Analyzing the Structure of U.S. Patents Network

Vladimir Batagelj[1], Nataša Kejžar[2], Simona Korenjak-Černe[3], and Matjaž Zaveršnik[1]

[1] Department of Mathematics, FMF, University of Ljubljana,
Jadranska 19, SI-1000 Ljubljana, Slovenia
[2] Faculty of Social Sciences, University of Ljubljana,
Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia
[3] Faculty of Economics, EF, University of Ljubljana
Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia

Abstract. The U.S. patents network is a network of almost 3.8 millions patents (network vertices) from the year 1963 to 1999 (Hall et al. (2001)) and more than 16.5 millions citations (network arcs). It is an example of a very large citation network.

We analyzed the U.S. patents network with the tools of network analysis in order to get insight into the structure of the network as an initial step to the study of innovations and technical changes based on patents citation network data.

In our approach the SPC (Search Path Count) weights, proposed by Hummon and Doreian (1989), for vertices and arcs are calculated first. Based on these weights vertex and line islands (Batagelj and Zaveršnik (2004)) are determined to identify the main themes of U.S. patents network. All analyses were done with `Pajek` – a program for analysis and visualization of large networks. As a result of the analysis the obtained main U.S. patents topics are presented.

1 Introduction

Patents are a very good source of data for studying the innovation development and technical change because each patent contains information on innovation, inventors, technical area, assignee etc. Patent data also include citations to previous patents and to scientific literature, which offer the possibility to study linkages between inventions and inventors. On the other hand we have to be aware of the limitations when using such datasets, since not all inventions are patented, the patent data are not entirely computerized, and that it is hard to handle very large datasets.

The database on U.S. patents (Hall et al. (2001)) was developed between 1975 and 1999. It includes U.S. patents granted between January 1963 and December 1999. It counts 2,923,922 patents with text descriptions and other 850,846 patents represented with scanned pictures, altogether 3,774,768 patents. There are 16,522,438 citations between them. Since it is a legal duty for the assignee to disclose the existing knowledge, a citation represents previously existing knowledge contained in the patent.

The idea of using patent data for economic research originated from Schmookler (1966), Scherer (1982), and Griliches (1984). Hall et al. (2001) included more information about patents in the analyses and also demonstrated the usefulness of citations.

The idea of our work was to look at the patents data as a large network. In the network patents are represented by vertices. Two patents (vertices) are linked with a directed link, an *arc*, when one cites the other one. We used the SPC method to obtain the weights of patents and their citations. Weight of a particular patent or particular citation can be interpreted as a relative importance of that patent or that citation in the network. We used weights to determine islands – groups of 'closely related' vertices.

Hall, Jaffe, and Trajtenberg aggregated more than 400 USPTO (United States Patent and Trademark Office) patent classes into 36 2-digit technological subcategories, and these are further aggregated into 6 main categories: Chemical, Computers and Communications, Drugs and Medical, Electrical and Electronics, Mechanical, and Others. We examined the constructed variable of technological subcategory and checked the titles of patents in order to confirm our hypothesis, that islands determine specific theme of patents.

2 Search path count method

Let us denote a network by $N = (V, L)$, where V is a set of *vertices* and L is a set of *arcs*. The arc (v, u) goes from vertex $v \in V$ to vertex $u \in V$ iff the patent represented by v cites the patent represented by u. This network is a *citation network*. Citation networks are usually (almost) *acyclic*. The cycles, if they exist, are short. Network can be converted to acyclic one by using different transformations – for example, by simply shrinking the cycles. Hummon and Doreian proposed in 1989 three arc weights to operationalize the importance of arcs in citation networks: (1) node pair projection count method, (2) search path link count method, and (3) search path node pair method.

Batagelj (1991, 2003) showed that the use of SPC (Search Path Count) method computes efficiently, in time $O(|L|)$, the last two (2) and (3) of Hummon and Doreian's weights. The SPC method assumes that the network is acyclic. In an acyclic network there is at least one *entry* – a vertex of indegree 0, and at least one *exit* – a vertex of outdegree 0. Let us denote with I and O the sets of all entries and all exits, respectively. The SPC algorithm assigns to each vertex $v \in V$ as its value the number of different I-O-paths passing through the vertex v; and similarly, to each arc $(v, u) \in L$ as its weight the number of different I-O-paths passing through the arc (v, u). These counts are usually normalized by dividing them by the number of all I-O-paths.

We calculated normalized weights of edges and vertices for the U.S. patents network using the SPC method in `Pajek`. The number of all paths through the network is 1,297,400,940,682. We multiplied the weights with one million since the normalized values of most of the weights were very small.

