Ukrainian Catholic University

Master Thesis

Concept embedding and network analysis of scientific innovations emergence

Author:
Serhii Brodiuk

Supervisor:
PhD. Vasyl Palchykov
Co-supervisor:
Prof. Dr. Yurij Holovatch

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2020

Declaration of Authorship

I, Serhii BRODIUK, declare that this thesis titled, “Concept embedding and network analysis of scientific innovations emergence” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Concept embedding and network analysis of scientific innovations emergence

by Serhii BRODIUK

Abstract

Novelty is an inherent part of innovations and discoveries. Such processes may be considered as the appearance of new ideas or as the emergence of atypical connections between existing ones. The importance of such connections hints for investigation of innovations through network or graph representation in the space of ideas. In such representation, a graph node corresponds to the relevant notion (idea), whereas an edge between two nodes means that the corresponding notions have been used in a common context. The question addressed in this research is the possibility to identify the edges between existing concepts where the innovations may emerge.

To this end, a well-documented scientific knowledge landscape has been used. Namely, we downloaded 1.2M arXiv.org manuscripts dated starting from April 2007 and until September 2019; and extracted relevant concepts for them using ScienceWISE.info platform. Combining approaches developed in complex networks science and graph embedding the predictability of edges (links) on the scientific knowledge landscape where the innovations may appear is investigated. We argue that the conclusions drawn from this analysis may be used not only to the scientific knowledge analysis but are rather generic and may be applied to any domain that involves creativity within.

Acknowledgements

I want to thank my family for supporting me all the time. The impact of their side is significant, and I appreciate it a lot.

Also, I would like to thank Oleksii Molchanovskyi for his efforts in the Data Science program management as well as Ukrainian Catholic University for organizing and hosting a variety of exciting events.

Finally, the foremost gratitude goes out to my research supervisors. I owe a considerable piece of thanks to Vasyl Palchykov, who, I bet, gave me more than usually supervisors provide. Special thanks to Yurij Holovatch for sharing his experience in the Complex Networks/System domain. With their encouraging constantly, this thesis has a much more presentable shape.

Contents

To my patient and loving and lovely family.

Chapter 1

Introduction

An idea of scientific analysis of science is not new. It is at least as old as the science itself, see, e. g. [1] and references therein. Contemporary studies in this domain share a common specific feature: besides traditional philosophical and culturological context, such analysis attains quantitative character. The questions of interest cover a wide spectrum, ranging from fundamental, such as: what is the structure of science? How do its constituents interact? How does knowledge propagate? [2, 3] to entirely practical ones: which fields of science deserve financial investments or how to rate scientists in a particular domain? [4, 5]. All these and many more questions constitute a subject of a science of science or logology [6].

The problem we consider in this thesis concerns an emergence of new scientific knowledge or the so-called scientific innovation. Quantitative investigation and modeling of innovations are not straightforward. On the one hand, one may think of innovation as an emergence of a new idea, see, e. g. [7]. Another approach considers innovation as an atypical combination of existing ideas, see, e. g. [8]. The goal of our work is to suggest a way to quantify analysis of scientific innovations emergence; and to propose an approach to identify edges on the graph of knowledge where innovations may emerge. We believe that such analysis, if successful, is useful both from the fundamental point of view, explaining properties of knowledge formation, as well as is of practical relevance, helping to detect innovation-rich fields.

To reach this goal, we will analyze a body of scientific publications (we take an `arXiv` repository of research papers [9]) and analyze its dynamics with a span of time. We will use a specially tailored software, ScienceWISE.info platform [10], to extract a set of concepts from all publications on an annual basis. These are the properties of this set of concepts that will serve us as a proxy of structural features and dynamics of human knowledge. In particular, we will use complex network theory [11–14] to track intrinsic connections between concepts that are contained in different papers. Using several completing each other approaches we will construct a complex network of concepts (as a proxy of a complex network of knowledge) and we will calculate its main topological characteristics, paying particular attention to the emergence of new links between existing concepts. This last may serve as a signal about the emergence of atypical combinations between existing ideas, i. e. about scientific innovations. We will refine our analysis by exploiting embedding technique [15, 16] to quantify a proximity measure between different concepts and in this way, we will establish a solid and falsifiable procedure to quantify an emergence of possible scientific innovations in certain fields of science.

The set up of the thesis is as follows. In the next Chapter 2, we establish a background behind this investigation by describing the dataset used in the analysis and by reviewing relevant literature. In Chapter 3, we represent the dataset as a network and investigate its topological properties. In the following Chapter 4, we introduce the concept embedding technique; and investigate the dynamics of link appearance and describes an approach that may be used to detect the edges where scientific innovations may emerge. The results are summarized in the last Chapter 5 that contains conclusions and outlooks.

Chapter 2

Review

In this chapter, we set up a background behind the performed investigation. We start with the data needed to perform the analysis, the way the data have been collected, and performed a quick overview of the methods used to perform the analysis.

2.1 Source of data: arXiv

To be able to answer the question of interest of the thesis, we are required to have at hands a well-documented collection that represents scientific knowledge, at least in a single domain of science. E-repository of preprints `arXiv.org` [9] is a good candidate for such source of data: at the moment of writing this thesis, there are about 1.6M manuscripts uploaded to the `arXiv`. Besides title, abstract, a list of authors `arXiv` allows full-text access to all manuscripts. Such full-text access enables one to extract scientific ideas/concepts from the text of manuscripts.

`arXiv` covers a variety of scientific fields such as physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. The average daily upload rate is 400 $\approx$ 12.5K new manuscripts per month (every next year there are $\approx$ 5000 more articles than the previous one, starting in 1991 - numbers are increasing, see Figure 2.1)

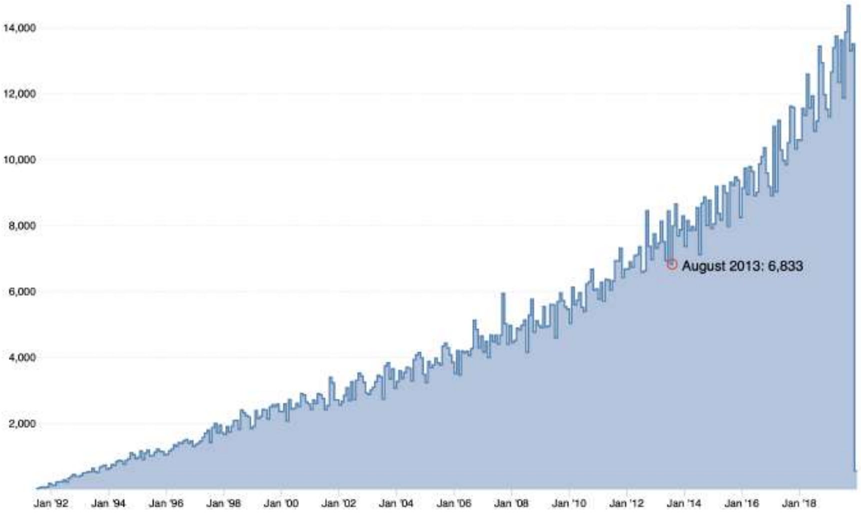

FIGURE 2.1: Monthly Submission Rates to `arXiv.org` for more than 25 year period: from 1991 till 2020. The number of submissions tends to increase continuously. The total number of submitted manuscripts is about 1.6M. The majority of the manuscripts have been assigned Physics as their primary category. Source (as on Jan 2020): arxiv.org/stats/monthly_submissions

Not every category is represented equally in the e-repository. arXiv has been initially designed to serve a community in a specific domain of physics (theoretical high-energy physics) but has grown significantly. Nowadays, the majority of papers from the physics domain are initially submitted to arXiv, but in the other domains, the coverage may be significantly smaller. Indeed, the combination of significant coverage of research publications and full-text access to the manuscript makes arXiv a proper source of data for our purpose.

As we have noted above, each paper submitted to the arXiv contains, besides the full-text, different metadata such as authors, subject category (categories), journal reference, and DOI (Digital Object Identifier, if any), submissions history with dates, etc. For the purpose of our study, we need to extract from all papers their main words, key terms, called concepts, that to some extent represent the content of the paper, both with respect to the subject of research and methods applied. To this end, we will use a ScienceWISE platform, specially tailored for such tasks. We describe it in more detail in the next Section 2.2.

