






Rapid Communication

A mechanism for evolution of the physical concepts network

V. Palchykov ², M. Krasnytska ^{1,2}, O. Mryglod ^{1,2}, Yu. Holovatch ^{1,2,3}

¹ Institute for Condensed Matter Physics, National Acad. Sci. of Ukraine, 79011 Lviv, Ukraine

² ⁴ Collaboration & Doctoral College for the Statistical Physics of Complex Systems, Leipzig-Lorraine-Lviv-Coventry, Europe

³ Centre for Fluid and Complex Systems, Coventry University, Coventry, CV1 5FB, United Kingdom

Received May 1, 2021, in final form June 1, 2021

We suggest an underlying mechanism that governs the growth of a network of concepts, a complex network that reflects the connections between different scientific concepts based on their co-occurrences in publications. To this end, we perform empirical analysis of a network of concepts based on the preprints in physics submitted to the arXiv.org. We calculate the network characteristics and show that they cannot follow as a result of several simple commonly used network growth models. In turn, we suggest that a simultaneous account of two factors, i.e., growth by blocks and preferential selection, gives an explanation of empirically observed properties of the concepts network. Moreover, the observed structure emerges as a synergistic effect of these both factors: each of them alone does not lead to a satisfactory picture.

Key words: *complex systems, complex networks, network of concepts, evolutionary model, preferential attachment*

Networks of concepts, i.e., semantic networks that reflect the relations between concepts in a certain domain are ubiquitously met in different spheres of modern life [1]. Their importance is both due to the fundamental reasons and numerous applications, ranging from ontologies in computer and information science [2] to visual knowledge maps that serve as an aid showing where to look for a certain knowledge [3]. Such networks are of particular interest for the logology — ‘science of science’, that aims at quantitative understanding of the origins of scientific discovery and creativity, its structure and practice [4, 5]. Scientific papers are an ideal source to investigate such processes, providing validated and open results of scientific creativity that are recorded in text formats and supplied by numerous supporting pieces of information. A common approach to the quantitative description of the knowledge structure is via the analysis of its projections to semantic spaces for different domains, see e.g., [6] and references therein. The latter can be modelled as complex networks based on topic-indicating labels. To give a few examples, one can mention here the networks of papers in physics that co-used PACS¹ numbers [7, 8], biomedical papers that co-mentioned the same chemical entities [9], papers in cognitive neuroscience [10] and in quantum physics [6] with co-occurrence of predefined concepts, Wikipedia pages devoted to mathematical theorems [11], etc. In all the above cases, complex network formalism enables quantitative analysis of similarities between different entities which are typically considered as indicators of topical relatedness and, therefore, as projections of knowledge.

Besides, the networks discussed above rise as an outcome of a dynamical process at which a new knowledge is acquired. Innovations themselves can be interpreted as an emergence of new concepts or new relations between the existing ones [12–14]. Modelling such process is a challenging task both for its fundamental relevance and numerous practical implementations. The process of a scientific discovery itself is governed by the structure of scientific knowledge. At the same time, it leads to changes in

¹Physics and Astronomy Classification Scheme

Table 1. Some features of networks of concepts addressed in our study. An empirically observed network (first line) is compared with three different models discussed in the paper: Erdős-Rényi, Barabási-Albert, and growth by blocks with preferential selection (GBPS). The following features are shown: the number of nodes N , number of links L , density of links ρ , average node degree $\langle k \rangle$, its standard deviation σ and maximal value k_{\max} , assortativity mixing by degrees r , mean clustering coefficient $\langle c \rangle$ and global transitivity T .

