PREDICTION of EMERGING TECHNOLOGIES BASED on ANALYSIS of the U.S. PATENT CITATION NETWORK

Péter Érdi[1,2]

[1] *Center for Complex Systems Studies, Kalamazoo College, Kalamazoo, Michigan*

[2] *Dept. Biophysics, KFKI Res. Inst. Part. Nucl. Phys. Hung. Acad. Sci. Budapest, Hungary*

KALAMAZOO COLLEGE

PROSPECTIVES ALUMNI STUDENTS & STAFF VISITORS

Center for Complex Systems Studies

HOME

last revised: March 13, 2008

web@kzoo.edu

Comments for CCSS specific pages

BUDAPEST CNS GROUP

Content

1. Data, Rules, Prediction:
   Lessons from Tycho de Brahe, Kepler and Newton

2. The Rules Behind the Development of Patent Citation Network

3. Prediction of Emerging Technologies based on Co-citation Clustering

Data, Rules, Prediction: Lessons from Tycho de Brahe, Kepler and Newton

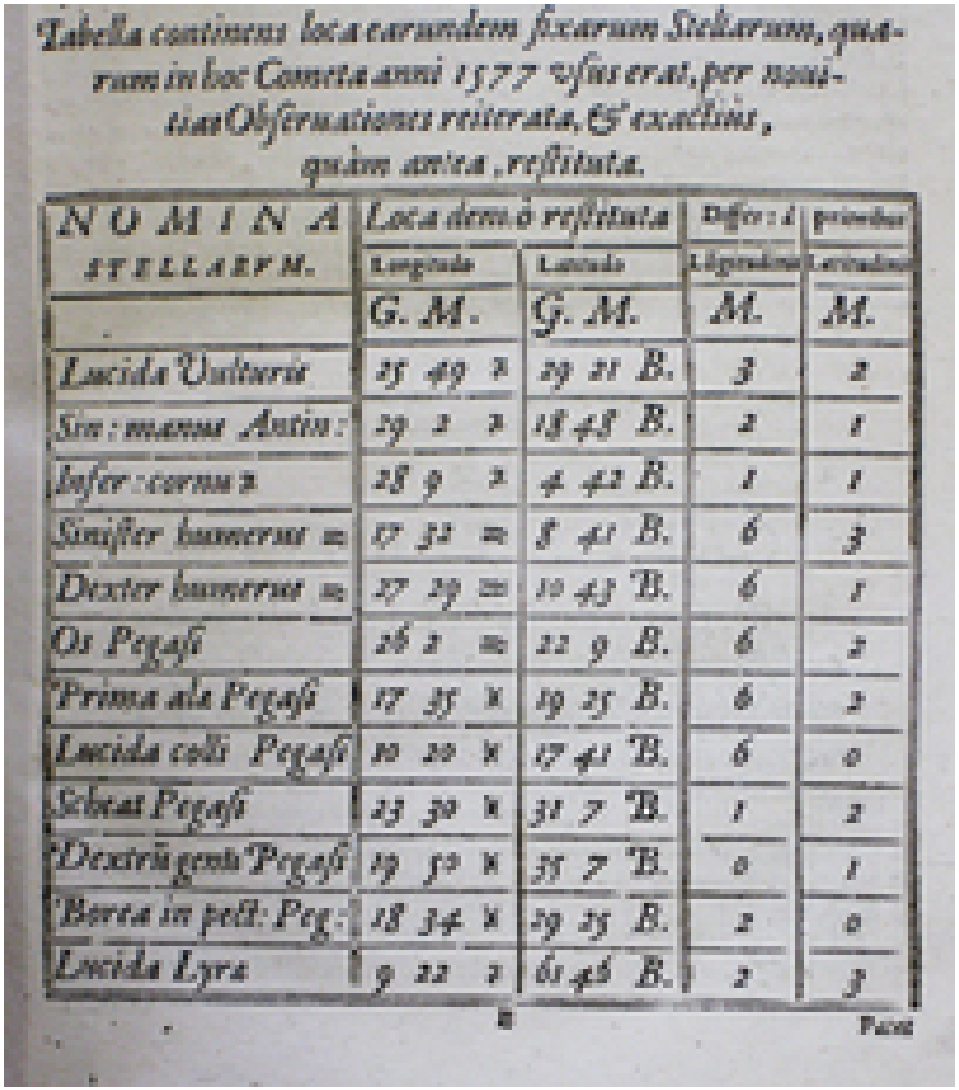


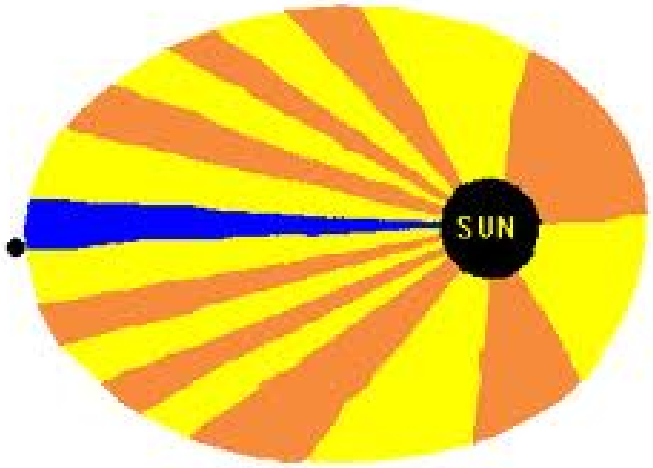
Tabella continens loca earundem fixarum Stellarum, quarum in hoc Cometa anni 1577 vsus erat, per nouitias Obseruationes restrata, & exactius, quàm antea, restituta.