2.2 Concept extraction: ScienceWISE.info platform

To extract specific words or combination of words that carry a specific scientific meaning (concepts) from each manuscript we use ScienceWISE.info platform [10, 17]. The platform has been built to support the daily activities of research scientists. The goal of the platform is to ”understand” the interests of its users and to recommend them relevant newly submitted manuscripts. For this purpose, arXiv serves as one of the datasource of new submissions. In order to understand the research interests of the users, the platform extracts scientific concepts from the texts of the manuscripts and compares the concept vector of the manuscript and the corresponding concept vector of the user's research interest.

Concept extraction approach implemented into this platform has two phases: i) automatic key phrase (concept candidate) extraction and ii) crowd-sourced validations of scientific concepts. During the first phase, each manuscript is scanned by the KPEX algorithm [18]. The algorithm extracts key phrases from the text of the manuscript, and these key phrases serve as concept candidates. Then, during the second step, the concept candidates are reviewed by the registered users of the platform who are permitted to validate the concepts. The described procedure arrived at approximately 20,000 concepts as of the date when this thesis was written. About 500 of them have been marked as generic concepts assuming their generic meaning (the ones like Energy, Mass or Temperature).

A user of the platform is allowed to navigate over all identified concepts within each manuscript of the platform. To get the concepts for an arXiv manuscript a user may find this manuscript on the arXiv.org webpage and then press on the ScienceWISE icon in the bookmark menu. The user will be redirected to the webpage on ScienceWISE.info portal that corresponds to the given arXiv manuscript. This page, besides metadata (title, abstract, authors, assigned categories, and subcategories), contains a set of concepts identified within the manuscript together with the usage count. Moreover, if the concept has been marked as generic in the dataset, the formatting allows us to unveil this information. In this thesis, we will make use of this portal to extract available arXiv manuscript and the concepts extracted within. The details of the resulting dataset follow below.

2.3 Dataset

Navigating over ScienceWISE.info platform at the end of September of 2019, we accessed a collection with near 1.2M arXiv manuscripts with metadata and concepts list for each one (from April of 2007 till September of 2019). As data is publicly available (anyone with access to the internet could get it), we scraped it to storage on our side with a convenient structure for further manipulations. A detailed data parsing approach could be found at GitHub repository [1]. Similar dataset (can be considered as a small subset of the described above) of 36386 Physics domain articles have been previously investigated in [19–21].

Once the dataset is downloaded and prepared for the analysis, the first step of our investigation is to analyze the topological properties of the resulting concept network using the tools of Complex network theory. The next Section 2.4 describes the basis of this theory.

2.4 Complex networks

Complex network theory, see e.g. [11, 12, 22–25], had evolved from the Graph theory into a new field in the late 1990s when the WWW allowed access to the networks that have not been available before and the computational resources allowed to analyze data of the size that was impossible beforehand. As a result of such analysis it has been found that properties of many networks like WWW, Internet, Public transport networks, Power grids, etc. can not be explained by the existing models in graph theory, especially by Erdös-Rényi random graph [2]. Among these properties, one may highlight scale-free feature (a special functional form of the node degree distribution, defined in the following Chapter 3), the small-world property, tolerance to random failures, and vulnerability to targeted attacks.

To understand the reasons behind such properties, network theory adopted methodology from Complex System theory [26] and proposed a number of generative models. Complex systems denote systems that are composed of many interacting parts, often called agents, which display collective behavior that does not follow trivially from the behaviors of the individual parts, see, e. g. [26]. Among the models developed within network theory, one may mention Barabási-Albert model [3] that can reproduce scale-free networks and Watts-Strogatz model [4] of small-world network.

In the first step of our analysis, we will represent a set of the manuscript and related concepts as a network and investigate its topological features. The details of the network representation and features to be analyzed will be found in Chapter 3. After performing an analysis of concept network, where the link between two concepts indicates the relation between them and the weight of the link proxy the strength of such connection, we will employ an alternative measure of the proximity between scientific concepts, taken from embedding techniques.

[1] github.com/sergibro/concept-graphs

[2] en.wikipedia.org/wiki/Erdos-Renyi_model

[3] en.wikipedia.org/wiki/Barabási-Albert_model

[4] en.wikipedia.org/wiki/Watts-Strogatz_model

2.5 Graph (concept) embeddings

When we talk about embedding, we mean a set of tools and techniques that allow to embed members of multidimensional space into another space that usually consists of a smaller number of dimensions, assuming that these new ”aggregate” dimensions contain proper features combinations of old dimensions. Such dimensionality reduction techniques include Singular vector decomposition (SVD) and Principal component analysis (PCA).

To estimate the similarity between different concepts/nodes in our analysis, we have decided to use the word2vec approach, see [15]. It has been specifically designed to deal with texts and may be easily modified to used concepts instead of words. Then, the list of concepts for articles will form ”sentences”. However, if we have the graph representation, other approaches may be applied, which deal with graphs directly, e. g. node2vec [27]. More specifically, we have selected PyTorch-BigGraph tool implementation [16], which is available from the second quarter of 2019. It provides all necessary functionality to cover the part about building concept embedding. Moreover, its implementation allows to use multi-core processors, which implies high computational speed. This is essential with extended graph sizes.

Once concepts have been embedded into multidimensional space, we may measure the proximity between them using, e. g. cosine similarity between the corresponding vectors. The links between concepts within the network representation of dataset and context similarity between the corresponding concepts are two alternative measures of the proximity between the considered pairs of concepts. Below we will investigate how do the two measures affect the predictability of the emergence of a connection between pairs of concepts.

Chapter 3

Topological features of concept networks

In this chapter, we will determine the main topological characteristics of the network of concepts extracted from `arXiv` for the years 2013 and 2015. We will be interested in general network characteristics such as node degree distribution, shortest path length, clustering coefficient, global transitivity, etc. Besides, we will define the weights of graph links. They will serve us as a proxy for the link importance.

The chapter is organized as follows: In the first Section 3.1, we explain how to represent sets of concepts extracted from different manuscripts in the form of a complex network. Section 3.2 is about our approaches to chose statistically significant links. Subsequently, in Section 3.3, we give definitions of typical network characteristics and determine these characteristics for the complex networks of concepts.

3.1 Network construction

As mentioned in Chapter 2 by navigating over ScienceWISE.info platform we were able to collect data for about 1.2M manuscripts submitted to `arXiv` between 2007 and 2019. For each manuscript, a set of concepts found within its text has been recorded. The total number of unique extracted concepts is 19,446, and the number of concepts per manuscript varies in range 0-1164. It has a bimodal distribution with two maxima located around 4 and 32 concepts per article, see Figure 3.1 for details.

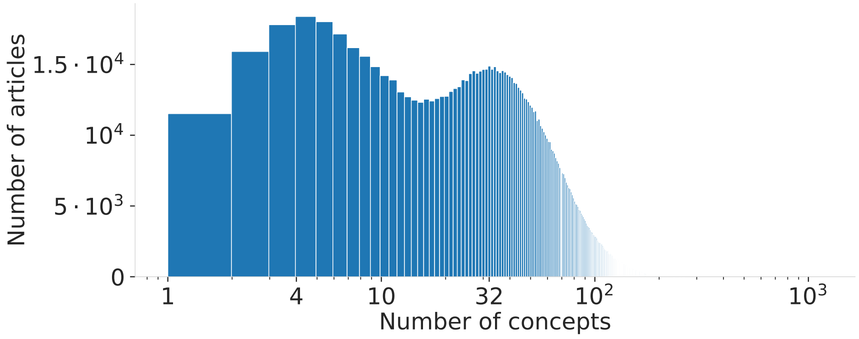

FIGURE 3.1: The number of concepts identified within articles. The distribution is bimodal indicating two maxima: around 4 and around 32 concepts identified within the texts of the manuscripts.

While the maxima around 32 concepts per article may really indicate mode for the number of concepts identified within the text, 4 concepts per article are a quite small number and may highlight parsing issues. In particular, problems with parsing certain pdf documents. Some articles may contain as many as 1000 concepts.

These may refer to review articles: besides being long (number of pages for review articles is usually essentially larger than for the research ones), these articles should cover a number of research questions that lead to a wide vocabulary of concepts used. Indeed, the article for which the highest number of concepts (1164) have been identified in our dataset is *Astrophysics in 2006* by V. Trimble, M. J. Aschwanden, and C. J. Hansen [1]. This is a review article, as supposed many articles with a lot of concepts.