	N	$L, \times 10^6$	ρ	$\langle k \rangle$	σ	k_{\max}	r	$\langle c \rangle$	T
empirical	11853	5.38	7.66%	908	1146	9970	-0.28	0.74	0.38
Erdős-Rényi	11853	5.38	7.66%	908	29	1023	0.00	0.08	0.08
Barabási-Albert	11853	5.38	7.66%	908	568	3875	0.01	0.15	0.15
GBPS	11554	1.50	2.25%	260	788	7603	-0.62	0.95	0.12

this structure: in other words, they dynamically update each other. Presence of such a co-evolution is a typical feature of any complex system [15, 16] and is reflected, in particular, in the growth dynamics of the underlying complex networks of terms, keywords, labels or tags that become co-chosen from some predefined semantic space. Modelling such complex networks, along with their empirical analysis, is a challenging task that provides a deeper understanding of their growth mechanisms [12, 17, 18].

In this Letter, we suggest an underlying mechanism that governs the growth of a network of concepts originating from the texts of preprints in physics submitted to e-print repository arXiv [19]. First, we perform an empirical analysis of this network and calculate its topological characteristics. We discuss the main network features and show that simultaneous account of two factors i.e., growth by blocks and preferential selection, gives an explanation of empirically observed properties. A detailed account of our analysis is to be published elsewhere [20].

We used the vocabulary of scientific concepts in the domain of physics that has been collected by the ScienceWISE.info platform [21] and refined by continuous updates by expert evaluations. The resulting ontology includes such concepts as Ferromagnetism, Quantum Hall Effect, Renormalization group, and thousands of others. To our knowledge, currently such a vocabulary is the most comprehensive vocabulary of this type in the domain of physics. The sample of articles we analysed consists of 36,386 entities submitted to arXiv during a single year 2013 that have been assigned to a single category during submission process and is in one-to-one correspondence with the data set being analyzed in [14, 22, 23]. For each of the articles, a set of its inherent concepts has been defined using the above mentioned vocabulary of concepts. In this way, we arrived at the data that are conveniently described as a bipartite network consisting of the nodes of two types: articles A_1, A_2, \dots, A_N and concepts C_1, C_2, \dots, C_N , each A -node is linked to those C -nodes that represent its inherent concepts. While the properties of the bipartite network and its one-mode projection into the space of articles were analysed in [22, 23], here we concentrate on its one-mode projection into the space of concepts. Now, all C -nodes that were connected to the same A -node enter the network as a complete graph or clique. Hereafter, such a one-mode projection is called a *network of concepts* and is a subject of empirical analysis and modelling.

The main characteristics of the network of concepts constructed based on the data described in the former paragraph are given in the first line (denoted as ‘empirical’) of table 1. There, out of many network indicators, we display those that describe the most typical features addressed below. In particular, the empirically observed network of concepts is very dense: the density of links $\rho = 2L/N(N-1) = 7.66\%$. This number indicates that concepts are densely connected within a considered discipline: the authors who conduct research in physics, extensively use common terminology. One of the consequences is the high value of the mean node degree. Standard deviation of the node degree distribution indicates a high level of inhomogeneity among concept co-occurrence statistics. This can be also observed from the skewed shape of the histogram of node degree values $N(k)$ as shown in figure 1a by grey discs. The tail of the histogram may be visually compared with a power-law function $k^{-\gamma}$ with an exponent close to $\gamma = 1$. While this empirical network cannot be formally classified as the so-called *dense network* [24–27], it is much denser compared to other real networks [20]. Similar shapes of node degree distributions were found and declared to be robust for a few other analogous empirical networks [17, 18]. Negative value

