Modelling students' knowledge organisation: Genealogical conceptual networks

Ismo T. Koponen *, Maija Nousiainen

Department of Physics, P.O. Box 64, FI-00014 University of Helsinki, Finland

Abstract

Learning scientific knowledge is largely based on understanding what are its key concepts and how they are related. The relational structure of concepts also affects how concepts are introduced in teaching scientific knowledge. We model here how students organise their knowledge when they represent their understanding of how physics concepts are related. The model is based on assumptions that students use simple basic linking-motifs in introducing new concepts and mostly relate them to concepts that were introduced a few steps earlier, i.e. following a genealogical ordering. The resulting genealogical networks have relatively high local clustering coefficients of nodes but otherwise resemble networks obtained with an identical degree distribution of nodes but with random linking between them (i.e. the configuration-model). However, a few key nodes having a special structural role emerge and these nodes have a higher than average communicability betweenness centralities. These features agree with the empirically found properties of students' concept networks.

Key words: Concept networks, Directed networks, Modelling

1 Introduction

Learning scientific knowledge and its structures involves understanding what its key concepts are, and how these concepts are related and connected as part of a system of other concepts. Such interconnections between concepts have also an essential role in establishing their meaning. Scientific knowledge thus forms a system of networked concepts and higher conceptual structures (e.g. laws and models). Learning such knowledge is also learning to construct and map possible conceptual connections in that system [1–3]. The structure

* Corresponding author
Email address: ismo.koponen@helsinki.fi (Ismo T. Koponen).

of the knowledge system also affects how concepts are introduced in teaching scientific knowledge and how they are acquired in formal teaching and learning [4–8]. Scientific knowledge may appear to be complex and its concepts entangled in many ways with other concepts, but in learning the basic forms of knowledge organization and acquisition may be based on the use of simple patterns. Recent cognitively oriented research on learning suggests that procedures of knowledge construction and processing are often reducible to simple basic patterns of diverse types of hierarchies, cliques, transitive patterns and cycles [9–13].

The relational structure of scientific knowledge and knowledge as a networked system is effectively and transparently illustrated by using complex network methods as a kind of cartography of knowledge [14–16]. Network methods have also been applied to model scientific discovery [17,18]. Such approaches are also well adapted to the related problems of knowledge retrieval and acquisition, and of learning about knowledge [19–21]. Linguistic and lexical structures and their learning has also been approached from a viewpoint of networks, with results that show the importance of relational connections between words in learning their meaning [22,23]. Learning scientific knowledge, its terms and concepts and their syntactic structures apparently share many similarities with these recently explored fields of knowledge processing and acquisition. These notions have encouraged the idea that paying attention to such patterns may help in understanding the cognitive processes behind knowledge construction and its acquisition, and may also lead to the development of computational models for cognitive processes in learning [11–13].

Empirical studies of students' knowledge of physics concepts, investigated through using concept maps and networks and related techniques, have revealed that students' declarative (i.e. expressible in terms of writing or symbolically) knowledge is structured. The structure, however, is not hierarchical as often assumed [4] but instead web-like and contingent [24–27]. Such web-like concept networks drawn by students are locally tightly connected and have highly clustering cliques and globally long paths that connect several concepts [24–27]. The networks that are at the same time coherent and contingent provide diverse ways for students to conceptualise the system of scientific knowledge. Two properties of much interest in that context are the relatively high local clustering of concept-nodes in the networks, and the role of communicability and the communicability betweenness centrality. The high local clustering appears to be related to how students use auxiliary concepts while the high communicability centrality is typical for globally important key concepts that provide the overall coherence of the concept network [25,27].

Here we concentrate on the problem of learning scientific knowledge and its concepts, and ask how the relational structure of physics concepts, as it is found in empirical studies, can be modelled as a network. Generative models that produce high clustering operate typically by explicitly controlling the addition of triangular subgraphs in the networks [28], or in terms of controlling the rewiring that favours formation triangular subgraphs [29]. The generative mechanism is also at the core of the present model. In addition, the contiguous paths to introduce new concepts sequentially in the network are of importance. The contiguity of the real concept networks is modelled by assuming that students introduce new concept in the networks by following a simple genealogical strategy, relating new concept preferably to concepts introduced only a few generation steps before, and by using a handful of basic linking-motifs; new knowledge that becomes added to the network is processed by using the recently processed knowledge and simple patterns. The model shares similarities with ordered (or directed) acyclic graphs [21,28,30,31] and models used to describe networks of science and scientific knowledge [15–18]. The results of the model suggest that knowledge ordering and processing by human learners may indeed take place in terms of very simple patterns, while more extensive and complex structures are outgrowths of combinations of these simple patterns. This simplicity of knowledge organisation strategies opens up interesting possibilities for computational, network based modelling of knowledge acquisition in learning.

2 Concept networks: Empirical sample

The concept networks taken here as empirical examples to test the model are constructed by physics students, based on their thinking how concepts in electricity and magnetism (including electromagnetic induction) are related. Our sample of 12 different concept maps, each one made by a different student, comes from a 7-week teacher preparation course, where topics in electricity and magnetism were discussed at the level of a first-year introductory university course. The task was to represent how concepts, principles and laws could be introduced as part of a growing conceptual network by using the key experiments and models, as they are discussed in textbooks and known to students. The number of nodes and links students included in their concept networks was not restricted. The empirical sample of networks to which we compare our model consist only of 12 networks, where the smallest network has 44 nodes (concepts) and 64 links, while the most extensive has 69 concepts and 129 links. On average, the networks have 60 nodes and 95 links. The analysis of the structure of the 12 concept networks has revealed that students tend to connect a concept to a small number (from 2 to 4 most often, with an average value of 3) of other concepts [27]. An example of the concept network made by students is shown in Figure 1. Some of the most important concepts in the 12 networks are provided in Table I.

