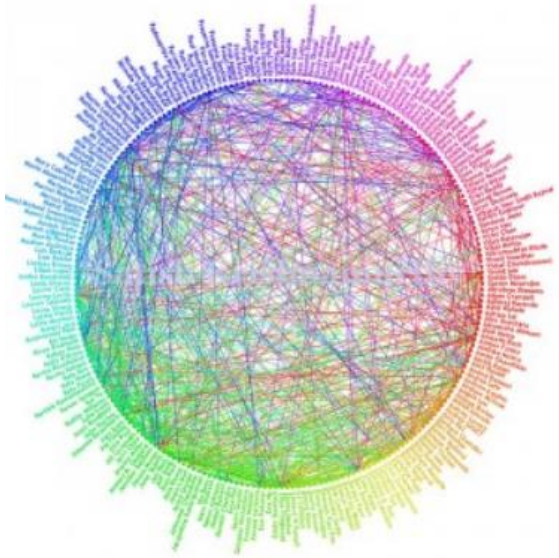
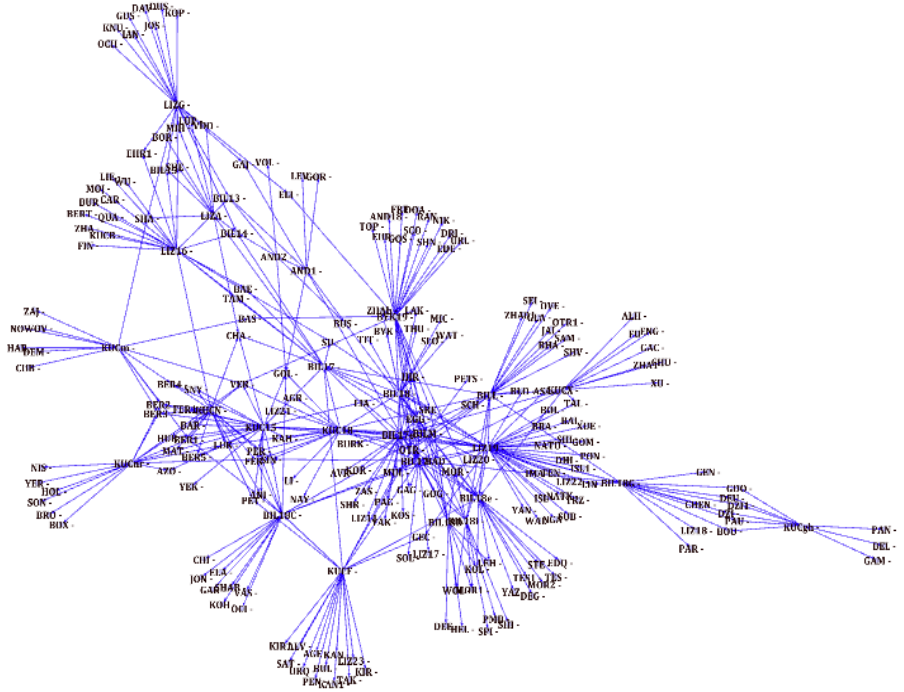
Scientometric study of research productivity

Oleksandr Kuchanskyi,
Doctor of Technical Sciences,
Professor of Taras Shevchenko National University of Kyiv

Information space of scientific activity subjects





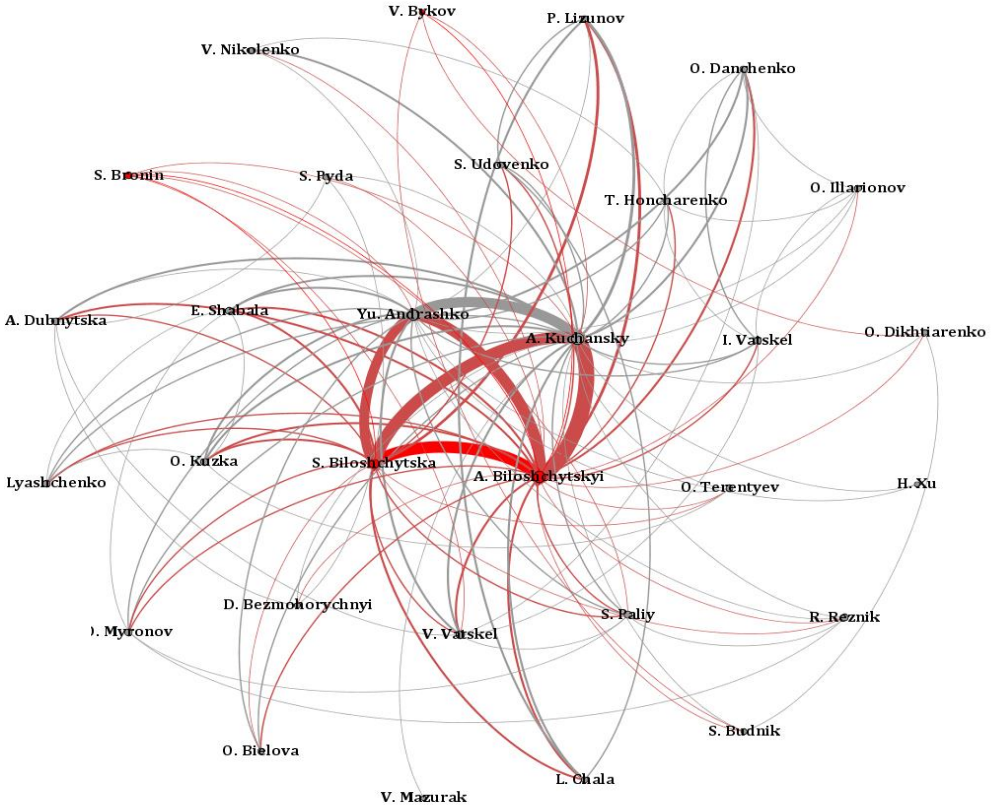


Figure 1. - Network of scientific cooperation and network of citations of scientific publications

