RESEARCH ARTICLE

Time evolution of the hierarchical networks between PubMed MeSH terms

Sámuel G. Balogh[ID][1]*, **Dániel Zagyva**[1], **Péter Pollner**[2], **Gergely Palla**[2]

1 Dept. of Biological Physics, Eötvös University, Budapest, Hungary, **2** MTA-ELTE Statistical and Biological Physics Research Group, Budapest, Hungary

* balogh@hal.elte.hu

Abstract

Hierarchical organisation is a prevalent feature of many complex networks appearing in nature and society. A relating interesting, yet less studied question is how does a hierarchical network evolve over time? Here we take a data driven approach and examine the time evolution of the network between the Medical Subject Headings (MeSH) provided by the National Center for Biotechnology Information (NCBI, part of the U. S. National Library of Medicine). The network between the MeSH terms is organised into 16 different, yearly updated hierarchies such as "Anatomy", "Diseases", "Chemicals and Drugs", etc. The natural representation of these hierarchies is given by directed acyclic graphs, composed of links pointing from nodes higher in the hierarchy towards nodes in lower levels. Due to the yearly updates, the structure of these networks is subject to constant evolution: new MeSH terms can appear, terms becoming obsolete can be deleted or be merged with other terms, and also already existing parts of the network may be rewired. We examine various statistical properties of the time evolution, with a special focus on the attachment and detachment mechanisms of the links, and find a few general features that are characteristic for all MeSH hierarchies. According to the results, the hierarchies investigated display an interesting interplay between non-uniform preference with respect to multiple different topological and hierarchical properties.

Introduction

In the recent decades the network approach has become fundamental in the studies of various phenomena in nature and society, ranging from the level of interactions within cells to the level of the Internet, economic networks, and the society [1, 2]. A very important topic in this field is related to the hierarchical organization of networks [3–7]. Grasping the signs of hierarchy in networks is a non-trivial task with a number of possible different approaches, including the statistical inference of an underlying hierarchy based on the observed network structure [4], and the introduction of various hierarchy measures [8–13]. Examples of empirical studies on hierarchical networks are including the transcriptional regulatory network of Escherichia coli [14], the dominant-subordinate hierarchy among crayfish [15], the leader-follower network of pigeon flocks [16, 17] and harems of Przewalski horses [18], the rhesus macaque

kingdoms [19], neural networks [20] and technological networks [5], scientific journals [21], social interactions [22–25], urban planning [26, 27], on-line news content [28], ecological systems [29, 30], and evolution [31–33]. In addition, hierarchical organisation is also related to the non-normality of networks [34], the topological properties of various scientific and techno-scientific fields [21, 35–39] (usually depicted by citation networks), and the optimal performance of interacting agent groups [40–42]. Hierarchies are usually depicted as directed acyclic graphs, in which the links are not allowed to form directed cycles, and where a pair of nodes connected by a link are assumed to be in some sort of asymmetric relationship with each other such as parents and children, leaders and followers, etc.

Networks representing real systems are subject to constant evolution in most of the cases, and some relevant aspects of the laws forming the shape of networks changing over time have already been uncovered in the scientific literature. Probably most famous is the preferential attachment rule for growing scale-free networks, which is one of the key concepts of the Barabási-Albert model [43], and was detected also by empirical studies of network data [44–47]. Another notable example is provided by the studies of the various statistical features of community evolution in networks [48]. Along the same line, in the present paper our aim is to examine the statistical properties of time dependent networks with a hierarchical structure.

Our study is based on the data provided by the NCBI about the MeSH terms, which were introduced for helping the search in the PubMed publication database of the NCBI (comprising more than 29 million citations for biomedical literature) at various levels of specificity. The MeSH terms are hierarchically organized: At the most general level of the hierarchical structure we find very broad headings such as "Organisms" or "Information Science", whereas more specific headings are found at deeper (more narrow) levels. Due to the rapidly developing nature of the medical-, biochemical- and biological sciences, the set of available MeSH terms are yearly updated by the curators of PubMed. This provides a fascinating empirical data-set for the study of time dependent hierarchical networks. A few previous studies on this data-set have already been published, approaching the development of the MeSH term hierarchies from an ontological perspective [49–52]. The main focus of these results was on the growth of the system, concentrating on how are the newly introduced MeSH terms categorized and linked under already existing older MeSH terms. In our present study we show that restructuring plays an equally important role in forming the structure of the MeSH hierarchies.

Our goal is to examine the statistical features of the time evolution in the observed hierarchies. One of the central questions we are interested in is how do the different topological- and hierarchical properties of the nodes influence the attachment and detachment of links during the restructuring. Understanding the nature of these processes can help the creation of hierarchy evolution models that can predict which part of the hierarchy is most likely to be rewired in the future, and what is the expected change in the overall features of the hierarchy.

Data and methods

Basic properties of the MeSH hierarchies

The directed networks we consider are based on the classifications provided by PubMed, specifying at least one parent for any available MeSH term, except for the roots of the hierarchies. The raw data we use is publicly available on the link provided in Ref. [53]. There are altogether 16 different roots, and the total number of descendants of the individual roots (the sizes of the hierarchies) varies roughly between a 1,00 and a 10,000 nodes, whereas the time span of our analysis is 14 years. In Table 1 we list a few basic properties of these networks, including the minimum and maximum sizes, the maximum level depth and the average fraction of changed

Table 1. Basic hierarchy data.

	Root name	size range	max. depth	average change
A	Anatomy	1350–1826	10	4.66%
B	Organisms	2252–3815	13	6.49%
C	Diseases	3975–4799	8	4.23%
D	Chemicals and Drugs	6902–9934	11	6.22%
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment	2040–2924	9	4.89%
F	Psychiatry and Psychology	807–1083	7	3.60%
G	Phenomena and Processes	1733–2259	10	15.18%
H	Disciplines and Occupations	334–537	8	12.07%
I	Anthropology, Education, Sociology and Social Phenomena	449–641	9	5.23%
J	Technology, Industry, Agriculture	254–582	10	8.92%
K	Humanities	152–200	7	3.93%
L	Information Science	322–476	9	5.82%
M	Named Groups	174–290	7	5.71%
N	Health Care	1072–1795	10	4.94%
V	Publication Characteristics	137–163	6	3.44%
Z	Geographicals	369–402	6	1.94%

The 1st column lists the hierarchy ID, the 2nd gives the name of the root, the 3rd column provides the minimum and maximum sizes during the time evolution, the 4th contains the maximum level depth, and finally the 5th column lists the average fraction of changed links under one year.

links under one time step (one year). In S1 Text we also provide more detailed tables listing the yearly size of the hierarchies, together with the number of added and deleted nodes and links.