The design of the concept networks was based on simple rules so that the nodes in the map were required to be: 1) concepts or quantities; 2) laws; 3) models or 4) experiments. The links, on the other hand, were required to be actions or procedures, e.g., change in a quantity, setting the value of a quantity, or determining the value of a quantity. The students constructed the concept networks sequentially, adding one concept at a time to form a network. The starting concepts of the networks vary, but most often the concepts of electric charge, force and electric current are found among the starting concepts. In constructing the networks, students followed simple genealogical strategies so that they relate new concepts preferably to concepts introduced only a few generation steps before and simple linking-motifs in introducing the new concepts [24,26,32]. In introducing new concepts as part of the already existing network, students based the introduction of new concepts either on textbook experiments (i.e. known to students as discussed in textbooks) to operationalise the concept (i.e. make it measurable) or to models to introduce them deductively [24–26]. In a a typical case of a textbook experiment that illustrates how a new concept can be added as a part of an existing network students introduce a concept C by using two previous concepts A and B, and quite frequently then also add a connection between A and B if such a connection does not exist; this produces transitive, triangular linking-motifs. A very similar procedure is involved in using models to introduce new concepts and laws (relational connection between concepts) by deducing them on the basis of existing theoretical knowledge [24–26,32]. The resulting concept networks are thus ordered and directed having a kind of genealogical ordering where all new concepts added to the network are related to some (at least one) previous ones. The number of steps between concepts in the order they were introduced is termed a genealogical step. The meaning and content of the nodes in the empirical sample of real student networks is discussed in more detail elsewhere [25,32] and are not reproduced here where the modelling of the structural features are in focus.

Table I
The key concepts appearing in the students' concept maps. The numbers given in bold text are the best substantiated concepts. Some concepts appear twice, either theoretically (t) or empirically (e) substantiated.

Concept	Concept	Concept
2. Electric charge	**38.** Electric potential	**71.** Magnetic flux density (e)
8. Coulomb's law	**44.** Electric field (t)	72. Electric current
14. Displacement current	51. Gauss' law	**83.** Magnetic force
15. Electric field lines	**57.** Magnetic interaction	**91.** Magnetic field
27. Superposition of fields	**63.** Magnetic moment	**100.** Induction law
28. Electric field (e)	**66.** Magnetic flux density (t)	**109.** Rotational electric field
33. Mechanical work	69. Magnetic flux	**113.** Ampere-Maxwell law

The concept networks and many of their properties (like small cycles and dif-

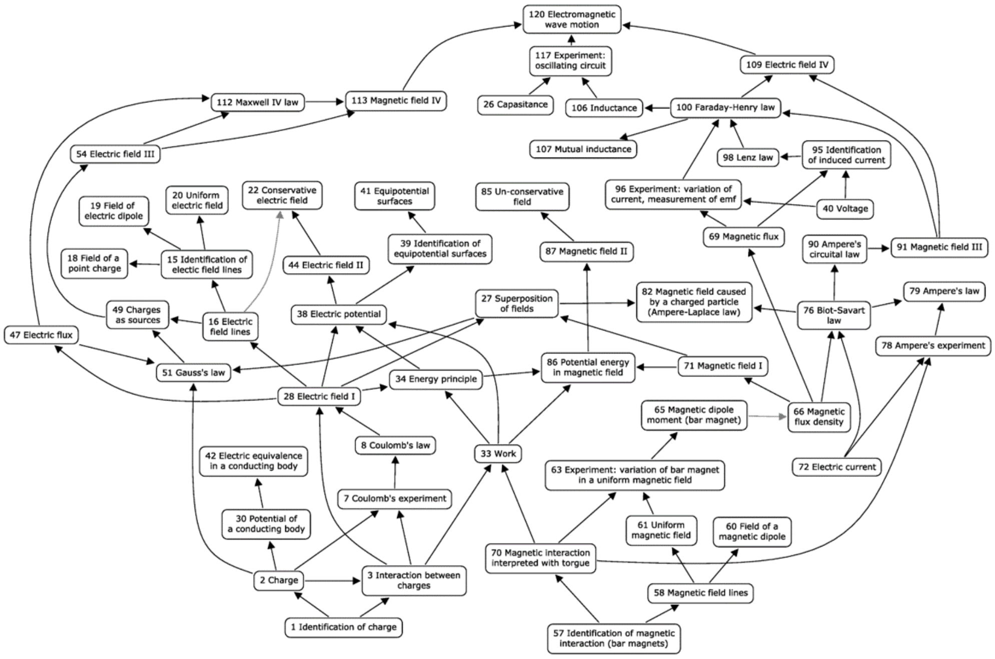
120 Electromagnetic wave motion
117 Experiment: oscillating circuit
109 Electric field IV
112 Maxwell IV law
113 Magnetic field IV
26 Capasitance
106 Inductance
100 Faraday-Henry law
54 Electric field III
107 Mutual inductance
98 Lenz law
95 Identification of induced current
19 Field of electric dipole
20 Uniform electric field
22 Conservative electric field
41 Equipotential surfaces
85 Un-conservative field
96 Experiment: variation of current, measurement of emf
40 Voltage
18 Field of a point charge
15 Identification of electic field lines
44 Electric field II
39 Identification of equipotential surfaces
87 Magnetic field II
69 Magnetic flux
90 Ampere's circuital law
91 Magnetic field III
49 Charges as sources
16 Electric field lines
38 Electric potential
27 Superposition of fields
82 Magnetic field caused by a charged particle (Ampere-Laplace law)
76 Biot-Savart law
79 Ampere's law
47 Electric flux
78 Ampere's experiment
51 Gauss's law
34 Energy principle
86 Potential energy in magnetic field
71 Magnetic field I
28 Electric field I
65 Magnetic dipole moment (bar magnet)
66 Magnetic flux density
8 Coulomb's law
33 Work
42 Electric equivalence in a conducting body
63 Experiment: variation of bar magnet in a uniform magnetic field
72 Electric current
7 Coulomb's experiment
30 Potential of a conducting body
61 Uniform magnetic field
60 Field of a magnetic dipole
70 Magnetic interaction interpreted with torgue
2 Charge
3 Interaction between charges
58 Magnetic field lines
1 Identification of charge
57 Identification of magnetic interaction (bar magnets)

