Prediction of Emerging Technologies Based on Analysis of the U.S. Patent Citation Network

Péter Érdi[1,2], Kinga Makovi[1,2,4], Zoltán Somogyvári[2],
Katherine Strandburg[5], Jan Tobochnik[1],
Péter Volf[2,3], László Zalányi[2,1]

[1]Center for Complex Systems Studies,
Kalamazoo College, Kalamazoo, Michigan
Kalamazoo, MI 49006, USA

[2]Department of Biophysics
KFKI Research Institute for Particle and Nuclear Physics,
Hungarian Academy of Sciences, Budapest, Hungary
H-1525 Budapest, P.O. Box 49

[3]Department of Measurement and Information Systems
Budapest University of Technology and Economics

[4]Department of Sociology,
Columbia University

[5]New York University School of Law
40 Washington Square South New York, NY 10012

Correspondence to Péter Érdi, address: Kalamazoo College, 1200 Academy Street, Kalamazoo, MI 49006, USA; phone: +(269) 337-5720; fax: +(269) 337-7101; e-mail: perdi@kzoo.edu

Abstract

The network of patents connected by citations is an evolving graph, which faithfully represents the innovation process. A patent citing another implies that the cited patent reflects a piece of previously existing knowledge that the citing patent builds upon. A methodology presented here (i) identifies actual clusters of

patents: i.e. technological branches, and (ii) gives predictions about the temporal changes of the structure of the clusters. A predictor, called the **citation vector**, is defined for characterizing technological development to show how a patent cited by other patents belongs to various industrial fields. The clustering technique adopted is able to detect the new emerging recombinations, and predicts emerging high-impact technology clusters. The predictive ability of our new method is illustrated on the example of USPTO subcategory 11, Agriculture, Food, Textiles. We have determined a cluster of patents based on citation data up to 1991, which show significant overlap of the class 442 formed at the beginning of 1997. These new tools of predictive analytics will support policy decision making processes in science and technology, and helps to formulate recommendations for action.

Keywords: patent citation; network; co-citation clustering; technological evolution

1 Introduction

In this paper we present a conceptual and computational framework for making predictions about technological development. The framework is based on patent data, which long has been recognized as a rich and potentially fruitful source of information about innovation and technological change. Besides describing and claiming inventions, patents cite previous patents and other references to identify any "prior art." Patents, as **nodes**, and citations between them, as **edges**, form a directed network.

Complex networks have garnered much attention in the last decade. The application of complex network analysis to innovation networks has provided a new perspective from which to understand the innovation landscape [42]. The **patent citation network**, a huge, growing directed graph, is the result of a social game played by governmental institutions, universities, individual inventors, private firms, research institutes, patent lawyers, patent examiners and attorneys. Given that the citation network encapsulates information about technological relationships and progress provided by those players, understanding its development could help to inform policy makers as to how to allocate resources optimally to research and development.

Our approach belongs to the field of **predictive analytics**, which is a branch of data mining concerned with the prediction of future trends. The technique we develop and implement is based on mining the patent citation network. The evolution of the patent citation network reflects (if imperfectly) technological evolution. Both patentees and patent examiners have incentives to cite materially related prior patents. Patent applicants are legally required to list related patents of which they are aware. Patent examiners seek out the most closely related prior patents so that they can evaluate whether a patent should be granted. Consequently, citation of one patent by another represents a technological connection between them and the patent citation network reflects information about technological connections known to patentees and patent examiners.

The central element of predictive analytics is the predictor, a mathematical object that can be defined for an individual, organization or other entity and employed to predict its future behavior. Here we define a **citation vector** for each patent to play the role of a predictor, i.e., to characterize the temporal change of technological fields. Each

coordinate of the citation vector represents how frequently the patent has been cited by other patents in a particular technological category. Changes in this citation vector over time reflect the changing role that a particular patented technology is playing as a contributor to later technological development.

To track the development of technological clusters, we employ clustering algorithms based on a measure of similarity defined using the citation vectors. We hypothesize that patents with similar citation vectors will belong to the same technological field. Thus, the formation over time of new clusters should correspond to the emergence of new technological directions. Using this approach we test whether past changes in the clusters can be detected, and future changes can be predicted.

To summarize, we present a new computational algorithm for recognizing emerging fields of technology based on the temporal evolution of patent clusters defined by patterns in the citations they receive. Our methodological approach is to identify the community structure of non-assortative patents - those which receive citations from outside their own technological areas - and to make predictions about the near future by describing the evolution of that structure in terms of elementary events such as the birth, merging, and splitting of clusters.

When a new predictive method is constructed, it should be able to ”predict” evolution from the ”more distant” past to the ”more recent” past, a process called backtesting. To illustrate the potential of our approach and demonstrate the emergence of new technological fields from the patent citation data, we illustrate in this paper the ”prediction” of an emerging new technological area by comparing our clustering results with the recognition of a new technological class by the US Patent and Trademark Office (USPTO). In other words, we validated our method by looking at real historical changes. We believe that the method will be useful for analyzing the historical evolution of patent technology and predicting possible near-future changes.

2 Literature review

There is a huge literature on making predictions on emerging technologies based on the analysis of patents or scientific papers, and we can review a small fraction of the most important ones from the perspective of our goal.

2.1 Citation analysis

Citation analysis has a big tradition to evaluate research performance [14, 33]. Co-citation analysis goes back to the now classical works of Small [52, 15]: ”...A new form of document coupling called co-citation is defined as the frequency with which two documents are cited together. The co-citation frequency of two scientific papers can be determined by comparing lists of citing documents in the Science Citation Index and counting identical entries. Networks of co-cited papers can be generated for specific scientific specialties, and an example is drawn from the literature of particle physics. Co-citation patterns are found to differ significantly from bibliographic coupling patterns, but to agree generally with patterns of direct citation. Clusters of

co-cited papers provide a new way to study the specialty structure of science. They may provide a new approach to indexing and to the creation of SDI profile..."