The links in our network representation are pointing from the parents to their children. Since a part of the MeSH terms have multiple parents, the studied networks are not strictly tree-like, instead they correspond to a directed acyclic graphs. Due to the yearly updates, the structure of these networks is subject to constant evolution: New MeSH terms can appear, terms becoming obsolete can be deleted or be merged with other terms, and also already existing parts of the network may be rewired. To illustrate these processes, in Fig 1. we show two snapshots from subsequent years, depicting the changes in a small subgraph from the hierarchy A (Anatomy). According to the picture, a relatively large variety of modifications can occur already in a single time step. E.g., 'Cranial Fossa Anterior' is a newly appearing MeSH term, which is classified under 'Skull Base' in Fig 1b. This type of process can be viewed in general as the growth of the hierarchy. Another intuitive process is rewiring, when both the source and the target of a newly appearing link are actually already existing ('old') nodes, such as e.g., the new link between 'Head' and 'Scalp' in Fig 1b. Naturally, links becoming obsolete can also become deleted, as e.g., the link from 'Body regions' to 'Skin' in Fig 1a. There are also somewhat less intuitive change types as well, such as the insertion of a new node into the middle part of a branch, as e.g., the link from 'Upper Extremity' to 'Arm' in Fig 1b, or the appearance of a new link between two new nodes. A detailed classification of the possible change types is given in the Results section.

Measuring preference during attachment or detachment

Our main focus in this paper is on the examination of possible preference with respect to various node properties during the attachment and detachment of the links. The method we use for detecting whether the attachments/detachments are uniform with respect to a given
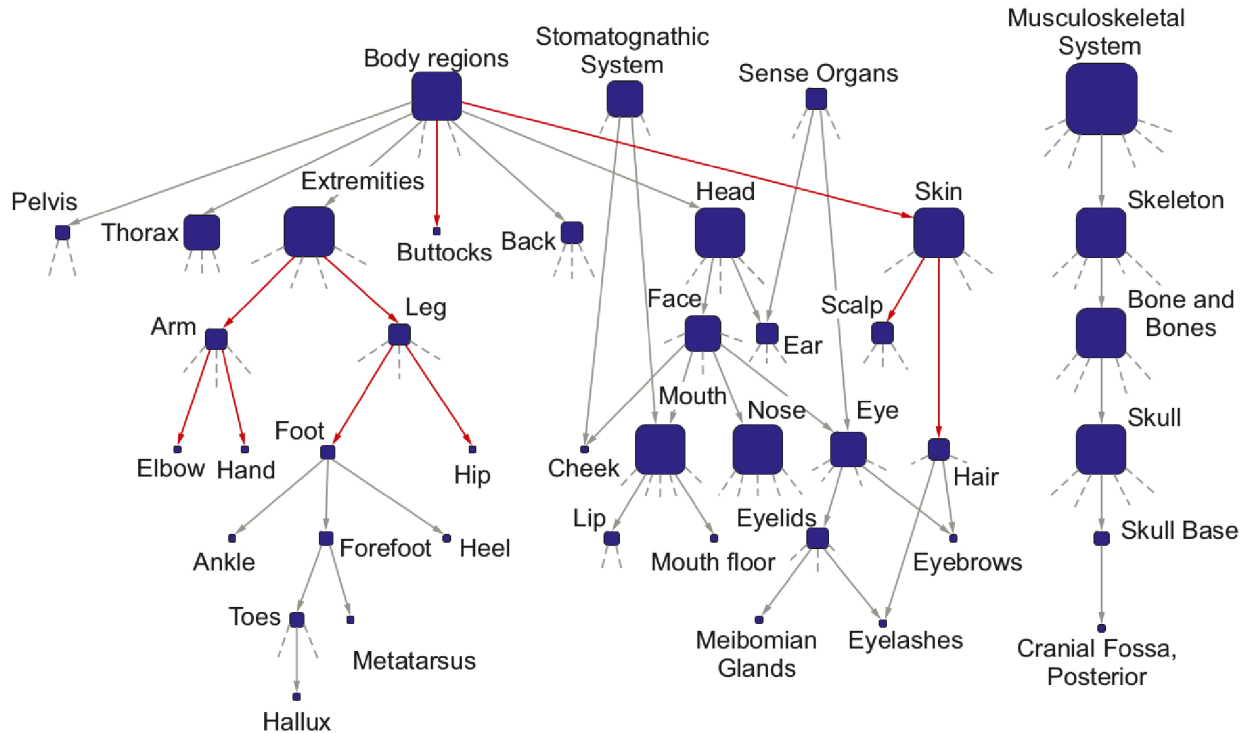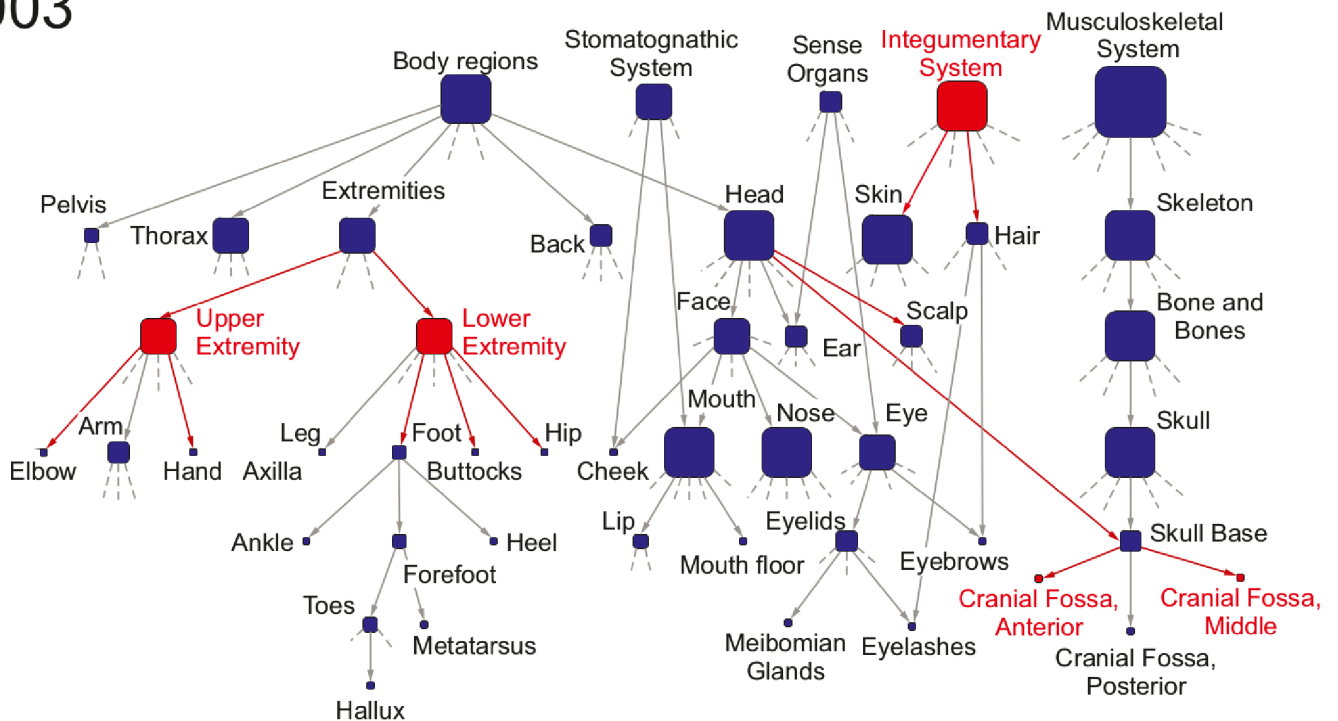
Fig 1. Changes between subsequent time steps in a MeSH hierarchy. a) A small part of the hierarchy 'A' (Anatomy) in 2002. Red links are deleted in the next time step b) The corresponding part of the same hierarchy in 2003. Nodes and links colored red are newly appearing elements.

property x, or instead show preference towards high (or low) values of x is based on comparing the distribution of x for the chosen nodes during the change event and the distribution of x amongst the available nodes [23].