Fig. 1. An example of a concept network drawn by a student (redrawn for clarity). The map is shown only to illustrate the overall appearance of the maps. The content of the links is not essential here.

ferent centralities of nodes) are close to what one could expect on the basis of the configuration-model, where the degree sequence of nodes is fixed but the nodes are otherwise linked randomly. A handful of nodes, however, have properties that are significantly different from what could be expected on the basis of the configuration-model [27]. These few nodes (concepts) provide long contiguous paths throughout the network and connect otherwise unconnected or poorly connected parts of the network [24–26,32]. Such nodes can be effectively and reliably discerned through the communicability betweenness centrality [33–35]. The reason why the communicability betweenness centrality is a relevant centrality measure is related to the fact that, in concept networks, the information, contained in already existing parts of the network, is effectively channelled to support the introduction of new nodes through contiguous paths that connect the nodes. The concepts which have a high betweenness centrality can be thus identified as kinds of key concepts in the networks, not only from a viewpoint of structure but also of content [27,32]. In Table I the concepts that have higher than average communicability betweenness centrality are shown in bold text. It is easy to recognise that these concepts are also central in regard to the content. It should be noted that the structural analysis based on the communicability betweenness centrality and content analysis are independent forms of analysis, but conclusion drawn from both are mutually supportive and strongly suggest the conclusion that for concept networks global, contiguous connections are typical feature of key concepts. Some of the key concepts are also a part of long cyclical paths and thus contribute appreciably to subgraph centralities [27]. In addition, the local clustering coefficient of many nodes in the rule-based concept networks is substantially higher (from 0.3 to 0.6) than expected in case of network of identical degree sequence but random linking (configuration-model). The concepts with high clustering only, however, if their communicability betweenness centrality is not high, are not central from the viewpoint of content but are only auxiliary concepts [24,25]. Consequently, the model introduced here attempts to reproduce and explain the following structural features of rule-based concept networks:

- Degree distribution that is peaked around small values of degree (typically a degree of 3), but where also larger degrees are found (typically up 12-14).
- Local clustering coefficients of the order of 0.3-0.6 of nodes occurring with substantially higher probability than in a network with identical degree sequence but random linking (configuration-model).
- Communicability betweenness centralities [34,35] and subgraph centralities which are relatively high for some nodes, occurring with substantially higher probability than in the configuration-model.

In addition to these features the real networks are modular, consisting of three modules. Modularity is here simply a pre-determined property because the networks describe the three phenomenologically distinct but connected areas: electricity, magnetism and electromagnetic induction. This modularity, however, must be taken into account in modelling the network.

3 The model

The networks that simulate the construction process of real concept networks and their empirical properties are based on simple generative rules and the use of a few basic linking-motifs introducing new concepts. These simulated networks are then compared with the real networks through the distribution of node degrees D, local clustering coefficient C, communicability betweenness subgraph centralities B and S, respectively. The statistical significance of the results is evaluated based on Z-scores.

3.1 Network generation

The basic assumptions we make about how the learners process and represent knowledge are as follows:

(1) New concepts are introduced on the basis of the old concepts, which are

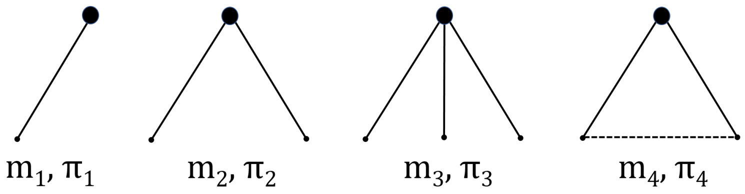

Fig. 2. The four linking-motifs m_i, $i = 1, \dots, 4$ of the genealogical model. The new node (large black dot) is added to ancestral nodes (tiny black dots) with a probability π_i. The linking-motif 4 also connects (dotted line) two ancestral nodes.

already part of the network. Relatively few (from two to four) of the old concepts are used as the basis of introducing the new concepts. The connection between new and old ancestral concepts are thus close, usually from three to seven genealogical steps in the ordered tree of ancestors.

(2) Concepts are recognized on the basis of the phenomenological meaning (here concepts related to electricity and concepts related to magnetism). This gives rise to the modularity. Within the module conceptual connections are close (three to seven step), while between modules the connections of nodes are separated by a greater number of genealogical steps (more than ten).

(3) Four basic linking-motifs are used in adding new nodes to the network. Each new node is linked either to: one pre-existing node (linking-motif m_1); two pre-existing nodes thus forming a 2-star (linking-motif m_2); three pre-existing nodes thus forming a 3-star (linking-motif m_3); and to two pre-existing nodes which also become connected, thus forming a triadic pattern (linking-motif m_4). Each of these linking-motifs m_k appear with the probability π_k with $\sum \pi_k = 1$.