NOMINA STELLARVM.	Loca denuò restituta Longitudo		Latitudo		Differentiæ primæ Longitudinis	Latitudinis
	G.	M.	G.	M.	M.	M.
Lucida Vulturis	25	49 ♑	29	21 B.	3	2
Sin: manus Antin:	29	2 ♑	18	48 B.	2	1
Infer: cornu ♑	28	9 ♑	4	42 B.	1	1
Sinister humerus ♒	17	33 ♒	8	41 B.	6	3
Dexter humerus ♒	27	29 ♒	10	43 B.	6	1
Os Pegasi	26	2 ♒	22	9 B.	6	2
Prima ala Pegasi	17	35 ♓	19	25 B.	6	2
Lucida colli Pegasi	20	20 ♓	17	41 B.	6	0
Scheat Pegasi	23	30 ♓	31	7 B.	1	2
Dextrũ genu Pegasi	19	50 ♓	35	7 B.	0	1
Berta in poſt: Peg:	18	34 ♓	29	25 B.	2	0
Lucida Lyræ	9	22 ♊	61	46 B.	2	3

Pacat

The world of Tycho Brahe: **DATA COLLECTION**

Data, Rules, Prediction: Lessons from Tycho de Brahe, Kepler and Newton



Kepler: **MATHEMATICAL** but not predictive

Newton's Law of Universal Gravitation

$$\vec{F} = \frac{-GMm\hat{r}}{r^2}$$

Newton's 2nd Law

$$\vec{F} = d/dt\,(m\vec{v})$$

Figure 11.0

Newton's laws: **PREDICTIVE**

The Rules Behind the Development of Patent Citation Network

Analysis at the level of individual patents

- Patents: nodes; Citation: edges
- Very large data set (about 5 million patents between 1975 and 2011)
- Data available electronically (USPTO + NBER dataset)
- Its evolution reflects technological changes
- Relevance to patent policy

from the beginning to the end: United States Patent 7,930,766
Kley April 19, 2011
fluid delivery for scanning probe microscopy

The Rules Behind the Development of Patent Citation Network

Citation networks

- special directed networks
- edges and vertices are never deleted from the network
- all outgoing edges of a vertex are added right after the vertex itself
- we will assume that a single vertex is added to the citation network in each time step
- there are no loops

The Rules Behind the Development of Patent Citation Network

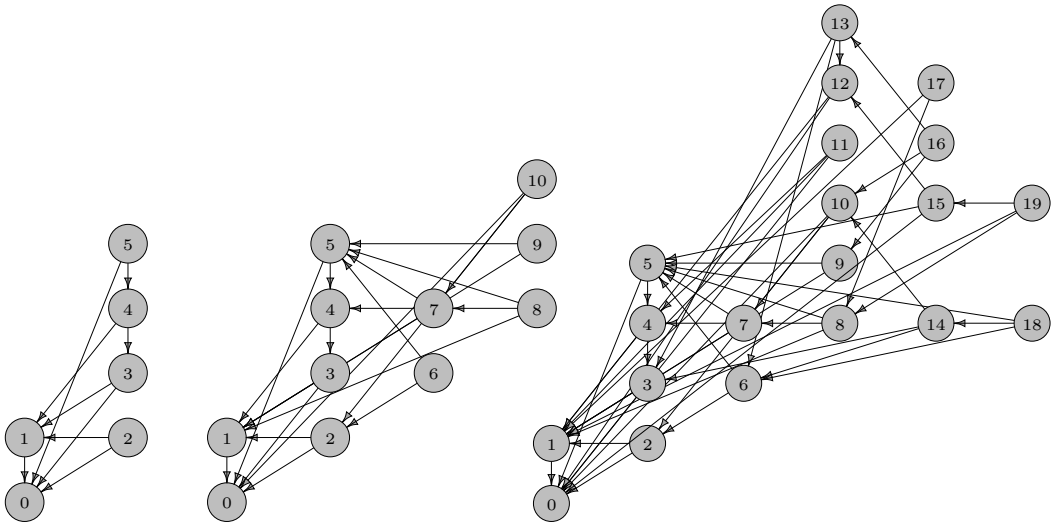


Figure 3.1: Some snapshots for a citation network. Since the new vertices are always added to the right and to the top of old vertices, all edges go to the left or downwards. Here we show three snapshots, vertices 6-10 are added *between* the first and the second and 11-19 *between* the second and the third. Notice that all outgoing edges of a vertex are added with the vertex itself.