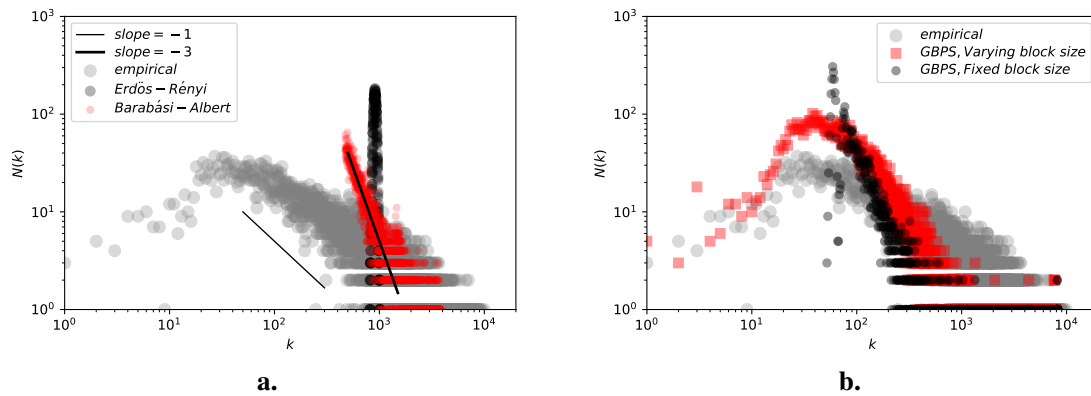


Figure 1. Node degree histograms $N(k)$ of networks of concepts addressed in our study. An empirically observed network (grey discs) is compared with those generated by Erdős-Rényi and Barabási-Albert models (panel **a**, black and red discs, correspondingly) and growth by blocks with preferential selection, GBPS (panel **b**, black discs: fixed block size, red squares: varying block size).

of the assortative mixing by degrees $r = -0.28$, defined as Pearson correlation coefficient between node degrees on both ends of links in the network, indicates that in the network of concepts, the high-degree nodes attract low-degree ones of a high extent. The presence of connectivity patterns is featured by comparatively high values of the mean clustering coefficient $\langle c \rangle$ and global transitivity T (cf. $\langle c \rangle = T = 1$ for the complete graph and $\langle c \rangle = T = 0$ for a tree). For a node i of degree $k_i > 1$, the clustering coefficient is a ratio of existing links m_i between its neighbouring nodes to all possible connections between them, $c_i = 2m_i[k_i(k_i - 1)]^{-1}$. In turn, the global transitivity T is defined as a ratio between the number of closed triplets in the network and the total number of network triplets [28]. The difference between the two values, $\langle c \rangle$ and T , indicates specific topological features of the network. With quantitative measures of basic network features at hand, let us proceed with modelling a growth process that results in network topology similar to the empirically observed one.

We start with the Erdős-Rényi random graph [29] and Barabási-Albert preferential attachment [30] models. Both models allow us to generate uncorrelated networks with the same number of nodes N and links L as the empirical one. Therefore, the density of links ρ and the average node degree $\langle k \rangle$ coincide too. The discrepancies become evident with more in-depth analysis. Results of the network characteristics calculated for an ensemble average over 100 realizations for each model are shown in the 2nd and 3rd lines of table 1. The Erdős-Rényi random graph is much more homogeneous than the empirical network: the standard deviation σ is almost 40 times smaller than that for the empirical concept network, the maximal node degree k_{\max} exceeds its average value $\langle k \rangle$ by 12% only. This may be observed in figure 1a, where the corresponding histogram $N(k)$ is shown by black discs. The Barabási-Albert model, that has growth and preferential attachment as key ingredients, better reproduces empirical network node degree heterogeneity: k_{\max} exceeds $\langle k \rangle$ by more than 300%, σ almost 20 times exceeds its value for Erdős-Rényi graph. However, the decay of $N(k)$ is much faster than in the empirical network (see the red discs in figure 1a and the solid line that corresponds to $N(k) \sim k^{-\gamma}$ with the Barabási-Albert model decay exponent $\gamma = 3$ [30]). The discrepancies are even more pronounced when one considers connectivity patterns between nodes of different degrees. Similar to the Erdős-Rényi graph, the Barabási-Albert network is neither assortative nor disassortative, indicating the feature of the empirical network of concepts that cannot be captured by the models. The other feature that is not captured by the models is the difference between the average clustering coefficient $\langle c \rangle$ and the global transitivity T , even though the values for the Barabási-Albert model are closer to those for the empirical network than the ones for the Erdős-Rényi network.