The network is generated by introducing nodes $1, 2, \dots i-1, i, i+1, \dots n$ sequentially so that each node $i+1$ is connected in a directed way to some of the preceding nodes $1, 2, \dots i$. The directionality is defined from ancestor nodes to new nodes. The probability distribution function (PDF) that $i+1$ connects to the given ancestor i' which is j steps away from it (i.e. to node $i' = i + 1 - j$) is assumed to follow a discrete gamma-distribution [36]

$$f_{i,j}(\alpha, \lambda) = \frac{1}{Z_i\,(\alpha, \lambda)}\, j^{\alpha - 1} \exp\left[-\lambda j\right] \ , \tag{1}$$

where parameters α and λ control the form of the distribution. The normalization $Z_i(\alpha, \lambda)$ is obtained in a closed form in terms of Lerch's transcendental ϕ-function [36]

$$Z_i(\alpha, \lambda) = \phi(\mathrm{e}^{-\lambda}, 1 - \alpha, 1) - \mathrm{e}^{-\lambda(\alpha - 1)}\ \phi(\mathrm{e}^{-\lambda}, 1 - \alpha, 1 + i) \tag{2}$$

In practice, the detailed functional form of the distribution is not crucial, as it is peaked. The discrete gamma-distribution is chosen because it is flexible and the cumulative distribution function (CDF) for PDF in Eq. (1) can be given in the form

$$F(i,j) = \sum_{j'=1}^{j} f_{i,j'} = \frac{Z_i(\alpha,\lambda)}{Z_j(\alpha,\lambda)} \tag{3}$$

This CDF is the basic distribution that defines the genealogical back-reference step lengths (i.e. how many steps separates the new node and the ancestral node). The CDF thus determines how new concepts are linked as part of the pre-existing network. The same rule of linking is used for intra-cluster linking of nodes within a given module, and between inter-cluster linking between the modules. The difference between intra- and inter-cluster linking is that in inter-cluster, linking operates with all four linking-motifs but with intra-cluster linking operates with m_1 only.

3.2 Simulation method

In all simulations, we used the roulette wheel -method, where events are realized in proportion to their probabilities [37]. The nodes are connected on the basis of CDF in Eq. (3). The simulation consists of three steps, which are:

(1) The module size N and its variation $\pm\Delta N$ are selected.
(2) The linking-motif is selected. In selecting the basic linking-motif used to add a new node a discrete set of 4 possible motifs k with probabilities π_k are arranged with cumulative probability $\Phi_k = \sum_{i=1}^{k} p_i / \sum_{i=1}^{4} p_i$. The linking-motif m_k is selected if the random number $0 < r < 1$ falls in the slot $\Phi_{k-1} < r < \Phi_k$. In simulations only the probability π_3 is varied. Varying π_3 affects also the probabilities of other events because of the normalisation $\sum \pi_k = 1$ but the ratios π_i/π_j, $i,j \neq 3$ remain unchanged.
(3) For each ancestral node appearing in the selected linking-motif, the genealogical distance is selected. For each link to an ancestor, a random number $r \in [0,1]$ is generated and the new node is connected to an ancestor at a distance j^* defined from $r = F(i,j^*)$.

In simulations, attempts to connect already connected nodes may occur, in which case no multiple connections are allowed. In practice, values of j^* corresponding to different values of r are tabulated in advance for each i, so that repeated inversions of $r = F(i,j^*)$ during the simulation is avoided. In the simulation model we have three modules. Within the modules we use the same parameters to connect the nodes, but the values of the parameters between the modules can be different from their values within the modules.

3.3 Network analysis

Connections between nodes in concept networks are described in terms of the adjacency matrix $\mathbf{a}$, where variables a_{ij} indicate the connections between nodes i and j. If nodes are connected, then $a_{ij} = 1$ and if there is no connection, then $a_{ij} = 0$. Here we analyse the networks as undirected ones with $a_{ij} = a_{ji}$, although the generation process is directed. This is because in the empirical sample the directionality is ambiguous and not easily interpretable (students use direction to indicate a cause or an effect). For a network of N nodes these elements form a $N \times N$ dimensional adjacency matrix $\mathbf{a}$. In the analysis of network properties, we focus on four observables: 1) the node degree D, 2) the local clustering coefficients C, 3) the communicability betweenness B, and 4) the subgraph centrality S. The definitions and formulas to obtain these observables from the adjacency matrix are provided in Appendix A.

The decision to focus on the clustering and on the communicability centrality and the subgraph centrality are motivated by empirical notions based on previous results on the analysis of students networks, which has shown that:

(1) The local clustering D is related to how students use auxiliary (not always central or key concepts) locally to substantiate (or support) the introduction of new concepts (nodes) in the network [24,26].
(2) The communicability betweenness B is related to how a given concept (node) is globally susbtantiated (or supported) by all other nodes in the network. The substantiation of the node is namely based on information which is passed from one node to another, thus making the introduction of new nodes contiguous to previously existing nodes. This is also the basic generative dynamics coded in the genealogical model introduced here. As shown previously, the nodes which have a high value of communicability betweenness in concepts networks are also central from the point of view of the content of the concept network, i.e. they are kinds of key concepts [27,32].
(3) The subgraph centrality S provides complementary information of the global substantiation of a given node by taking into account the cyclical paths that provide the overall coherence. The subgraph centrality S, although it gives information largely redundant to infomation already contained in B, is useful in ranking the importance of key concepts [27].

For these reasons we focus in the analysis of the genealogical model on observables D, C, B and S. The roles of these observables in content analysis and in finding the key concepts of the real concept networks, reported in more detail elsewhere [24,26,27,32], is not further discussed here.

The statistical significance of the observable values $O \in \{D, C, B, S\}$ as measured from the genealogical model (G-m) networks is estimated by comparing the results based on the genealogical-model to results obtained from a null-model, where the degree sequence is identical to the original network, but the links are established randomly [39,33]. In what follows we refer to the null-model as the configuration-model (C-m). The statistical significance of the given value of observable O is here assessed by calculating its Z-score (see Appendix A), which is a commonly used simple measure to assess the statistical significance of observables' values in networks [33,38]. The Z-score of a given observable O is particularly suitable given that the nodes of interest are those which have higher than average value of the local clustering C or the communicability betweenness B.

The similarity of the distributions corresponding to different model parameters is also of interest, although in the present study, the purpose is not to focus on the detailed form of the distributions or their detailed comparisons to the empirical sample. Empirical results in cases studied here are scarce, as they are based on a limited number of samples and all the values based on empirical data are tentative rather than conclusive. Moreover, the distributions are discrete and their estimation by smooth parametric distributions is not pursued here. For these reasons, we have chosen to base the similarity comparison of distribution p and q on the Kullback-Leibler divergence $\mathrm{KLD}(p|q)$ (see Appendix A), which is a nonparametric information theoretical measure for the similarity between distributions [39,40].