3.1.1 Bipartite network

The above described dataset may be naturally represented as a bipartite network: the network that consists of the nodes of two types, whereas links connect the nodes of different types only. In our case, the two types of nodes represent manuscripts and scientific concepts correspondingly. A link between a manuscript-node and concept-node exists if the concept has been found within the text of the corresponding manuscript. The illustration of the bipartite network reconstruction is shown in panels **a, b** of Figure 3.2.

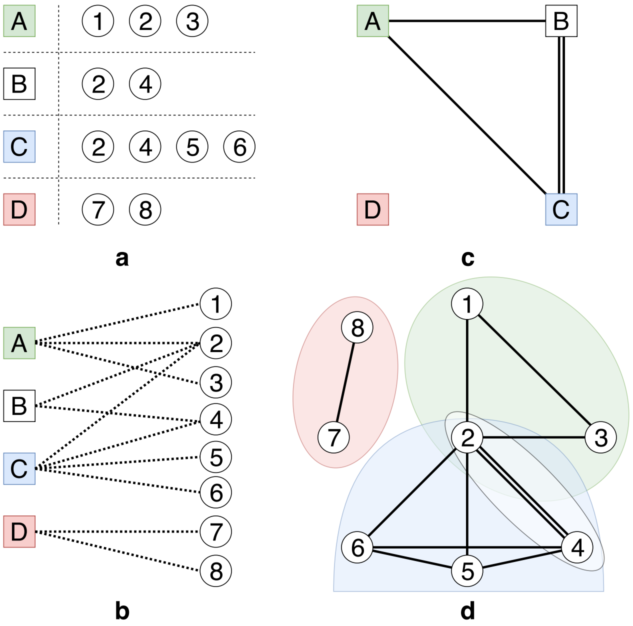

FIGURE 3.2: Illustration of the dataset and three network representations constructed from it. Squared nodes represent manuscripts and circles represent concepts. Panel **a** illustrates a dataset of four manuscripts and the concepts identified each of them. Panel **b** is a bipartite network representation of the dataset. A link connects a square and a circle if the corresponding concept has been identified within the text of the manuscript. Panels **c** and **d** illustrate single-mode projections of the bipartite network to the manuscript and concept spaces, correspondingly. The nodes on a panel **c** represent manuscripts that are connected to each other if the corresponding manuscripts share common concepts. Nodes of panel **d** represent scientific concepts. Two concepts are connected to each other if they have been identified within the text of at least one common manuscript.

[1] V. Trimble, M. J. Aschwanden, and C. J. Hansen, *Astrophysics in 2006*, Space Science Reviews, **132**, 1-182 (2007)

Figure 3.2 also contains two complementary network representations of the dataset that are the single-mode projections of the described above the bipartite network. Manuscript-to-manuscript network (see Figure 3.2 **c**) consists of nodes that represent manuscripts. The link connects two nodes if the corresponding manuscripts share at least one concept in common. This representation has been previously analyzed in [19], where its single year slices have been investigated. The analysis, in particular, highlighted the power of network representation of the data, especially by investigating meaningful discrepancies between the resulting clustering structure of the network and author-made categorical classification of the manuscripts.

3.1.2 Concept network

Another projection is a projection of the bipartite network to the concept space, see Figure 3.2 **d**. In this representation, the nodes correspond to scientific concepts. A link connects two nodes if the corresponding concepts have appeared together in the text of at least one manuscript. Below we will be investigating topological features of this network.

To proceed with the analysis, we will consider two slices of data. In the first slice, we consider the manuscripts submitted during the 2013 year only, while in the second subset, the manuscripts submitted only during the year 2015 will be considered. Such a set-up will allow us, in particular, to compare some properties of a concept network as they evolve in time. Thus the first subset (the year 2013) network consists of 16,229 nodes, which is a naturally smaller number than the number of concepts identified within the entire dataset covering 13 years of manuscript submissions. Similarly, for the year 2015 we arrived at 16,660 nodes. 15,431 concept-nodes are shared between representations across two subsets as their intersection.

Below we will consider topological features of both concept networks. However, before starting such analysis, let us mention that topological features considered below will not include information about the strength of each link. We will include this information by considering two additional networks that will be constructed from concept networks. Both these extra networks are subsets of the concept network. The idea behind is that in the network, some links may be considered as statistically significant, while the others may be considered as insignificant. Below we will use two approaches to distinguish between significant and insignificant links and will keep in the additional network only ”significant links”.

3.2 Choice of statistically significant links

3.2.1 Filtering by weight

As mentioned above, a link between two concept-nodes exists if the corresponding concepts have been found together in the texts of at least one manuscript. Let us assign a weight to the link w_{ij} such that it equals the number of manuscripts in the dataset that contains concepts i and j simultaneously. Then the simplest way to filter out insignificant links is to consider the hard threshold on the link weight. Below we will consider threshold value $\omega = 10$ and keep the links for which $w_{ij} > \omega$ ($\approx$ 16.5% of total links remain). If such procedure removes all links from a node, we will remove this node as well, so there are no isolated nodes in the network. The reduced number of links will be seen in the tables that summarize network topology characteristics below.

We are aware of the fact that such filtering approach is not the best possible way of identifying significant links, but it will serve us as a starting point of network filtering. A more sophisticated approach for identifying significant links is to consider the disparity filter proposed in [28].

3.2.2 Filtering by disparity

The idea behind the disparity filter is to take into account network inhomogeneity. Indeed different nodes play different roles in the network: they have a different number of links connected to them, these links have different weights, and it is not fair to judge on the link significance taking into account link weight only. For example, if node i has only one connection to node j and node j has only one connection to node i and the weight of the corresponding connection $w_{ij} = 5$, it may be still significant. However, a link between i and j with $w_{ij} = 10$ connecting two nodes i and j with $k_i = k_j = 100$ connections each (below we will call the number of links connected to node i as its degree k_i) may be insignificant given the other links to nodes i and j have significantly higher weights.

The key idea of the method is to calculate the probability that a given link has as many connections as observed or more in a random setting. In this random setting sum $s_i = \sum_j w_{ij}$ of weights of all links connected to node i is fixed. s_i will be referred below as strength of the node i. However, this total weight s_i is distributed randomly among all k_i links (k_i is fixed) connected to node i. This probability, known as p-value reads

$$\alpha_{ij} = 1 - (k_i - 1) \int_0^{p_{ij}} (1-x)^{k_i - 2} dx \tag{3.1}$$

where k_i is degree of i-th node, and $p_{ij} = w_{ij}/s_i$ is the normalized link weight. Then one may set a threshold for p-value, and only the links will small enough p-value are kept, meaning that a random process can not arrive at a link with such weight.

Note that the link between i and j may be considered significant if we look at it from node i and insignificant if we stand at node j. In our analysis we set a threshold for p-value $\rho = 0.1$, and keep a link between i and j if $\alpha_{ij} < \rho$ or $\alpha_{ji} < \rho$, i. e. if it is significant from at least on node standpoint. As a result of such procedure, $\approx 85\%$ of links will be removed as insignificant. The effect of removing these links on the other network characteristics will be considered below and summarized in Table 3.1.

In the next Section 3.3, we define a set of topological properties that will be calculated to characterize the topology of concept networks.

3.3 Network characteristics

In this Section, we will provide definitions for the main characteristics of a network. Once these characteristics are calculated for the network of concepts, we will be able to compare the topology of the network of concepts with the topology of the other real-world networks and to say about the evolution of the network of concepts over time (given different snapshots of the network over time).