To understand the possible mechanisms that lead to the concept network under consideration, let us develop a model that is capable of reproducing its empirically observed features. Doing so, we do not put as a primary goal reaching a high precision of reproducing the given set of metrics. Rather we are

interested in a qualitative description of the main tendencies in network structure and their explanation by network generation mechanisms. The model of the network evolution that we suggest is based on the simultaneous account of two factors: growth by blocks and preferential selection. Consider a process with discrete time $t = 1 \dots \mathcal{N}$. At each time step, a new article A_t that contains a block of n_t concepts is generated. It joins the concept network as a complete graph of n_t nodes. The article generation consists of two steps: (i) drawing the block size n_t and (ii) selecting particular concepts to populate the block. Below, we choose an option when n_t is drawn from the actual distribution of the number of concepts per article in the empirical data set while other options are discussed in [20]. Let us explain step (ii) more in detail. When a new article A_t is generated at time $t > 1$, the already existing data set consists of a set of $t - 1$ articles \mathbb{A}_{t-1} and a set of N_{t-1} different concepts \mathbb{C}_{t-1} . The new article A_t may contain some of the above N_{t-1} concepts as well as the *novel* concepts that are introduced for the first time. Within our model, we fix the probability of the i -th concept of article A_t to be a novel one, $\pi_{t,i}^{\text{novel}} = \nu$. Consequently, with probability $1 - \nu$ a concept of the generated article is one of the already existing N_{t-1} concepts. Moreover, let us consider that the already existing concepts have different chances to be selected to populate an article: the more popular the concept is (among the first $t - 1$ articles), it is more likely to be selected to populate the t -th one. We call such a process a *preferential selection*. The probability $\pi_{t,i}^{\text{exist}}(C_j)$ for the concept C_j to be selected is proportional to the number of articles $N_{t-1}(C_j)$ in which the concept C_j has appeared:

$$\pi_{t,i}^{\text{exist}}(C_j) = \frac{(1 - \nu)N_{t-1}(C_j)}{\sum_l N_{t-1}(C_l)}, \quad C_j \in \mathbb{C}_{t \setminus i-1}, \quad (1)$$

where $\mathbb{C}_{t \setminus i-1}$ is the subset of concepts \mathbb{C}_{t-1} excluding $i - 1$ concepts selected for article A_t and the denominator sums the number of times each concept C_l from the set $\mathbb{C}_{t \setminus i-1}$ has appeared in all articles.

By the above described evolution mechanism, the concept network grows by adding cliques to the existing graph. At each time t , once a new article A_t of n_t concepts is generated, it enters the concept network as a complete graph of n_t nodes and $n_t(n_t - 1)/2$ links. Thus, during its evolution, the following processes may be observed in a generated concept network: (i) addition of new nodes, (ii) appearance of links between new nodes and between new and already existing nodes, (iii) appearance of new links between previously unconnected existing nodes, which is important for generation of dense networks. We compare the main features of the network of concepts generated by the growth by blocks with preferential selection mechanism in the last line of table 1. As for the two previously described models, we display the values averaged over an ensemble of 100 network realizations. The number of articles generated in our simulations was set to be exactly the same as the number of articles ($\mathcal{N} = 36,386$) in the empirical data set. Fixing the number of articles does not ensure that the generated network will have the same number of nodes (concepts). The remaining free parameter of the model has been chosen $\nu = 8.8 \cdot 10^{-3}$ to give a reasonable value of the number of concepts N , see [20] where other concept selection mechanisms were considered. As one can see from the table, now the modeled network of concepts possesses two features that Erdős-Rényi and Barabási-Albert models failed to reproduce: it is disassortative ($r < 0$) and its mean clustering coefficient and global transitivity differ from each other. The fact that the growth by blocks and preferential selection mechanism correctly grasp the main features of the network of concepts is further supported by the form of the node degree histogram, as shown by red squares in figure 1b. Now one observes characteristic decays in the regions of small and large values of k . Black discs in the plot show an outcome of the modified model, when each block of concepts has a fixed size [20] that leads to an obvious sharp lower bond.

