GM	Ford
lower bound [%]	11	14.8	13.0	16.0	15.5	9.2
upper bound [%]	13.6	17.4	15.2	19.6	21.4	11.5

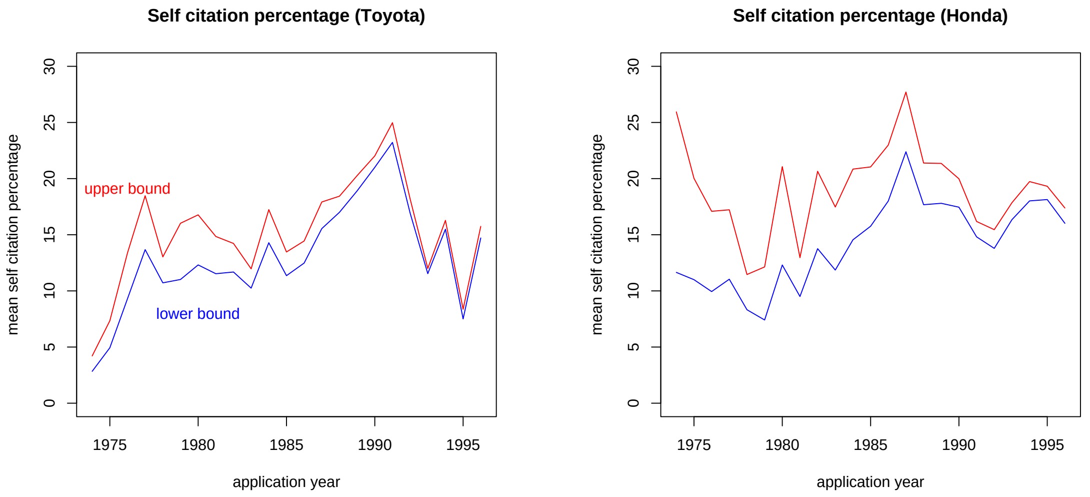

Figure 5: Development of self citations for Japanese companies.

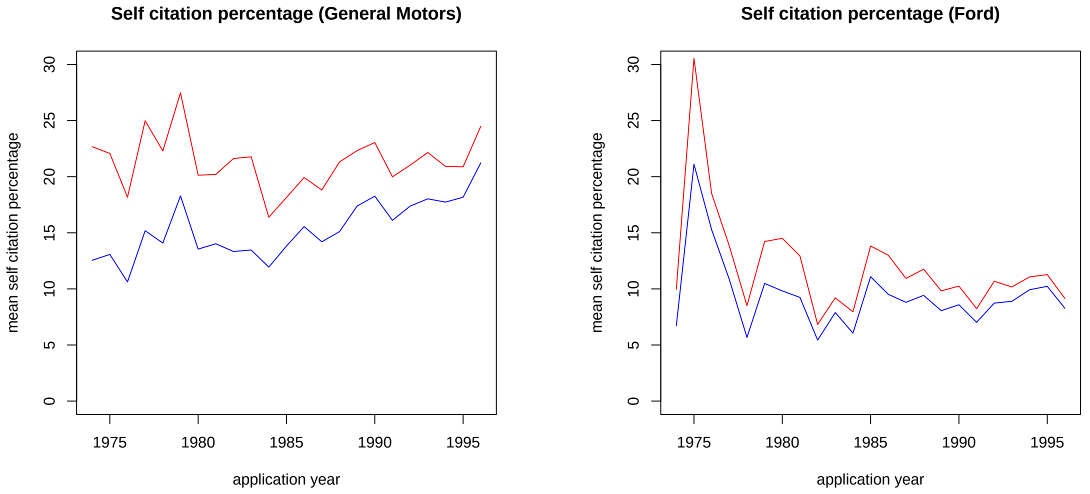

Figure 6: Development of self citations for American companies.

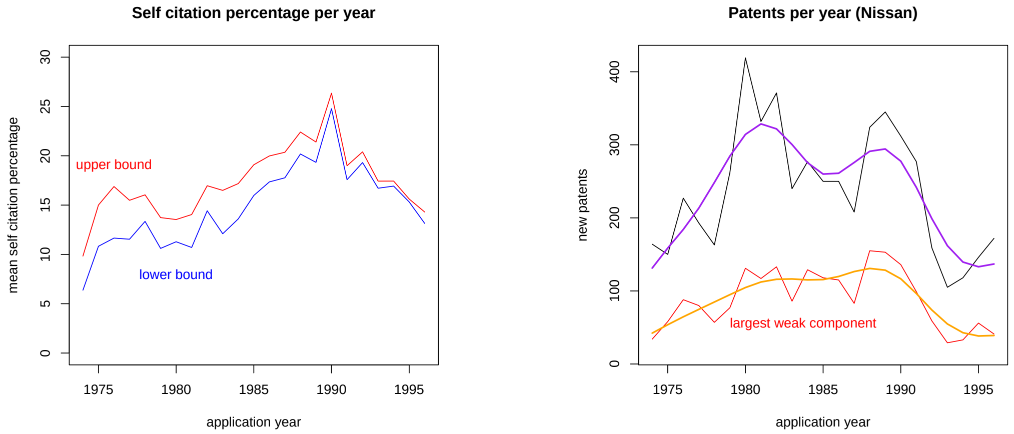

Figure 7: Development of self citations and growth of patents for Nissan company.