The Rules Behind the Development of Patent Citation Network

Kernel Function

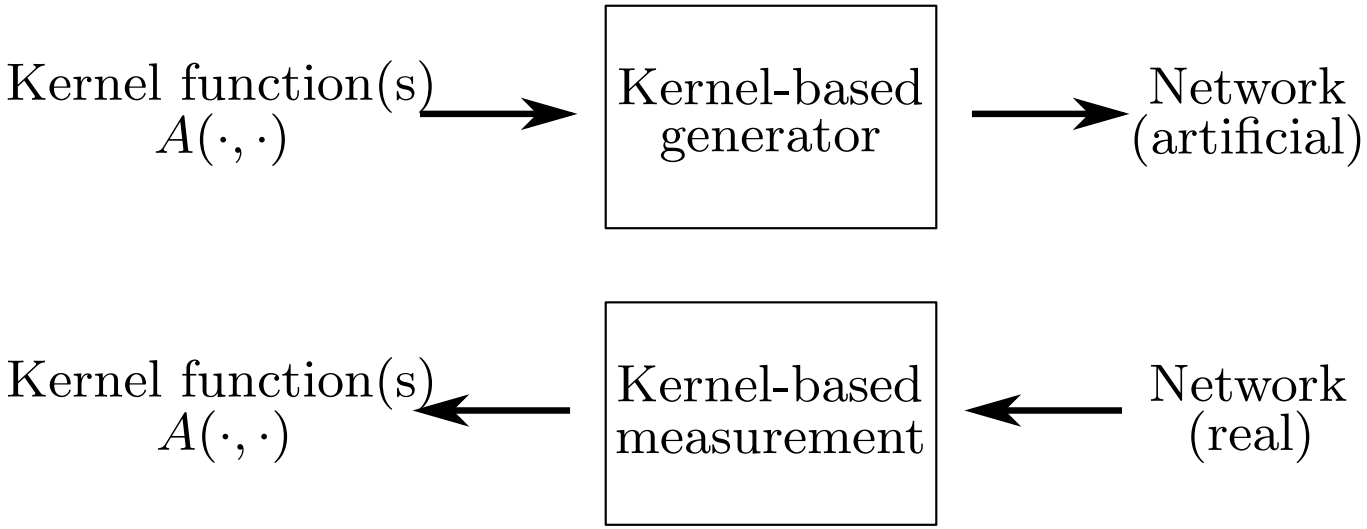
- Property vector: $(\mathbb{X})$, x-vertex: a vertex with property vector x
- Kernel function: $A : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$
- Higher kernel function value $\to$ more probable the realization of the given edge
- The probability that a given edge e connects an x-vertex to a y-vertex:

$$P[c(e,x,y)=1] = \frac{A(x,y)N(t(e),x,y)}{\sum_{(x',y')\in\mathbb{X}\times\mathbb{X}} A(x',y')N(t(e),x',y')}$$

[$c(e,x,y)$ are indicator random variables, (one for every edge–property vector triple. $c(e,x,y)=1$ if and only if edge e connects an x vertex to y-vertex; $t(e)$ is the time step *before* the addition of edge e; $N(t(e),x,y)$ is the number of possible x-y connections in time step $t(e)$.]

The Rules Behind the Development of Patent Citation Network

Direct and Inverse Problems



The Rules Behind the Development of Patent Citation Network

Solving the Inverse problem

The Frequentist Method

The Maximum Likelihood Method

The goal is to extract a kernel function from the network evolution data.

The function to be maximized is the probability that a kernel function generates exactly the observed network. i.e. (for citation networks):

$$\prod_e \frac{A(x_e)}{S(t(e))} = \prod_{i=1}^{n} A(i)^{M_i} \prod_e [\sum_{i=1}^{n} p_i^{t(e)} A(i)]^{-1}$$

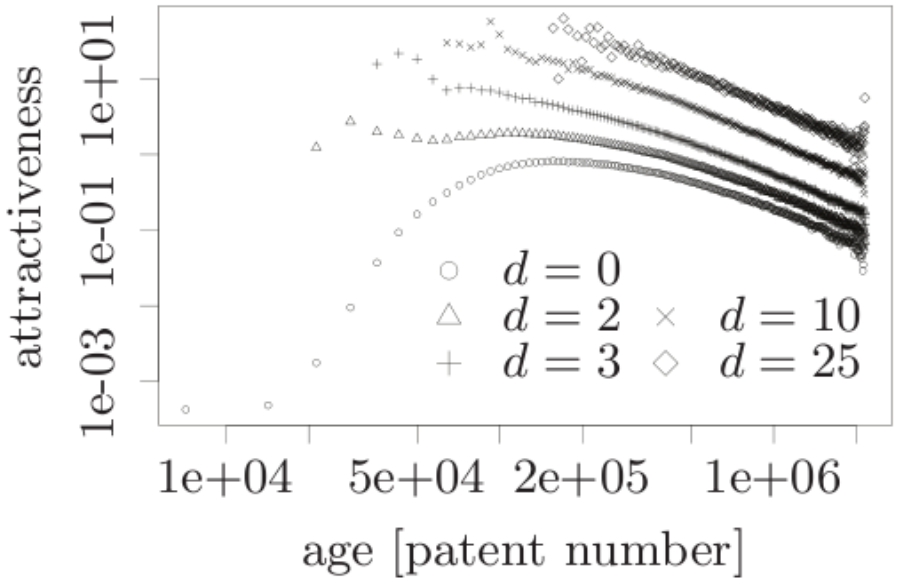
the S normalization factor is

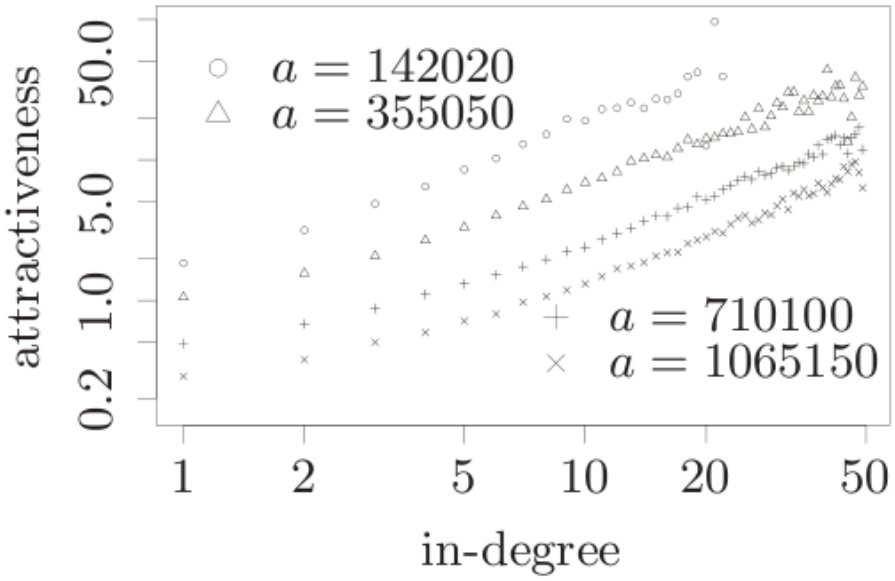
$$S(t(e)) = N^{t(e)} \sum_{i=1}^{n} p_i^{t(e)} A(i).$$