Attachment events.   We begin by discussing attachment events, where (previously non existing) new links appear in the system. For simplicity let us consider first only two consecutive time steps in the data set for a single hierarchy, where we would like to examine whether the choice of nodes in the initial state is preferential or not with regard to x. We denote the probability distribution of x at the initial state by $p(x)$, and the complementary cumulative distribution by $Q(x) = \sum_{x' \geq x} p(x')$, corresponding to the fraction of nodes in the hierarchy having a property value at least as large as x in the initial state. In case the attachment is independent of x, the number of nodes chosen having a property value x or larger is expected to be simply proportional to $Q(x)$. However, if larger values of x are preferred, then nodes having large x value are chosen at a higher frequency compared to what we would expect based on $Q(x)$, and similarly, if lower values of x are preferred, then nodes with large x values are chosen at a lower frequency compared to the expectation based on $Q(x)$. Therefore, by denoting the number of actually chosen nodes having a property value at least as large as x in the attachments as $w(x)$, and taking its ratio compared to $Q(x)$ as

$$W(x) = \frac{w(x)}{Q(x)}, \tag{1}$$

we obtain a function that is constant if the attachment is uniform in x, since in this case $w(x)$ and $Q(x)$ are simply proportional to each other for any x. However, if larger values of x are preferred, the shape of $W(x)$ becomes increasing as a function of x, whereas in the opposite case, when the attachment/detachment prefers lower values of x, the shape of $W(x)$ becomes decreasing.

A noteworthy property of $w(x)$ is that for any fixed value of x, it follows a binomial distribution,

$$P(w(x) = k) = \binom{A}{k} u(x)^k (1 - u(x))^{A-k}, \tag{2}$$

where A is the number of attachment events, and $u(x)$ denotes the probability for choosing a node having a property value at least as large as x. Simplest case is when choosing is independent of the given property, and therefore, $u(x) = Q(x)$. If instead we assume a linear preference with regard to the studied property, $u(x)$ can be expressed as

$$u(x) = \sum_{i:\, x_i \geq x} x_i p(x_i) \Big/ \sum_{j=1}^{A} x_j p(x_j), \tag{3}$$

where the summations run over the nodes in the hierarchy.

In any case, based on (2) the expected value and standard deviation of $w(x)$ can be given as $\langle w(x) \rangle = A \cdot u(x)$ and $\sigma(w(x)) = \sqrt{A \cdot u(x)(1 - u(x))}$, respectively. By moving from $w(x)$ to $W(x)$ we obtain that according to (1) the mean and standard deviation for $W(x)$ can be written as

$$\langle W(x) \rangle = \frac{A \cdot u(x)}{Q(x)}, \tag{4}$$

$$\sigma(W(x)) = \frac{\sqrt{A \cdot u(x)(1 - u(x))}}{Q(x)}, \tag{5}$$

which for an attachment process independent of x take the simple form of

$$\langle W(x) \rangle = A, \tag{6}$$

$$\sigma(W(x)) = \frac{\sqrt{A(1 - Q(x))}}{\sqrt{Q(x)}}. \tag{7}$$

We have tested the behavior of $W(x)$ by simulating $A = 10000$ attachment events on hierarchy D at year 2002, the results are shown in Fig 2. According to the plots, the measured $W(x)$ remained within the standard deviation around the analytically calculated average for both purely random attachments (orange color), and linear preferential attachment with an additive constant.

When applying the above method for measuring preference in the empirical data, for every time step t (except for the last) we can measure the complementary cumulative distribution $Q_t(x)$, and count how many nodes having a property value at least as large as x have been selected by the given attachment mechanism between t and $t + 1$, denoted by $w_t(x)$. By aggregating their ratio in analogy with (1), we can define

$$W_{\mathrm{emp}}(x) = \sum_{t=1}^{t_{\max}-1} \frac{w_t(x)}{Q_t(x)}. \tag{8}$$

This can be compared to e.g., the mean and standard deviation of the random variable corresponding to the sum of the supposed $W(x)$ under the assumption of independence from x, which according to (6 and 7) can be given as

$$\langle W_{\mathrm{rand}}(x) \rangle \quad = \sum_{t=1}^{t_{\max}-1} A_t, \tag{9}$$

$$\sigma(W_{\mathrm{rand}}(x)) \quad = \left[\sum_{t=1}^{t_{\max}-1} \frac{A_t(1 - Q_t(x))}{Q_t(x)} \right]^{\frac{1}{2}}, \tag{10}$$

where A_t denotes the number of attachment events between time steps t and $t + 1$.

Detachment events. An important difference between the addition of new links and link deletion events is that in the latter case, the natural assumption for the random choice (independent of any node property) is choosing a link uniformly at random from all existing links. The different nature of the two processes can be already seen in the domains over which they are defined, namely that attachment can occur between any pair of nodes, whereas detachment events can take place only over already existing links. A noteworthy consequence of this is that the degree of any node taking part in a detachment event is at least one before the event, which is not necessarily the case for attachment events.