Weng et al. [62] investigated the technological role and technological position of patents based on the concept of structural equivalence, a fundamental notion it the classical theory of social networks. Patents sufficiently close to each other in their citation patterns are considered as equivalents. The method was specifically adopted to insurance business methods patents. The analysis was done at the level of individual patterns. Recently OuYang and Weng [37] adopted, among others patent citation analysis for new product design. Their method is restricted, however, for two-step citations only. Patents and their citations were used as indicators for technology forecasting by Chang et al. [4]. They found a small set of "basic" patents and than to group them into clusters. Lee at al. [27] analyzed the patent citation network to study the case of electrical conducting polymer nanocomposite. A distance-based citation map was constructed mostly to visualize technology evolution.

It is very remarkable the extensive work of Kajikawa and his coworkers in using citation analysis and clustering techniques for predicting emerging technologies. The roadmap technique [24, 25] is applied to make links among such different concepts, as product, technology and science. Some directions in the sustainable energy industry (fuel cell, solar cell) were detected by citation analysis. Another citation-based method (analyzing the scientific publications) was used for biomass and bio-fuels [23]. Citation networks of scientific publications were analyzed to detect emerging knowledge domains (adopted to specific fields). Citation networks were divided into co-cited clusters [50]. Topological clustering is also was used to detect emerging fields in organic light-emitting diodes. A similar work addressed the emerging research fields in regenerative medicine [51]. A comparative study of the structures of the citation network of scientific publications with those of patents were given by [49]. The time-lag between scientific discovery and its commercialization needs more thorough investigation. Co-citation based clustering (in the scientific literature) related to the possibility of tracking modularity was offered [56].

Co-citation clusters were also derived by Wallace et al [29] for scientific specialties. Specifically, they adopted a method [2] relies on the topology of the weighted network. Several recent papers also used co-citation analysis [3], specifically patent co-citation analysis was adopted more than a decade ago [31, 30, 32].

2.2 Understanding the patent system

Understanding the interaction between innovation and the patent system is difficult for many reasons. Increased patenting, for example, can stem from various causes, including an increased pace of technological change, an increased range of patented technology due either to expansion of the scope of legally patentable subject matter or to the birth of new fields of technology, a growing perception of the usefulness of patents as business tools, or the issuance of lower quality patents. Empirical investigation of the patent system can play an important role in understanding how to maintain the appropriate balance. Some basic works are briefly reviewed here.

The use of patent titles for identifying the topics of invention and forecasting trends was offered by Courtial et al [22] using co-word analysis. Ernst [11] used patent data

for technological forecasting of some technology categories in the machine tool industry.

Lai and Wu [26] offers an approach to develop a patent classification system based on patent similarities to assist patent manager in understanding the basic patents for a specific industry, the relationships among categories of technologies and the evolution of a technology category.

In two papers [54, 55] our goal was to explain legal scholar why and how to use the methodologies offered by the modern network science and related fields. Patents and their citations form a directed network, meaning that citations go from later patents to earlier patents and not in the opposite direction, in which patents are the network nodes and citations are directed links. Citations convey valuable information about the relationships between the technologies covered by the citing and cited patents. One can thus view the patent citation network as a kind of map of the space of patented technology, indicating the relationships between various pieces of" property" in that space.

2.3 Emerging technologies

Emerging technologies can be identified by technical innovations which represent progressive developments within a field for competitive advantage. A recurring theme for having a conceptual framework of this emergence is the relationship between biological and technological evolution [43]. It is interesting to realize that both fields are characterized by two extreme approaches: the first emphasizes the gradual, incremental nature of changes, while the other sees rapid, often discontinuous transitions, in the spirit of the theory of punctuated equilibrium [17].

As Adner and Levithal [43] writes: "Just as biological speciation is not a genetic revolution - the DNA of the organism doesn't suddenly mutate - technological speciation is not usually the result of a sudden technological revolution. The revolution is in the shift of application domain. The distinct selection criteria and new resources available in the new application domain can result in a technology quite distinct from its technological lineage. Framing technology evolution in terms of speciation leads us to differentiate between a technology's technical development and a technology's market application. This distinction is useful in understanding broad patterns of technological change, and leads to specific strategic implications for technology management."

"The continuous emergence of new technologies and the steady growth of most technologies suggest that relying on the status quo is deadly for any firm..." [53]. Day and Schoemaker argued [8]: "...The biggest dangers to a company are the ones you don't see coming. Understanding these threats -and anticipating opportunities- requires strong peripheral vision."

2.4 The Role of Metrics in Science and Innovation Policy

Understanding the development of the patent citation network has both scientific interest and practical potential from the policymakers' point of view. Developing a new technology is a highly risky and costly activity. Often, economically significant patents are the result of considerable basic research and product development. Basic research

is funded by the government and, to a lesser extent, by large firms such as those found in the medical and pharmaceutical industries. Product development is carried out by a range of players, including start-up companies spending large fractions of their revenues on innovation. Not surprisingly research and development (R&D) costs have risen rapidly in the past few decades. For example, in 2002, the United States spent 2.5% of its national income on R&D and was 7^{th} in a World Bank ranking of countries by R&D investment.

