Journal of Big Data

RESEARCH Open Access

Internal dynamics of patent reference networks using the Bray–Curtis dissimilarity measure

József Baranyi[1], Szilveszter Csorba[2], Zsuzsa Farkas[2], Tünde Pacza[3] and Ákos Józwiak[2*]

*Correspondence:
Jozwiak.Akos@univet.hu

[1] Institute of Nutrition, University of Debrecen, Debrecen, Hungary
[2] Digital Food Institute, University of Veterinary Medicine, Budapest, Hungary
[3] Doctoral School of Food and Nutrition Science, University of Debrecen, Debrecen, Hungary

Abstract

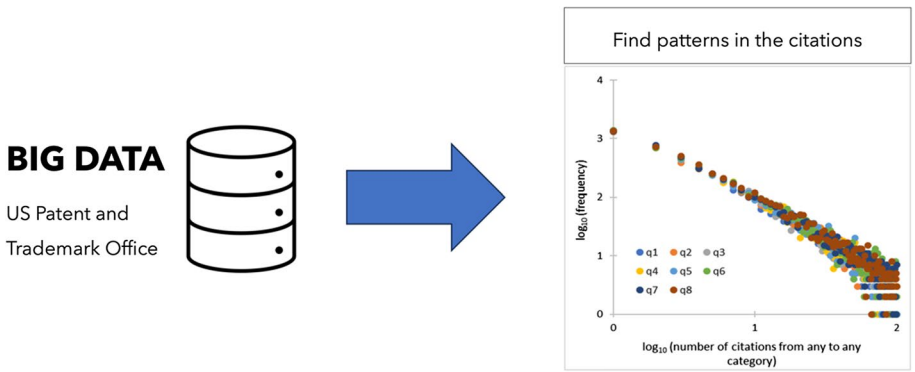
Background: Patents are indicators of technological developments. The science & technology categories, to which they are assigned to, form a directed, weighted network where the links are the references between patents belonging to the respective categories. This network can be conceived as a kind of intellectual ecology, lending itself to mathematical analyses analogous to those carried out in numerical ecology. The non-metric Bray–Curtis dissimilarity, commonly used in quantitative ecology, can be used to describe the internal dynamics of this network.

Results: While the degree-distribution of the network remained stable during the studied years, that of the sub-networks of with at least k links showed that $k=5$ is a critical number of citations: this many are needed that the bias towards already highly cited works come into effect (preferential attachment). Using the d_{ij} Bay-Curtis dissimilarity between nodes i and j, a surprising pattern emerged: the log-probability of a change in d_{ij} during a quarter of year depended linearly, with a negative coefficient, on the magnitude of the change itself.

Conclusions: The developed methodology could be useful to detect emerging technological developments, to aid decisions, for example, on resource allocation. The pattern found on the internal dynamics of the system depends on the categorisation of the patents, therefore it can serve as an indicator when comparing different categorisation methods.

Keywords: Patents, Intellectual ecology, Technological development, Bray–Curtis dissimilarity, Network analysis, Emerging patterns

Graphical Abstract



Introduction

With the advent of "Big Data", the bottleneck in predictive sciences is not *finding relevant data* but *how to make sense of them*. Studying patent databases may reveal patterns in technological advances that can be used to aid decision making, for example during resource allocations.

Patents are one of the most important indicators of human technological advancements. They are connected by references, thus generating links between the technological categories, too, to which the patents are assigned. This network can be conceived as a kind of intellectual ecology, lending itself to mathematical analyses analogous to those carried out in numerical ecology [1].

One of the most important concepts in trend- and pattern-recognition is the dissimilarity between two objects. The choice how to quantify it is not obvious and does not necessarily result in a distance-concept, for which the criteria are:

(i) $d(A,B) \geq 0$ and $d(A,A) = 0$ (i.e. the dissimilarity between two different objects A and B is a non-negative number and it is zero if $A = B$).