When assuming a uniform random choice between the existing links, high degree nodes appear to be involved in the link deletion events with higher probability compared to low degree nodes simply because they have a higher number of connections. To take this into account, we have to redefine the formula of $Q(x)$ for detachment events. First let us consider the case, where we are interested in whether some property of a node x has an effect on the
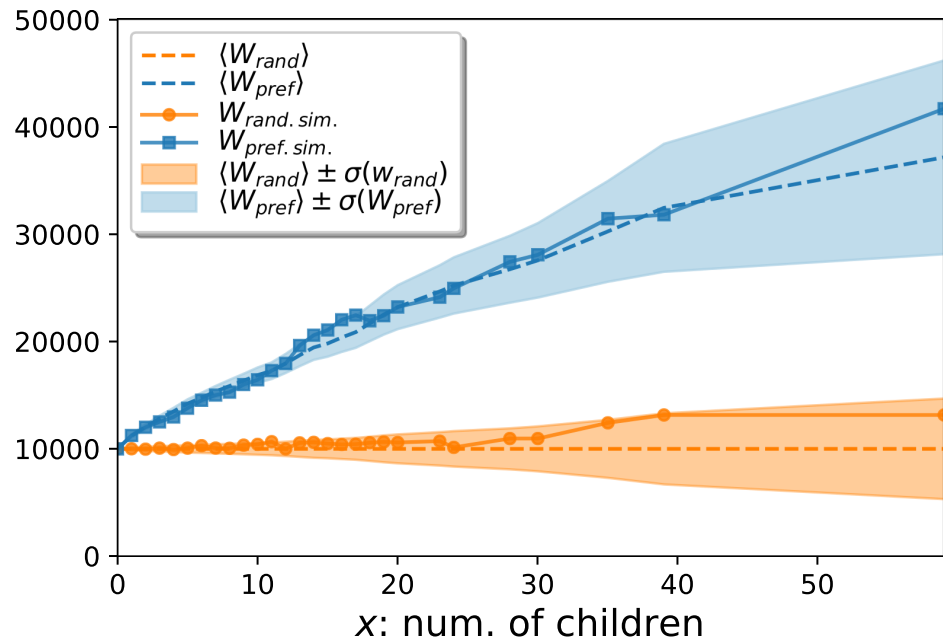
Fig 2. Testing $W(x)$ by simulated attachments. The property x here corresponds to the number of children, and the full symbols connected by continuous lines show the measured $W(x)$ for random attachment (independent of x) in orange (circles), and for preferential attachment with an additive constant (i.e. when a newly added node connects to node i with a probability $\frac{k_i+a}{\sum_i k_i+a}$ where a is an arbitrary constant) in blue (squares). Dashed lines correspond to the analytic mean for $W(x)$, whereas the shaded areas indicate the standard deviation around the mean.

likelihood that an out link is detached from it (also meaning that from the point of view of the deleted link, this node plays the role of source node). If we choose at random from all possible links, the probability that we pick an out link from a node with out degree k_{out} is given by k_{out} $p(k_{out})/\langle k_{out}\rangle$ (also referred to as access degree), where $p(k_{out})$ denotes the out degree distribution, and $\langle k_{out}\rangle$ is the average out degree (which is the same as the average in degree). Combining the previous arguments with the law of total probability, the probability distribution for property x on the source node of randomly selected links can be written as

$$p_{out}(x) = \frac{1}{\langle k_{out}\rangle}\sum_{k_{out}} p(x \mid k_{out})k_{out}p(k_{out}), \qquad (11)$$

where $p(x \mid k_{out})$ denotes the conditional probability that the property value is x, given that the out degree of the node is k_{out}. Note that in the special case of a sharply peaked degree distribution around a characteristic value $k_{out} \approx \langle k_{out}\rangle$, the above expression for $p_{out}(x)$ coincides with the analogous formula of $p(x)$ for attachment events. Based on $p_{out}(x)$, the complementary cumulative distribution $Q_{out}(x)$ can be calculated as usual,

$$Q_{out}(x) = \sum_{x'\geq x} p_{out}(x'). \qquad (12)$$

If in contrast to out links, we are interested in the deletion of incoming links and the possible effect on the likelihood of such events by some node property x, we can formulate analogous formulas to the above using the in degree distribution $p(k_{in})$. In this case the probability

distribution for property x on the target node of randomly selected links can be written as

$$p_{\mathrm{in}}(x) = \frac{1}{\langle k_{\mathrm{in}} \rangle} \sum_{k_{\mathrm{in}}} p(x \mid k_{\mathrm{in}}) k_{\mathrm{in}} p(k_{\mathrm{in}}), \tag{13}$$

where $p(x \mid k_{\mathrm{in}})$ denotes the conditional probability that the property value is x, given that the in degree of the node is k_{in}, and the corresponding complementary cumulative distribution is given by

$$Q_{\mathrm{in}}(x) = \sum_{x' \geq x} p_{\mathrm{in}}(x'). \tag{14}$$

Otherwise, the analysis for the link deletion events is the same as in case of the attachment events: We can calculate $Q_t(x)$ using either (12) or (14), and by plugging the result together with the observed $w_t(x)$ into (8) we obtain $W_{\mathrm{emp}}(x)$. To decide whether we can speak about a possible preference or anti-preference with respect to the chosen property, $W_{\mathrm{emp}}(x)$ has to be compared to the $W(x)$ expected based on neutral behaviour, calculated using (9 and 10).

Results

We applied the methodology outlined in the previous section to study the time evolution of the hierarchies listed in Table 1 where the system size exceeds 1000 nodes during the whole recorded time period, corresponding to hierarchies A, B, C, D, E, G, and N. Before actually showing the results, first we need to specify the different possible attachment and detachment event types. In terms of the changing links we have two large categories: added (new) links and deleted links. When examining the endpoints of added links, both the source and the target can be either an already existing (old) node, or a new node, thus, there are altogether 4 types of added links. The case of deleted links is much simpler in this respect, as both endpoints must correspond to old nodes. Therefore, there are in total 5 different possibilities for changes in the connections. However, when examining the possible effect of a given node property on the likelihood that the node is going to take part in an attachment/detachment event, we also have to specify whether the node is the source or the target of the involved link. Thus, for any node property of interest we can examine 10 different scenarios over the time evolution of the hier-archies. Naturally, when interested in the possible effect of a node property of an old node, the value of the property is always measured before the link change event (e.g., if the change occurs between time steps t and $t + 1$, then it is recorded at t), whereas for new nodes we can only measure their properties at the time point of their appearance (i.e., at $t + 1$ for link change events between t and $t + 1$). We list the yearly frequencies of the different event types for the studied hierarchies in Tables A-G in S1 Text.