Because innovation is unpredictable, R&D investment is often risky. The risks inherent in R&D investment are deemed worthwhile as new technologies are the token of economic development and new products are a means of keeping or increasing market share [46, 45, 47]. Patent citation analysis directly offers metrics for characterizing innovative developmental projects. In the long term, understanding the emergence of new technological fields could help to orient investment and reduce risk, resulting in improved economic efficiency.

Daim et al. [7] combined bibliometrics, and system dynamics to make predictions. While the dynamic modeling based on patent number is very interesting, patent citation was not really used. It is very important to see what emerging technologies could change the game?

Performance metrics (the quantitative evaluation of research activity, outputs, impacts) is used by science and technology decision-makers who may not be technical experts, but are interested in having objective credible measures of quality that could support resource allocation decisions. It is far from trivial how to choose "appropriate metrics", for the unintended consequences of metrics in technology evaluation, see [44].

Our general approach and the specific proposed method provide improved understanding of the evolution of innovation. Specifically, the computational technique presented here seems to be able to identify emerging high-impact technology clusters as they emerge.

The literature reviewed here very briefly (and similar other papers not mentioned here) support our view that clustering of citation networks is an efficient tool of predicting emerging fields. Our approach is unique, since we analyzed much larger subset of citation network, than other works did before. Accordingly, the identified objects of the development are large clusters of patents, thus our approach is a more systematic one.

3 The USPTO Technology classification system

The US patent system is not only the largest but also the best documented patent citation dataset. Thus, we have chosen this as a primary field of our investigation, keeping in mind, that our analysis could be applied to other patent databases as well.

The USPTO has developed a classification system of about 450 **classes**, and over 120,000 **subclasses**. The system is used by patent examiners and by applicants and their attorneys and agents as a primary resource for assisting them in searching for relevant prior art. Classes and sub-classes are subject to ongoing modification, reflecting the USPTO's assessment of technological change. Not only are new classes added to

the system, but also patents can be reclassified. As we discuss later, that reclassification provides us with a natural experiment, which offers an opportunity to test our methodology for detecting emerging new fields [19]. Within the framework of a project sponsored by the National Bureau of Economic Research (NBER), a higher-level classification system was developed, in which the 400+ USPTO classes were aggregated into 36 **subcategories**[1], which were further lumped into six **categories** (Computers and Communications, Drugs and Medical, Electrical and Electronics, Chemical, Mechanical and Others). As with any classification system, this system also reflects ad hoc decisions on what constitutes a category or a subcategory, however the classifications appear to show sufficient robustness.

4 Lessons from network theory

4.1 Patent citation analysis: from microscopic to mesoscopic and macroscopic description

The structure and dynamics of the patent citation network, like those of many other complex networks, can be studied at different levels. We have previously studied the growth of the patent citation network at the "microscopic" level of individual patents [6, 9, 54, 5, 55]. The USPTO-defined classes and NBER database subcategories and categories can be seen as the essential structural units at higher levels of abstraction. Recently Ref. [18] analyzed the net flows of citations between NBER categories and subcategories in an attempt to determine the relative influence of different fields on technology growth at the "macroscopic" level. In this paper we focus our attention on phenomena at what might be termed the "mesoscopic" level to look at structures within the network on scales intermediate between individual patents and categories containing large numbers of patents.

4.2 Evolving clusters

As detailed below, in searching for evolving technology clusters we make use of patterns of citation based on NBER subcategories to group patents that are cited similarly together. The dynamics of the community structure of the patents allow us to make predictions about the near future by describing the evolution of clusters in terms of birth, death, growth, shrinking, splitting and merging, which are analogous to the cluster dynamical **elementary events** found in Ref. [38]. Figure 1 illustrates these events.

[Figure 1 about here.]

[1] 11 – Agriculture,Food,Textiles, 12 – Coating, 13 – Gas, 14 – Organic Compounds, 15 – Resins, 19 – Miscellaneous-Chemical, 21 – Communications, 22 – Computer Hardware&Software, 23 – Computer Peripherials, 24 – Information Storage, 31 – Drugs, 32 – Surgery&Med Inst, 33 – Biotechnology, 39 – Miscellaneous-Drgs&Med, 41 – Electrical Devices, 42 – Electrical Lighting, 43 – Measuring&Testing, 44 – Nuclear&X-rays,45 – Power Systems, 46 – Semiconductor Devices, 49 – Miscellaneous-Electric, 51 – Mat.Proc&Handling, 52 – Metal Working, 53 – Motors&Engines+Parts, 54 – Optics, 55 - Transportation, 59-Miscellaneous-Mechanical, 61 – Agriculture,Husbandry,Food, 62 – Amusement Devices, 63 – Apparel&Textile, 64 – Earth Working&Wells, 65 – Furniture,House,Fixtures, 66 – Heating, 67 – Pipes&Joints, 68 – Receptacles, 69 – Miscellaneous-Others.

It is important to study whether recombination of preexisting technologies to create new innovations - reflected in the patent citation data - can be identified as combinations of such elementary events.

Because the patent citation network is a social system, the potential scope and limitations of prediction are different from those in the natural sciences. Patent laws, habits of patent examiners, the pace of economic growth and many other factors influence the development of the patent network, and, correspondingly, the creation of patents changes the innovative environment. Our methodology relies heavily on assessments of technological relationships made by grass root participants in the system. This limits the scope of predictions: we can predict only relatively short term patterns of behavior.

5 Research methodology

5.1 Definition of a predictor for the technological development

We constructed a quantity which we call a **citation vector** and devised a method that enables us to capture the time evolution of technological fields at the level of subcategories.