Each network consists of nodes connected by links. The size of the network may refer either to the number of nodes N or to the number of links L in the entire network. Each node i is characterized by its degree k_i defined as the number of links connected to it. While k_i describes individual nodes, its average value $\langle k \rangle$ is a global characteristic defined as

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{3.2}$$

where i runs over all N nodes in the network. Average node degree $\langle k \rangle$ and their maximal value $k_{\max}$ for all networks considered in this thesis are summarized in Table 3.1.

network	N	$L, \times 10^6$	$\rho, \%$	$\langle k \rangle$	$k_{\max}$	l	$l_{\max}$	$\langle c \rangle$	C	r
g-2013	16,229	11.1	8.46	1,373	15,345	1.92	3	0.77	0.37	-0.324
g-2015	16,660	12.7	9.12	1,520	15,935	1.91	4	0.77	0.38	-0.325
w-2013	9,999	1.8	3.69	369	8,856	2.00	4	0.89	0.28	-0.390
w-2015	10,770	2.2	3.84	414	9,661	2.00	4	0.89	0.28	-0.382
d-2013	13,358	1.6	1.84	246	11,665	2.01	4	0.90	0.14	-0.375
d-2015	13,969	1.9	1.92	268	12,367	2.00	5	0.89	0.14	-0.368

TABLE 3.1: Aggregated characteristics of concepts networks constructed by several criteria: g-2013 and g-2015 denote entire network constructed from articles submitted to arXiv during years 2013 and 2015, correspondingly. w-2013 and w-2015 denote subgraphs of the above networks where link of weight $w_{ij} \leq \omega = 10$ have been removed and, afterwards, isolated nodes have been removed as well. d-2013 and d-2015 denote the subgraphs of g-2013 and g-2015 networks, correspondingly, where links with p-value $\alpha_{ij} \geq \rho = 0.1$ have been removed, subsequently removing isolated nodes. The table contains the number of nodes N, the number of links L, density of links $\rho = 2L/N(N-1)$, an average $\langle k \rangle$ and maximal $k_{\max}$ node degrees, average l and maximal $l_{\max}$ shortest path lengths, average clustering coefficient $\langle c \rangle$, global transitivity C, and assortativity mixing by degree r.

The table also contains the number of nodes N, the number of links L and density of links ρ, i. e. the ratio $2L/N(N-1)$ between the number L existing links in the network and the total possible number of links $N(N-1)/2$. Original networks (before applying any filtering techniques) are denoted by a g- suffix followed by the manuscript submission year. For example, g-2013 network indicates the concept network constructed from a set of all manuscripts submitted to arXiv during the year 2013. Prefix w- indicates that the concept has been generated from the corresponding g- network by removing links whose weight $w_{ij} > \omega$. After removing the low-weight link, some nodes could become isolated, i. e. without any links connected to them. Such nodes will be removed from the network too. Similarly, the network denoted by name with prefix d- indicate concept networks constructed from the g- network by removing the links for which p-value $\alpha_{ij} \geq \rho$. Subsequently, the resulting isolated nodes have been removed.

Table 3.1 indicates that the number of nodes (concepts) for original networks constructed for years 2013 and 2015 is similar (16,229 versus 16,660). The number of links L and density of links ρ is a bit higher for the network constructed for the year 2015. After applying filtering techniques, about $15\% - 35\%$ of nodes become isolated and are removed from the filtered networks. The density of links decreases in about 2 - 4 times once the networks are filtered ($2\% - 4\%$ in filtered networks vs approximately 9% in the original ones).

Both original (not filtered) network are characterized by a similar value of an average node degree $\langle k \rangle \approx 1,500$. In both networks, $\langle k \rangle$ is about 10 times smaller than the maximal node degree $k_{\max}$ observed in the corresponding networks. Taking into account that node degrees vary from their minimal value $k_{\min} \approx 1$ to $k_{\max}$, such significant difference between $\langle k \rangle$ and $k_{\max}$ may indicate skewed shape of the node degree distribution. We will investigate this question below. An average degree $\langle k \rangle$ of the resulting networks and its maximal value $k_{\max}$ decreases after applying filters on link weights. Note that the ratio between the two becomes more pronounced. Instead of 10 times difference for the original networks, $k_{\max}$ exceeds $\langle k \rangle$ for filtered networks in 20 - 50 times. This may indicate that filtered networks become even more heterogeneous than the original ones.

To investigate network heterogeneity in more detail, let us plot the node degree distribution $P(k)$ for the original network and for its filtered version. $P(k)$ is defined as the probability that the randomly selected node has degree k for a network of infinite size. Figure 3.3 shows these unnormalized distributions (instead of probabilities, the plots show the number of nodes) for the original network, and the network filtered by link weight criterion for the 2013 year dataset.

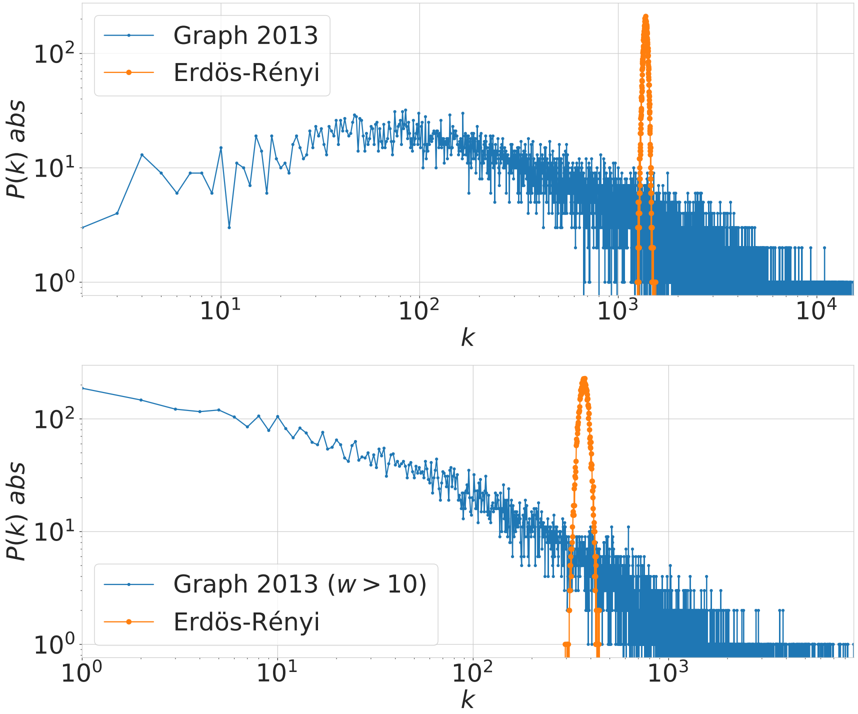

FIGURE 3.3: Unnormalized node degree distribution $P(k)$ for 2013 year graphs (indicated by blue color). The top panel represents the unfiltered network, while the bottom panel represents a network filtered by the link weight criterion. Yellow color represents degree distribution for the Erdös-Rényi random graph with the same number of nodes and links as the corresponding concept network. The distributions show that concept networks are much more heterogeneous than one would expect by a random Erdös-Rényi like scenario.

The comparison of the plots in Figure 3.3 reveals a difference between the shapes of distributions for unfiltered and filtered networks. While in the unfiltered network, there is a tendency for $P(k)$ to increase with a small k and then to decrease with large values of k, $P(k)$ has a continuous tendency to decrease for filtered networks. This means that in the unfiltered network, the probability of finding a node with a small k is relatively small while in the filtered network, on the opposite, this probability is quite high. Indeed, the nodes that have degree $k = 1$ in an unfiltered network represent the concepts that have co-occurred with one another concept only in the entire dataset. Such cases, in particular, are quite rare due to the distribution in Figure 3.1. Here the number of concepts per article set a minimum value for the node degree k. Such restrictions lead to relatively small values of $P(k = 1)$ for unfiltered networks. In contrast, the number of concepts identified in the article does not strongly affect node degrees for filtered networks. As a result, $k = 1$ is the most probable degree in the filtered network.

Besides the degree distribution for the concept networks, Figure 3.3 shows the degree distribution of the Erdös-Rényi random graphs with the same number of nodes as the corresponding concept network. This random graph has been generated by creating N isolated nodes and connecting these nodes by L links such that each link connects a pair of randomly selected nodes. The plots show that concept networks are much more heterogeneous than Erdö-Rényi random graphs, meaning that there is a much wider spread of node degrees than one would expect in a simple random process.

Having defined node degrees, one may calculate assortativity mixing by degrees r, defined as Pearson correlation coefficient between node degrees on both ends over existing link [11, 12, 24]:

$$r = \frac{\sum_{ij} ij(e_{ij} - q_i q_j)}{\sum \sigma_q^2} \tag{3.3}$$

where σ_q is the standard deviation of the distribution q_k. The networks with $r > 0$ are referred to assortative networks with typical examples being social networks. In assortative networks, high-degree nodes are more likely to be connected to other high-degree nodes, while low-degree nodes have a tendency to be connected to other low-degree nodes more likely than one would expect by chance. Disassortative networks (for which $r < 0$) have an opposite tendency: high-degree nodes tend to be connected to low-degree ones more often than one would expect by chance and vice versa. Examples of disassortative networks are Internet and WWW, see e. g. [11, 12, 24] and references therein. As indicated in Table 3.1 concept networks are disassortative. Moreover, filtered networks look to be more disassortative than unfiltered ones. The reasons behind the negative r may be explained if we assume that generic (high degree) concepts and specific (low degree) ones have different functions in the context of research publications. E. g. high-degree concepts glue pieces of the research together.