In our studies we focused on the following properties: number of children (out degree), number of parents (in degree), total number of descendants, total number of ancestors. As an illustration, in Fig 3. we show parts of the results obtained for hierarchies D, C and G. In Fig 3a the $W_{\mathrm{emp}}(x)$ is plotted for hierarchies C and D, obtained from events where a new link point-ing to a new node is attached to an old node, and x is corresponding to the total number of descendants of the source node. The curves indicate strong preference for large values of x, as they clearly exceed $W_{\mathrm{rand}}(x) + \sigma(W_{\mathrm{rand}}(x))$ by an order of magnitude. Interestingly somewhat the opposite can be seen in Fig 3b, showing the results for the same hierarchies in case of inser-tion of new links between pairs of already existing nodes, where x is corresponding to the total number of ancestors of the source node. The fact that $W_{\mathrm{emp}}(x)$ is way below $W_{\mathrm{rand}}(x) - \sigma(W_{\mathrm{rand}}(x))$ indicates that the probability for the attachment of an incoming link to a node with higher number of ancestors is lower than what we would expect at random. In Fig 3c we
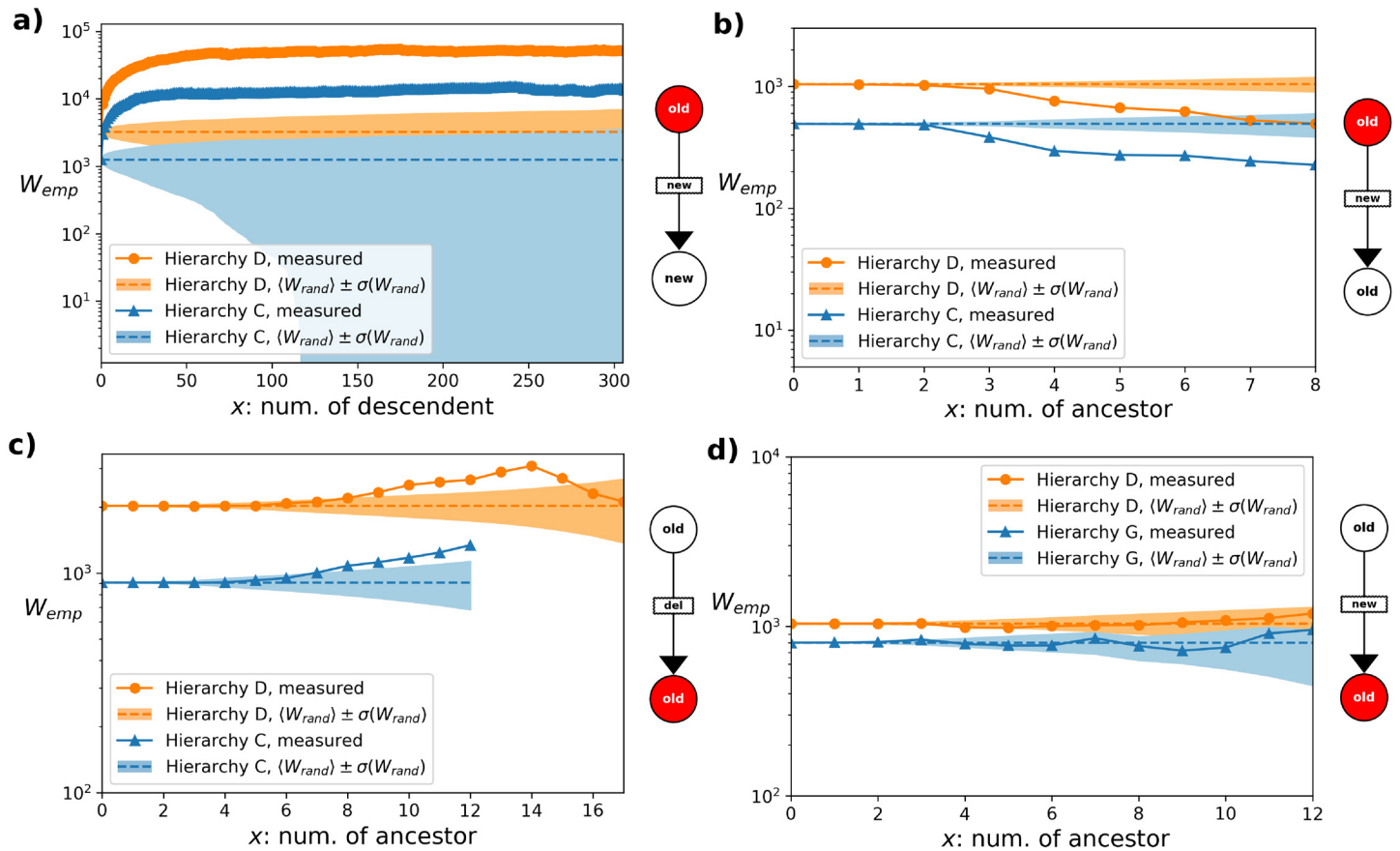
Fig 3. Measuring preference in attachment and detachment events. In each panel we compare $W_{emp}(x)$ defined in (8) to the mean and standard deviation of $W(x)$ for random events, given in (9 and 10) and indicated by dashed lines in shaded areas. The pictograms beside the panels show the type of the studied attachment/detachment events and highlight in red whether the given property x was measured on the source or on the target of the links involved in the events. a) Results for the total number of descendants of source nodes in attachments of new links pointing from old nodes to new nodes in hierarchies D (orange) and C (blue). b) $W_{emp}(x)$ for the number of ancestors of source nodes on new links appearing between old nodes, measured in hierarchies D (orange) and C (blue). c) The same plots when x is equal to the number of ancestors of the target nodes in link deletion events for hierarchies D (orange) and C (blue). d) $W_{emp}(x)$ in case x is corresponding to the number of ancestors of the target node in attachment of new links between old nodes.

considered link deletion events, and according to the results $W_{emp}(x)$ shows a non-monotonous behaviour as a function of x for the total number of ancestors of the target node in case of hierarchy D and a weak preference in case of hierarchy C. The peak in $W_{emp}(x)$ for hierarchy D is suggesting that there is a preferred value of x, where the likelihood of the node taking part in the given type of detachment event is maximal. Finally, in Fig 3d we show the results for the insertion of new links between old nodes (similarly to Fig 3b), but this time we depict $W_{emp}(x)$ for the total number of ancestors of the target node in case of hierarchies D and G. According to the results $W_{emp}(x)$ runs within the range of the standard deviation around $\langle W_{rand}(x) \rangle$, thus, this type of attachment does not show any preference with respect to the number of ancestors of the target node.

Similar plots for the rest of the attachment/detachment types and for the other hierarchies are given in S1 Text. Based on the seen behaviour of $W_{emp}(x)$ we can categorise the observed behaviour as follows:

- Strong indication of preference (s+): $W_{emp}(x)$ shows a monotonous increasing behaviour, and exceeds $\langle W_{rand}(x) \rangle + \sigma(W_{rand}(x))$ by a large amount, (as e.g., in case of Fig 3a).

- Weak indication of preference (w+): $W_{\mathrm{emp}}(x)$ shows a monotonous increasing behaviour, exceeds $\langle W_{\mathrm{rand}}(x)\rangle + \sigma(W_{\mathrm{rand}}(x))$, but only by a small amount.

- Strong indication of no preference (s0): $W_{\mathrm{emp}}(x)$ remains within the standard deviation around $\langle W(x)\rangle$, (as e.g., in case of Fig 3d).

- Weak indication of anti-preference (w-): $W_{\mathrm{emp}}(x)$ shows a monotonous decreasing behaviour, and falls under $\langle W_{\mathrm{rand}}(x)\rangle - \sigma(W_{\mathrm{rand}}(x))$, by a small amount.