Specifically, we define the citation vector for every patent at any given time in the following way:

1. For each patent, we calculate the sum of the citations **received** by that patent in each of the technological subcategories defined in Ref. [19] not including its own subcategory. The coordinate corresponding to each patent's own subcategory is set to zero to concentrate on the recombination of different technologies. This gives us 36 sums for each patent, which we treat as entries in a 36-component vector.

2. For each patent, we normalize the 36-component vector obtained in the previous step using a Euclidean norm to obtain our citation vector. The citation vector's components may be interpreted as describing the relative influence that a patent has had on different technological areas. Patents that have not received any citations are assigned a vector with all 0 entries.

The **impact** of a patent on future technologies changes over time, and thus the citation vector evolves to reflect the changing ways in which a patented invention is used in different technological fields.

Similarly to the method of the pioneer of co-citation analysis, Henry Small [52], we hypothesize that a group of patents that are cited by patents from the same set of technological areas with similar proportion have similar roles in the patent universe. We seek to group patents into **functional clusters** based on their roles in the space of technologies. To do this, we define the similarity between two patents as the scalar product of their citation vectors and apply clustering algorithms based on this similarity measure.

Our focus is on those innovations that were influential in industries other than their own. In other words we are concentrating on those patents which received **non-**

assortative citations [34]. Because a high number of patents with only intra subcategory citations tend to mask the recombinant process, citations within the same subcategory have been eliminated from the **citation vector** to highlight those innovations which took part in the recombination of technologies. Finally, we eliminated all the patents with a 0 citation vector i.e. all the patents which either did not receive any citations or received citations only from their own subcategory.

Our algorithm for predicting for the technological development consists of the following steps:

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).

The discussion thus far leaves us with two key issues: (i) What algorithms should be chosen to cluster the patents? (ii) How should we link the clustering results from consecutive time steps? We briefly discuss each in the following subsections.

[Table 1 about here.]

5.2 Identification of patent clusters

Several clustering and graph partitioning algorithms are reasonable candidates for our project. An important pragmatic constraint in choosing clustering algorithms is their time complexity. Given the fact that we are working on a huge database, we face an unavoidable trade-off between accuracy and time-consumption. Because we do not know **a priori** the appropriate number of clusters, hierarchical methods are appropriate, since they do not require the specification of the number of clusters in advance. Available clustering methods include the k-means and the Ward methods, which are point clustering algorithms [60]. Graph clustering algorithms, such as methods using edge-betweenness [16, 36], random walks [41] and the MCL [58] method are also available. The otherwise celebrated clique-percolation method [38] employs a very restrictive concept of a k-clique, making it difficult to mine communities from the patent database. Spectral methods [35] are not satisfactory due to their extreme time-consumption, because we have to calculate the eigenvalues of a relatively **dense** matrix. In the application presented here we adopted the Ward method.

5.3 Detection of structural changes in the patent cluster system

The structure of dendrograms resulting from hierarchical clustering methods such as the Ward method 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. Thus, all types of temporal changes in the cluster structure can be divided into four elementary events: 1) increase or 2) decrease in the height of an existing branching point, 3) insertion of a new and 4) fusion of two existing branching points. To find these substantial, structural changes, we 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.

To test where our clusters are meaningful we compare the emergence of new clusters to the introduction of new classes by the USPTO. Potential new classes can be identified in the clustering results 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 (and we have shown that there exist) significant branches which are **not** identified by the current system. we test our approach by seeing whether clusters that emerge at a particular time are later identified as new classes by the USPTO.

6 Results and model validation

[Figure 2 about here.]

We have chosen the NBER subcategory 11, Agriculture, Food, Textiles as an example, to demonstrate our analysis methods. The rationals of our choice are:

1. Subcategory 11 (SC 11) has moderate size (comparing to other subcategories), which was appropriate to the first test of our algorithm.

2. SC 11 is heterogeneous enough to show non-trivial structure.

3. A new USPTO class, the class 442 was established recently within the subcategory 11.

Note, that restriction of the field of investigation does not restrict the possibility of cross-technological interactions, since the citation vector remained 36 dimensional, including all the possible interactions between the actually investigated and all the other technological fields.

6.1 Patent clusters: existence and detectability

We begin by demonstrating the existence of local patent clusters based on the citation vector. Such clusters can be seen even with the naked eye by perusing a visualization of the 36 dimensional citation vector space projected onto two dimensions, or can be extracted by a clustering algorithm. See Fig. 2.

6.2 Changes in the structure of clusters reflects technological evolution

Temporal changes in the cluster structure of the patent system can be detected in the changes of dendrograms. We present the dendrogram structure of the subcategory 11 at two different times (Fig. 3). Comparing the hierarchical structure in 1994 and 2000, we can observe both quantitative changes, when only the height of the branching point (branch separation distance) changed, and qualitative changes, when a new branching point has appeared.

[Figure 3 about here.]

6.3 The emergence of new classes: an illustration

[Figure 4 about here.]

The most important preliminary validation of our methodology is our ability to "predict" the emergence of a new technology class that was eventually identified by the USPTO. As we mentioned earlier, the USPTO classification scheme not only provides the basis for the NBER subcategories that define our citation vector, it also provides a number of natural experiments to test the predictive power of our clustering method. When the USPTO identifies a new technological category it defines a new class and then may reclassify earlier patents that are now recognized to have been part of that incipient new technological category. (Recall that there are many more USPTO classes than NBER subcategories – within a given subcategory there are patents from a number of USPTO classes.) If our clustering method is sensitive to the emergence of new technological fields, we might hope that it will identify new technological branches before the USPTO recognizes their existence and defines new classes.