Individual and collective scientific activity subjects

Individual scientific activity subjects

Collective scientific activity subjects

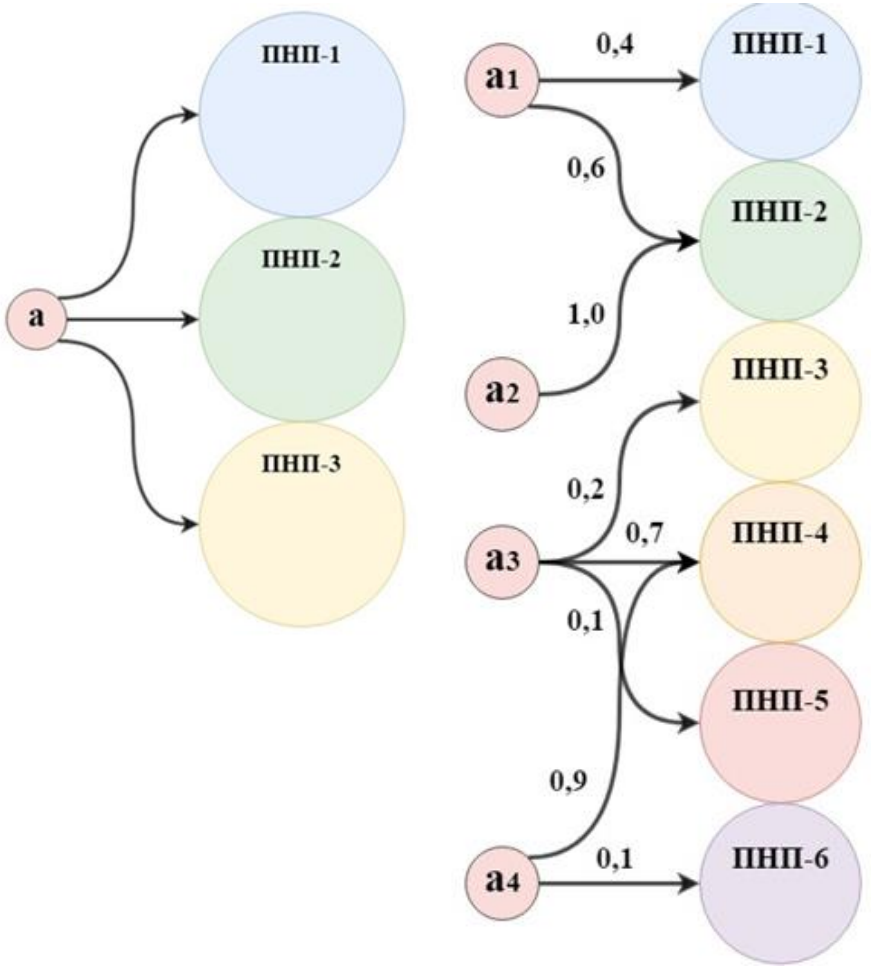


Figure 2. - Belonging of scientists to several subject scientific spaces

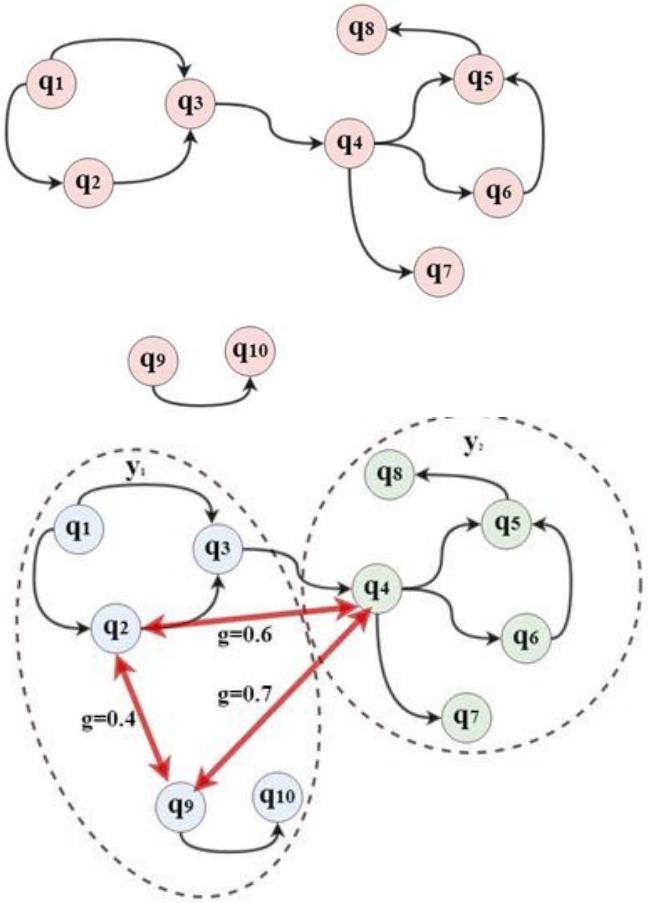


Figure 3. - Publication clusters

Set models of identification of scientific activity subjects

Let $A=\{a_1, a_2, ..., a_n\}$ be a set of scientific activity subjects or scientists who are active representatives of the educational and scientific space, that is, they are engaged in active publication activities, n is the number of scientists. Let $Q=\{q_1, q_2, ..., q_m\}$ be a set of scientific publications published by scientists from set A, m is the number of publications.