- Strong indication of anti-preference (s-): $W_{\mathrm{emp}}(x)$ shows a monotonous decreasing behaviour, and falls under $\langle W_{\mathrm{rand}}(x)\rangle - \sigma(W_{\mathrm{rand}}(x))$, by a larger amount, (as e.g., in case of Fig 3b)

- Indication of preference with a peak (p+): $W_{\mathrm{emp}}(x)$ shows a non-monotonous behaviour, and has a maximum exceeding $\langle W_{\mathrm{rand}}(x)\rangle + \sigma(W_{\mathrm{rand}}(x))$ by a significant amount, (as e.g., in case of Fig 3c).

- Indication of anti-preference with a peak (p-): $W_{\mathrm{emp}}(x)$ shows a non-monotonous behaviour, and has a minimum falling under $\langle W_{\mathrm{rand}}(x)\rangle - \sigma(W_{\mathrm{rand}}(x))$ by a significant amount.

- Insufficient statistics (i.s): in a number of cases it is not possible to draw a conclusion based on the empirical data. This may be due to the fact that the given type of attachment/detachment occurs rarely, or because that the distribution of the given node property is extremely narrow, resulting in a very limited range for x.

In Table 2 we give a summary overview of the results for the largest hierarchy (corresponding to hierarchy D), where the table is organised as follows: rows are corresponding to the 4 studied node properties, measured either on the source node (top 4 rows) or the target node (bottom 4 rows) of the changing links, and the table columns indicate the attachment/detachment types. In each cell we provide the category of the observed behaviour based on the corresponding plot. For example, the 3rd cell in the 3rd row is based on the orange curve in Fig 3a, the 4th cell in the 4th row is connected to Fig 3b, the 4th cell in the last row is corresponding to Fig 3d, etc.

The overall pattern of the different preference types in Table 2 is highly non-trivial. E.g., all possible link change event types show preference with respect to the number of children of the source node (first row in Table 2), and all except for two (addition of new links between new nodes and deletion of old links) show anti-preference with respect to the total number of

Table 2. Summary of the results for hierarchy D. The columns of the table correspond to the studied different link types, and the rows indicate the studied node property on either the source (top 4 rows) or the target (bottom 4 rows). The 3rd, 4th and 5th columns correspond to impossible link types, therefore, are left empty. The entries in the cells correspond to the following abbreviations: 's+', 's0' and 's-' for strong indication of preference, no preference and anti-preference, 'p+' and 'p-' for indication of preference or anti-preference with a peak, and 'i.s' for insufficient statistics.

D		link: add				link: del			
		source: new		source: old		source: new		source: old	
		target: new	target: old	target: new	target: old	target: new	target: old	target: new	target: old
source	child.	s+	s+	s+	s+				s+
	par.	i.s.	i.s.	s−	i.s.				i.s.
	desc.	p−	s+	s+	s+				p+
	anc.	s0	s−	s−	s−				s0
target	child.	s−	i.s.	s−	s0				s0
	par.	i.s.	i.s.	s+	s0				s+
	desc.	s−	i.s.	s−	s0				s0
	anc.	s+	s0	s+	s0				p+

ancestors of the source node (4th row in Table 2). Interestingly, in the 3rd row of Table 2 (corresponding to the number of descendants of the source node) both preference and anti-preference is occurring among the cells corresponding to the different link change types. Seemingly the properties of the target nodes (bottom 4 rows) have a smaller effect compared to the properties of the source nodes (top 4 rows), indicated by the higher number of cells falling into the category of evidence for no-preference (s0). Nevertheless, preference with respect to the number of parents and number of ancestors, and anti-preference with respect to the number of children and number of descendants can be seen for a couple of the link change types.

Summary tables analogous to Table 2 for the other hierarchies are listed in the S1 Text. In order to be able to draw conclusions on the general features of the evolution of the studied hierarchies, we also provide an aggregated table with the same cell structure, in which the contribution from the individual tables were averaged in a simple manner, as shown in Table 3. According to that we can make the following observations about the presence of preference or anti-preference with respect to the different node properties during the growth and restructuring of the studied hierarchies:

- We can see strong signs of preference with respect to the number of children of the source node for both the addition of new links pointing from old nodes to new ones, and for the deletion of already existing links between old nodes.

- These link change events together with the addition of new links between already existing links clearly show preference with respect to the total number of descendants of the source node as well.

- All possible link change types show anti-preference with respect to the total number of ancestors of the source node. This effect is strong in case of addition of new links with an old source node, and for adding new links pointing from new nodes to old ones, whereas can be considered somewhat less pronounced for new links between two new nodes, and relatively weak for link deletions.

- We can see both preference, neutral behaviour and anti-preference with respect to the total number of ancestors of the target node: the addition of new links pointing to new nodes and link deletions seem to display a weak preference, the addition of new links between old

Table 3. Aggregated summary results. Based on Table 2. and Tables H-N in S1 Text, the contribution to a given cell is counted according to 's+' = 1, 'w+' = 'p+' = 0.5, 's0 = 0', 'w−' = 'p−' = -0.5, 's−' = -1, and the obtained sum is divided by the number of tables contributing to the given cell. Aggregated cells become 'i.s' if more than 3 out of the 7 tables has 'i.s.' as well.

		link: add				link: del			
s− neutral s+									
	Σ	source: new		source: old		source: new		source: old	
		target: new	target: old	target: new	target: old	target: new	target: old	target: new	target: old
source	child.	i.s.	i.s.	1.0,	i.s.				0.67,
	par.	i.s.	i.s.	i.s.	i.s.				i.s.
	desc.	i.s.	i.s.	0.89,	0.83,				0.61,
	asc.	-0.44,	-0.75,	-0.90,	-0.78,				-0.28,
target	child.	i.s.	i.s.	i.s.	i.s.				i.s.
	par.	i.s.	i.s.	i.s.	i.s.				i.s.
	desc.	i.s.	i.s.	i.s.	i.s.				i.s.
	asc.	0.24,	-0.11,	0.29,	0.0,				0.29,

nodes displays neutral behaviour, whereas in case of the addition of new links pointing from new nodes to old ones, we can observe a weak anti-preference.

- The attachment/detachment processes seem to be more influenced by the properties of the source node of the changing links, compared to the influence of the properties of the target nodes. This is supported by the fact that the top 4 row in Table 3 contains much higher number of cells with values (other than 'i.s.'), and the magnitude of these is larger on average compared to cells in the bottom 4 rows.

An important further point to note is that the different hierarchies showed consistency in the sense that both preference and anti-preference was never observed simultaneously when comparing the same cells across the different summary tables.

Discussion

We studied the change mechanisms of time evolving hierarchies between the PubMed MeSH terms using statistical methods. Although previous research has already shown interesting results regarding the growth of these networks [49–52], an important conclusion we can make based on our analysis is that deletion events and rewiring between already existing parts of the system are equally important in shaping the form of these hierarchies. This is supported by Tables A-G in S1 Text, according to which the number of deleted links together with the number of new links between already existing nodes under one time step is usually of the same magnitude as the number of new links connected to newly appearing nodes.