(ii) $(A,B) = d(B,A)$ (symmetry)

(iii) $d(A,B) < d(A,C) + d(C,B)$, for any third C (triangle inequality)

The last requirement does not necessarily hold for many commonly used dissimilarity measures. It is difficult to achieve, for example, for transport routes, due to their typical hub-centred organisations, if the distance is measured by the duration of the travel between nodes, as a journey to a hub is generally faster than between non-hubs.

A record in a patent database contains certain attributes of a given patent, among others the scientific-technological category to which it has been assigned. Patents typically refer to other patents, which can be conceived as links between the respective technological categories. If a patent assigned to category c_j refers to patents assigned to categories $c_1 \ldots c_n$, then we say that the latter categories influence c_j. If we quantify the weight of this influence by the number of the respective references, then an 'influence-vector' can be assigned to each category, showing how much it is affected by the other categories.

Patent data have been used by several authors [2–4] to describe, possibly predict, technological changes. Érdi et al. [2] defined a citation vector $\mathbf{v} = [v_1 \ldots v_n]$ for each patent,

where an entry v_i represented how many times patents from category i, were cited in the patent. After suitable normalisation, v_i was seen as a weight to what extent the category c_i influenced the patent. The authors grouped the patents according to the Euclidean distance between their citation vectors, then, by means of the Ward-algorithm [5], produced a dendrogram, the temporal dynamics of which were used to predict significant changes in the system.

A critical point in this approach is the use of the Euclidean metric, i.e., the distance concept was derived from the scalar product of vectors. We demonstrate, in what follows, the advantages of considering a non-metric dissimilarity measure instead. We show how the state and internal dynamics of technological developments can be described by network science methods and using the dissimilarity measure of Bray–Curtis [6]. The transformation of Gower [7] will be used to visualize the state and temporal behaviour of the system in two dimensions.

Method

The Bray–Curtis (BC) dissimilarity measure

Let the $\{c_i\}$ $(i=1...n)$ be technological categories and define the $\mathbf{S}=[s_{ij}]_{\mathrm{nxn}}$ citation matrix as follows: let s_{ij} be the total number of patents that was put in the c_j category and cited patents assigned to c_i where $i \neq j$. This s_{ij} number, a non-negative integer, can be conceived as a quantification of the extent to which the c_i category influenced c_j, i.e., the weight of the $c_i \rightarrow c_j$ directed edge. As these can change with time, in fact we can also use the $\mathbf{S}(t)=[s_{ij}(t)]_{\mathrm{nxn}}$ notation. Our rule sets the diagonal entries of the matrix $\mathbf{S}(t)$ to zero.

Quantify the dissimilarity between the c_j and c_k technological categories by means of their respective influence vectors following the idea of Bray and Curtis [6]:

$$d_{jk} = \frac{\sum_{i=1}^{n} \left| s_{ij} - s_{ik} \right|}{\sum_{i=1}^{n} s_{ij} + \sum_{i=1}^{n} s_{ik}} \left(j = 1 \ldots n; k = 1 \ldots n \right) \tag{1}$$

where, for the $i=1...n$ index of the summation, $i \neq j$, *and* $i \neq k$.

This way, again, we excluded the references of the categories to each other, as we want to see how similar category c_j is to category c_k, in terms of the composition of their influence vectors, from which we left out the direct link between them. We call $d_{jk} \in [0,1]$, the *BC-dissimilarity* assigned to the (c_j, c_k) category-pairs. The focus of our analysis is the temporal variation of the $\mathbf{D_{BC}}=[d_{jk}]_{\mathrm{nxn}}$ dissimilarity matrix.

A subset of the categories will be called *contracting* in a time interval, if the above dissimilarity between any two members of the subset consistently decreases during that interval. This means that the composition of the respective influence vectors of the categories belonging to this subset is becoming more and more similar to each other. Note that the direct link between two categories do not affect their BC-dissimilarity, as the summations in Eq. (1) excludes the $i=j$ and $i=k$ cases. In other words, if the (c_j, c_k) categories get closer to each other, that does not necessarily mean that they would refer to each other at higher probability, but it is the composition of their two respective

influence-vectors (that excludes the direct $c_j \rightarrow c_k$ and $c_k \rightarrow c_j$ references) that is becoming similar with time.