4 Results

In the simulations, the parameters were varied and the set of parameters which appeared best to correspond to the empirical results for the distribution of degrees D were chosen for closer scrutiny. No exact agreement, however, with the empirical sample was attempted since the empirical sample consists of only 12 cases, which is too few for comprehensive comparative analysis. Four models A-D with different set of parameters, given in Table II, were chosen for closer scrutiny. The parameters were chosen so that variations are limited to the relevant range of average degree $\bar{D}$ roughly from a value of 3 to 4. In what follows we discuss the effects of parameter variations on the distributions for the degree D, the local clustering coefficient C, and for the subgraph and communicability betweenness centralities S and B, respectively. In all simulations the number of modules is three corresponding to the number of modules in the empirical sample. In what follows, only the probability π_3 of linking-motif m_3 and back-reference step length L are varied. In addition, the fixed average size $N = 20$ of modules is varied by $\pm\Delta N$ to check the robustness of results against relative changes of the order of 10%-50% in module size. All results are for 1000 sample of the given parameterization of a genealogical-model (models A-D). For each realization in the sample, 10 repetitions for corresponding configuration-models with identical degree sequences were done. Simulations, configuration-model generation and all analysis are carried out by using the graph and network package of Wolfram Mathematica 11.

Table II
The parameters used in simulations of the genealogical-model. The parameters α and λ are for inter-cluster linkage, corresponding average genealogical step length L. The parameters for intra-cluster linkage are $\alpha = 13$ and $\lambda = 1.3$ corresponding $L = 18$ and they remain the same for all simulations. For linking-motifs only $\pi_3 \equiv \pi$ is provided since only it is varied while the ratios $\pi_2/\pi_1 = 0.17$, $\pi_4/\pi_1 = 0.25$ and $\pi_4/\pi_2 = 0.67$ of the other linking-motifs remain unchanged. For intra-cluster linking only linking-motif 1 with $\pi_1 = 0.80$ is active. In all cases the average module size is $N = 20$ with $\Delta N = 5$ for random variation in module size. The average values of the node degree ($\bar{D}$),the local clustering ($\bar{C}$), the communicability centrality ($\bar{B}$) and the subgraph centrality ($\bar{S}$) are also provided.

Model	α	λ	L	π	$\bar{D}$	$\bar{C}$	$\bar{B}$	$\bar{S}$
A	9	1	9	0.25	3.3	0.15	0.065	0.017
B	4	1.5	6	0.40	3.4	0.16	0.068	0.018
C	9	1	9	0.20	3.7	0.14	0.061	0.016
D	4	1.5	6	0.60	3.9	0.15	0.063	0.016

The genealogical-model A with parameters as given in Table II provides the closest match with the empirical networks. A close match by parameter optimization, however, is not pursued here because instead of quantitative agreement the focus is on the qualitative similarities between the empirical networks and the networks generated by the genealogical model. Simulations are used to explore ensembles of networks generated by the genealogical model: how measurable properties of these networks are distributed within ensembles with similar degree distribution of nodes, and how the thus obtained measurable values compare to the empirical observations. The networks generated by the genealogical model (G-m) as well as the empirical networks (Emp) are compared against the configuration-model (C-m) networks, which have identical degree sequences with the networks they are compared to.

The distributions of the degree D, the local clustering C and the communicability betweenness and the subgraph centralities B and S, respectively, for each node are calculated for all simulation cases. In Fig. 3 are shown results for distributions of these observables as the pairwise histograms, where on the left is shown the empirical distribution (Emp) compared with the result corresponding to configuration-model (C-m), and on the right are the results of genealogical model A (G-m) compared with the corresponding configuration-model (C-m). The scatter-plots of values of observables are compared for the genealogical-model (G-m) and the configuration-model (C-m) on the upper right diagonals, while comparisons for the empirical (Emp) and genealogical-model (G-m) are shown in the lower left diagonals. The results in Fig. 3 show that qualitatively the genealogical-model A provides results close to real networks, but for local clustering both the genealogical-model A and the real networks have local clustering different from the configuration model. For the communicability betweenness and subgraph centralities B and S the differences are not so clear, provided that a few outliers in B in the case of the empirical results are ignored.

The results in Fig. 3 suggest that for most nodes the values for B and S agree with the values as they emerge in the corresponding configuration-model networks. This means that the centralities are predictable on the basis of the configuration-model provided that the degree sequence is known. On closer inspection, a small subset of nodes have values of B and S which occur with clearly higher probabilities than could be expected on the basis of configuration-model. Such nodes are few, but the nodes with exceptional high B and S centralities are the most interesting ones, and supposedly also the most important ones that we should pay close attention to. These nodes have important roles in providing the network global connectivity and in connecting substantial portions of the network. In the case of empirical networks just these kinds of nodes can be identified as key concepts in the network [27].

The results for the local clustering C of the genealogical model are close to the empirical results for students' real networks, while the results for both of them are clearly different from the results obtained for configuration models with identical degree distributions. Also the communicability betweenness centrality B shows deviations from the configuration model results, while values of the subgraph centrality S appears to be closer to the configuration-model. These differences are clearly revealed by the Z-values of the observables shown in Fig. 4. The Z-values are calculated for binned observables, for each of the 12 networks and each generated genealogical network against the 10 configuration samples. The resulting Z-values of empirical networks and genealogical networks are displayed as box-whisker plots. The extreme medians and the upper- and lower-quartiles of the Z-values are shown in Table III. For comparisons, the Kullback-Leibler divergences are also provided in Table III.