The main focus of our studies was on measuring preference during attachment and detachment events with respect to four different node properties characterising the hierarchy members. By setting up a general framework for this sort of analysis we could show that the likelihood for nodes to take part in restructuring events can be effected by their properties under quite a number of different circumstances. We found that when new links appear pointing from already existing nodes to newly appearing ones, the nodes with larger number of children (larger out degree) are chosen as source nodes for this type of links with significantly larger probabilities compared to uniform random choice. This effect is analogous to the preferential attachment rule of the Barabási–Albert network model [43], which was also observed empirically in different growing network systems [44–47]. However, in our case a larger number of children also increases the likelihood of loosing an out link (corresponding to a link deletion event).

Another property for which we observed similar behaviour is the total number of descendants, where in addition to the above two effects we could also detect preference during the addition of a new links pointing to other already existing nodes. In parallel, we observed anti-preference with respect to the number of ancestors of the source node for all possible link change types. Interestingly, if the node acts as the target of the changing link, we can observe both preference and anti-preference with respect to the number of ancestors for the different link change types. Since the number of descendants and the number of ancestors are defined only in case of hierarchies, the related results have no previously observed analogy in general time dependent networks.

From the practical point of view, the above findings also imply that based on the topological and hierarchical properties of the MeSH terms, it is possible to locate nodes in the present state of a MeSH hierarchy for which the likelihood of taking part in link change events in the future is higher compared to other nodes in the system. For the larger hierarchies, this sort of 'flagging' of the MeSH terms might help the curators carrying out the yearly update by providing candidates who's neighborhood should be checked. However, it is also quite plausible that

the inclusion of further properties not included in the present study (such as e.g., the frequency of the queries by the users) may turn out to be vital in further studies along this direction.

Finally, we note that according to Table 2 and Tables H-N in S1 Text one can observe a mild variance across the different hierarchies in terms of whether a given link change type displays some sort of preference with respect to a given property, or we see a neutral behaviour (or insufficient statistics) instead. Nevertheless, the results across the different hierarchies are consistent in the sense that we cannot observe both preference in case of one hierarchy, and anti-preference in case of another hierarchy for the same link change type and node property. This consistency is encouraging from the point of view of further research focusing on building network models for time evolving hierarchies. In addition, we note that although the empirical studies in this work are restricted to the networks between MeSH terms, it is quite plausible that a part of these features are more universal and occur in time evolution of networks with a hierarchical structure in general.

Conclusion

In summary, our findings show that the growth and rewiring of the examined hierarchies are governed by non-trivial preference in the attachment mechanisms of the links. According to our results, the attachment is non-uniform with respect to multiple different topological and hierarchical node properties, and among the different possible link change scenarios we could observe both preferential and anti-preferential attachments, depending on the given node property of interest. These facts indicate that time evolution of these systems is far more complex compared to simple preferential attachment models, providing very interesting future challenges for modelling and further statistical analysis.

Supporting information

S1 Text. Supporting information.
(PDF)

Author Contributions

Conceptualization: Sámuel G. Balogh, Gergely Palla.

Data curation: Dániel Zagyva, Péter Pollner.

Formal analysis: Sámuel G. Balogh, Dániel Zagyva, Péter Pollner, Gergely Palla.

Funding acquisition: Péter Pollner, Gergely Palla.

Investigation: Sámuel G. Balogh, Dániel Zagyva, Péter Pollner, Gergely Palla.

Methodology: Sámuel G. Balogh, Gergely Palla.

Project administration: Péter Pollner.

Visualization: Sámuel G. Balogh, Dániel Zagyva, Gergely Palla.

Writing – original draft: Sámuel G. Balogh, Péter Pollner, Gergely Palla.

Writing – review & editing: Sámuel G. Balogh, Péter Pollner, Gergely Palla.

References

1. Albert R, Barabási AL. Statistical mechanics of complex networks. Rev Mod Phys. 2002; 74:47–97. https://doi.org/10.1103/RevModPhys.74.47

2. Mendes JFF, Dorogovtsev SN. Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford: Oxford Univ. Press; 2003.

3. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical Organization of Modularity in Metabolic Networks. Science. 2002; 297:1551—1555. https://doi.org/10.1126/science.1073374 PMID: 12202830

4. Clauset A, Moore C, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. Nature. 2008; 453:98–101. https://doi.org/10.1038/nature06830 PMID: 18451861

5. Pumain D. Hierarchy in Natural and Social Sciences. vol. 3 of Methodos Series. Dodrecht, The Netherlands: Springer Netherlands; 2006.

6. Corominas-Murtra B, Rodríguez-Caso C, Goñi J, Solé R. Measuring the hierarchy of feedforward networks. Chaos. 2011; 21:016108. https://doi.org/10.1063/1.3562548 PMID: 21456850

7. Zafeiris A, Vicsek T. Why We Live in Hierarchies? A Quantitative Treatise. Berlin: Springer; 2018.

8. Trusina A, Maslov S, Minnhagen P, Sneppen K. Hierarchy measures in complex networks. Phys Rev Lett. 2004; 92:178702. https://doi.org/10.1103/PhysRevLett.92.178702 PMID: 15169201

9. Mones E, Vicsek L, Vicsek T. Hierarchy Measure for Complex Networks. PLoS ONE. 2012; 7:e33799. https://doi.org/10.1371/journal.pone.0033799 PMID: 22470477

10. Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C. On the origins of hierarchy in complex networks. Proc Natl Acad Sci USA. 2013; 110:13316–13321. https://doi.org/10.1073/pnas.1300832110 PMID: 23898177

11. Czégel D, Palla G. Random walk hierarchy: What is more hierarchical, a chain a tree or a star? Scientific Reports. 2015; 5:17994. https://doi.org/10.1038/srep17994 PMID: 26657012

12. Gupte M, Shankar P, Li J, Muthukrishnan S, Iftode L. Finding hierarchy in directed online social networks. In: Proceedings of the 20th international conference on World wide web. ACM; 2011. p. 557–566.

13. Letizia E, Barucca P, Lillo F. Resolution of ranking hierarchies in directed networks. PLOS ONE. 2018; 13(2):1–25. https://doi.org/10.1371/journal.pone.0191604

14. Ma HW, Buer J, Zeng AP. Hierarchical sructure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics. 2004; 5:199. https://doi.org/10.1186/1471-2105-5-199 PMID: 15603590

15. Goessmann C, Hemelrijk C, Huber R. The formation and maintenance of crayfish hierarchies: behavioral and self-structuring properties. Behav Ecol Sociobiol. 2000; 48:418–428. https://doi.org/10.1007/s002650000222

16. Nagy M, Akos Z, Biro D, Vicsek T. Hierarchical group dynamics in pigeon flocks. Nature. 2010; 464:890–893. https://doi.org/10.1038/nature08891 PMID: 20376149

17. Nagy M, Vásárhelyi G, Pettit B, Roberts-Mariani I, Vicsek T, Biro D. Context-dependent hierarchies in pigeons. Proc Natl Acad Sci USA. 2013; 110:13049–13054. https://doi.org/10.1073/pnas.1305552110 PMID: 23878247