Let the set of all pairs between the elements of the sets A and Q be assigned

$$A\times Q=\{(a,q)|a\in A\wedge q\in Q\}$$

and the binary relationship $U\subset A\times Q$, which defines the authorship of publications $q\in Q$. In addition, we determine a binary relationship that defines the citation of all publications $q\in Q$,

$$Q\times Q=\{(q_i,q_j)|q_i,q_j\in Q, i\neq j\}.$$

A set of all publications by a scientist $a\in A$ is denoted as

$$Q(a_i,t)=\{q_j\in Q|(a_i,q_j)\in U, t\in T\},\quad i=\overline{1,n},\quad j=\overline{1,m}.$$

A set of scientific publications cited by each scientific activity subject at the time t is denoted through $\overline{C}(a_i,t)$, the set of publications in which the publications of the scientific activity subject are cited at the time t – through $C(a_i)$, thus:

$$\overline{C}(a_i,t)=\left\{\begin{matrix} q_j\in Q|(q_y,q_j)\in C,\\ q_y\in Q(a_i),\\ t\in T, y=\overline{1,m}\end{matrix}\right\},\quad i=\overline{1,n},\quad j=\overline{1,m}.$$

$$C(a_i,t)=\left\{\begin{matrix} q_j\in Q|(q_j,q_y)\in C,\\ q_y\in Q(a_i),\\ t\in T, y=\overline{1,m}\end{matrix}\right\},\quad i=\overline{1,n},\quad j=\overline{1,m}.$$

For each publication $q\in Q$, determine the set of its authors

$$A(q_j)=\{a_i\in A|(a_i,q_j)\in U\},\quad i=\overline{1,n},\quad j=\overline{1,m}.$$

as well as a set of scientific publications quoted by a given publication $q\in Q$ at the time $t-\overline{C}(q_j)$, and a set of publications quoting the publication $q\in Q$ at the time $t-C(q)$:

$$\overline{C}(q_j,t)=\left\{\begin{matrix} q_j\in Q|(q_j,q_y)\in C,\\ t\in T, y=\overline{1,m}\end{matrix}\right\},\quad j=\overline{1,m},$$

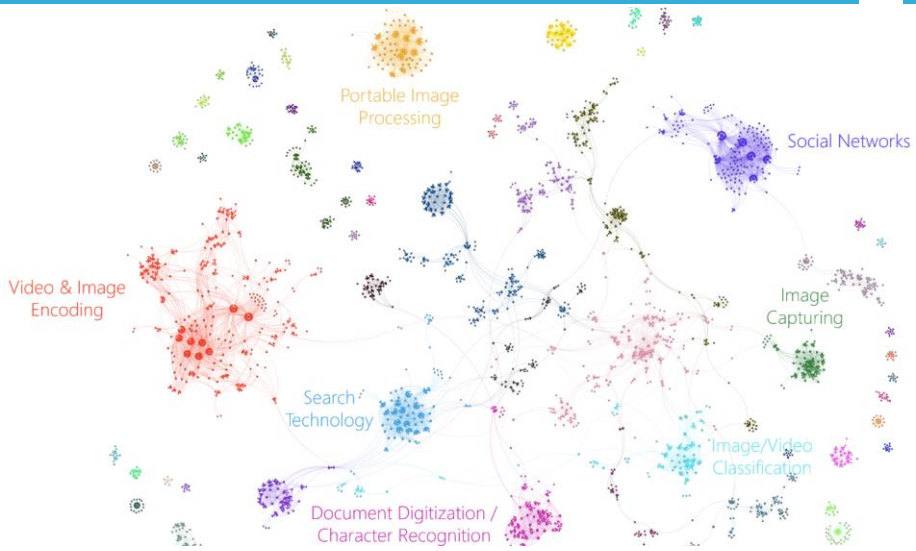
$$C(q_j,t)=\left\{\begin{matrix} q_j\in Q|(q_y,q_j)\in C,\\ t\in T, y=\overline{1,m}\end{matrix}\right\},\quad j=\overline{1,m}.$$

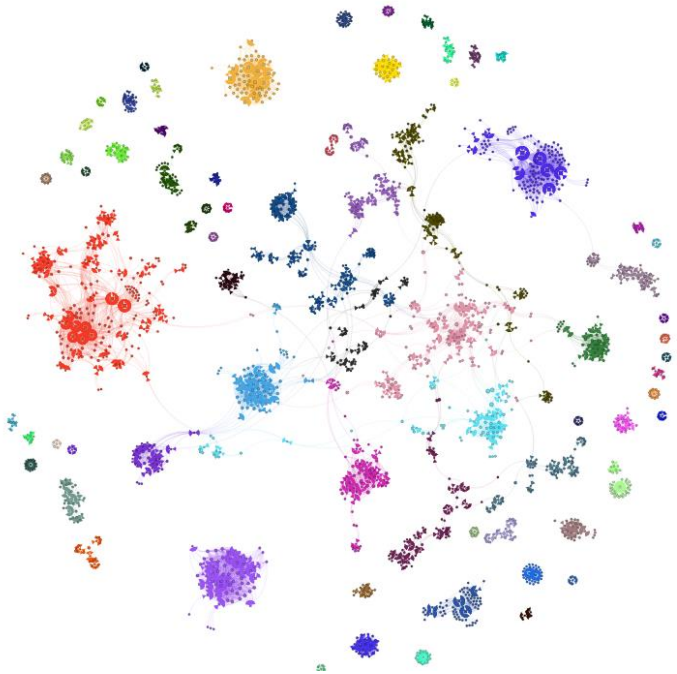
Each publication $q\in Q$ is matched with its abstract $q^A\in Q^A$.