While node degree k_i characterize "popularity" of node i, clustering coefficient c_i of node i describes the level of connectivity among i-th neighbours:

$$c_i = \frac{2m_i}{k_i(k_i - 1)}, \quad k_i > 1, \tag{3.4}$$

where $k_i(k_i - 1)$ is the doubled number of all possible connections between k_i neighbours of node i and m_i is the number of existing connections among these k_i nodes. Mean value $\langle c \rangle$ of c_i, averaged over all nodes in the network, characterizes the local density of neighborhood links in the entire network

$$\langle c \rangle = \frac{1}{N} \sum_{i=1}^{N} c_i, \tag{3.5}$$

and will be referred below as an average clustering coefficient. The values of $\langle c \rangle$ for all networks are shown in Table 3.1. The values are quite high reaching about

$\langle c \rangle \approx 77\%$ for unfiltered networks and about $\langle c \rangle \approx 89\%$ for filtered ones. These values are one order of magnitude higher than one would expect in an Erdös-Rényi random graph for which $\langle c \rangle$ equals to the probability that any two nodes are connected to each other. Indeed, given the same number of nodes and links, one arrives at $\langle c \rangle = \rho$: an average clustering coefficient equals the density of links in the corresponding network, see Table 3.1. The other observation is that the average clustering coefficient $\langle c \rangle$ for filtered networks is higher than $\langle c \rangle$ for unfiltered networks independent of the submission year. This is quite natural, assuming that filtering procedure cuts weak links between different communities. Such links may be important for information spreading through the network, and they contribute negatively to the value of the clustering coefficient.

The clustering coefficient serves to measure specific correlations present in the network structure. Another alternative way to quantify such correlations for the entire network is to define network global transitivity C. Instead of calculating the average value of local measurements, C is defined as a ratio between the total number of connected triplets in the network and the number of all possible triangles:

$$C = \frac{\text{number of closed triplets (clique of 3 nodes)}}{\text{number of all triplets}}. \tag{3.6}$$

Comparing the values of average clustering coefficient $\langle c \rangle$ and global transitivity C for the considered networks, see Table 3.1, one observes that C is smaller than $\langle c \rangle$ for all networks considered. Such differences may indicate the community structure of the concept networks. Indeed for the same reasons as described above, C for the Erdös-Rényi graph (without community structure) equals $\langle c \rangle$. On the other side, if we imagine a graph that consists of several large enough cliques (fully connected subgraphs) connected by a few links, its average clustering coefficient will reach 1 or close value, while C will be significantly smaller due to the existence of open triangles between cliques. Our results indicate that the difference between $\langle c \rangle$ and C is more pronounced for filtered concept networks than for the original ones. This means a more pronounced clustering (community) structure in filtered networks.

Figure 3.4 shows the dependence between the node clustering coefficient and node degree for unfiltered and filtered concept networks.

Negative correlations between k_i and c_i indicate the hierarchical organization of the concepts network. This is quite natural assuming that some concepts may have related (children) sub-concepts like `Friction Force` or `Electromagnetic Force` for `Force` concept.

For measuring the distances between two different nodes i and j the shortest path length l_{ij} is used. It is defined as the minimal number of edges one has to pass to reach node i starting at node j. For the entire network, one may calculate the average shortest path length $\langle l \rangle$ defined as an average value of l_{ij} over all ij pairs for which such path exists. $\langle l \rangle$ can tell how far different parts of the graph are located within the network. The diameter of the network $l_{\max}$ is defined as the maximal value of the shortest paths l_{ij} found in the network. The values of both $\langle l \rangle$ and $l_{\max}$ for all networks considered here are shown in Table 3.1. First one may observe that the lengths of shortest paths are quite small with average value $\langle l \rangle \approx 2$ and $l_{\max} = 3 \div 5$. So, concept networks are quite compact. Finally, comparing original and filtered concept networks, one sees that the average shortest path length for filtered networks are larger than for the original concept networks. This is expected behavior if we assume that filtering procedure cuts weak links that bridge strongly connected clusters of concepts. Then, there are fewer bridges in filtered networks than in an original one, resulting in an increase in path lengths.

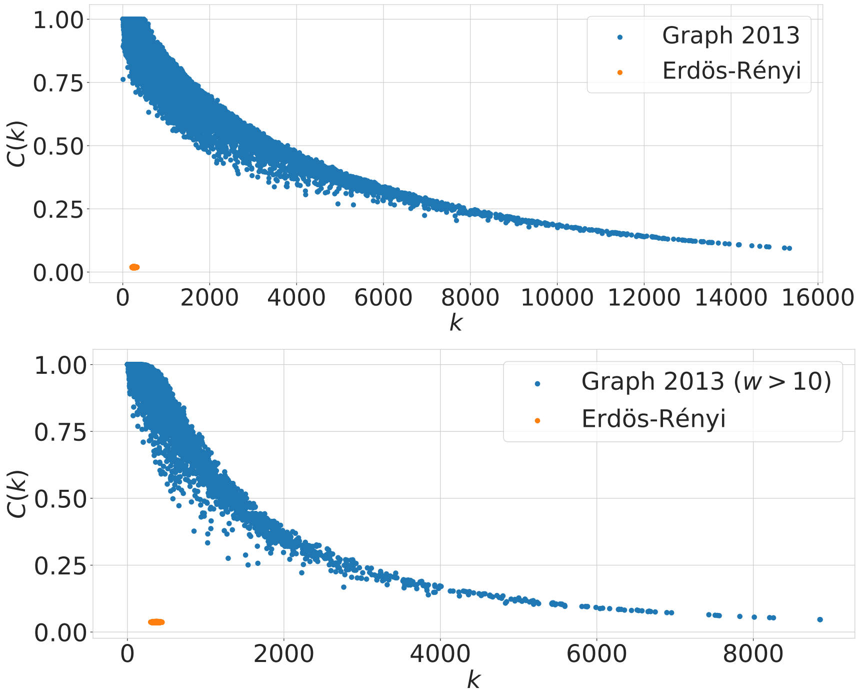

FIGURE 3.4: Node clustering coefficient c_i as a function of node degree k_i for unfiltered concept network (top panel) and filtered one using link weight criteria (bottom panel) for year 2013. Blue points show the dependencies for concept network and orange points correspond to the Erdös-Rényi random graphs of the same size.

To conclude, our analysis of the topology of the concept network indicates that observed concept networks are heterogeneous graphs that obey internal clustering (community structure) and hierarchical organization. These properties of the concepts network are independent of the subset of data used (constructed from 2013 and 2015 year data). These features, however, are more pronounced once weak links have been removed. As is follows from the comparison of data obtained for different years and via different procedures of relevant link determination, cf. Table 3.1, complex networks under consideration attain a range of universal features that do not change with time and characterize the system of concepts as a whole. In particular, they are the small world networks [11, 12, 22–25] characterized by a small size (mean shortest path and maximal shortest path values) and large value of clustering coefficient. The last also brings about the presence of strong correlations. Moreover, an essential difference between the clustering coefficient and global transitivity serves as evidence of possible community structure. In turn, the negative value of assortativity suggests that they are disassortative networks where a group of central nodes (hubs) serves as common attraction points for nodes with lower degree values.

Chapter 4

Scientific innovations and concept embedding

In this chapter, we return to the main question of interest in this study: is it possible to detect in advance fields where scientific innovations may emerge? In particular, we are interested in the questions of the prediction power of concept embedding (Section 2.5).

The Chapter is organized as follows. We start with our view on scientific innovations in Section 4.1, then in Section 4.2 we introduce embedding dimension to our analysis and in Section 4.3, we finish this chapter with the investigation of the predicting abilities of concept embedding.

4.1 Scientific innovations

Investigation of scientific innovation emergence is not straightforward. The simplification adopted in frames of this thesis considers innovations as the appearance of a new statistically significant link between nodes that previously were not linked to each other. In this way, the emergence of such a link is treated as a novelty introduced into the graph of scientific concepts.