3 Determining islands

The following table

size	1 & 2	3 & 4	5 & 6	7 & 8	9 & 10	11 & 12	13 & 14	15 & 16	19	3,764,117
number	2830	583	276	72	35	12	6	2	1	1

shows the (not fully detailed) distribution of the size of *weak components.* A weak component is a subnetwork of vertices that are connected when disregarding the arcs direction. There exist several small weak components and one huge one (3,764,117 vertices). This implies that most of the patents are somehow connected to almost all other patents. Patents in small weak components might be the early ones (granted before the year 1975), which would be densely connected if the citation data were available or there might exist patents that started a very specific topic which is indeed just locally connected.

Based on the calculated weights more informative connected subnetworks can be also determined. For this purpose we used line and vertex islands. Islands (Batagelj and Zaveršnik (2004), Zaveršnik (2003)) are connected subnetworks (groups of vertices) that locally dominate according to the values of vertices or lines.

Let $N = (V, L, p)$ be a network with vertex property $p : V \to \mathbb{R}$. Nonempty subset of vertices $C \subseteq V$ is called a *vertex island* of network N if the corresponding induced subnetwork is connected and the weights of the neighboring vertices $N(C)$ are smaller or equal to the weights of vertices from C

$$\max_{u \in N(C)} p(u) \leq \min_{v \in C} p(v).$$

The line islands are defined similarly. Let $N = (V, L, w)$ be a network with line weight $w : L \to \mathbb{R}$. Nonempty subset of vertices $C \subseteq V$ is called a *line island* of network N if there exists a spanning tree T in the corresponding induced subnetwork, such that the lowest line of T has larger or equal weight than the largest weight of lines from C to the neighboring vertices

$$\max_{(u,v) \in L, u \notin C, v \in C} w(u, v) \leq \min_{e \in L(T)} w(u, v).$$

Let us look at values $p(v)$ of vertices as *heights of vertices.* The network can be seen as some kind of a *landscape*, where the vertex with the largest value is the highest peak. Eliminating the vertices (and the corresponding lines) with height lower than t, we obtain a group of internally connected subnetworks – islands called a *vertex cut* at *cutting level* t. Unfortunately this does not give a satisfying result. We are usually interested in subnetworks with specific number of vertices – not smaller than k and not larger than K – trying to embrace single theme clusters. To identify such islands we have to determine vertex cuts at all possible levels and select only those islands of the selected size. Batagelj and Zaveršnik (2004) developed an efficient algorithm for determining such islands. It is implemented in `Pajek`.

We determined vertex islands of sizes $[1, 300]$. When determining line islands of sizes $[2, 300]$ we excluded all 'weak, submerged' lines (lines in the line island with weights lower than the largest value of the line linking island to the rest of network). We obtained 24,921 vertex islands on 36,494 vertices and 169,140 line islands on 424,191 vertices.

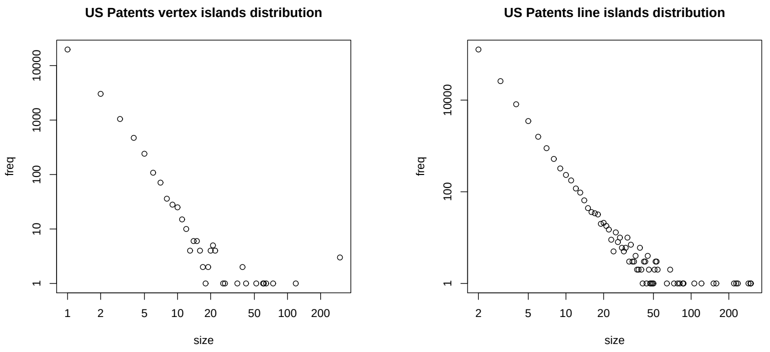

Fig. 1. Distributions of the islands based on their size.

Figure 1 shows the size distribution of islands (vertex and line islands respectively). The x axis represents the size of islands and the y axis represents the frequency of islands. It can be seen that the relation in the log-log scale is almost linear. With the increase of the size of islands, their frequency decreases with power-law.

4 Line islands

In this section some of the most interesting line islands will be presented. We chose them with respect to their size and to the size of the weights on the lines. We scanned through all the islands that have at least 21 vertices and islands that are smaller but with the minimal line weight 10. There are 231 such islands. We were interested in the technological subcategory, title of the patent and grant year for each vertex (each patent) of the island.

Titles of the patents were obtained from the website of The United States Patent and Trademark Office. We automatized their extraction using statistical system R and its package XML.

We found out that the patents in an individual island are dealing with the same (or very similar) topic which is very likely in the same technological subcategory. We noticed that islands with smaller weights are mainly of categories Others (category code 6) and Chemical (code 1). With increase in the minimal island weight, categories first change in favor of Drugs and Medical category (code 3), then Electrical and Electronics (code 4) and last to Computers and Communications (code 2). Interestingly Mechanical category was not noticed throughout the scan.