Scalar assessment of the productivity of a scientific activity subject is a certain functional representation Φ:

$$\Phi: A\to R,$$

where R is the set of real numbers.





Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Omirbayev, S., Mukhatayev, A., Faizullin, A., Toxanov, S. (2021). Development of the set models and a method to form information spaces of scientific activity subjects for the steady development of higher education establishments. Eastern-European Journal of Enterprise Technologies, 3 (2 (111)), 6–14. doi: https://doi.org/10.15587/1729-4061.2021.233655

https://www.kenedict.com/

Set models of identification of scientific activity subjects

$$\left(\aleph^I(a_i,t_0),\aleph^I(a_i,t_1),\ldots,\aleph^I(a_i,t_{x-1}),\aleph^I(a_i,t_x)\right).$$

The individual scientific activity subject at the time t is represented by a set of identifiers that determine it:

$$\aleph^I(a_i,t)=\left\langle\begin{matrix}Q(a_i,t),\overline{C}(a_i,t),C(a_i,t),\\ Q^A(a_i,t),\Phi(a_i,t),M(a_i,t)\end{matrix}\right\rangle,$$

where $Q(a_i,t))$ is the set of publications by an individual scientific activity subject a_i at the time t;

$\overline{C}(a_i,t)$ – the set of scientific publications cited by the scientific activity subject a_i at the time t;

$C(a_i,t)$ – the set of publications that quote the publications by the scientific activity subject a_i at the time t;

$Q^A(a_i,t)$ – the set of abstracts of the publications by the scientific activity subject a_i at the time t;

$\Phi(a_i,t)$ – the evaluation of the performance of the scientific activity subject a_i at the time t, for example, the Hirsch index;

$M(a_i,t)$ – the number of international projects in which the scientific activity subject a_i participates at the time t.

In addition, the model of identification of an individual scientific activity subject could include the impact factor of his/her publications and other parameters.

$$\left(\aleph^C(a_i,t_0),\aleph^C(a_i,t_1),\ldots,\aleph^C(a_i,t_{x-1}),\aleph^C(a_i,t_x)\right).$$

The collective scientific activity subject is represented by a set of identifiers that determine it:

$$\aleph^C(a_i,t)=\left\langle\begin{matrix}\Phi_1(a_i,t),\Phi_2(a_i,t),\Phi_3(a_i,t),\\ \Phi_4(a_i,t),\Phi_5(a_i,t)\end{matrix}\right\rangle,$$

where $\Phi_1(a_i,t)$ is the normalized assessment of the international activity by the scientific activity subject a_i at the time t, for example, the number of internships of employees of a collective scientific activity subject abroad, the number of projects with foreign funding, etc.;

$\Phi_2(a_i,t)$ is the normalized assessment of the cohort of higher education applicants of the collective scientific activity subject, a_i, at the time t, if the subject renders educational services, for example, as a higher education establishment (HEE). The assessment of HEE should take into consideration the educational component, the cohort of students: the average EIT score of those enrolled, passing minimum, etc., which are important estimates that determine the potential of HEE;

$\Phi_3(a_i,t)$ is the normalized assessment of the scientific or scientific-pedagogical composition of the collective scientific activity subject a_i at the time t; it can be an average quantitative assessment of the productivity of employees over a certain period of time;

$\Phi_4(a_i,t)$ is the normalized quantitative estimation of research activity of the collective scientific activity subject at the time t;

$\Phi_5(a_i,t)$ is the normalized assessment of available resource support at the time t, in particular the material and technical support of the collective scientific activity subject.

Potential of the development of scientific activity subjects

Collective scientific activity subjects

The potential of the development of scientific activity subjects is determined by the rate of change in the indicators of identifiers of these subjects, which is calculated as a percentage comparison of the current value of each identifier with the value of one of the previous periods. For collective scientific activity subjects:

$$S^C(a_i,t_x,t_b)=\frac{1}{5}\sum_{j=1}^{5}\frac{\Phi_j(a_i,t_x)-\Phi_j(a_i,t_{x-b})}{\Phi_j(a_i,t_{x-b})}\cdot 100\,\%, \qquad (14)$$

where $S^C(a_i, t_x, t_b)$ is the rate of change in the indicators of identifiers of collective scientific activity subjects a_i at the time t_x relative to the time t_b, that is, the potential of collective scientific activity subjects a_i.

Individual scientific activity subjects

Similarly, for individual scientific activity subjects, a rate of change in the indicators is determined from the following formulas:

$$Q^N(a_i,t_x)=\frac{|Q(a_i,t_x)|-\min\limits_{j=1,n}(|Q(a_j,t_x)|)}{\max\limits_{j=1,n}(|Q(a_j,t_x)|)-\min\limits_{j=1,n}(|Q(a_j,t_x)|)},$$

$$C^N(a_i,t_x)=\frac{|C(a_i,t_x)|-\min\limits_{j=1,n}(|C(a_j,t_x)|)}{\max\limits_{j=1,n}(|C(a_j,t_x)|)-\min\limits_{j=1,n}(|C(a_j,t_x)|)},$$