To confirm that a high self citation rate is connected with high research and development we examined the growth of patents in 1 year time slices. Figure 7 shows the citations and growth graph for Nissan patents. It can be seen that the second highest peak in growth highly correlates in time with the highest peak of self citations.

The lower curve in the second graph represents the growth of patents in the largest weak component. Both increases were smoothened by Spencer 15-point moving average filter where cubic polynomials pass undistorted (Brockwell, Davis, 1991).

4 Second approach: Shrinking of the network

In the previous section we tried to determine the general trends in research and development for the selected companies over the past 37 years. In this section we propose another approach for studying the patent network as a temporal network. Figure 8 presents the increase in all patents through time. The left graph shows the growth of patents in one year time slices and the right graph the cumulative growth. It can be seen that the growth of patents as a whole increases from the year 1985 on. This can be noticed on the right graph as the upward deviation from the line in the cumulative growth of patents from the late 1980s on.

Is there any difference in patent growth when separating patents according to their category? How are the relationships among categories changing over time? To answer these questions we shrunk all vertices from the same category in the same time slice into one vertex. We then studied the obtained smaller networks over time.

First we had to consider how to choose a sliding time window. We used the knowledge about *backward citation lags* (Hall et al., 2001), that is the time difference between the grant year of the citing patent and that of the cited patents. The highest number of cited patents were granted 3 and 4 years before the citing patent. The number of even older patents that were cited drastically decreases with age. Since application year and grant year somehow correlate (right graph of figure 8), we used this information. We chose time slices of 4 years with no history. All the citations with a lag of more than 4 years

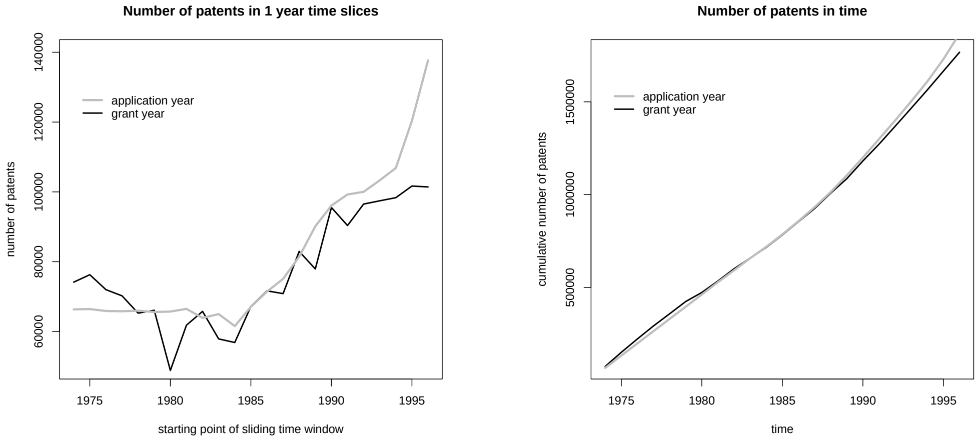

Figure 8: Growth of number of patents in one year time slices and cumulative growth of patents according to grant year and application year.

were excluded. We could interpret our decision as less lagged citations as part of the research and development at the *current time* and the other citations as used as *references to well known methods* patented earlier. Figure 9 shows four shrunken networks in a 4 year sliding time window.

We looked closely at the growth of the number of patents for every single category and the relative growth of citations within each category (Figure 10). The findings are similar to the overall patents. The growth of patents drastically increases from the year 1980 on. There are two fast developing fields (Computers & Communication and Drugs & Medical). From the relative growth of citations it can be seen that number of citations per patent is slightly increasing. The only exception is in the category Computer & Communication, where the growth is immense (the number of citations is now more than twice as high as it was thirty years ago).