We proceed by considering a network of scientific concepts built upon manuscripts submitted to `arXiv` during the year 2013. Let us consider a pair of concepts i and j. In terms of link existence, these concepts may be either connected by a link or disconnected, meaning no link between i and j. The fraction of pairs connected by links equal to the density of links in the corresponding concept network and has a value $\rho = 8.46\%$. Some of the links that carry low weight may be considered as spurious links rather than statistically significant, meaning that they could arise as a result of noise rather than a real coupling between the corresponding concepts. In this thesis, we consider two alternative ways to filter out such spurious links: i) naive filtering by setting up a link weight threshold and ii) disparity filtering that employs statistical significance testing. By setting a link weight threshold $\omega = 10$ (see Section 3.2.1 for details) we arrive at 1,844,453 significant links (i. e. only 1.4% pairs of concepts are connected by significant links). Alternatively, using a disparity filter with the p-value threshold $\rho = 0.1$ (see Section 3.2.2) the number of statistically significant links equals to 1,642,958, covering 1.25% of all pairs of concepts.

Consequently, about 98.5% of pairs of nodes are either disconnected or are connected by rather spurious links. Some of these pairs may become connected in the future by statistically significant connections. The emergence of such connections is referred in this thesis as scientific innovations. The questions we are asking in this research are related to the predictability of such innovations. In particular, we are interested in the power of concept embedding technique to distinguish between the pairs of concepts that will become connected vs the pairs that will stay disconnected in the future. In this chapter setting, by the present we mean a collection of manuscripts submitted to arXiv during 2013 and by the future – the collection submitted to arXiv during 2015 (corresponding graphs of concepts).

In the analysis below, we will have more than 100M potential connections between concepts. The tables 4.1, 4.2 below show how many of these pairs are connected by strong links and how many of them are disconnected or are weakly connected (Table 4.1); or are connected by statistically significant links / statistically insignificant links or are disconnected (Table 4.2). Pairs of nodes i and j that are either connected by a link with weight $w_{ij} \leq \omega$ or are not connected by a link in network will be referred below as `weakly connected nodes` and `weak/missing link`, while pairs of nodes i and j that are connected by a link with weight $w_{ij} > \omega$ will be referred below as `strongly connected nodes` and `strong links`, correspondingly.

	Number	Percentage
`weak/missing links`	129,837,653	98.6%
`strong links`	1,844,453	1.4%

TABLE 4.1: The numbers and percentages of pairs of concepts with different link weight during year 2013. The table shows that majority of the pairs are either loosely connected ($w \leq 10$) or disconnected in year 2013.

Similarly, a pair of nodes i and j that are either connected with p-value $\alpha_{ij} \geq \rho$ or are not connected by a link in network will be referred below as statistically `insignificant link`, while a pair of nodes i and j that are connected by a link with p-value $\alpha_{ij} < \rho$ will be referred below as statistically `significant links`.

	Number	Percentage
`insignificant links`	130,039,148	98.75%
`significant links`	1,642,958	1.25%

TABLE 4.2: The numbers and percentages of pairs of concepts with different p-value during year 2013. The table shows that majority of the pairs are either statistically insignificant or disconnected in year 2013.

The tables show that the majority of the pairs are either disconnected or weakly/ insignificantly connected to each other. In what follows below, we will look for an emergence of new links between these pairs: such links may be considered as atypical combinations of existing ideas. Therefore one may expect innovations in the papers where such pairs appear [8].

Thus the questions of our interest are related to forecasting the pairs where such innovations may emerge given the number (fraction) of such connections is known. These numbers (fractions) we calculate empirically comparing dataset of 2013 and 2015. The numbers are summarized in Tables 4.3 and 4.4.

Table 4.3 indicates that only 564,330 pairs (0.43%) became strongly connected ($w_{ij} > 10$) in 2015 out of 129,837,653 weakly connected/disconnected pairs in 2013. At the same time, it shows that 1,662,432 (90.13%) out of 1,844,453 strongly connected pairs in 2013 remained strongly connected in 2015.

Disparity filter arrives at a similar picture as weight filtering did. Only 475,788 pairs (0.37%) became statistically significant in 2015 out of 130,039,148 insignificant/disconnected pairs in 2013. At the same time, 1,389,652 pairs (84.58%) out of

	Number of strong links in 2015
weak/missing links in 2013	564,330
strong links in 2013	1,662,432

TABLE 4.3: The numbers of strong links between nodes of 2015 year network, divided by weight of the links between them during year 2013. The total number of links of weight $w_{ij} > 10$ in year 2015 equals 2,226,762 (after intersection with pairs from g-2013). The division into two groups has been performed by choosing the link weight threshold of $w_{ij} = 10$. The table shows that majority of the "strongly connected links" in year 2015 have been strongly connected in year 2013.

	Number of significant links in 2015
insignificant links in 2013	475,788
significant links in 2013	1,389,652

TABLE 4.4: The numbers of statistically significant links (with $\alpha_{ij} < 0.1$) between nodes of 2015 year network, divided by statistical significance of links between them during year 2013. The total number of statistically significant links $\alpha_{ij} < 0.1$ in year 2015 equals to 1,865,440 (after intersection with pairs from g-2013). The table shows that majority of the statistically significant links in year 2015 have been statistically significant in year 2013.

1,642,958 statistically significant links in 2013 remained statistically significant in the year 2015.

To conclude, less than 0.5% of disconnected/weakly connected/statistically insignificant links between concepts in 2013 became strong/significant in the year 2015. In the next section, we will analyze the ability of concept similarity obtained using the embedding technique to discriminate between the pairs that will become statistically significant vs remain insignificant in 2015.

4.2 Embedding similarity

The key assumption is that i) concepts that appear in a similar context will have close enough vectors in embedded space and ii) that the concepts that carry similar content are more likely to become connected in the future.

For this reason we use concept co-occurrence matrix for year 2013 and embedded each concept vector in 100 dimensional space using PyTorch-BigGraph [16], see Section 2.5. The whole detailed pipeline we used for described graphs and embeddings formulation could be found at GitHub repository [1]. As a result, each concept i becomes associated with a vector $\vec{v}_i$ in the embedded space. The similarity s_{ij} between a pair of concepts i and j is then calculated as a cosine similarity [2] between the corresponding vectors $\vec{v}_i$ and $\vec{v}_j$. In general, one may expect positive correlations between weight w_{ij} of the link between nodes i and j and the similarity s_{ij} between the corresponding concept vectors. Indeed, such behavior is observed in Figure 4.1, where concept embedding similarity s_{ij} between pairs of nodes i and j is shown as a function of weight w_{ij} of a link between the corresponding nodes.

[1] github.com/sergibro/concept-graphs

[2] en.wikipedia.org/wiki/Cosine_similarity

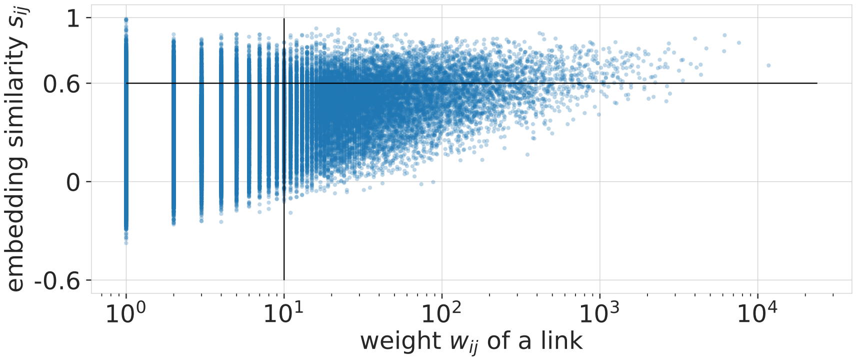

FIGURE 4.1: Concept embedding similarity s_{ij} as a function of a weight w_{ij} of a link connecting nodes i and j for corresponding concepts. X-axis shows $\log(w_{ij})$ for links with non-zero weights only. 2013 year data has been considered. The figure indicates positive correlations between the two characteristics with a significant level of fluctuations. The thresholds for both characteristics ($\omega = 10$ and $\zeta = 0.6$) are shown by solid lines.