Figures 4 and 5 illustrate the emergence of class 442, which was not defined by the USPTO until 1997. Figure 4 shows how patents that will eventually be reclassified into

class 442 can be seen to be splitting off from other patents in subcategory 11 as early as 1991. The visually recognizable cluster of patents in Fig. 4 that will later be reclassified into class 442 can be identified by the Ward method with cutoff at 7 clusters in 1991, as is shown in Fig. 5. The histogram in Fig. 5 shows the frequency of patents with a given cluster number and USPTO class. Patents that will eventually be reclassified into class 442 are already concentrated in cluster 7. The Pearson-correlation between the class 442 and the correspondig clusters in our analysis resulted in high values: 0.9106 in 1991; 0.9005 in 1994; 0.8546 in 1997 and 0.9177 in the end of 1999. This example thus demonstrates that the citation vector can play the role of a predictor: emerging patent classes can be identified.

Based on this and other historical examples of new class formation, we have **tested** and **validated** our clustering methods. In future work, we will seek to answer the question: what is the time difference between the detection of the first signs of the splitting and the official formation of the new class? The characteristic time of new class formation could be field-specific, and hence we will compare results between categories.

[Figure 5 about here.]

7 Discussion

The patent citation network can be viewed as a **time-evolving complex system**. Historically, scholars have sought to understand technological change using evolutionary analogies, describing it as a process of recombination of already existing technologies [48, 57, 21, 61, 20]. Inventions are often described as combinations of prior technologies. "... For example, one might think of the automobile as a combination of the bicycle, the horse carriage, and the internal combustion engine" [12, 39, 40, 13]. This feature of technological advance is well-recognized in patent law and has been the subject of recent Supreme Court attention. See KSR Int'l Co. v. Teleflex, Inc., 550 U.S. 398 (2007).

We assume that social systems are **causal systems** – complex systems with **circular causality** and **feedback loops** [9, 10] – and we also assume that their statistical properties may allow us to uncover **rules** that govern their development. (For similar attempts see Refs. [28] and [1]). "... Analogously to what happened in physics, we are finally in the position to move from the analysis of the "social atoms" or "social molecules" (i.e., small social groups) to the quantitative analysis of social aggregate states ..." [59]. Our study of the specific example of the patent citation network will help to advance the study of how **complex social systems** evolve.

As scientific predictions in general, our work is also based on the analysis of past events and and present state. We assumed, that co-citation clusters represents the species of the technological evolution, thus the recognition of new clusters could be used for predicting new emergent technologies.

The main limitation of our work is the time lag between the birth of a new technique and its appearance in the patent databases as the accumulation of new citations. Csárdi et al. [5] showed, that patents receive the majority of their citations around 15 month

after their issuing and this time lag show little variance across different fields. This time constant set the horizon of our method. However, the specific example we showed, that we were able to identify a new class well before its official introduction in spite of the limitations of the database.

The presented methodology oversimplifies the patent system in many ways. The technological fields are not homogeneous in respect to their propensity to patents, average number of citations per patents, etc. For the sake of simplicity, these differences were not taken into account in this work, but should be included in the future work, to refine the current approach.

Since obviously there is no proper method to determine the appropriate number clusters, the method suggested here is able to provide only necessary but not sufficient criterion for the identification of a new technological branch. To put it another way, we offer a decision support system: we are able to identify the candidates of the hot spots of the technological development, which worth attention.

In future work, we will also investigate the specific mechanisms of new class formation. New technological branches can be generated either by a single (cluster dynamical) elementary event or by combinations of such events. For example, a new cluster might arise from a combination of a merging and a splitting. By examining historical examples, we will clarify how the elementary events **interact** to build the recombination process and identify the typical "microscopic mechanisms" underlying new class formation.

Finally, we will scan the database to identify "hot spots" that may reflect the incipient development of new technological clusters. In this way, we hope to come up with predictions for the near future and give answers to the question: which might be the technologies of tomorrow?

Acknowledgments

PE thanks the Henry Luce Foundation for its support. KJS acknowledges the generous support of The Filomen D'Agostino and Max E. Greenberg Research Fund. Thanks for Fülöp Bazsó, Mihály Bányai, Judit Szente, Balázs Ujfalussy for discussions.

References

[1] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis. Mining graph evolution rules. In W. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5781 of *European Conference on Machine Learning and Knowledge Discovery in Databases.*, pages 115–130. Springer, 2009.

[2] Lambiotte R Blondel VD, Guillaume J-L and Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P10008, 2008.

[3] J Hou C. Chen, F. Ibekwe-SanJuan. The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology*, 61:1386–1409, 2010.

[4] S B Chang, K K Lai, and S M Chang. Exploring technology diffusion and classification of business methods: using the patent citation network. *Technol. Forecast. Soc. Change*, 76:107–117, 2009.

[5] G. Csárdi, K.J. Strandburg, J. Tobochnik, and P. Érdi. Chapter 10. The inverse problem of evolving networks - with application to social nets. In B. Bollobás, R. Kozma, and D. Miklós, editors, *Handbook of Large-Scale Random Networks*, pages 409–443. Springer-Verlag, 2009.

[6] G. Csárdi, K.J. Strandburg, L. Zalányi, J. Tobochnik, and P. Érdi. Modeling innovation by a kinetic description of the patent citation system. *Physica A*, 74(1–2):783–793, 2007.

[7] T. U. Daim, G. Rueda, H. Martin, and P. Gerdsri. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technol. Forecast Social Change*, 73:981–1012, 2006.

[8] GS Day and Schoemaker PJH. Scanning the periphery. *Harvard Business Review*, pages 1–12, 2005.

[9] P. Érdi. *Complexity Explained.* Springer Verlag, 2007.