The Z-values for the clustering in Fig. 4 show that the median of Z-value increases with increasing observable value for the clustering C and the communicability betweenness centrality B. For clustering $0.24 < C < 0.45$ the median of Z scores for clustering is 4-5, with the highest value $Z = 5.5$ for empirical networks and $Z = 4.3$ for the genealogical-model A. These indicate statistically significant deviations (p-values $p < 0.001$) from clustering in the configuration-model. Clustering, however, is not indicative of being a key concept (i.e. important in the global sense) in students' real networks, as is

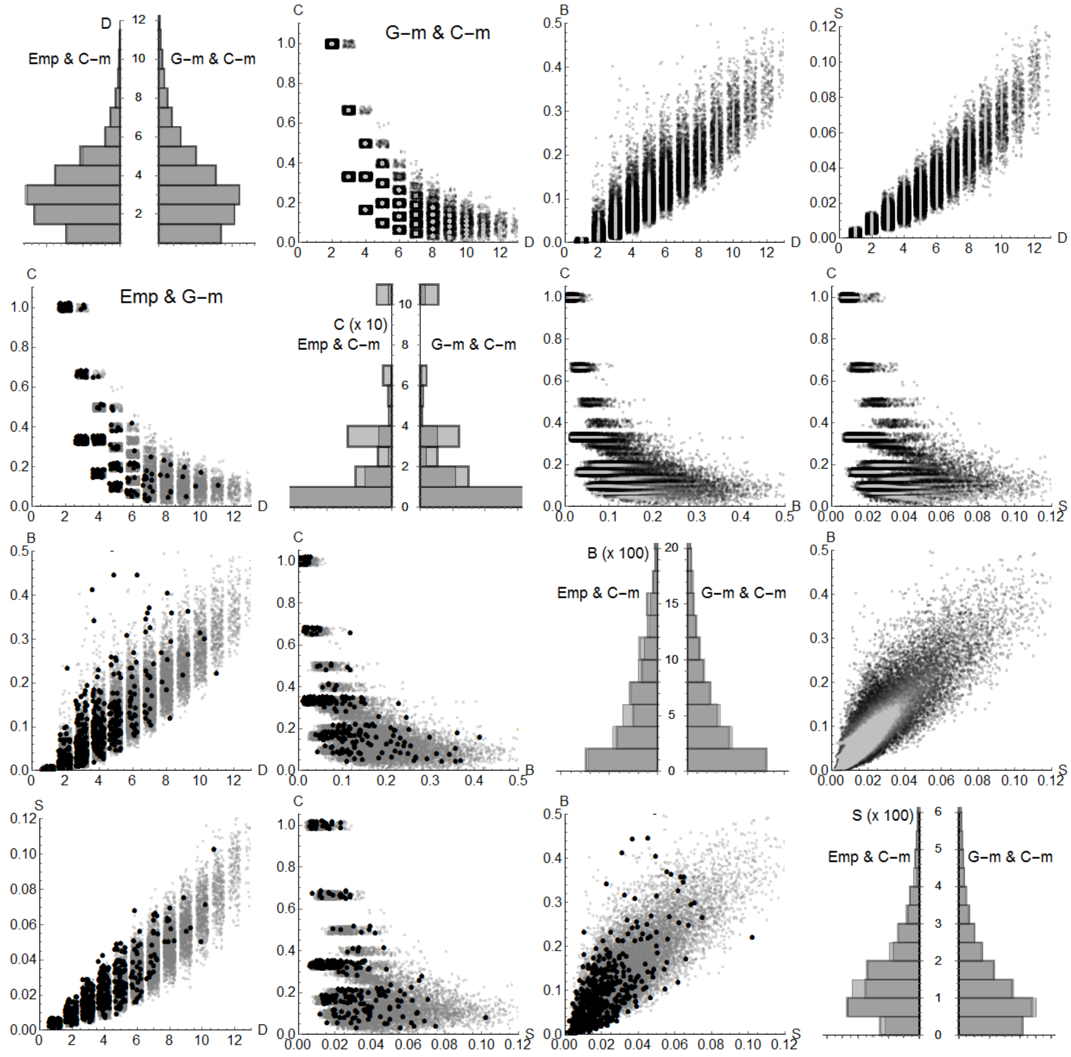

Fig. 3. The distribution and scatter-plots of observables $O \in \{D, C, B, S\}$ in real, empirical networks (Emp) and genealogical-model A (G-m) networks compared with configuration-model (C-m) networks with equivalent degree sequences. The distribution of the degree D, local clustering coefficient C, communicability betweenness centrality B and subgraph centrality S are shown as pairwise histograms in the diagonals. The scatter-plots of the observables are shown in the lower left and upper right diagonals. In the lower left diagonal (denoted by Emp & G-m) the distributions of students' networks (Emp, black markers) and model A networks (G-m, grey markers) are compared. In the upper right diagonal (denoted G-m & C-M) the model A (G-m, black markers) and corresponding configuration-model (C-m, grey markers) networks are compared. All histograms are normalized to unity and can be thus compared although the vertical scales are not shown.

known from analysis of empirical concept networks[27]. Rather, high clustering reflects the important role of local 2- and 3-star linking-motifs, m_2 and m_3, respectively, in substantiating the addition of nodes that ultimately often remain as auxiliary nodes.

The Z-values of the communicability betweenness and subgraph centralities B and S, respectively, have a rather similar form of distribution in G-m and C-m. The values of $Z \approx 2.2$ for high enough values $B > 0.2$ reveal that B in the empirical sample and in the genealogical-model differs significantly from the values found in the configuration-model, at the significance level of $p < 0.05$. The subgraph centralities, on the other hand, are not statistically significantly different from values found in the configuration-model, as is shown by Fig. 4 and extreme values provided in Table III. This means that in regard to the subgraphs, for the most part, the genealogical-model networks are like configuration-model networks. The differences between the empirical to configuration-model networks, and the genealogical- to configuration-model networks are summarised by the quantile-quantile plots shown in the lowest row in Fig. 4. The quantile-quantile -plots show that for $Z < -2.5$ and $Z > 2.5$ the quantiles of the data are quite different from the quantiles expected on the basis of normal distributions. The interpretation that clustering is significantly different (higher) in empirical and genealogical networks than in configuration-model is complemented by the Kullback-Leibler divergencies (KLD) provided in Table III.