Existence and uniqueness were proved by Gábor Csárdi, and the whole procedure has been generalized for non-citation networks.

Variables: In-Degree and Age

$A(d, l) = A_d(d)A_l(l)$: linear preferential attachment times double Pareto age-dependent part





Sections from the *in-degree* and *age* based maximum likelihood fitted kernel function for the US patent citation network. Both plots have logarithmic axes. From Gábor Csárdi.

in-degree dependent kernel function can be very well fitted with:
$A(d) = d^{\alpha} + a$; α exponent is close to unity (may lead to scale-free networks)

$$A_l(l) = \begin{cases} (l/t_p)^{\beta_p - 1} & \text{if } l \leq t_p, \\ (l/t_p)^{-\alpha_p - 1} & \text{if } l > t_p. \end{cases}$$

Some lessons learned from the "microscopic" analysis

- "number of citations received' and "age" are relevant variables
- the functional forms of the "attractiveness" of the patents on these variables were found
- "stratification" – more and more nodes have very few citations and less and less nodes have many citations
- "sleeper patents" matter: it may happen that old patents gain new significance in light of later advances
  (Upjohn's 1969 patent $\#3,461,461$ for minoxidil: initially developed to treat hypertension, but it was later noticed that one of its side effects was hair growth. Although the patent was issued in 1969, the bulk of its citations came in the 1980s and 1990s, when inventors started developing hair loss treatments based on minoxidil)
- changes in the laws of the patent review process and in the level of rigorousness of the patent examinations over-accelerated the process

Prediction of Emerging Technologies based on Co-citation Clustering

- General Plan
- Background and Significance
- Methodology
- Results so far
- Conclusions and Plans

General Plan

Conceptual frameworks

- to develop, validate and test a *new technique* about new directions of technological development
- patent citation network
- predictive analytics

Working hypotheses

1. the evolution of the patent citation network reflects (if imperfectly) technological evolution
2. a quantity, the *citation vector*, can be defined appropriately to play the role of a predictor, i.e., to characterize the temporal change of technological fields
3. clusters of patents, which are the signature of new developmental directions, can be identified based on patterns of similarity in the citations they receive

General Plan

Technology classification systems

- USPTO: 450 *classes*, and over 120,000 patent *subclasses*
- new classes added; patents can be reclassified
- NBER: 36 *sub-categories* further lumped into
- six *categories*: Computers and Communications, Drugs and Medical, Electrical and Electronics, Chemical, Mechanical and Others