In order to take the weight (strength) of *connections* among categories into account we computed *hubs and authorities* (Batagelj, Ferligoj, 2003) on the shrunken networks. The method works for directed networks. Let denote a directed network as $N(V, L)$ where V is a set of all vertices and L is a set of all arcs. Every vertex $v \in V$ gets two weights (x_v and y_v). Vertex is a good hub (has large weight x_v) if it points to many good authorities

$$x_v = \sum_{u:(u:v)\in L} y_u$$

and it is a good authority (has large weight y_v) if it is pointed to by many good hubs

$$y_v = \sum_{u:(v:u)\in L} x_u.$$

The corresponding vectors x and y are the principal eigenvectors of matrices W^TW and WW^T, where W is the adjacency matrix of the network. The iterative method to get vectors x and y is implemented in `Pajek`.

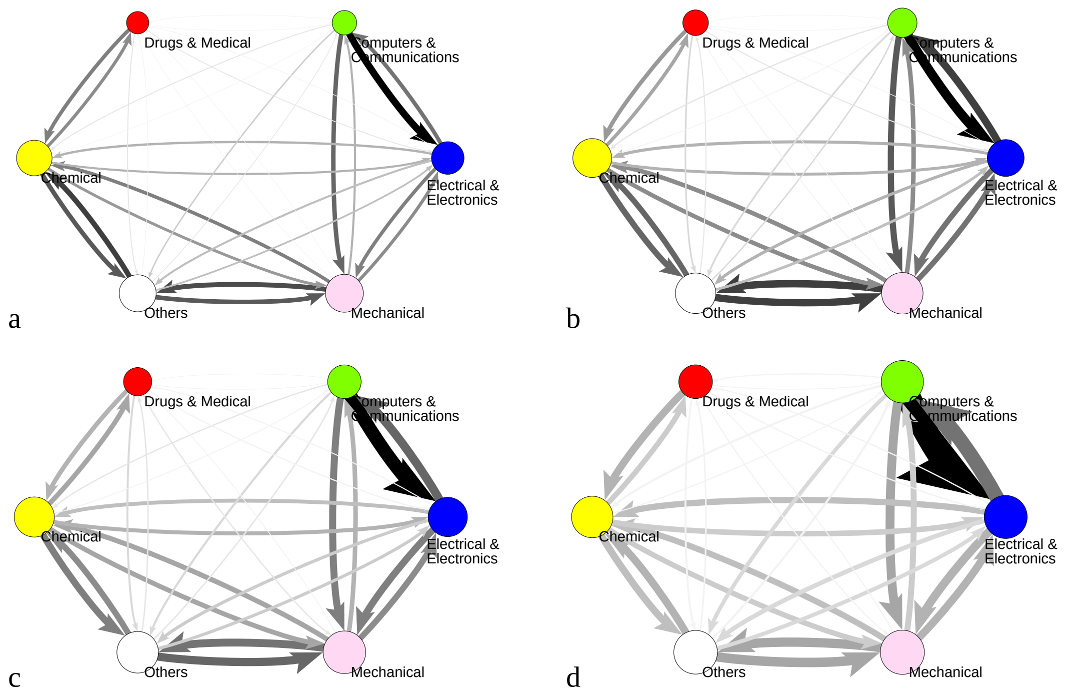

Figure 9: Four examples of network in 4 year time window shrunk by categories. Starting years: figure a – 1984, figure b – 1987, figure c – 1990 and figure d – 1993.

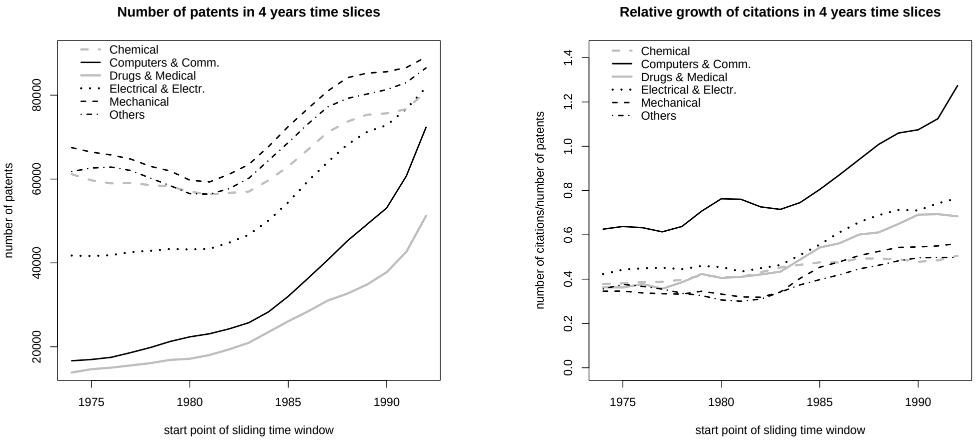

Figure 10: Growth of number of patents and relative growth of citations within category.