Gower-visualisation of the *Bray–Curtis* dissimilarity between categories

Consider the $\mathbf{G} = [g_{kl}]_{\text{nxn}}$ Gower-transformation of a $\mathbf{D_{BC}} = [d_{kl}]_{\text{nxn}}$ dissimilarity matrix:

$$g_{kl} = d'_{kl} - \frac{\sum_{j=1}^{n} d'_{kj}}{n} - \frac{\sum_{j=1}^{n} d'_{jl}}{n} + \frac{\sum_{i=1,j=1}^{n,n} d'_{ij}}{n^2} \; (k = 1 \ldots n; l = 1 \ldots n) \tag{2}$$

where $d'_{kl} = -0.5\, d_{kl}.^2$

The eigenvectors of **G** form a basis, with properties that can be utilized for visualization [1]. Namely, as the **G** matrix has non-negative and different eigenvalues (λ_1, λ_2 ….), consider the **V** matrix of its $\mathbf{v_k}$ column-eigenvectors and take the **W** matrix of the $\mathbf{w_k} = \text{sqr}(\lambda_k) \cdot \mathbf{v_k}$ modified (column-) eigenvectors ($k = 1, 2\ldots$). As proved by Gower [7], the scalar product of any two row-vectors of the $\mathbf{W = [w_1, w_2 \ldots]}$ matrix will be equal to the dissimilarity between the two respective categories that can be visualized by these row-vectors with ordinary Euclidean distance between them. Besides, their first 2–3 components are much greater than the rest, which can therefore be omitted, for the sake of representing the categories by these first few components of the row vectors, i.e., in 2 or 3 dimensions.

The temporal evolution of the obtained points demonstrates how the dissimilarities increase or decrease with time, with the potential of identifying what combinations of technological areas emerge or tend to form clusters (Fig. 4.).

Data

We analysed a freely available dataset on patents registered in 2018–19 of the United States Patent and Trademark Office (USPTO). The database is updated quarterly on the PatentsView Web portal. Accordingly, in what follows, we use the notation q_1 = quarter 1 of 2018; … q_8 = quarter 4 of 2019. Data were downloaded from the 'Data Downloads Tables' (tables named 'ipcr' and 'uspatentcitation'). Relevant data from the two tables were merged to establish the International Patent Classification (IPC) category of each patent. Each record of the compiled table represented a patent with (i) a unique ID of the actual patent; (ii) the IDs of the patents *cited by* the actual one; and (iii) the ID of the patents *citing* the actual one. Besides, the record also contained the IPC categories to which the actual patent was assigned and the date when the patent was approved. As we found different levels of patent classification, for the sake of simplicity, only the first level categories have been exploited. If the category or the approval date was not available, then the record was omitted. Altogether 115 categories (c_i, where $i = 1 \ldots 115$) were distinguished and linked by reference lists of the patents as described above. The categories were divided, in the database, into 8 groups denoted by A-H, and we followed this notation.

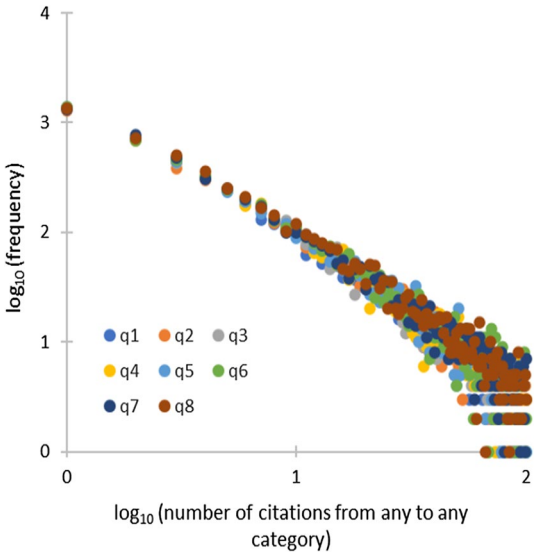


Fig. 1 The distribution of the weighted degrees of nodes (categories) throughout 2018–2019, on the log–log scale, follows the power-law