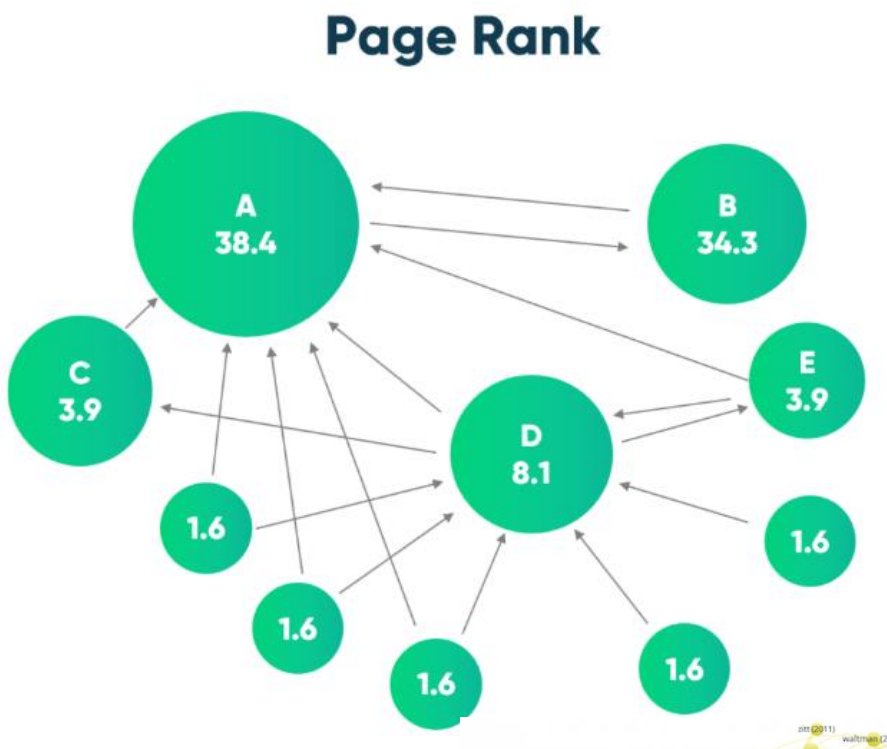
$$M^N(a_i,t_x)=\frac{|M(a_i,t_x)|-\min\limits_{j=1,n}(|M(a_j,t_x)|)}{\max\limits_{j=1,n}(|M(a_j,t_x)|)-\min\limits_{j=1,n}(|M(a_j,t_x)|)},$$

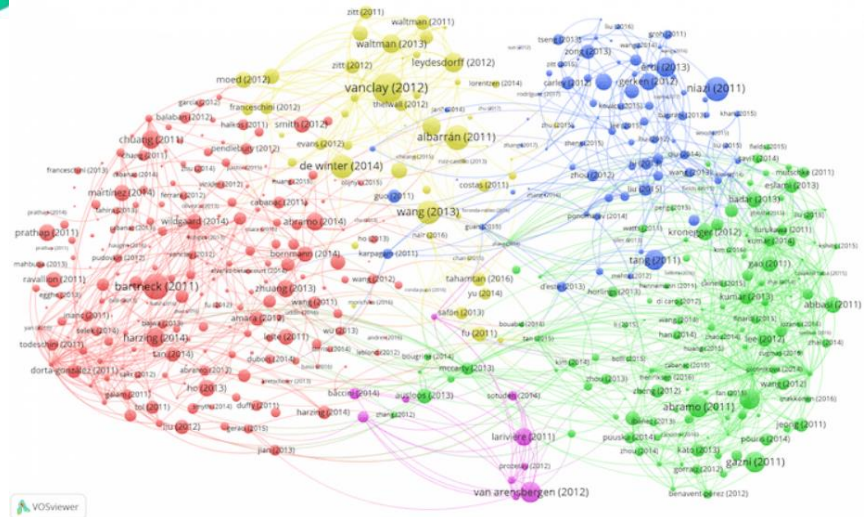
$$\Phi^N(a_i,t_x)=\frac{\Phi(a_i,t_x)-\min\limits_{j=1,n}(\Phi(a_j,t_x))}{\max\limits_{j=1,n}(\Phi(a_j,t_x))-\min\limits_{j=1,n}(\Phi(a_j,t_x))},$$

$$S^I(a_i,t_x,t_b)=$$

$$=\frac{1}{4}\cdot\left(\begin{array}{l}\dfrac{Q^N(a_i,t_x)-Q^N(a_i,t_{x-b})}{Q^N(a_i,t_{x-b})}+\\ +\dfrac{C^N(a_i,t_x)-C^N(a_i,t_{x-b})}{C^N(a_i,t_{x-b})}+\\ +\dfrac{M^N(a_i,t_x)-M^N(a_i,t_{x-b})}{M^N(a_i,t_{x-b})}+\\ +\dfrac{\Phi^N(a_i,t_x)-\Phi^N(a_i,t_{x-b})}{\Phi^N(a_i,t_{x-b})}\end{array}\right)\cdot 100\,\%,$$

PageRank Method





Let $U=\{U_1,U_2,\ldots,U_s\}$ is the set of collective subjects,
$A=\{a_1,a_2,\ldots,a_d\}$ is the set of individual subjects.
Certain individual subjects are affiliated with each collective subject:

$$f: A\times U\to\{0,1\}.$$

Let us denote by $A^h=\{a_1^h,a_2^h,\ldots,a_{d_h}^h\}$ the set of individual subjects affiliated with the collective subject U_h, $h=\overline{1,s}$, $a_j^h\in A$, $A^h\subset A$, $j=\overline{1,d_h}$, d_h is the number of individual subjects affiliated with the corresponding collective subject U_h. Let $P=\{p_1,p_2,\ldots,p_n\}$ is the set of all scientific publications. Let $p^t(a_i^h)$ is the number of scientific publications published by an individual subject a_i^h at time t, $t\in T$, $T=\{t_1,t_2,\ldots,t_N\}$. Let $c_i^t(a_i^h)$ is the number of citations of a scientific publication p_i at time t of an individual subject a_i^h affiliated with a collective subject of scientific activity U_h, $i=\overline{1,g_i^h}$, g_i^h is the number of scientific publications of an individual subject a_i^h.