General Plan

Evolving clusters

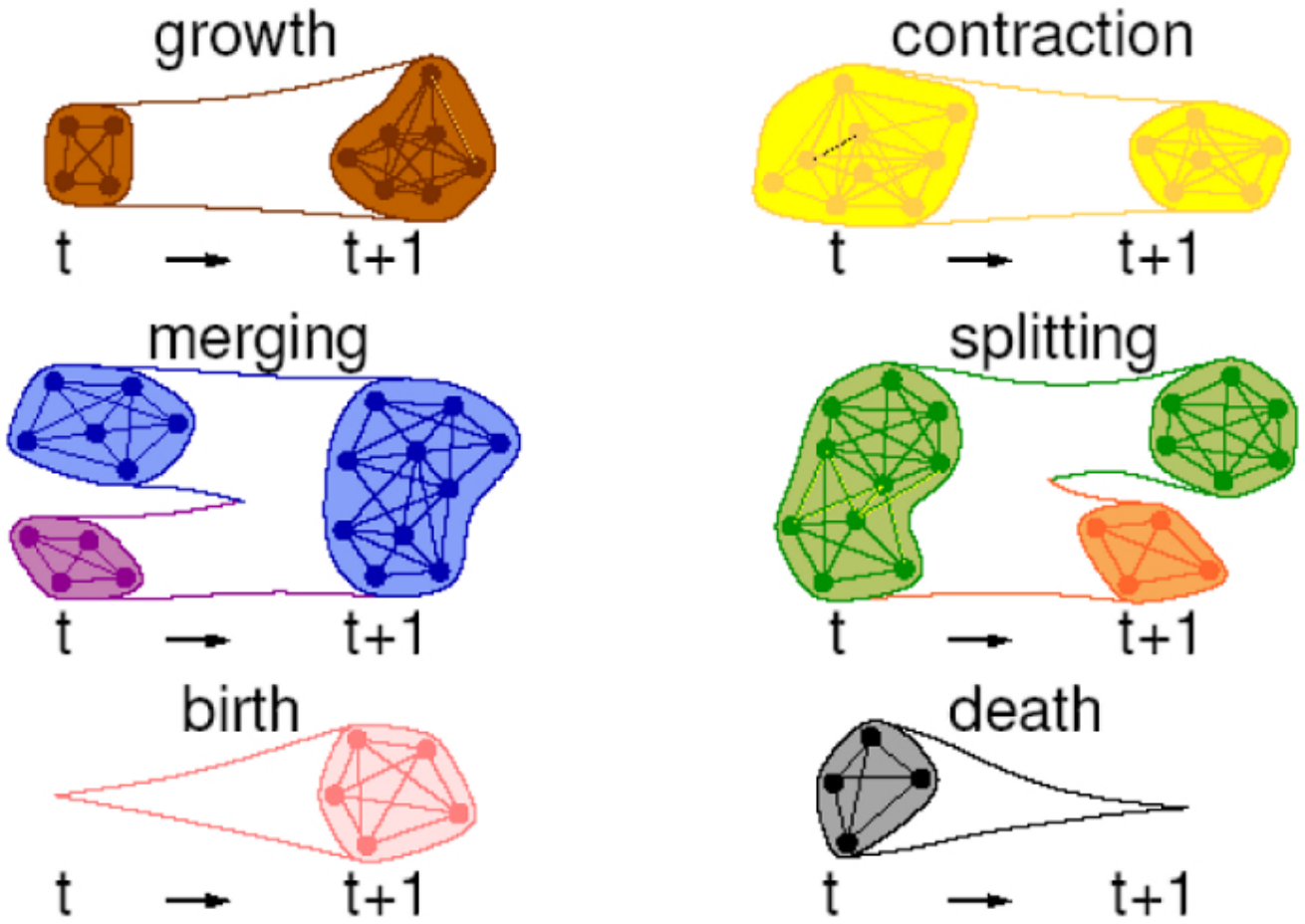


Figure 1: Possible elementary events of cluster evolution. Based on Palla et al. (2007

General Plan

Specific aims

1. to provide a general predictive analytic methodology, which is able **to identify structural changes** in the patent cluster system and reveal *precursors* of emerging new technological fields

2. to test and validate the predictive force of the new methodology based on **historical examples** of new class formation

3. to identify **specific mechanisms of the recombination process** and formation of new classes

4. to scan the database to identify "**hot spots**" that may reflect incipient development of new technological clusters

Methodology

Definition of a predictor for the technological development

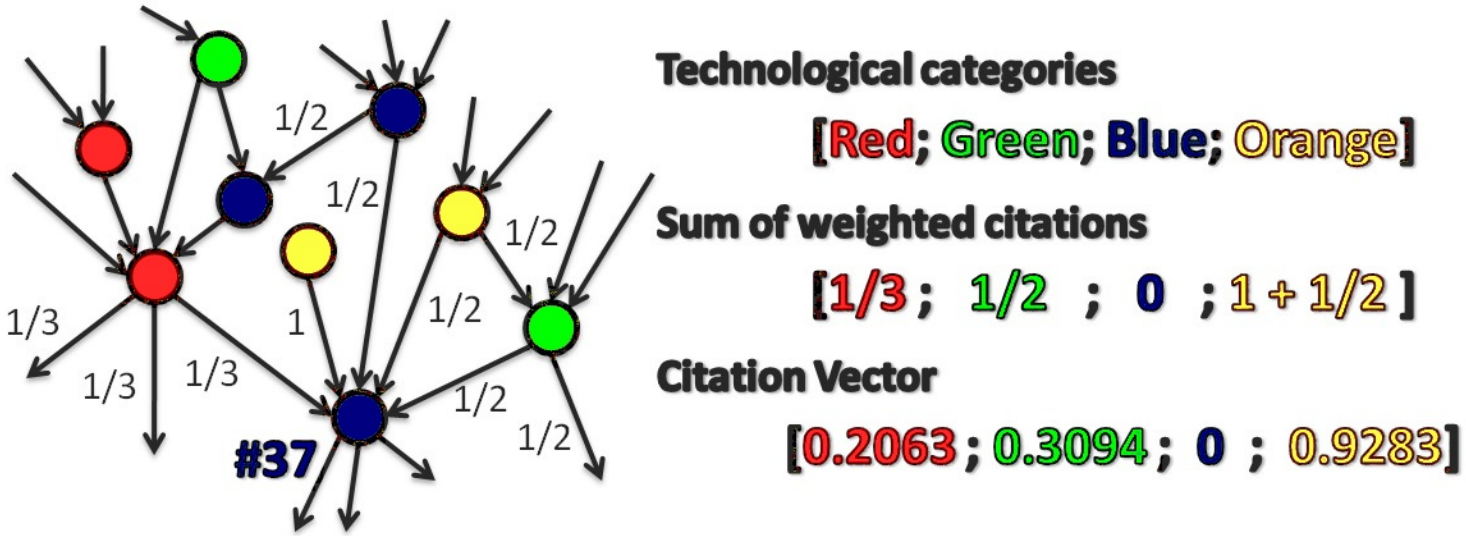


Figure 2: Illustration of citation vector calculation in case of four technological categories denoted by the four different colors. The outgoing citations are weighted by the out-degree of their source. The citations originating from the same category (blue in this case) are excluded from the citation vector and the corresponding vector component is set to zero. The received weighted citations are summed and normalized in order to obtain the citation vector.

Methodology

Algorithm for predicting for the technological development

1. Select a time point t_1 between 1975 and 2007 and drop all patents that were issued after t_1.

2. Keep some subset of subcategories: $c_1, c_2, \ldots, c_n$ – to work with a reasonably sized problem.

3. Compute the citation vector. Drop patents with assortative citation only.

4. Compute the similarity matrix of patents by using the scalar product between the corresponding citation vectors.

5. Apply a hierarchical clustering algorithm to reveal the functional clusters of patents.

6. Repeat the above steps for several time points $t_1 < t_2 < \cdots < t_n$.

7. Compare the dendrogram obtained by the clustering algorithm for different time points to identify structural changes (as emergence and/or disappearance of subcategories).