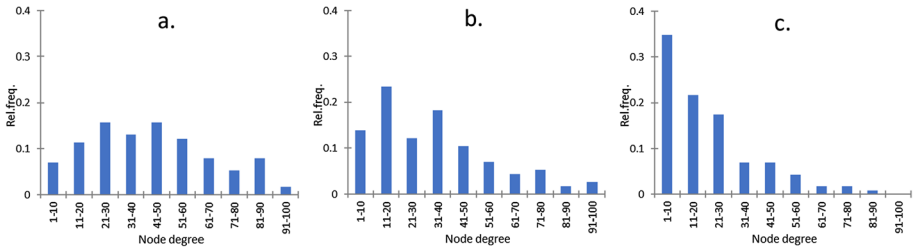


Fig. 2 In-degree distributions in the *q1* quarter-year if only those edges are considered whose weight is exactly *k*, where (**a**) $k=1$; (**b**) $k=2$; (**c**) $k=5$

Results

In each quarter year of 2018-19, more than 80,000 patents were processed and each of them was assigned to one of the 115 categories, so the number of nodes in the network is $N=115$. Their total number of references were more than 300,000 in each quarter-year, about third of which was omitted because they cited patents that were assigned to the same category.

This way, we constructed a weighted, directed network of categories, where the direction of an edge is from a cited category to a citing one and its weight is the number of respective citations. As follows from above, the sum of the weights (the entries of our S matrix) is ca. 200,000. The weights span from one to several hundreds, the smaller numbers being much more frequent (the lowest number, 1, occurring >1000 times), than the higher ones. The degree-distribution of the network nodes (i.e., the frequency diagram of the number of citations in the categories) follows the power-law as shown by Fig. 1. This indicates that the overall weighted citation network of

categories falls in the commonly observed scale-free group [8], throughout all the 8 quarter of years.

Note that the non-weighted version of the network would have ca. 5-6000 edges, out of the possible 13,000, so it is relatively dense. In this network, the average in-degree of a node (i.e. how many other categories it cited at least once) would be ca. 45–50, out of the possible maximum 114.

For the weighted network, which is our focus, we studied the properties of some weight-defined subnetworks. Figure 2 shows the distribution of the in-degrees of the nodes (i.e. the proportion, or relative frequency, of the incoming edges, with weights, within the total number of weighted edges), if only those edges count that represent exactly k citations, where $k = 1, 2, 5$. The distribution strongly depends on how many citations are needed to form an edge. With k increasing, the in-degree distribution transforms from unimodal to close-to-exponential.

That this finding is valid through the studied two years is demonstrated in Fig. 3. The in-degree distribution of the subnetwork with edges representing exactly 1 citation (Fig. 3a) is close to Poissonian for all the eight quarter-years. Recall that, while the degree distribution of an Erdős-Rényi random graph is Poissonian, that of a scale-free network follows the power-law [8]. This suggests that citing a patent from another category only once can be just random, resulting in the Erdős-Rényi option. However, when the edges represent several (at least ca. 5 citations; see Fig. 3b), then the pattern is more reminiscent to that generated by the power-law. A logical explanation for this is that, if a category cites another one via at least five patents, then the principle of preferential attachment [8] is more detectable and the influence of the cited category increases according to this bias. The "preferential attachment" can be translated for our case as: the probability of citing a category is proportional with the number of citations that this category already has. It has been proven [8], that this mechanism leads to the linear pattern on the log–log scale shown by Fig. 1, towards which the exponential distribution shown in Fig. 3b is an intermediate step.