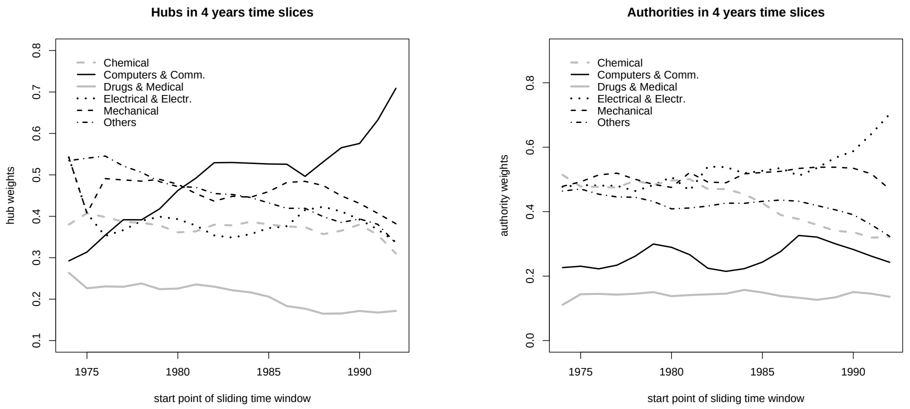

Figure 11: Hubs and authorities weights for categories over time.

Because the strength of connections *in* a category is a few orders more powerful than the strength of connections *among* categories, we decided to ignore the connections in a category, so we deleted all loops in the shrunken networks. We computed hubs and authorities on the new networks.

Figure 11 shows the changing of weights of categories for hubs and authorities over time. We can see that categories with large values of hubs up until 1980 are Mechanical and Others. They are completely overtaken by Computers & Communication later on. Hubs could represent categories which *combine knowledge* from other important technological categories the most. Drugs & Medical has the lowest weights, which can be explained by the fact that it depends almost solely on the category Chemical.

Categories with large values of authorities are Electrical & Electronic and Mechanical (which has been diminishing in relative influence from 1990 on). Authorities can be interpreted as categories that play a very important role in setting the foundations – *basic knowledge* that is therefore often cited by other patents. The lowest weight has, again, Drugs & Medical. Computers & Communication has the second lowest weight. They both belong to *applied sciences*.

5 Conclusion

We looked at U.S. patents data as a large network. Patents represent its vertices and the citations among them represent the links. For every patent its application year (the year a patent licence was applied for for an invention) was used to look at the network over time.

Our first approach to study the patent network as a temporal network was to identify an important part of the network and according to that extract some significant subnetworks. To obtain these subnetworks we searched for the most important themes (Kejžar et. al, 2004) in the entire network. We used Search Path Count Method (Batagelj, 2003) to gain weights of every patent and every citation and then used the algorithm for determining islands (Zaveršnik, Batagelj, 2004). The important part of the network represented the most powerful line island. We extracted subnetworks of five different companies with respect to the main category of this island and its largest assignee (company). We tried to determine the general trends in research and development for the selected companies over the past thirty-seven years.

As well, we proposed another approach for studying patent network as a temporal network, namely by shrinking all vertices from the same category in the same time slice to one vertex and then observing over time the obtained smaller networks. With the approach of shrinking vertices from the same category in the same time slice we obtained a very useful dataset of temporal networks. In order to look more closely at a specific segment of the network subcategories or assignee numbers can be then used to shrink vertices.

Acknowledgement

The author wishes to thank professor Vladimir Batagelj for his help and useful comments.

References

[1] Albert, R. and Barabási, A. L. (2002): Statistical mechanics of complex networks. *Reviews of Modern Physics*. **74**, 47, 1–54. http://arxiv.org/abs/cond-mat/0106096

[2] Batagelj, V. (2003): Efficient algorithms for citation network analysis. http://arxiv.org/abs/cs.DL/0309023

[3] Batagelj, V. and Ferligoj, A. (2003): Analiza omrežij (Network Analysis, lectures). http://vlado.fmf.uni-lj.si/vlado/podstat/AO.htm

[4] Batagelj, V. and Mrvar, A.: Pajek. http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[5] Brockwell, P. J. and Davis, R. A. (1991): *Time Series: Theory and Methods*. Berlin: Springer.

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

[7] Hall, B.H., Jaffe, A.B., and Trajtenberg, M. (2001): *The NBER U.S. Patent Citations Data File*. NBER Working Paper 8498. http://www.nber.org/patents/

[8] Kejžar, N., Korenjak-Černe, S., and Batagelj, V.: Analysis of U.S. patents network: Determining main themes. Presented at Sunbelt XXIV Conference, Portorož, May, 2004. In preparation.

[9] The R Project for Statistical Computing. http://www.r-project.org/

[10] The United States patent and trademark office. http://patft.uspto.gov/netahtml/srchnum.htm

[11] Zaveršnik, M. and Batagelj, V.: *Islands – identifying themes in large networks*. Presented at Sunbelt XXIV Conference, Portorož, May, 2004. In preparation.