Even though for the majority of links, the actual weight equals 0 since the two concepts have not appeared together in the same manuscripts, the similarity s_{ij} between their concept vectors in embedded space is expected to be non-zero. The reason behind such behavior/expectation is the following: since we reduce the dimensionality in embedded space from about 15,000 to 100, each dimension of embedded space may be considered as a combination of a number of concepts. Such aggregation leads to a much higher likelihood that two unconnected concepts have a non-zero similarity in an embedded space. Also, the chosen metric (cosine) play the role here. Indeed, our expectations are confirmed in Figure 4.2.

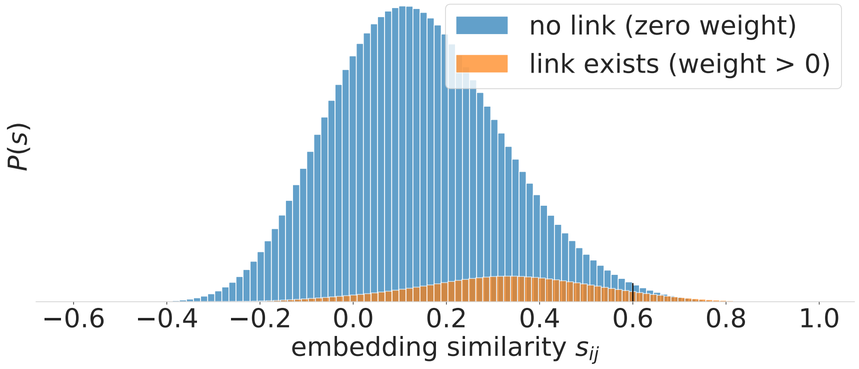

FIGURE 4.2: Histogram of embedding similarity for all possible pairs of nodes for the unfiltered 2013 network. Two groups are shown: pairs that are connected by a link (yellow bars) and pairs that are not connected by a link (blue bars). The solid line shows the considered threshold for embedding similarity ($\zeta = 0.6$).

The histogram shows the number of concept pairs for which concept embedding similarity falls into a corresponding bin, separately for pairs connected by a link (yellow bars) and for the pairs that are not connected by any link in the year 2013 (blue bars). The figure shows that the distribution for the disconnected pairs of nodes is shifted towards lower values of embedding similarities as compared to that for connected pairs of nodes. At the same time, the Figure 4.2 shows that there are pairs of disconnected nodes that are characterized by strong enough embedding similarity. There are pairs where one may expect the appearance of statistically significant links in the future.

Once similarities s_{ij} between concept vectors in embedded space have been calculated for each pair of concepts i and j, we divide all pairs of concepts into two groups: i) `Strong embedding similarity group`; ii) `Weak embedding similarity group`. To distribute pairs of nodes/concepts among the groups, we put an arbitrarily selected threshold of $\zeta = 0.6$, see also Figures 4.1 and 4.2. The pairs of concepts for which embedding similarity $s_{ij} \leq \zeta$ are assigned to `Weak embedding similarity group`, for convenience, we will refer to the corresponding pairs as `dissimilar concepts`. Instead, if the embedding similarity between concepts i and j $s_{ij} > \zeta$, the corresponding pair is assigned to a `Strong embedding similarity group` and will be referred below as `similar concepts`. We are aware of the fact that the selection of the other value of ζ threshold could change the distribution of concept pairs between the groups. However, we expect that such modification will not change the qualitative results of our analysis. Especially, because pairs on both extremes of embedding similarity will eventually be assigned to different groups.

The tables below show the number of concepts that fall into each of the defined above embedding similarity groups.

	`dissimilar concepts`	`similar concepts`
`weak link`	128,263,658	1,573,995
`strong links`	1,424,215	420,238

TABLE 4.5: The number of pairs of concepts that fall into each embedding similarity group for year 2013. In addition the pairs have been divided into different link weight categories ($w_{ij} > 10$ – `strong links` and $w_{ij} \leq 10$ – `weak link`).

Table 4.5 shows the number of pairs of concepts that fall into `Strong embedding similarity group` (`similar concepts`) and `Weak embedding similarity group` (`dissimilar concepts`), separately for concept pairs that are strongly connected in unfiltered network and for the pairs that are weakly connected. The table indicates significant differences in the allocation of pairs of concepts among embedding similarity groups for weakly and strongly connected pairs of nodes in the network. While only 1.2% (1,573,995) of node pairs connected by weak link (or disconnected pairs) in `g-2013` falls into `similar concepts` group, this fraction for strongly connected pairs of nodes reaches 22.8%. Thus, positive correlations between embedding similarity grouping and strong link existence, one may expect that 1.2% of mentioned above disconnected/weakly connected nodes are higher chances to become connected in the future than the remaining 98.8%. Similar results have been observed if one uses a disparity filter instead of link weight threshold filter, see Table 4.6.

	`dissimilar concepts`	`similar concepts`
`insignificant links`	128,412,677	1,626,471
`significant links`	1,275,196	367,762

TABLE 4.6: The number of pairs of concepts that fall into each embedding similarity group for year 2013, separately for statistically significant ($\alpha_{ij} < 0.1$) and statistically insignificant links ($\alpha_{ij} \geq 0.1$).

Thus, we expect that the grouping of pairs of nodes using embedding similarity may give us a value in predicting the pairs of concepts where statistically significant links will be established in the future. In the next Section 4.3, we will analyze the forecasting abilities of the concept embedding similarity.

4.3 Forecasting power of embedding similarity

With the data about the concept network for the year 2013 at hand, let us now consider the network of scientific concepts constructed from manuscripts submitted to arXiv during the year 2015. Below we perform preliminary analysis rather than propose a predictive model.

If we use link weight filtering, assuming that strong links are the ones with $w_{ij} > \omega$ ($\omega = 10$) we arrive at 2,227,384 (1.6%) strong links between $N(2015) = 16660$ nodes for year 2015. Table 4.7 shows how are these 2,227,384 links distributed among link weight groups and concept embedding groups in year 2013. 622 (less than 0.03%) pairs are not included as at least one of concept from them exists only in "future" g-2015 graph.

	dissimilar concepts in 2013	similar concepts in 2013
weak links in 2013	514,178	50,152
strong links in 2013	1,265,962	396,470

TABLE 4.7: Allocation of strong links in 2015 year concept network among two groups of corresponding concepts pairs in 2013: strong versus weak link weight in 2013 and strong versus weak concept embedding similarity group.

While the majority of concept pairs were either disconnected or connected by a weak link in 2013, the majority of strongly connected concept pairs in 2015 were connected by strong links in 2013 too.

Table 4.8 represents the numbers from Table 4.7 as the fractions of the number of concept pairs that belong to the combination of categories in 2013, shown in Table 4.5.

	dissimilar concepts during 2013	similar concepts during 2013
weak links during 2013	0.4%	3.19%
strong links during 2013	88.89%	94.34%

TABLE 4.8: Percentage of concept pairs that belong to a specific combination of link weight group and embedding similarity group in 2013 that either remained or became strong in 2015. For example, 3.19% in the table means that out of 1,573,995 pairs of nodes that were either disconnected or connected by a weak link in 2013 and belonged to Strong embedding similarity group (see Table 4.5), 50,152 became strongly connected in year 2015, see Table 4.7.

The table shows that about 90% of strong links in 2013 remained strong in the year 2015. If we take into account grouping by concept embedding similarity, we observe additional segregation: strong links with low embedding similarity in the year 2013 remained strong in the year 2015 in almost 89% of cases, while strong links with strong embedding similarity in the year 2013 remained strong in the year 2015 for more than 94% of cases. These results lead us to the following conclusions. First, if a link between two concept-nodes exists and this is a strong link, then it is likely that the link will exist in the future, and it will remain the strong one. In other words, the strength of a link is a good predictor for a link to belonging to the same category in the future. Second, strong links with high concept embedding similarity have higher chances to remain strong in the future than strong links that are characterized by low embedding similarity.

On the other side, weak links evolve to strong links quite rarely. Only 0.4% of weak links in 2013 evolved to strong links in year 2015 (compare numbers in Tables 4.1 and 4.3). However, classification of concepts pairs by their embedding similarity allowed us to identify a subgroup of these pairs for which the probability of becoming strong connections raises to 3.19%, i. e. in about 8 times. Even though the concept embedding similarity does not point the ”future” emergence of a new strong link in the network exactly, the results of our analysis indicate its power as one of the features to be used in such predictions.