Methodology

Identification of patent clusters

- to select and test clustering and graph partitioning algorithms to produce sufficiently good results for comparing and validating the clustering results

- time complexity: an unavoidable trade-off between accuracy and time-consumption

- the appropriate number of clusters are not known *a priori*: use hierarchical methods, which do not require that the number of clusters to be specified in advance

- k-means and the Ward method, which are point clustering algorithms

- graph clustering algorithms: edge-betweenness random walks and the MCL method

Methodology

Interplay: a new clustering algorithm in near linear time

- a new global graph clustering algorithm ("Risk": under testing, generalization and formal mathematical studies), to produce sufficiently good results for comparing and validating the clustering results (Péter Volf)

- near linear time

Methodology

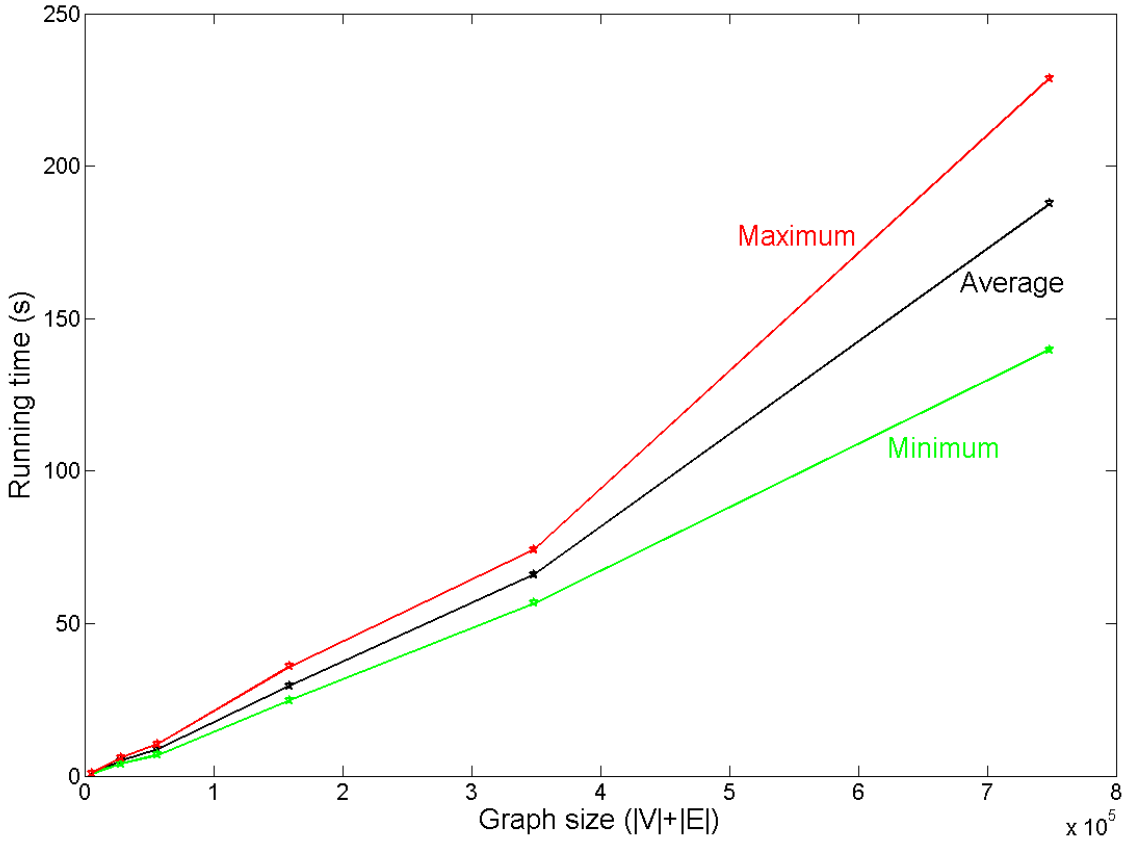


Figure 3: We used the LFR (Lancichinetti-Fortunato-Radicchi) benchmark to create the test graphs. Every graph has approximately the same community structure ($\tilde{1}4$ communities with different sizes), the same average degree (9-13), and the same degree distribution (negative exponent: 2). The number of vertices grow from 1000 to 100000. Based on these measurements, the time complexity of the algorithm is approximately $O((|V| + |E|)^{1.1})$.

Methodology

Algorithm	Rand index scores on three test graphs (Base of comparison: LFR benchmark)		
Label Propagation	0.999321	0.977393	0.973002
Walktrap	0.989369	0.969537	0.951077
FastGreedy	0.968807	0.951924	0.968126
Risk	**0.999706**	**0.989957**	**0.987758**

Figure 4: The algorithm reaches as high modularity scores as other well-known algorithms. Since the community structures of the graphs are known, the results can be compared directly to the actual community structures using Rand index. The table shows that our algorithm reaches notably higher similarity scores (Rand index) on every graph, than other well-known methods.

Methodology

Detection of structural changes in the patent cluster system

- ASSUMPTION: the structure of dendrograms REFLECTS the structural relationships between patent clusters

- In this hierarchy, each branching point is binary and defined only by its height on the dendrogram, corresponding to the distance between the two branches.