The largest island is surprisingly mostly of category Chemical. It has exactly 300 vertices and its minimal weight is the highest (3332.08). The patents are mostly from the category code 19 (Miscellaneous-Chemical). 254 vertices or approximately 84.7% are from this category. The second largest category code is 14 (Organic Compounds), which counts 27 vertices or 9%. When examining the titles of the patents we found out that this island is about *liquid crystals*, that could be used for computer displays. This somehow connects the theme to the category Computers and Communications and makes the theme in the largest island less surprising.

The second largest island (298 vertices) is even more homogenous in topic than the first one. Its theme is about *manufacturing transistors and semiconductor devices*, which is classified in category code 46 (Semiconductor Devices). There are 270 (approx. 90.6 %) vertices in even more specific classification group code 438 (USPTO 1999 classification class Semiconductor device manufacturing process).

We also observed small size islands with large minimal weights. The topic of 5 islands within 31 islands with the largest minimal weights deals with the *internal combustion engine for vehicles and its fuel evaporation system.* This very likely implies that there is a huge island (more than 300 vertices) about this theme, but due to our maximum island size restriction (300 vertices in the island) there are only its peaks (*subislands*) captured in our result. We verified this hypothesis by determining line islands of a larger size. When calculating islands of size $[2, 1000]$ there were 2 islands of 1000 vertices with the largest minimal weights. The theme of one of them is about internal combustion engines (for details see Kejžar 2005). It contains the small islands captured with the initial calculation of the line islands. This shows that this theme is much broader than most of other themes and hence it was not possible to embrace it completely with an island of maximum 300 vertices.

Figure 2 shows an island with 229 vertices and the seventh largest minimal weight, that has 3 strong theme branches. Through the years patents were granted (the oldest patents are at the bottom of the network) these 3 different topics became connected. In the Figure 3 the title of every patent in the first branch is shown separately. We can see that the topic of the longest branch is about *television market research with video on demand.* The second branch is about the *identity verification apparatus*, and the third about the *computer security system.* The three branches are thematically not far away, so the

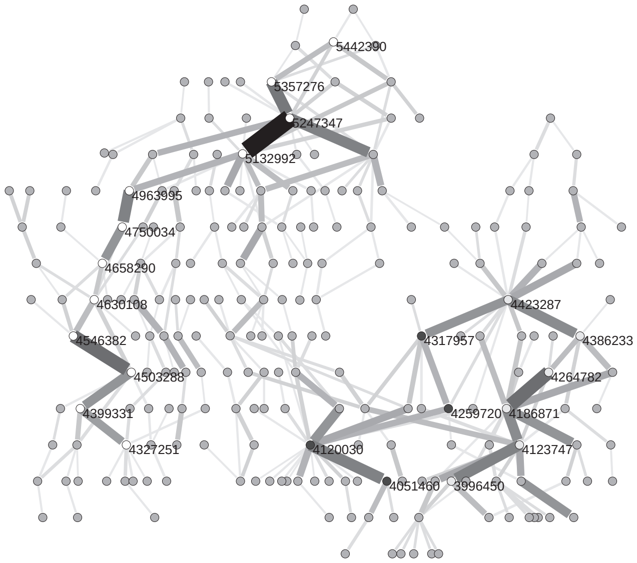

Fig. 2. An island with 3 theme branches.

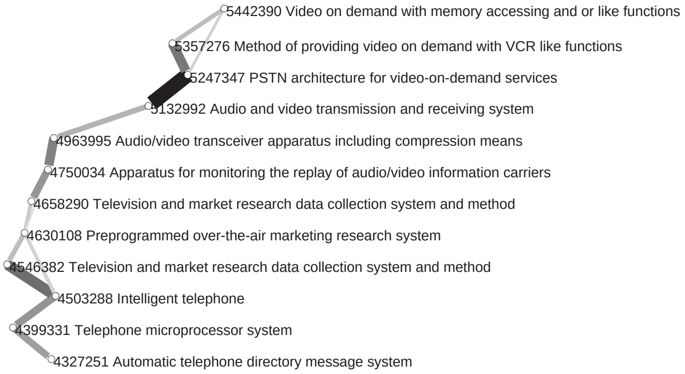
5442390 Video on demand with memory accessing and or like functions
5357276 Method of providing video on demand with VCR like functions
5247347 PSTN architecture for video-on-demand services
5132992 Audio and video transmission and receiving system
4963995 Audio/video transceiver apparatus including compression means
4750034 Apparatus for monitoring the replay of audio/video information carriers
4658290 Television and market research data collection system and method
4630108 Preprogrammed over-the-air marketing research system
4546382 Television and market research data collection system and method
4503288 Intelligent telephone
4399331 Telephone microprocessor system
4327251 Automatic telephone directory message system

Fig. 3. First branch of the island.

findings of a patent that mainly belongs to one of them use (or are used) also in the other two branches. This shows in the island the connections among branches that are strong enough to merge the three branches together in one (a bit more diverse) island.