Let's set the Markov matrix that determines the citation between publications through $M=\{c_{ij}\}_{i,j=\overline{1,n}}^n$, where n is the total number of scientific publications, $c_{ij}\in[0,1]$ is the probability of transition from one state to another, i.e. in the context of the PR method is the probability of citing one scientific publication in another, $M\geq 0$, $\sum_{i=1}^{n}c_{ij}=1$, $j=\overline{1,n}$.

All other coefficients are calculated iteratively according to the formula:

$$r_i^{k+1}=\alpha M r_i^k+\frac{1-\alpha}{n}\mathbf{E},$$

where $\mathbf{E}$ is the unit matrix, α is the damping factor.
The coefficients r_i^k are calculated iteratively:

$$r_i^{k+1}=\sum_{j=1}^{i-1}r_j^{k+1}c_{ij}+\sum_{j=i}^{n}r_j^k c_{ij}.$$

Evaluating Academic Papers by PageRank
https://delante.co/definitions/pagerank/

Time-Weighted PageRank Method with citation intensity

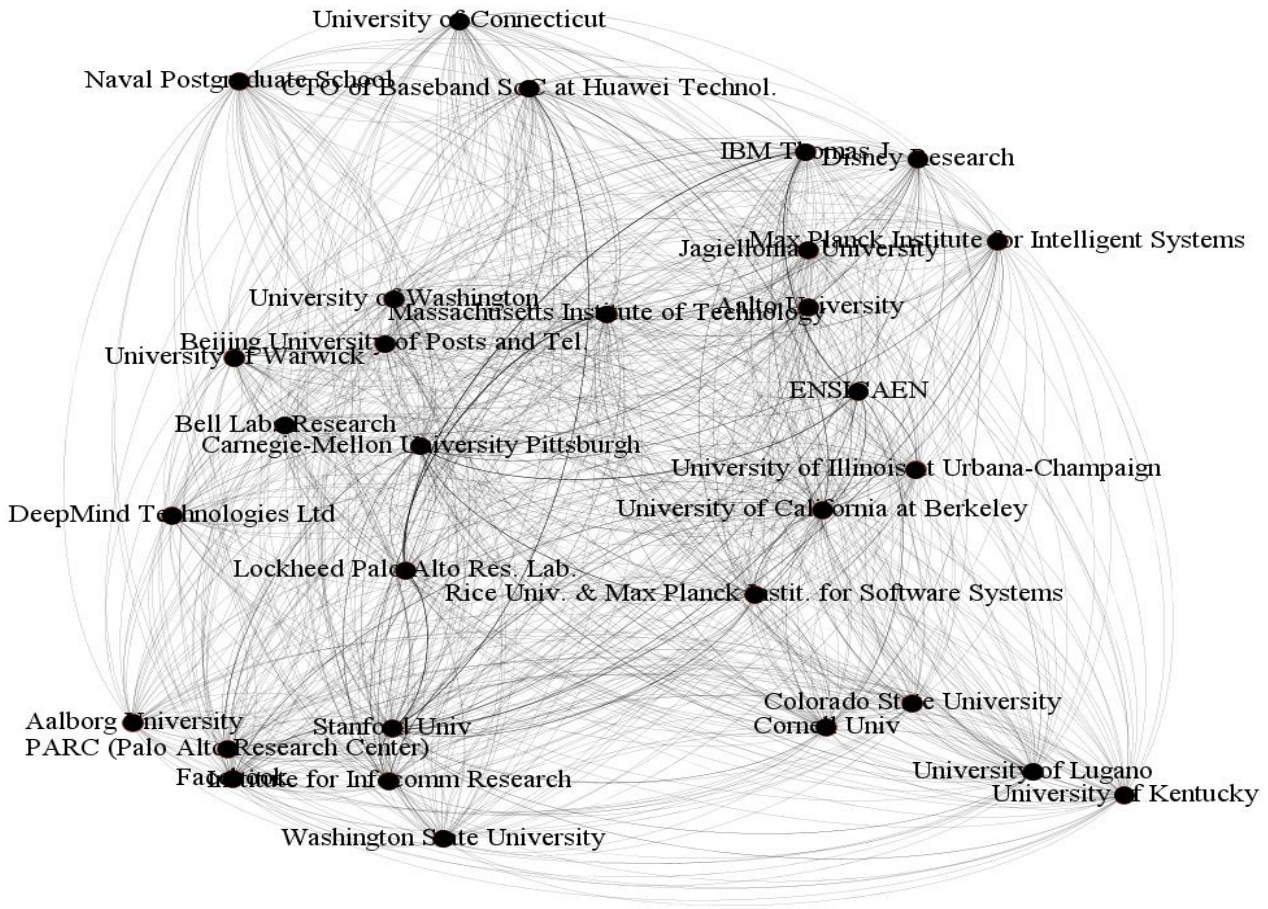


Figure 4. - Graph of citations of scientific publications for 30 selected collective subjects