Similar results have been obtained if we use classification of links between pairs of concepts using statistical significance testing instead of link weight threshold, see Table 4.9.

	`dissimilar concepts` during 2013	`similar concepts` during 2013
`insignificant links` during 2013	0.3%	3.01%
`significant links` during 2013	82.71%	91.06%

TABLE 4.9: Percentage of concept pairs that belong to a specific combination of link significance group and embedding similarity group in 2013 that either remained or became strong in 2015.

Thus, independent of the method used to classify pairs of concepts, either using link weight threshold or statistical significance testing, the results of our analysis indicate the ability of concept embedding similarity in predicting scientific innovations, i. e. the emergence of strong or statistically significant links in knowledge graphs.

Chapter 5

Conclusions and Outlooks

The goal of our work was to analyze the possibilities of innovation emergence in the course of knowledge generation. To this end, we have investigated the structure and dynamics of connections between scientific concepts that constitute a body of research papers, as recorded in the arXiv repository [9]. We have applied two methods, concept embedding and network analysis, to quantify properties of sets of concepts and to predict the emergence of new links (innovations) between different concepts. We have shown that whereas each of the above methods is a powerful tool to define certain features of a system of concepts, it is the combination of these two methods that leads to a synergetic effect and allows to forecast dynamics of new links creation and evolution of a system as a whole. The main results obtained in the course of our analysis include the following:

- We have represented a system of concepts of scientific papers in the form of a complex network. Different nodes in this network correspond to different concepts, and a link between two nodes-concepts means that they were exploited in the same paper. We have determined the quantitative characteristics of a complex network of concepts and their evolution with time, and the data is given in Table 3.1.

- We have used two complementary approaches to define the presence of a strong link between two nodes, i. e. of a link that serves as evidence of a relevant connection. In one approach, the criterion is given by a link weight. The second method takes into account subtle information about network intrinsic structure [28]. Corresponding data is shown in Table 3.1.

- As is follows from a comparison of data obtained for different years and via different procedures of relevant link determination, see Table 3.1, complex networks under consideration attain a range of universal features that do not change with time and characterize the system of concepts as a whole. In particular, they are the small world networks characterized by small size (mean the shortest path and maximal shortest path values) and large value of the clustering coefficient. The last also brings about the presence of strong correlations. Moreover, an essential difference between the clustering coefficient and global transitivity serves as evidence of possible community structure. In turn, the negative value of assortativity suggests that they are disassortative networks where a group of central nodes (hubs) serves as common attraction points for nodes with lower degree value.

- Concept embedding technique enabled us to find out proximity (by context, by subject, or related in any other way) between different concepts. With a measure of proximity at hand, we were in a position to compare it with the dynamics of new links emergence between different concepts. In turn, this enables one to reveal groups of concepts (subsequently – fields of knowledge) where innovations are probable to emerge. Corresponding statistical analysis is summarized in Tables 4.8 and 4.9.

The results obtained in this study may be useful both from the fundamental point of view, contributing to our understanding of how the knowledge is formed, as well as they may have the practical implementation. In particular, the methodology elaborated in the course of our analysis can be used to detect fields where innovations have a higher probability of appearing. A natural way to continue the analysis presented here is to evaluate practical outcomes (i. e. impact) of papers, where the higher probability of innovation is predicted. With the scientometric data at hand, such a task is not much time consuming and will be a subject of future work.

Another way to follow-up the analysis is to propose a model to predict the emergence of statistically significant links between already existing concepts. The model should take as input i) concept co-occurrence matrix, which will enable us to project it to concept network and perform embedding analysis, and ii) the fraction of insignificant links or disconnected concept pairs that will be connected in the future. The model should identify the pairs of concepts that are about to receive a statistically significant connection in the future. In the scope of this thesis, we have shown that concept embedding similarity is among the features to be used by the model. We expect that combining concept embedding similarity with the network-based proximity measure (link weight, a fraction of common neighbors, etc.), the model will be able to establish a good baseline for predictability of the emergence of novelties in concept networks.

Bibliography

[1] L. Zhmud. *The origin of the History of Science in Classical Antiquity*. Vol. 19. Walter de Gruyter, 2008.

[2] T. Lewens. *The meaning of science: An introduction to the philosophy of science*. Hachette UK, 2016.

[3] O. Mryglod, Y. Holovatch, R. Kenna, and B. Berche. "Quantifying the evolution of a scientific topic: reaction of the academic community to the Chornobyl disaster". In: *Scientometrics* 106.3 (2016), pp. 1151–1166.

[4] L Leydesdorff and S Milojević. *Scientometrics In: Lynch Micheal, editor. International Encyclopedia of Social and Behavioral Sciences*. 2015.

[5] B. Berche, Y. Holovatch, R. Kenna, and O. Mryglod. "Academic research groups: evaluation of their quality and quality of their evaluation". In: *Journal of Physics: Conference Series*. Vol. 681. 1. IOP Publishing. 2016, p. 012004.

[6] A. Zeng et al. "The science of science: From the perspective of complex systems". In: *Physics Reports* 714 (2017), pp. 1–73.

[7] I. Iacopini, S. Milojević, and V. Latora. "Network dynamics of innovation processes". In: *Physical review letters* 120.4 (2018), p. 048301.

[8] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. "Atypical combinations and scientific impact". In: *Science* 342.6157 (2013), pp. 468–472. ISSN: 0036-8075. DOI: 10.1126/science.1240474. eprint: https://science.sciencemag.org/content/342/6157/468.full.pdf. URL: https://science.sciencemag.org/content/342/6157/468.

[9] *arXiv*. arXiv gives an open access to 1,640,097 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics. Accessed: 2020-01-03. URL: https://arXiv.org.

[10] *ScienceWISE*. The ScienceWISE project aims to develop a scientist-generated on-line knowledge base fully integrated into the physics ArXiv.org. Accessed: 2020-01-06. URL: http://ScienceWISE.info.

[11] Y. Holovatch et al. "Complex networks". In: *Journal of Physical Studies* 10 (2006), pp. 247–289.

[12] M. Newman. *Networks: An Introduction*. OUP Oxford, 2010. ISBN: 9780191500701. URL: https://books.google.com.ua/books?id=LrFaU4XCsUoC.

[13] A.-L. Barabási et al. *Network science*. Cambridge university press, 2016.

[14] Y. Holovatch et al. "Statistical Physics of Complex Systems in the World and in Lviv". In: *Journ. Phys. Stud. (in Ukrainian)* 22.2801 (2018), p. 21.

[15] T. Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[16] A. Lerer et al. "PyTorch-BigGraph: A Large-scale Graph Embedding System". In: *arXiv preprint arXiv:1903.12287* (2019).

[17] A. Martini et al. "Sciencewise: Topic modeling over scientific literature networks". In: *arXiv preprint arXiv:1612.07636* (2016).

[18] A. Constantin. "Automatic structure and keyphrase analysis of scientific publications". PhD thesis. The University of Manchester (United Kingdom), 2014.

[19] V. Palchykov, V. Gemmetto, A. Boyarsky, and D. Garlaschelli. "Ground truth? Concept-based communities versus the external classification of physics manuscripts". In: *EPJ Data Science* 5.1 (2016), p. 28.

[20] V. Palchykov and Y. Holovatch. "Bipartite graph analysis as an alternative to reveal clusterization in complex systems". In: *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE. 2018, pp. 84–87.

[21] V. Palchykov and Y. Holovatch. "Modeling innovations and scientific discoveries through novel combinations of ideas". In: *In progress* (2019).

[22] R. Albert and A.-L. Barabási. "Statistical mechanics of complex networks". In: *Reviews of modern physics* 74.1 (2002), p. 47.

[23] S. N. Dorogovtsev and J. F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford, 2013.

[24] M. E. Newman. "The structure and function of complex networks". In: *SIAM review* 45.2 (2003), pp. 167–256.

[25] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[26] Y. Holovatch, R. Kenna, and S. Thurner. "Complex systems: physics beyond physics". In: *European Journal of Physics* 38.2 (2017), p. 023002.

[27] A. Grover and J. Leskovec. "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2016, pp. 855–864.

[28] M. Á. Serrano, M. Boguná, and A. Vespignani. "Extracting the multiscale backbone of complex weighted networks". In: *Proceedings of the national academy of sciences* 106.16 (2009), pp. 6483–6488.