In the forthcoming publication [20] we will give a more detailed account of the suggested network evolution mechanism along with the analysis of its various modifications. Several conclusions are at hand to finalize this brief report. First of all, one should not go too far in trying to reach a one-to-one mapping between the features of the empirically observed network of concepts and the modeled one. Indeed, the model which selects new concepts at random, completely neglects their content-related characteristics. Rather, the goal is to reveal which processes in the network evolution are relevant to its generic features. As we show in this report, these are the growth by blocks and preferential selection. Moreover, our analysis shows that the observed network structure emerges as a synergetic effect of both of these factors: each of them alone does not lead to a satisfactory picture. The model suggested in this paper may be also of relevance in analysing the generating mechanisms for dense networks which are the subject of

ongoing interest [24–27].

This work was supported in part by the National Academy of Sciences of Ukraine, project KPKBK 6541030 (O.M. & Yu.H) and by the National Research Foundation of Ukraine, project 2020.01/0338 (M.K.).

References

1. Sowa J. F., *Semantic Networks*, In: *Encyclopedia of Cognitive Science*, Wiley, 2006, doi:10.1002/0470018860.s00065.
2. Da Fontoura Costa L., *Phys. Rev. E*, 2006, **74**, 026103, doi:10.1103/PhysRevE.74.026103.
3. Van Eck N. J., Waltman L., *Scientometrics*, 2010, **84**, No. 2, 523–538, doi:10.1007/s11192-009-0146-3.
4. Zeng A., Shen Z., Zhou J., Wu J., Fan Y., Wang Y., Stanley H. E., *Phys. Rep.*, 2017, **714-715**, 1–73, doi:10.1016/j.physrep.2017.10.001.
5. Barabási A.-L., Wang D., *The Science of Science*, Cambridge University Press, Cambridge, 2021, 308, doi:10.1017/9781108610834.
6. Krenn M., Zeilinger A., *Proc. Natl. Acad. Sci. USA*, 2020, **117**, No. 4, 1910–1916, doi:10.1073/pnas.1914370116.
7. Herrera M., Roberts D. C., Gulbahce N., *PLoS ONE*, 2010, **5**, No. 5, e10355, doi:10.1371/journal.pone.0010355.
8. Pan R. K., Sinha S., Kaski K., Saramäki J., *Sci. Rep.*, 2012, **2**, 551, doi:10.1038/srep00551.
9. Foster J. G., Rzhetsky A., Evans J., *Am. Sociol. Rev.*, 2015, **80**, No. 5, 875–908, doi:10.1177/0003122415601618.
10. Beam E., Appelbaum L. G., Jack J., Moody J., Huettel S. A., *J. Cognit. Neurosci.*, 2014, **26**, No. 9, 1949–1965, doi:10.1162/jocn_a_00604.
11. Silva F. N., Travençolo B., Viana M. P., Costa L. F., *J. Phys. A*, 2010, **43**, No. 32, 325202, doi:10.1088/1751-8113/43/32/325202.
12. Iacopini I., Milojević S., Latora V., *Phys. Rev. Lett.*, 2018, **120**, No. 4, 048301, doi:10.1103/PhysRevLett.120.048301.
13. Uzzi B., Mukherjee S., Stringer M., Jones B., *Science*, 2013, **342**, No. 6157, 468–472, doi:10.1126/science.1240474.
14. Brodiuk S., Palchykov V., Holovatch Yu., In: *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, 2020, 366–371, doi:10.1109/DSMP47368.2020.9204220.
15. Thurner S., *43 Visions for Complexity*. World Scientific, Singapore, 2017, doi:10.1142/10360.
16. Holovatch Yu., Kenna R., Thurner S., *Eur. J. Phys.*, 2017, **38**, 023002, doi:10.1088/1361-6404/aa5a87.
17. Cattuto C., Barrat A., Baldassarri A., Schehr G., Loreto V., *Proc. Natl. Acad. Sci. USA*, 2009, **106**, No. 26, 10511–10515, doi:10.1073/pnas.0901136106.
18. Rzhetsky A., Foster J. G., Foster I. T., Evans J. A., *Proc. Natl. Acad. Sci. USA*, 2015, **112**, No. 47, 14569–14574, doi:10.1073/pnas.1509757112.
19. arXiv, URL <https://arxiv.org>.
20. Palchykov V., Krasnytska M., Mryglod O., Holovatch Yu., 2021, (in preparation).
21. ScienceWISE, [Online; accessed 20-Jun-2020], URL <http://sciencewise.info>.
22. Palchykov V., Gemmetto V., Boyarsky A., Garlaschelli D., *EPJ Data Sci.*, 2016, **5**, 28, doi:10.1140/epjds/s13688-016-0090-4.
23. Palchykov V., Holovatch Yu., In: *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, 2018, 84–87, doi:10.1109/DSMP.2018.8478505.
24. Seyed-Allaei H., Bianconi G., Marsili M., *Phys. Rev. E*, 2006, **73**, No. 4, 046113, doi:10.1103/PhysRevE.73.046113.
25. Bonifazi P., Goldin M., Picardo M. A., Jorquera I., Cattani A., Bianconi G., Represa A., Ben-Ari Y., Cossart R., *Science*, 2009, **326**, No. 5958, 1419–1424, doi:10.1126/science.1175509.
26. Zhou T., Medo M., Cimini G., Zhang Z.-K., Zhang Y.-C., *PLoS ONE*, 2011, **6**, 1–7, doi:10.1371/journal.pone.0020648.
27. Courtney O. T., Bianconi G., *Phys. Rev. E*, 2018, **97**, No. 5, 052303, doi:10.1103/PhysRevE.97.052303.
28. Luce R. D., Perry A. D., *Psychometrika*, 1949, **14**, No. 2, 95–116, doi:10.1007/BF02289146.
29. Erdős P., Rényi A., *Publ. Math. Debrecen*, 1959, **6**, 290–297, https://www.renyi.hu/~p_erdos/1959-11.pdf.
30. Barabási A.-L., Albert R., *Science*, 1999, **286**, No. 5439, 509–512, doi:10.1126/science.286.5439.509.