- Temporal changes in the cluster structure can be divided into four elementary events: 1) increase or 2) decrease of the height of an existing branching point, 3) insertion of a new and 4) fusion of two existing branching points.

- To find these structural changes, we will identify the corresponding branching points in the dendrograms representing the consecutive time samples of the network and follow their evolution through the time period documented in the database.

- Specifically, potential new classes can be found by comparing the dendrogram structure with the USPTO classification.

- While some of the branching points of the dendrogram are reflected in the current classification structure, there could be significant branches which are not identified by the current system.

Results so far

Local densities of patents exist in the citation space and can be found with clustering

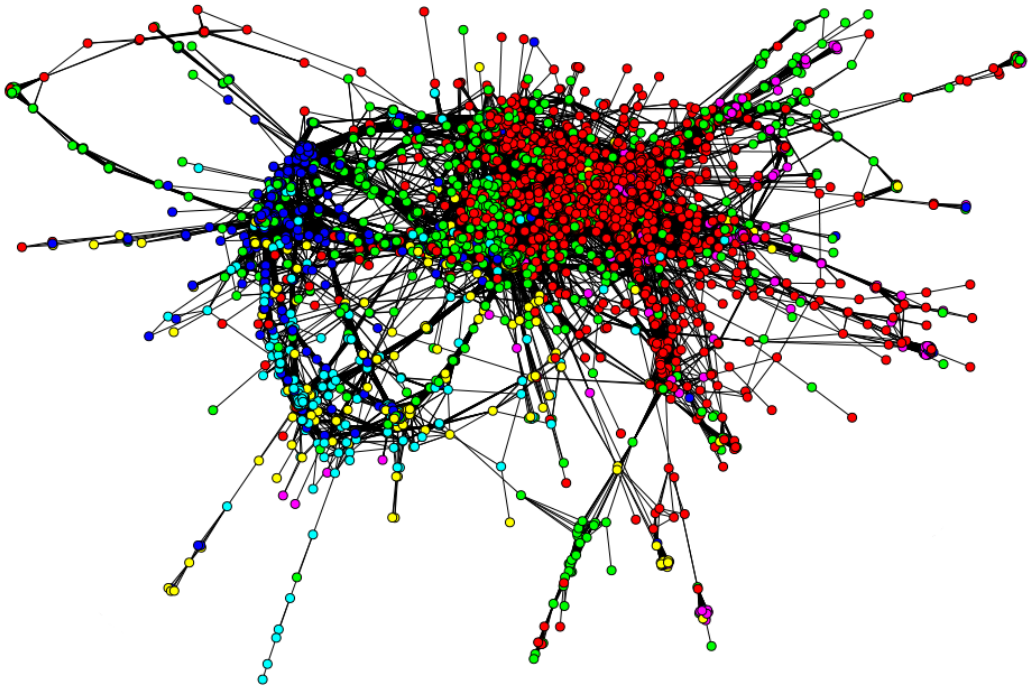


Figure 5: *Cluster structure of patents in the citation space.* Two-dimensional representation of patent similarity structure in the sub-category 11 by using the Fruchterman-Reingold algorithm. Local densities corresponding to technological areas can be recognized by naked eye or identified by clustering methods. The colors encode the US patent classes.

Results so far

Changes in the structure of clusters reflects technological evolution

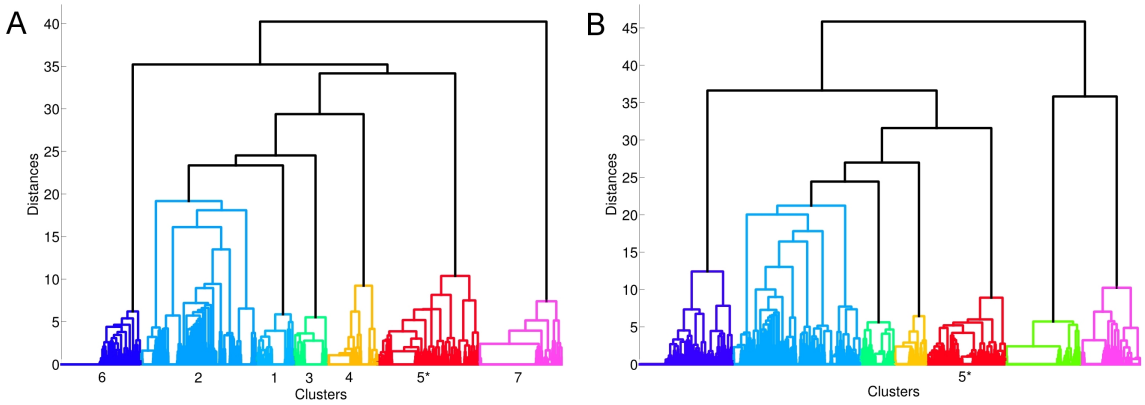


Figure 6: *Temporal changes in the cluster structure of the patent system.* Dendrograms representing the results of the hierarchical Ward clustering of patents in the sub-category 11, based on their citation vector similarity in 1994 (graph A) and 2000 (graph B). The x axis denotes a list of patents in sub-category 11, while the distances between them as defined by the citation vector similarity, are drawn on the y axis. (Patents separated by 0 distance form thin lines on the x axis.) The 7 colors of the dendrogram correspond to the 7 most widely separated clusters. While the overall structure similar in 1994 and 2000, interesting structural changes emerged in this period. The cluster **5** marked with the red color approximately corresponds to the new class 442, which was established in 1997, **but was clearly identifiable by our clustering algorithm as early as 1994.**