Table III
The Kullback-Leibler divergences KLD(p|q) for distributions of the degree D, the local clustering C, the communicability betweenness centrality B and the subgraph centrality S. The KLD is provided for the empirical sample (E) compared to the genealogical (G) and configuration (C) models, and genealogical compared to the configuration-model. The ordering of comparisons of $p, q \in \{E, G, C\}$ is as denoted in KLD($p|q$). The KLD values are given in units of 100xnats, which for present purposes is roughly interpretable as % of new information needed to infer p from q. For the Z-scores the median value (Med) and values of 75% upper quantile (Q_U) and 25% lower quantile (Q_L) are provided, for empirical and genealogical models, respectively. The p-values corresponding to the given median values of Z-scores are denoted by * for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$. The KLD of empirical degree distribution compared to genealogical is 6.6 100xnats.

	KLD($p|q$) in 100xnats			$Z_{\text{empirical}}$			$Z_{\text{geneal.}}$		
	$E|G$	$E|C$	$G|C$	Med	Q_U	Q_L	Med	Q_U	Q_L
C	2.3	21.8	12.9	5.5***	6.3	4.7	4.3***	5.5	2.5
B	2.7	3.8	0.4	2.2**	4.4	0.9	2.2*	3.4	0.7
S	3.8	1.5	0.1	1.6	2.4	0.3	0.4	1.5	-0.5

The sensitivity of results to changes in model parameters were tested by varying: 1) the size of modules by ±ΔN; 2) the back-reference step length L (by changing α and λ) of genealogical relatedness; and 3) the probability π_3 of 3-star linking-motif. Parameter variations were chosen so that the average degree

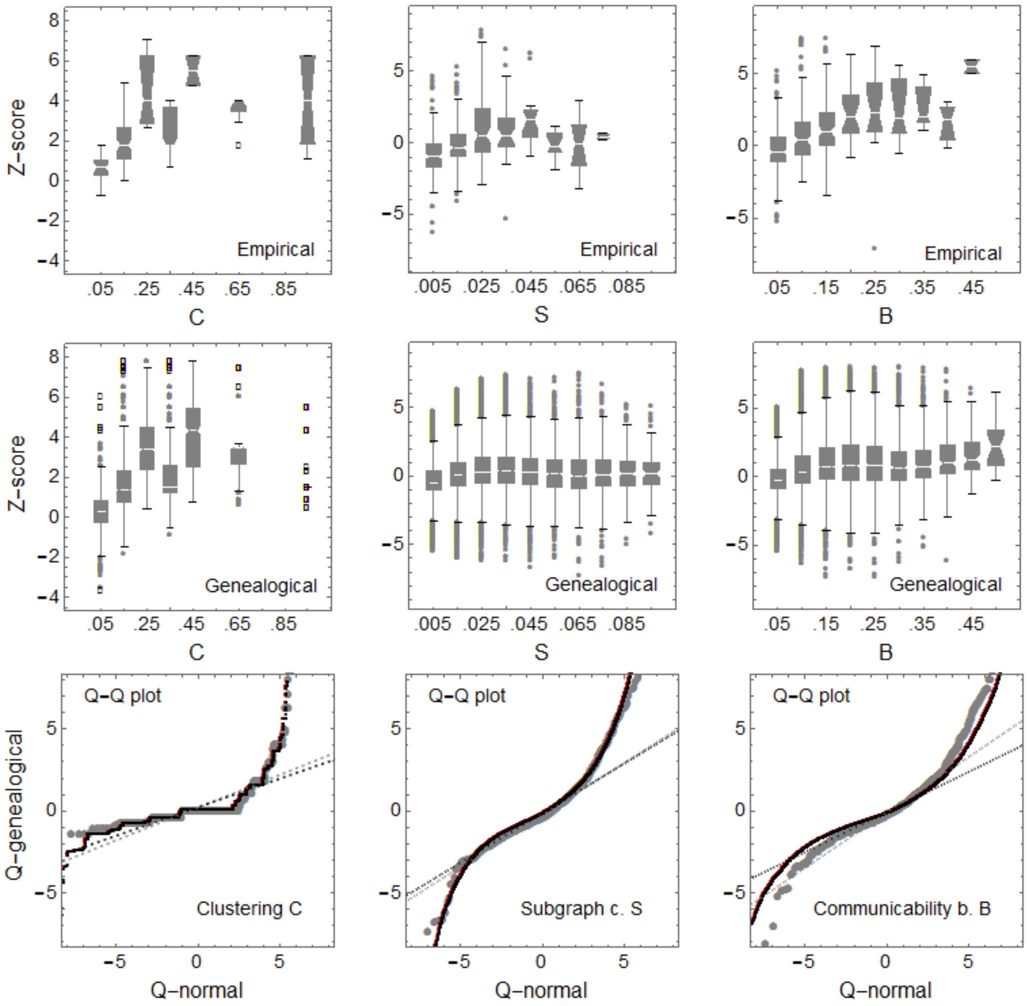

Fig. 4. The box-whisker plots of Z-values of the clustering C, and the communicability betweenness and subgraph centrality B and S, respectively, in the case of the empirical networks (upper row) and genealogical model A (middle row). The notched boxes in the box-whisker plots show the median values (white, middle bar in the box) of the observables and their upper (75%) and lower (25%) quartiles (upper and lower boundaries of the boxes, respectively). The whiskers show the limits beyond which the data-points are considered as outliers (the first set of outliers are shown as dots, for clustering also the second set are shown as boxes). The corresponding quantile-quantile plots (lower row) are shown. In the quantile-quantile plot, the dotted line represents the quantiles of normally distributed data.