5 Vertex islands

Some of the most interesting vertex islands were obtained by restricting the minimal vertex island's size to 10, or the minimal vertex island's size to 5 and weights larger than 10. There are 123 such islands on $2,971$ vertices.

Three of them have size 300. The one of them with the largest weights includes patents mostly from the category Chemical. The main theme in this island is the *liquid crystals* which is also the main theme in the main line island.

The next largest vertex island beside the largest three is the island on 119 patents. It is very homogenous – all patents belong to the Electrical and Electronics subcategory Semiconductor Devices (code 46), and all patents are classified into USPTO Patent Class 438.

Large and powerful vertex islands show a very similar structure of themes as in line islands. This is not surprising since weights (from the SPC method) on lines and neighboring vertices are highly correlated. It can be noticed that significantly less vertex islands than line islands are obtained when the same size range is considered.

There are also some small vertex islands with very high weights. Some of them are presented in the Table 1. The meaning of the codes for technical subcategories can be obtained from Hall, Jaffe and Trajtenberg article about the patents data.

Table 1. Some of the small vertex islands with the highest weights (minw > 200)

Island No.	**Size**	**Weights** **minw** **maxw**	**Subcategory**	**Theme**
24900	5	1018.58 1846.06	53	controlling an ignition timing for an internal combustion engine
24878	5	632.32 1039.10	46 19	fabricating monocrystalline semiconductor layer on insulating layer by laser crystallization
24874	8	590.82 1043.82	24	multiprocessor cache coherence system
24811	10	357.48 901.61	22 21, 49	area navigation system including a map display unit
24806	10	343.46 562.06	22	programmable sequence generator for in-circuit digital testing
24797	10	322.46 966.49	53	valve timing control system for engine
24796	10	318.69 1818.92	24 12, 19	track transverse detection signal generating circuit

6 Conclusion

An approach to determine main themes in large citation networks is presented, which can be viewed as a kind of network clustering. A very large network of U.S. patents was used as an example. We used the SPC (Search Path Count) method to get vertex and line weights. Vertices and lines with higher weights represent more important patents and citations in the network. We used them to determine vertex and line islands of the network. Islands are non overlapping connected subsets of vertices. Due to the citation links between the vertices, vertices have similar characteristics (similar topics in our case). Therefore islands can be viewed as thematic clusters.

The characteristics of patents in more than 300 islands were examined. The islands that were examined were selected by their size and minimal line or vertex weight. The results confirmed the hypothesis that an island consists of vertices with similar features (in our case themes). Due to the limited space in this paper we could only present the most important and the most interesting vertex and line islands. There are some differences between the vertex and the line islands, but the main themes and the main islands remain roughly the same.

References

ALBERT, R. and BARABÁSI, A.L. (2002): Statistical Mechanics of Complex Networks. *Reviews of Modern Physics, 74, 47* http://arxiv.org/abs/cond-mat/0106096

BATAGELJ, V. (2003): Efficient Algorithms for Citation Network Analysis. http://arxiv.org/abs/cs.DL/0309023

BATAGELJ, V. and FERLIGOJ, A.(2003): Analiza omrežij. (Lectures on Network analysis.): http://vlado.fmf.uni-lj.si/vlado/podstat/AO.htm

BATAGELJ, V. and MRVAR, A.: Pajek. Home page: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

BATAGELJ, V. and MRVAR, A. (2003): Pajek – Analysis and Visualization of Large Networks. In: Jünger, M., Mutzel, P., (Eds.): *Graph Drawing Software.* Springer, Berlin, 77-103.

BATAGELJ, V. and ZAVERŠNIK, M.: Islands – identifying themes in large networks. Presented at Sunbelt XXIV Conference, Portorož, May 2004.

HUMMON, N.P. and DOREIAN, P. (1989): Connectivity in a Citation Network: The Development of DNA Theory. *Social Networks, 11, 39-63.*

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/

KEJŽAR, N. (2005): Analysis of U.S. Patents Network: Development of Patents over Time. *Metodološki zvezki, 2, 2,195-208.* http://mrvar.fdv.uni-lj.si/pub/mz/mz2.1/kejzar.pdf

ZAVERŠNIK, M. (2003): Razčlembe omrežij. (Network decompositions). PhD. Thesis, FMF, University of Ljubljana.

The United States Patent and Trademark Office. http://patft.uspto.gov/netahtml/srchnum.htm

The R Project for Statistical Computing. Home page: http://www.r-project.org/