The value for the collective subject U_h:

$$\theta_{t_0^h} = \operatorname{arctg}\left(\frac{\sum_{j=1}^{d_h^t} c^t\left(a_j^h\right)}{\lambda\left(t_N - t_0^h\right)\sum_{j=1}^{d_h^t} p^t\left(a_j^h\right)}\right),$$

where $\theta_{t_0^h}$ is the intensity of citation of scientific publications by authors who are affiliated with the collective subject U_h at the moment of time t, $t \in T$, d_h^t is the number of individual subjects affiliated with the collective subject U_h at the moment of time t, $p^t\left(a_j^h\right)$ is the number of scientific publications published by the individual subject a_j^h at the moment of time t, $t \in T$, $c^t\left(a_j^h\right)$ is the number of citations of the scientific publications of the authors, which are affiliated with the collective subject U_h at the moment of time t, λ is the coefficient, $\lambda > 1$, t_0^h is the moment from which the calculation of the intensity of citations of scientific publications for the collective subject U_h. For new collective subjects, the number of scientific publications and citation indicators is zero until the first publications of authors affiliated with them appear.

Let's determine the coefficient q_h^0 taking into account the age and intensity of citations as follows:

$$q_h^0 = \beta \cdot \theta_{t_0^h} + (1-\beta) \cdot \sum_{k=t_0^h}^{N} \frac{\left(k - t_0^h + 1\right) \cdot c^{t_k}\left(a_j^h\right)}{x_i},$$

where $x_i = \sum_{k=t_0^h}^{T_h} \left(k - t_0^h + 1\right)$, q_h^0 is the value of the coefficient taking into account the age and intensity of citations for the collective subject, U_h, $h = \overline{1,s}$, $\beta \in [0,1]$.

Table 1. Distribution of collective subjects by classes

Class	Number of collective subjects
N	16544
WK	236
O	4928
NC	5791

Time-Weighted PageRank Method with citation intensity

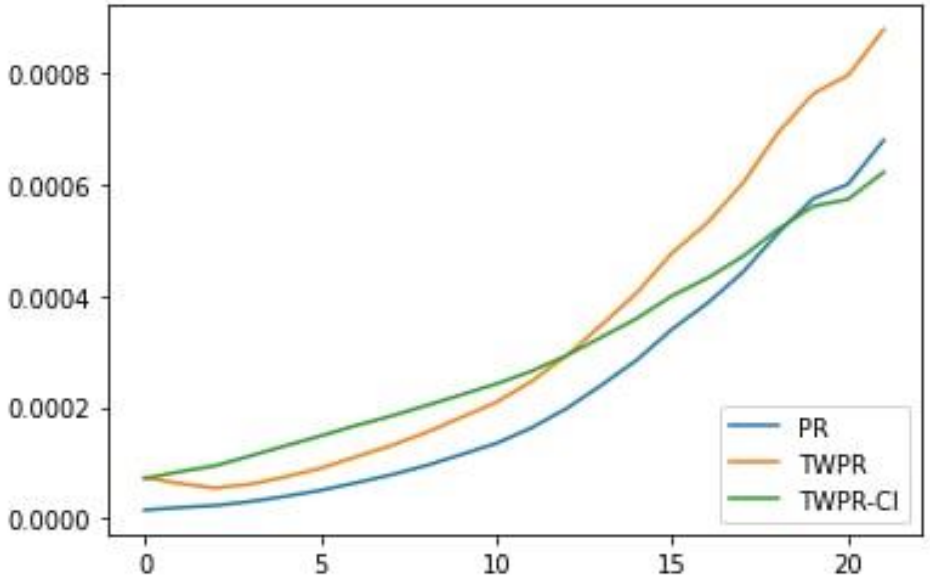


Figure 5. - Graph of comparison of scientific productivity estimates of collective subjects belonging to class N

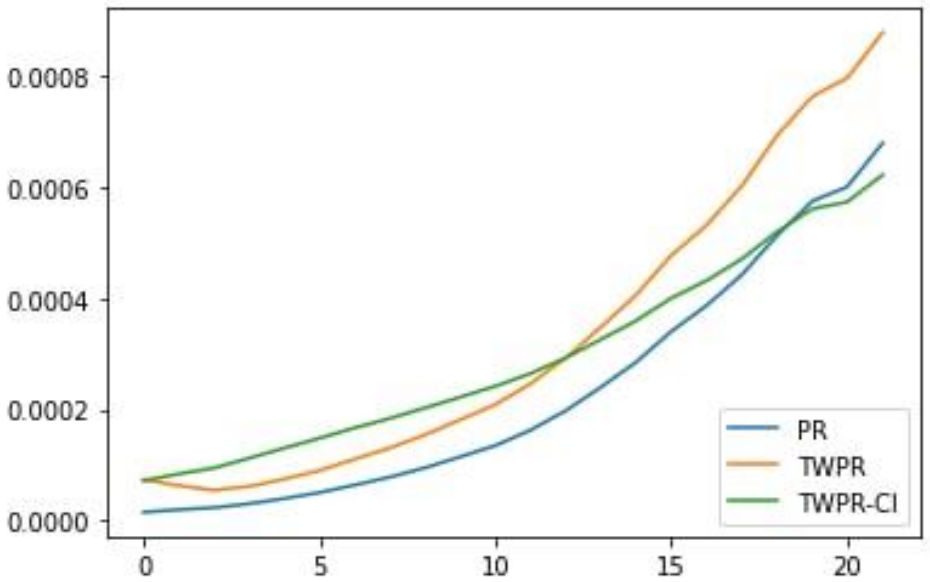


Figure 6. - Graph of comparison of scientific productivity estimates of collective subjects belonging to class WK