For each quarter-year, we calculated the Bray–Curtis dissimilarities between the category-pairs, thus creating the $\mathbf{D_{BC}}(q_i) = [\mathbf{d}_{BC}(\mathbf{q}_i)]$ dissimilarity matrices as a function of the quarter-years and their differences with the $\Delta d_{BC}(q_i)$ entries:

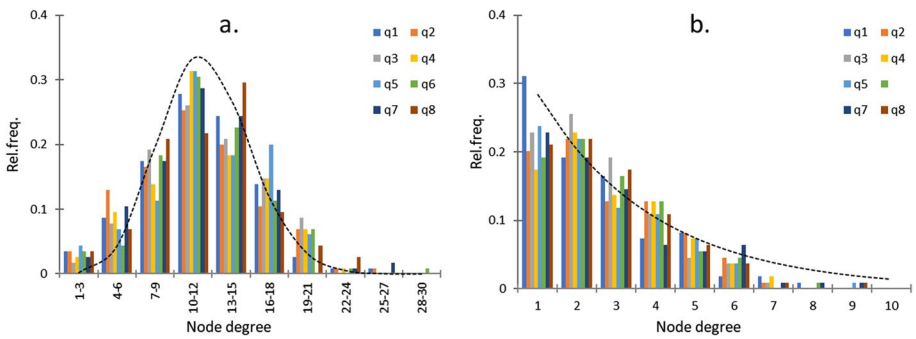


Fig. 3 In-degree-distribution of the categories (nodes) for all the eight quarter-years (*q1-q8*). Only those edges were considered, which represent (**a**) exactly one citation; (**b**) exactly five citations. The dotted lines represent the (**a**) Poisson- (**b**) exponential distribution, each with a mean-representing parameter that was estimated by the average of the respective dataset

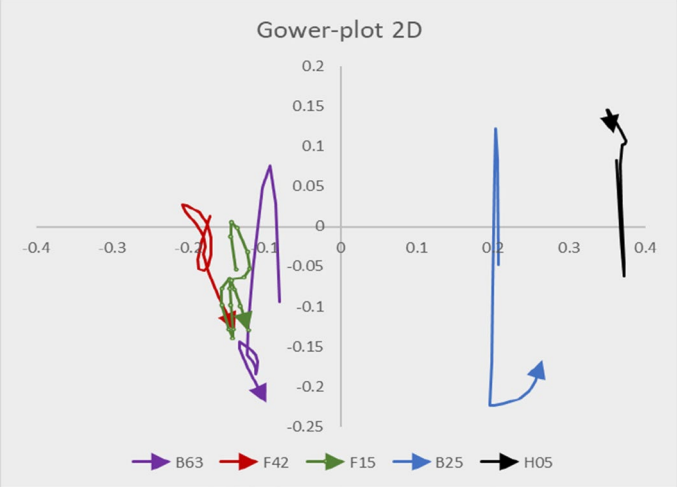


Fig. 4 The internal dynamics of five categories, when their dissimilarity is quantified by the Bray–Curtis measure, during the studied period (8 quarter-years). The Bray–Curtis space is projected onto a two-dimensional, Euclidian, Gower-space

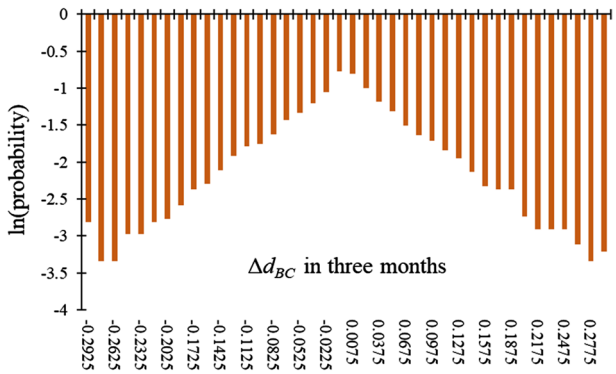


Fig. 5 Log-probability, or log(relative frequency), histogram of the Δd_{BC} changes in dissimilarities between categories in the patent database from the q_1 to the q_2 quarter-years. The figure shows that the further away two categories are in the BC-space, the more probable that they move closer to each other. So far this is expectable, but the linearity between the logarithm of the relative frequency of a change in the BC-dissimilarity and the BC-distance itself is surprising

$$\Delta d_{BC}(q_i) = d_{BC}(q_i) - d_{BC}(q_{i-1}) \quad (i = 2 \ldots 8) \tag{3}$$

For demonstration, Fig. 4. shows the movement of two contracting subgroups, with two and three members respectively. The movement is projected onto a two-dimensional Gower-space. The dynamics of the points can be followed in 3D, too, on Additional file 1, as a time-lapse simulation.