18. Ozogány K, Vicsek T. J Stat Phys. 2015; 158:628. https://doi.org/10.1007/s10955-014-1131-7.

19. Fushing H, McAssey MP, Beisner B, McCowan B. Ranking network of captive rhesus macaque society: A sophisticated corporative kingdom. PLoS ONE. 2011; 6:e17817. https://doi.org/10.1371/journal.pone.0017817 PMID: 21423627

20. Kaiser M, Hilgetag CC, Kötter R. Hierarchy and dynamics of neural networks. Front Neuroinform. 2010; 4:112. https://doi.org/10.3389/fninf.2010.00112 PMID: 20844605

21. Palla G, Tibély G, Mones E, Pollner P, Vicsek T. Hierarchical networks of scientific journals. Palgrave Communications. 2015; 1:15016. https://doi.org/10.1057/palcomms.2015.16

22. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. Phys Rev E. 2003; 68:065103. https://doi.org/10.1103/PhysRevE.68.065103

23. Pollner P, Palla G, Vicsek T. Preferential attachment of communities: The same principle, but a higher level. Europhys Lett. 2006; 73:478–484. https://doi.org/10.1209/epl/i2005-10414-6

24. Valverde S, Solé RV. Self-organization versus hierarchy in open-source social networks. Phys Rev E. 2007; 76:046118. https://doi.org/10.1103/PhysRevE.76.046118

25. Tóth BJ, Palla G, Mones E, Havadi G, Páll N, Pollner P, et al. Emergence of Leader-Follower Hierarchy Among Players in an On-Line Experiment. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2018. p. 1184–1190.

26. Krugman PR. Confronting the mystery of urban hierarchy. J Jpn Int Econ. 1996; 10:399–418. https://doi.org/10.1006/jjie.1996.0023

27. Batty M, Longley P. Fractal Cities: A Geometry of Form and Function. San Diego: Academic; 1994.

28. Tibély G, Sousa-Rodrigues D, Pollner P, Palla G. Comparing the Hierarchy of Keywords in On-Line News Portals. PLoS ONE. 2016; 11:e0165728. https://doi.org/10.1371/journal.pone.0165728 PMID: 27802319

29. Hirata H, Ulanowicz R. Information theoretical analysis of the aggregation and hierarchical structure of ecological networks. J Theor Biol. 1985; 116:321–341. https://doi.org/10.1016/S0022-5193(85) 80271-X

30. Wickens J, Ulanowicz R. On quantifying hierarchical connections in ecology. J Soc Biol Struct. 1988; 11:369–378. https://doi.org/10.1016/0140-1750(88)90066-8

31. Eldredge N. Unfinished Synthesis: Biological Hierarchies and Modern Evolutionary Thought. New York: Oxford Univ. Press; 1985.

32. McShea DW. The hierarchical structure of organisms. Paleobiology. 2001; 27:405–423. https://doi.org/10.1666/0094-8373(2001)027%3C0405:THSOOA%3E2.0.CO;2

33. Mengistu H, Huizinga J, Mouret JB, Clune J. The Evolutionary Origins of Hierarchy. PLOS Computational Biology. 2016; 12(6):1–23. https://doi.org/10.1371/journal.pcbi.1004829

34. Asllani M, Lambiotte R, Carletti T. Structure and dynamical behavior of non-normal networks. Science Advances. 2018; 4(12). https://doi.org/10.1126/sciadv.aau9403 PMID: 30547090

35. Katchanov YL, Markova YV. The "space of physics journals": topological structure and the Journal Impact Factor. Scientometrics. 2017; 113(1):313–333. https://doi.org/10.1007/s11192-017-2471-2

36. Chen C. Science Mapping: A Systematic Review of the Literature. Journal of Data and Information Science. 2017; 2(2):1–40. https://doi.org/10.1515/jdis-2017-0006.

37. Chen C, Song M. In: Science Mapping Tools and Applications. Springer, Cham; 2017.

38. Alvial-Palavicino C, Konrad K. The rise of graphene expectations: Anticipatory practices in emergent nanotechnologies. Futures. 2018; https://doi.org/10.1016/j.futures.2018.10.008.

39. Shan W, Wang J. Mapping the Landscape and Evolutions of Green Supply Chain Management. Sustainability. 2018; 10(3). https://doi.org/10.3390/su10030597

40. Zafeiris A, Vicsek T. Group performance is maximized by hierarchical competence distribution. Nature Communications. 2013; 4:2484. https://doi.org/10.1038/ncomms3484. PMID: 24048260

41. Zamani M, Vicsek T. Glassy nature of hierarchical organizations. Scientific Reports. 2017; 7:1382. https://doi.org/10.1038/s41598-017-01503-y PMID: 28469242

42. Zamani M, Camargo-Forero L, Vicsek T. Stability of glassy hierarchical networks. New Journal of Physics. 2018; 20(2):023025. https://doi.org/10.1088/1367-2630/aaa8ca

43. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286:509–512. https://doi.org/10.1126/science.286.5439.509 PMID: 10521342

44. Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A. 2002; 311:590–614. https://doi.org/10.1016/S0378-4371(02)00736-7

45. Jeong H, Néda Z, Barabási AL. Measuring preferential attachment in evolving networks. EPL. 2003; 61:567. https://doi.org/10.1209/epl/i2003-00166-9

46. Newman MEJ. Clustering and preferential attachment in growing networks. Phys Rev E. 2001; 64:025102(R). https://doi.org/10.1103/PhysRevE.64.025102

47. Pollner P, Palla G, Vicsek T. Preferential attachment of communities: The same principle, but a higher level. EPL. 2006; 73:478. https://doi.org/10.1209/epl/i2005-10414-6

48. Palla G, Barabási AL, Vicsek T. Quantifying social group evolution. Nature. 2007; 446:664–667. https://doi.org/10.1038/nature05670 PMID: 17410175

49. Tsatsaronis G, Macari N, Torge S, Dietze H, Schroeder M. A Maximum-Entropy approach for accurate document annotation in the biomedical domain. Journal of Biomedical Semantics. 2012; 3:S2. https://doi.org/10.1186/2041-1480-3-S1-S2 PMID: 22541593

50. de Leenheer P. 5. In: Hepp M, de Leenheer P, Moor AD, Sure Y, editors. Ontology evolution. US: Springer; 2008. p. 131–176.

51. McCray AT, Lee K. In: Küppers BO, Hahn U, Artmann S, editors. Taxonomic Change as a Reflection of Progress in a Scientific Discipline. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 189–208.

52. Tsatsaronis G, Varlamis I, Kanhabua N, Nørvåg K. In: Gelbukh A, editor. Temporal Classifiers for Predicting the Expansion of Medical Subject Headings. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 98–113.

53. In this study we use publicly available data from the website of PubMed;. Available from: https://www.nlm.nih.gov/mesh/filelist.html.