[10] P. Érdi. Scope and limits of predictions by social dynamic models: Crisis, innovation, decision making. *Evolutionary and Institutional Economic Review*, 7:21–42, 2010.

[11] H Ernst. The use of patent data for technological forecasting: The diffusion of cnc-technology in the machine tool industry. *Small Business Economics*, 9(4):361–381, 1997.

[12] L. Fleming. Recombinant uncertainty in technological search. *Management Science*, 47(1):117–132, 2001.

[13] L. Fleming and O. Sorenson. Technology as a complex adaptive system: evidence from patent data. *Research Policy*, 30:1019–1039, 2001.

[14] E. Garfield. *Citation Indexing - Its Theory and Application in Science, Technology and Humanities*. Philadelphia:ISI Press, 1983.

[15] E. Garfield. Co-citation analysis of the scientific literature: Henry small on mapping the collective mind of science. *Current Contents*, 19(3-13, May 10), 1993.

[16] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[17] SJ Gould and N Eldredge. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*, 3(2):115–151, 1977.

[18] Bernard Gress. Properties of the USPTO patent citation network: 1963-2002. *World Patent Information*, 32(1):3–21, 2010.

[19] Bronwyn H. Hall, Adam B. Jaffe, and Manuel Trajtenberg. The NBER patent citation data file: Lessons, insights and methodological tools. Working Paper 8498, National Bureau of Economic Research, October 2001.

[20] A. Hargadon and R. Sutton. Technology brokering and innovation in a product development firm. *Administrative Science Quarterly*, 42:716–749, 1997.

[21] Rebecca M. Henderson and Kim B. Clark. Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, 35(1):9–30, 1990.

[22] M. Callon J. P. Courtial and A. Sigogneau. The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics*, 26:231–242, 1993.

[23] Y. Kajikawa and Y. Takeda. Structure of research on biomass and bio-fuels: A citation-based approach. *Technological Forecasting and Social Change*, 75:1349–1359, 2008.

[24] Y. Kajikawa, O. Usui, K. Hakata, Y. Yasunaga, and K. Matsushima. Sructure of knowledge in the science and technology roadmaps. *Technological Forecasting and Social Change*, 75:1–11, 2008.

[25] Y. Kajikawa, J. Yoshikawa, Y. Takeda, and K. Matsushima. Tracking emerging technologies in energy researc: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, 75:771–782, 2008.

[26] KK Lai and Wu S-J. Using the patent co-citation approach to establish a new patent classification system. *Information Processing and Management*, 41:313–330, 2005.

[27] Pei-Chun Lee, Hsin-Ning Su, and Feng-Shang Wu. Quantitative mapping of patented technology – the case of electrical conducting polymer nanocomposite. *Technological Forecasting and Social Change*, 77(3):466 – 478, 2010.

[28] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD 2005: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA, 2005. ACM.

[29] Russell J. Duhon Matthew L. Wallace, Yves Gingras. A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology*, page 240 246, 2009.

[30] RG Kolar ME Mogee. Patent citation analysis of allergan pharmaceutical patents. *Expert Opinion on Therapeutic Patents*, 8(10):1323–1346, 1998.

[31] RG Kolar ME Mogee. Patent citation analysis of new chemical entities claimed as pharmaceuticals. *Expert Opinion on Therapeutic Patents*, 8(3):213–222, 1998.

[32] RG Kolar ME Mogee. Patent co-citation analysis of eli lilly & co. patents. *Expert Opinion on Therapeutic Patents*, 9(3):291–305, 1998.

[33] H.F Moed. *Citation Analysis in Research Evaluation*. Dordrecht (Netherlands): Springer, 2005.

[34] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Oct 2002.

[35] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

[36] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69(026113), 2004.

[37] K. OuYang and C.S. Weng. A new comprehensive patent analysis approach for new product design in mechanical engineering. *Technol. Forecast Social Change*, pages 1–17, 2011.

[38] G. Palla, A-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446:664–667, 2007.

[39] Joel M. Podolny and Toby E. Stuart. A role-based ecology of technological change. *The American Journal of Sociology*, 100(5):1224–1260, 1995.

[40] Joel M. Podolny, Toby E. Stuart, and Michael T. Hannan. Networks, knowledge, and niches: Competition in the worldwide semiconductor industry, 1984-1991. *The American Journal of Sociology*, 102(3):659–689, 1996.

[41] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. of Graph Algorithms and Applications*, 10:191–218, 2006.

[42] A Pyka, A Scharnhost. *Innovation Networks. New Approaches in Modelling and Analyzing.* Springer Verlag., 2009.

[43] Adner R and Levinthal DA. The emergence of emerging technologies. *California Management Review*, 45:50–66, 2002.

[44] Kostoff RN and Geisler E. The unintended consequences of metrics in technology evaluation. *Journal of Infometrics*, 1:103–114, 2007.

[45] P.P. Saviotti. *On the co-evolution of Technologies and Institutions.* Berlin, Hidelberg, 2005.

[46] P.P. Saviotti, M.A. de Looze, and M.A. Maopertuis. Knowledge dynamics and the mergers of firms in the biotechnology based sectors. *International Journal of Biotechnology*, 5(3–4):371–401, 2003.

[47] P.P. Saviotti, M.A. de Looze, and M.A. Maopertuis. Knowledge dynamics, firm strategy, mergers and acquisitions in the biotechnology based sectors. *Economics of Innovation and New Technology*, 14(1–2):103–124, 2005.

[48] J. Schumpeter. *Business Cycles.* McGraw-Hill, New York., 1939.

[49] N. Shibata, Y. Kajikawa, and I. Sakata. Extracting the commercialization gap between science and technology - case study of a solar cell. *Technological Forecasting and Social Change*, 77:1147–1155, 2010.