Results so far

The emergence of new classes: an illustration

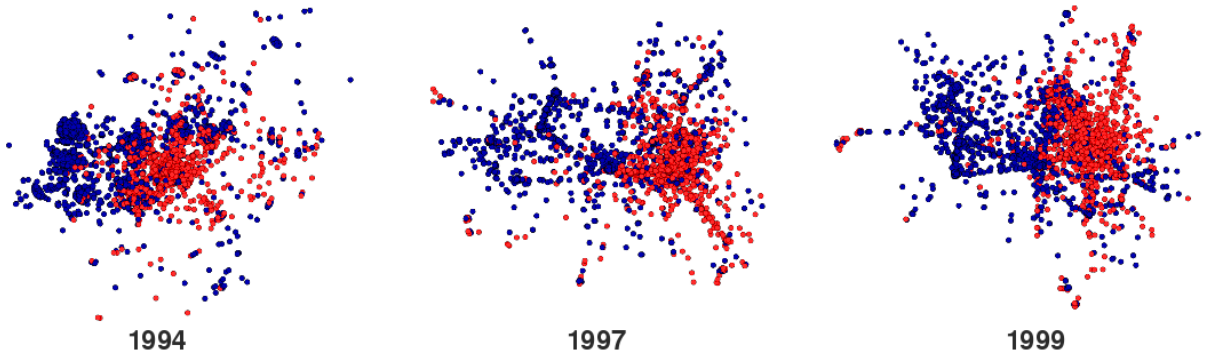


Figure 7: *An example of the splitting process in the citation space, underlying the formation of a new class.* In the 2D projection of the 36 dimensional citation space, position of the circles denote the position of the patents in subcategory 11 in the citation space in three different stages of the separation process (Jan. 1,1994, Jan. 1,1997, Dec. 31,1999). Red circles show those patents which were reclassified into the newly formed class 442, during the year 1997. The rest of the patents which reserved their classification after 1997 are denoted by blue circles. Precursors of the separation appear well before the official establishment of the new class.

The emergence of new classes: an illustration

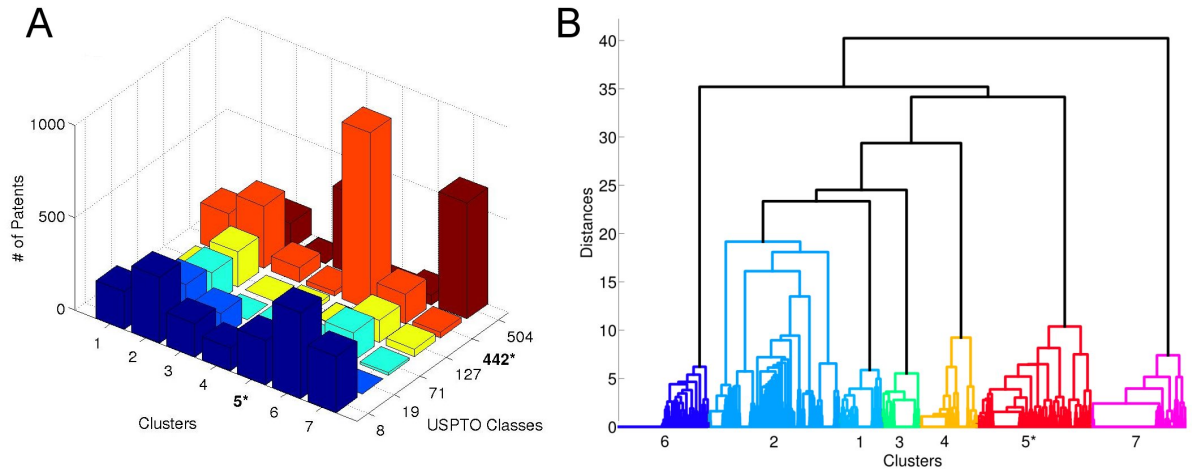


Figure 8: *Separation of the patents by clustering in the citation space, based on the Jan. 1,1994 data.* **A**: Distribution of the patents issued before 1994 in the subcategory 11, within the 6 official classes in 1997 on the class axis (also marked with different colors) and within the 7 clusters in the citation space. The clustering algorithm collected the majority of those patents which were later reclassified into the newly formed class **442** (red line) into the cluster **5** (both are marked with asterisk). Vice-verse, the cluster **5** contains almost exclusively such patents which were later reclassified. Thus, we were able to identify the precursors of the shaping new class by clustering in the citation space. **B**: The dendrogram belonging to the hierarchical clustering of the patents in the subcategory 11 in year 1994 shows that the branch which belongs to the cluster **5** is the fourth strongest branch of the tree. The coloring here refers to the result of the clustering, thus it is different from the colors in graph **A.**

Conclusions and Plans

- Patent citation network is a good source of information for making predictions for technological development
- Clustering methods should be tested and validated (help from patentologists would be appreciated!)
- Mechanisms of new class formations will be studied
- Database should be scanned to detect "hot spots" of emerging fields

Research group

Collaborators in the US:

Katherine Strandburg, Phd in Physics, Law Professor: NYU

Jan Tobochnik, KCollege, Computer Simulation Methods, AJP

Kinga Makovi, PhD student, Columbia University

Research group

Hungarian research group:

Zoltán Somogyvári, Péter Volf, László Zalányi; other researchers are also involved

Research group

Lionshare of the preliminary work: Gábor Csárdi: (now Harvard): direct and inverse problems of evolving networks

Kinga Makovi: concept formation

Péter Volf: lionshare of the clustering, data mining