Figures 1, 2 and 3 showed that the degree distribution of our network follow consistent patterns through 2018-19. However, this does not mean that the network is stationary. In fact, it hides significant internal dynamics as Fig. 5 shows. It suggests

that, in a quarter-year, the log-probability of a category getting closer to / further from another category by a given Δd_{BC}-measure is a linear function of that measure. The more intriguing this is because the BC-transformation is non-linear, and the BC-measure does not qualify for a distance concept.

It is an open question, answer to which is out of the scope of this paper, whether there is a mechanistic explanation behind the observed linear pattern. What we established was that the "significant" connections (i.e., edges representing at least five references) generated a close-to scale-free network through the 8 quarter-years, presumably driven by the mechanism of preferential attachment. The internal dynamics however is far from stationary and for all the seven histograms of the transitions made between the eight quarter-years showed the log-linear pattern of Fig. 5.

Discussion

Intellectual achievements are being built on each other, and we took the US Patent & Trademark Office database to analyse these interactions. Two critical simplifications were made when analysing the data: A/We only considered the first level of categorisation. B/We did not differentiate between the times of patent filing and patenting. These simplifications may affect the findings, but here our focus was the methodology rather than higher resolution analysis.

We constructed a weighted, directed network of categories where the weighted edges represent references between the patents belonging to the categories. For each node (category), an "influence vector" was assigned, composition of which characterising how other categories affect that node. The temporal changes in the (dis)similarity of the composition of these influence vectors were used to identify the dynamics of the constructed network, representing this way a sort-of evolving intellectual ecology.

A critical concept here is the measure of dissimilarity between categories. For this, we chose a non-metric dissimilarity measure, that of Bray–Curtis [6], which is commonly used in numerical ecology.

The developed methodology could be used for example to describe the emergence of new technological developments, or to support decisions on research and development resource allocation [9].

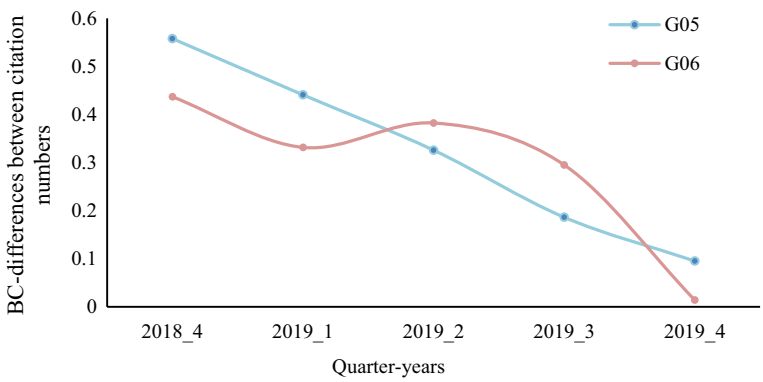


Fig. 6 Influence of the categories G05 and G06 on B25 and H05. Their component-wise BC-differences between the citation numbers show a decreasing trend over the studied five quarter-years

A method for this could start with identifying contracting subnetworks, as demonstrated in the 2D Gower-space in Fig. 4. To that end, the categories H05 ("Electric techniques not otherwise provided for") and B25 ("Hand-held tools; Portable, power-driven tools; Handles for hand implements; Workshop equipment; Manipulators") were further explored using a multi-step filtering method. We searched for the categories they referred to, then for patents within the search results that could explain the observed convergence over the two-year-long observation time.