[50] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima. Detecting emerging researc fronts based on topological measures in citation networks of scientific publications. *Technovation*, 28:758–775, 2008.

[51] N. Shibata, Y. Kajikawa, Y. Takeda, I. Sakata, and K. Matsushima. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change*, 78:274–282, 2011.

[52] H Small. Cocitation in scientific literature: New measure of relationship between two documents. *Journal of The American Society For Information Science*, 24:265–269, 1973.

[53] A. Sood and GJ. Tellis. Technological evolution and radical innovation. *Journal of Marketing*, 69:152–168, 2005.

[54] K. Strandburg, G. Csardi, J. Tobochnik, P. Érdi, and L. Zalányi. Law and the science of networks: An overview and an application to the "patent explosion". *Berkeley Technology Law Journal*, 21:1293, 2007.

[55] K. Strandburg, G. Csardi, J. Tobochnik, P. Érdi, and L. Zalányi. Patent citation networks revisited: signs of a twenty-first century change? *North Carolina Law Review*, 87:1657–1698, 2009.

[56] Y. Takeda, N. Shibata, Y. Kajikawa, I. Sakata, and K. Matsushima. Tracking modularity in citation networks. *Scientometrics*, pages 783–792, 2010.

[57] A. Usher. *A History of Mechanical Invention.* Dover, Cambridge, MA., 1954.

[58] S. van Dongen. A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, 2000.

[59] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.

[60] J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[61] Martin L Weitzman. Hybridizing growth theory. *American Economic Review*, 86(2):207–12, 1996.

[62] C.S. Weng, W.-Y. Chen, H.-Y. Hsu, and S.-H. Chien. To study the technologgical network by structural equivalence. *Journal of High Technology Management Research*, 21:52–63, 2010.

8 Biographies

Péter Érdi is the is the Henry R. Luce Professor of Complex Systems studies in Kalamazoo College, and also the head of the Department of Biophysics, KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences. He has been working on the fields of computational neuroscience and computational social sciences.

Kinga Makovi is a PhD student in the Department of Sociology of Columbia University, and holds an MS in mathematical economics from Corvinus University in Budapest (2010). Her interests include social networks, sociology of education, quantitative methods and simulation techniques in social sciences

Zoltán Somogyvári is a senior research fellow of the Department of Biophysics, KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences. He is an expertise in developing new methods of analysing large data sets,

Katherine Strandburg is a Professor of Law at New York University School of Law. Her teaching and research activities are in the areas of intellectual property law, cyberlaw, and information privacy law. Prior to her legal career, she was a research physicist at Argonne National Laboratory, having received her Ph.D. from Cornell University.

Jan Tobochnik is the Dow Distinguished Professor of Natural Science in Kalamazoo College. He also serves as the Editor, American Journal of Physics. His research involves using computer simulations to understand a wide variety of systems. In the last decade he was involved in investigating the structural properties of some social networks.

Péter Volf just get his MSc in the Department of Measurement and Information Systems of the Budapest University of Technology and Economics and works as a junior fellow at the Department of Biophysics, KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences. His main interest is now developing efficient clustering algorithms.

László Zalányi is the acting head of of the Department of Biophysics, KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences. His research areas are the application of stochastic methods to neural and social systems, and network theory.

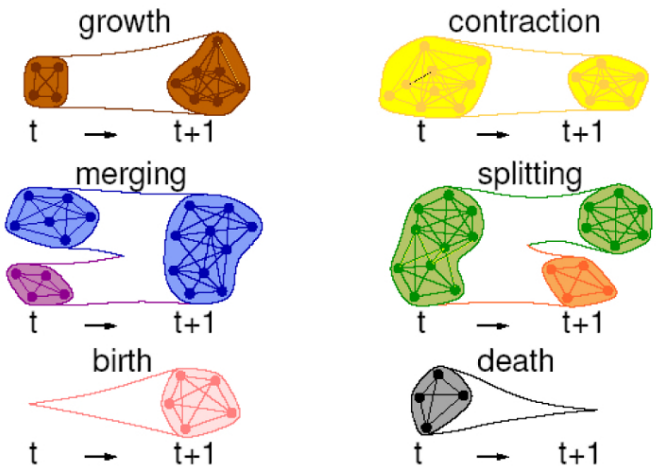

Figure 1: Possible elementary events of cluster evolution. Based on Ref. [38].

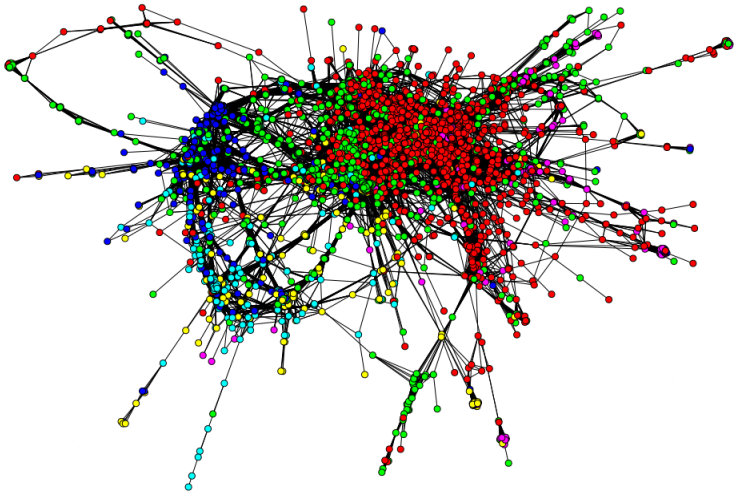

Figure 2: **Cluster structure of patents in the citation space.** Two-dimensional representation of patent similarity structure in the subcategory 11 in Dec. 31, 1999, by using the Fruchterman-Reingold algorithm. Local densities corresponding to technological areas can be recognized by naked eye or identified by clustering methods. The colors encode the US patent classes: red corresponds to class 8; green: 19; blue: 71; magenta: 127; yellow: 442; cyan: 504.