Механізм еволюції мережі фізичних концепцій

В. Пальчиков², М. Красницька^{1,2}, О. Мриглод^{1,2}, Ю. Головач^{1,2,3}

¹ Інститут фізики конденсованих систем НАН України, вул. Свенціцького, 1, 79011 Львів, Україна

² Співпраця \mathbb{L}^4 і Коледж докторантів "Статистична фізика складних систем",
Ляйпціг–Лотарингія–Львів–Ковентрі, Європа

³ Центр плинних та складних систем, Університет Ковентрі, Ковентрі, CV1 5FB, Велика Британія

Ми пропонуємо механізм, що визначає зростання мережі концепцій — складної мережі, що відображає взаємозв'язки між різними науковими концепціями, базуючись на даних про їх співпояву у публікаціях. З цією метою, ми виконуємо емпіричний аналіз мережі концепцій, оснований на препринтах з фізики, завантажених на сервер arXiv.org. Ми розраховуємо мережеві характеристики та показуємо, що вони не можуть бути отримані за допомогою кількох простих загальноновживаних моделей зростання мереж. В свою чергу, ми пропонуємо одночасне врахування двох факторів: зростання блоками та переважний вибір, — що дають пояснення емпірично отриманих характеристик мережі концепцій. Спостережувана структура виникає внаслідок синергетичного ефекту обох цих факторів — врахування кожного з них окремо не дає задовільної картини.

Ключові слова: *складні системи, складні мережі, мережа концепцій, еволюційна модель, переважне приєднання*