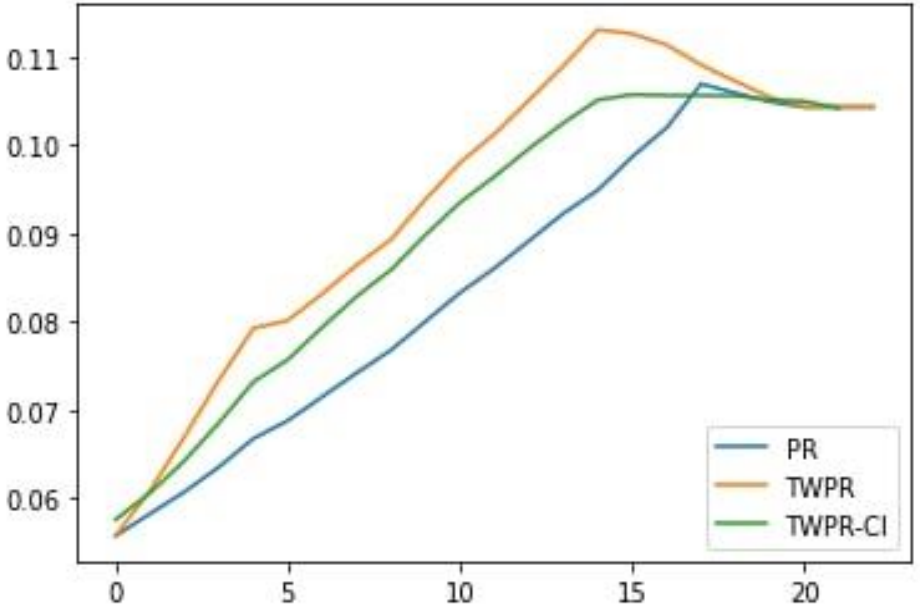
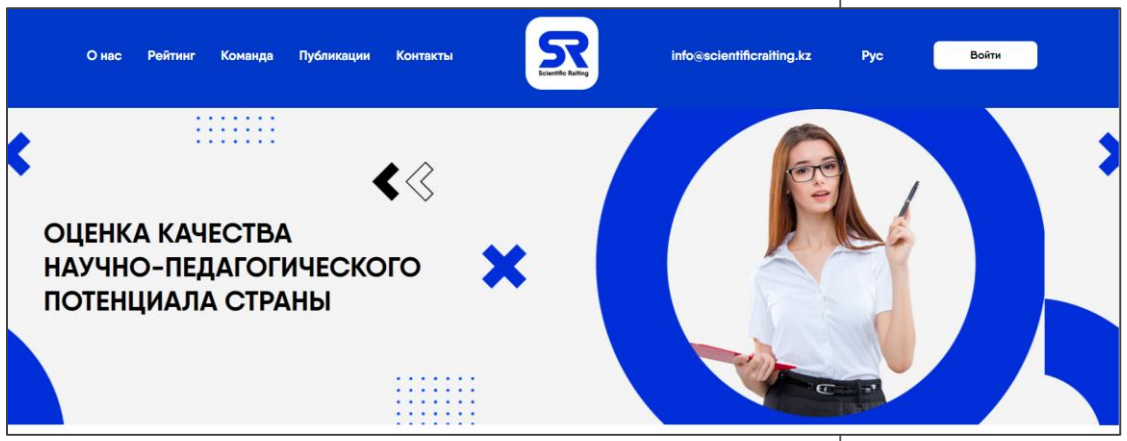


Figure 7. - Graph of comparison of scientific productivity estimates of collective subjects belonging to class O

Information system

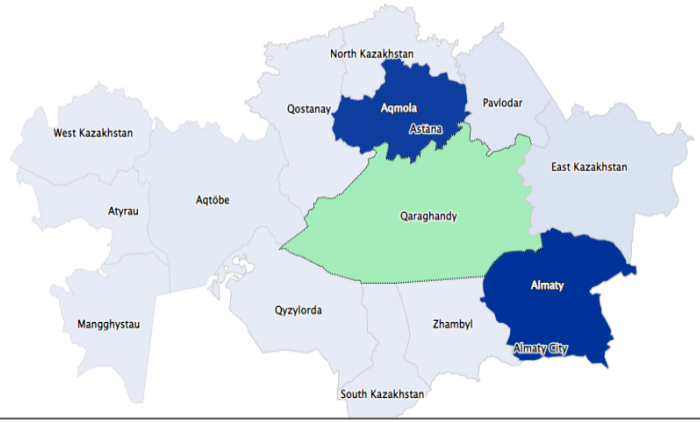


Критерии оценивания

- Контингент обучающихся
- Цитируемость статей сотрудников вуза
- Количество статей в журналах, рекомендованных ККСОН
- Количество ППС и научных работников
- Издание журналов, рекомендованных ККСОН
- Объем внешнего финансирования НИР
- Количество статей, опубликованных сотрудниками
- Издание журналов, входящих в международные базы Web of Science, Scopus
- Проведение научных конференций, форумов, труды которых индексируются в международных БД Web of Science, Scopus

Аналитика

Количество вузов в базе	Количество учёных в базе	Количество статей в базе
24	8920	48259



Публикационная деятельность	24878
Публикации из базы данных Google Scholar	5169
Цитирований по базе данных Google Scholar	19457
Публикации из базы данных EBSCO	52
Публикации в изданиях рекомендованных ККСОН РК	200

Thank you

E-mail: kuchansky@knu.ua