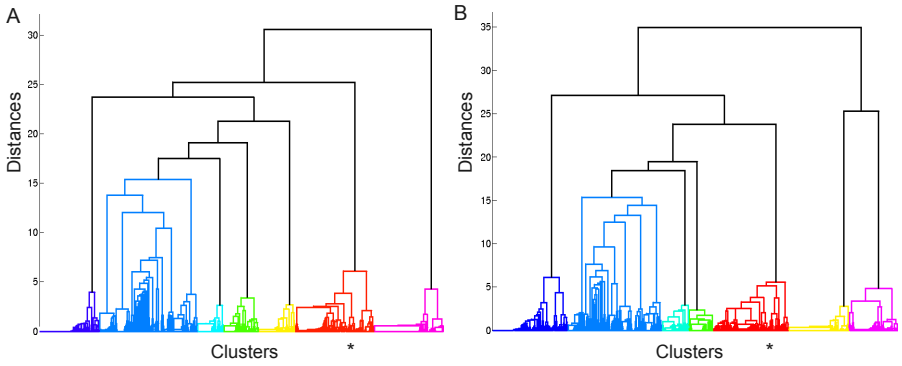

Figure 3: **Temporal changes in the cluster structure of the patent system.** Dendrograms representing the results of the hierarchical Ward clustering of patents in subcategory 11, based on their citation vector similarity on Jan. 1, 1994 (18833 patents in graph A) and Dec. 31, 1999 (25624 in graph B). The x axis denotes a list of patents in subcategory 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 is similar in 1994 and 1999, interesting structural changes emerged in this period. The cluster marked with the red color and asterix approximately corresponds to the new class 442, which was established in 1997, but was clearly identifiable by our clustering algorithm as early as 1991.

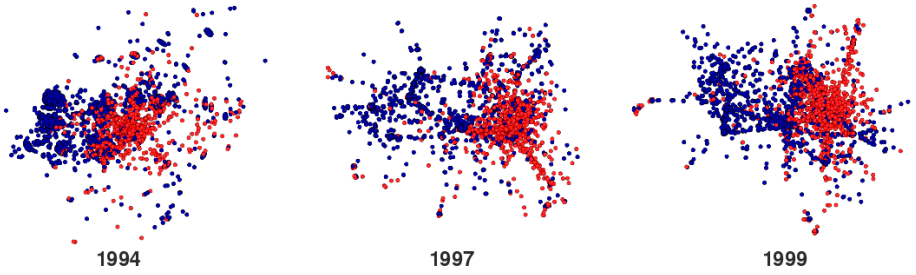

Figure 4: **An example of the splitting process in the citation space, underlying the formation of a new class.** In the 2D projection of the 36 dimensional citation space, position of the circles denote the position of the patents in subcategory 11 in the citation space in three different stages of 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.

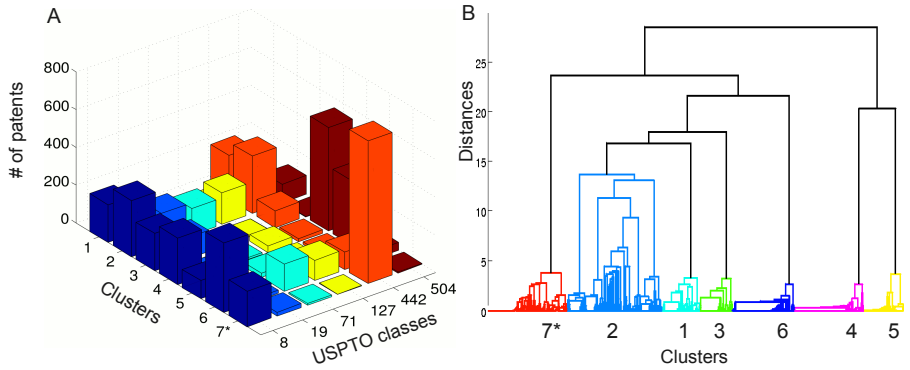

Figure 5: **Separation of the patents by clustering in the citation space, based on the Jan. 1,1991 data. A**: Distribution of the patents issued before 1991 in the subcategory 11, within the 6 official classes in 1997 on the class axis (also marked with different colors) and within the 7 clusters in the citation space. The clustering algorithm collected the majority of those patents which were later reclassified into the newly formed class **442** (orange line) into the cluster **7** (marked with an asterisk). Vice-verse, the cluster **7** contains almost exclusively those patents which were later reclassified. Thus, we were able to identify the precursors of the emerging 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 1991 shows that the branch which belongs to the cluster **7** is the most widely separated branch of the tree. The coloring here refers to the result of the clustering, unlike graph **A** where coloring marks the USPTO classes.

Number of	Jan. 1, 1991	Jan. 1, 1994	jan. 1, 1997	Dec. 31, 1999
patents in the whole database	4980927	5274846	5590420	6009554
patents in the subcategory 11	18833	21052	23191	25624
patents belong to class 442 from 1997	2815	3245	3752	4370
patents with non-zero citation vector in 11	7671	9382	11245	13217
citations connected to patents in SC 11	70920	92177	120380	161711

Table 1: Number of patents in the examined networks and subnetworks in different moments.