We found that the categories G05 (controlling, regulating) and G06 (computing, calculating, counting) had a significant impact on their H05 and B25 citing categories. For both cited categories (G05, G06), the BC dissimilarity measure from H05 and B25 showed a steady decrease during the studied period (Fig. 6). This trend my be expectable by an expert; however, the role of quantitative modelling is not only to find new patterns but also rank various scenarios, thus give an objective tool to technologically less experienced managers, who nonetheless may be responsible for making decisions, for example about investing in new areas.

When exploring the patent database, the linear pattern demonstrated by Fig. 5 is probably the most surprising finding. Intrigued, we downloaded the data of Microsoft Academic Graph containing—among others—scientific publication records, citation relationships and fields of study. We carried out the same analysis as in case of patent data, but no linear pattern was observed for the analogous distribution shown in 5. Therefore, our observation is not due to some properties of the BC-dissimilarity measure. The reason might be the way how the patents were categorised, though it is not clear why.

Nonetheless, as an application, the observed linearity could be a reference for other categorisation methods. However, while the preferential attachment principle [8] is an elegant explanation for the scale-free pattern shown by some of the previous results, analogous mechanistic reason for this last one seems neither straightforward nor intuitively predictable.

Abbreviations

BC	Bray–Curtis (BC) dissimilarity
IPC	International patent classification
USPTO	United States Patent and Trademark Office

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-024-00883-z.

Additional file 1: Annex 1. Category G05 patents most cited by category B25 patents. **Annex 2.** Category G06 patents most cited by category B25 patents.

Acknowledgements
Not applicable.

Author contributions
JB and ÁJ developed the study conception and design. Data collection and preparation were performed by Sz. Cs., data analysis was performed by all authors, and data visualization was performed by JB, ZsF and TP. The first draft of the manuscript was written by JB and ÁJ and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding
Open access funding provided by University of Veterinary Medicine.

Availability of data and materials
Data supporting the results reported in the article can be found here: We analysed a freely available dataset on patents registered in 2018-19 of the United States Patent and Trademark Office (USPTO). The database is updated quarterly on the PatentsView Web portal: https://patentsview.org/download/data-download-tables. Data were downloaded from the 'Data Downloads Tables' (tables named 'ipcr' and 'uspatentcitation'). We analysed data of Microsoft Academic Graph containing—among others—scientific publication records, citation relationships and fields of study, accessible here: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 12 May 2023 Accepted: 19 January 2024
Published online: 04 February 2024

References

1. Legendre PLL. Numerical Ecology. New York: Elsevier; 2012.
2. Érdi P, Makovi K, Somogyvári Z, Strandburg K, Tobochnik J, Volf P, et al. Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics. 2013;95(1):225–42. https://doi.org/10.1007/s11192-012-0796-4.
3. Bruck P, Réthy I, Szente J, Tobochnik J, Érdi P. Recognition of emerging technology trends: class-selective study of citations in the U.S. Patent Citation Network. Scientometrics. 2016;107(3):1465–75. https://doi.org/10.1007/s11192-016-1899-0.
4. Beltz H, Rutledge T, Wadhwa RR, Bruck P, Tobochnik J, Fülöp A, et al. Ranking algorithms: application for patent citation network. In: Bossé É, Rogova GL, editors., et al., Information quality in information fusion and decision making. Cham: Springer International Publishing; 2019. p. 519–38.
5. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236–44. https://doi.org/10.1080/01621459.1963.10500845.
6. Bray JR, Curtis JT. An ordination of the upland forest communities of southern wisconsin. Ecol Monogr. 1957;27(4):325–49. https://doi.org/10.2307/1942268.
7. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53(3–4):325–38. https://doi.org/10.1093/biomet/53.3-4.325.
8. Barabási A-L, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–12. https://doi.org/10.1126/science.286.5439.509.
9. Farkas Z, Országh E, Engelhardt T, Zentai A, Süth M, Csorba S, Jóźwiak Á. Emerging risk identification in the food chain—a systematic procedure and data analytical options. Innovat Food Sci Emerg Technol. 2023. https://doi.org/10.1016/j.ifset.2023.103366.

Publisher's Note