values remained between 3 to 4, which is within the scope of interest. Results for the distribution of D, C, B and S for models A-D with different parameters (see Table II) are shown in Fig. 5. The distribution for S is qualitatively so similar to B that it is not shown in what follows. The Kullback-Leibler divergences KLD($p|q$) are given in Fig. 5 for model A (distribution p) and for models B-D (distribution q), and for comparisons, also in cases where q corresponds to the configuration model. In all cases, with increasing probability of 3-star linking-motifs m_3 the distribution of D, C and B shift to larger values

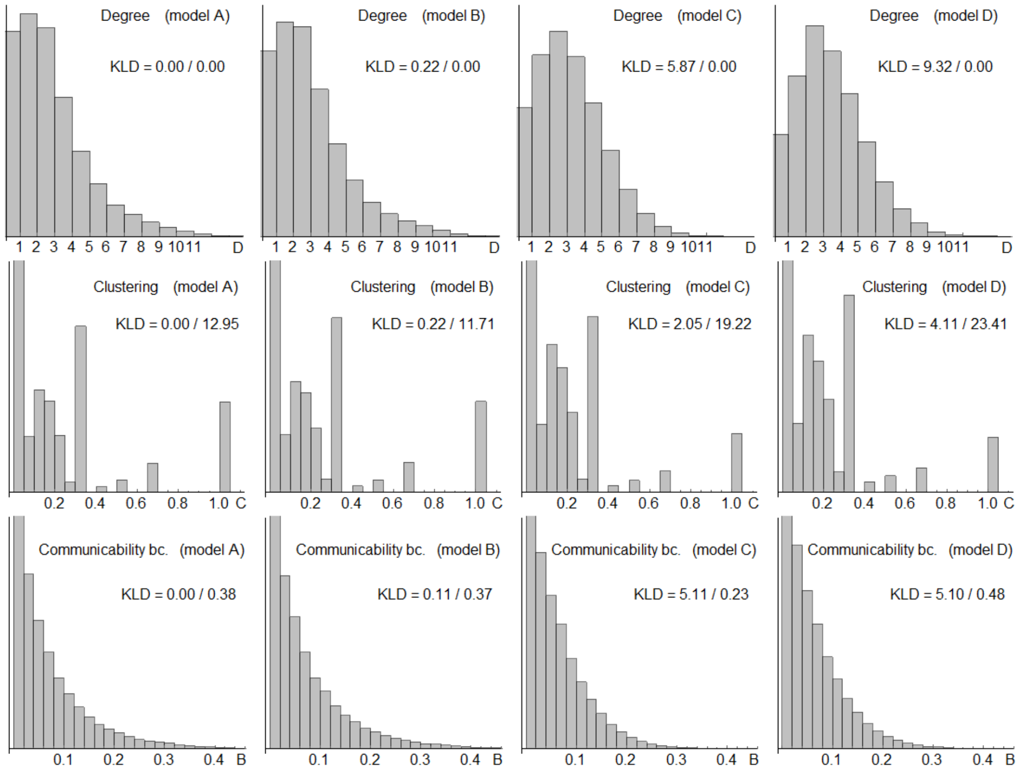

Fig. 5. Histograms of distributions of the degree D, clustering coefficient C and communicability betweenness centrality B, respectively, for models A-D corresponding to different parameters (see Table II). The Kullback-Leibler difference KLD$(p|q)$ between distributions p for model A and q for models B-D are given on the left side of the slash and corresponding KLD for the configuration model on the right side. The models A and C are with parameters $\alpha = 9.0$ and $\lambda = 1.0$ with average back-reference steps $L = 6$ and $\pi = 0.25$ and 0.40, respectively, while models B and D are with $\alpha = 4$ and $\lambda = 1.5$ with $L = 6$ and $\pi = 0.60$ and 0.25, respectively. In all cases $\Delta N = 5$. All the histograms are normalized to unity and can be thus compared although the vertical scale is not shown.

as expected but the changes are rather moderate as shown by the values of KDL. Only for the communicability betweenness centrality are the changes due to increasing probability of 3-star motifs larger than the KLD between the genealogical and configuration-models.

Within the chosen range of parameter variations the changes in average values of the observables are nearly linear in regard to changes and the average value of observables adequately characterizes the changes in the distributions. Therefore, instead of absolute changes we report in Table IV the rate of change $\Delta O/\Delta\pi$ of average values of observables $O \in \{D, C, B, S\}$ when the probability π of 3-stars changes. The rates of changes are reported for models with L=9.0 and 6.0 and for $\Delta N = 3, 5$ and 7, respectively. The total range of variation of π is 0.45, and values for rates of changes are obtained as averages for π=0.20, 0.25, 0.35, 0.45, 0.55, 0.60 and 0.65.

Table IV
Rate of change $\Delta O/\Delta\pi$ of observable mean values $O \in \{D, C, B, S\}$ when the probability π_3 of 3-stars changes. Simulation results are for models with Ł = 6 and 9 and for $\Delta N = 3, 5$ and 7.

	$L = 9$			$L = 6$		
	$\Delta N = 3$	$\Delta N = 5$	$\Delta N = 7$	$\Delta N = 3$	$\Delta N = 5$	$\Delta N = 7$
$\Delta D/\Delta\pi$	1.90	1.89	1.83	2.23	1.98	2.58
$\Delta C/\Delta\pi$	0.46	0.46	0.59	0.61	0.90	0.36
$\Delta B/\Delta\pi$	0.50	0.50	0.53	1.27	1.58	1.47
$\Delta S/\Delta\pi$	-1.31	-0.13	-0.13	0.06	0.20	-0.42

The results show that when the probability of π_3 increases, the average values of observables D, C, B also increase but that of S decreases for large back-reference step lengths. The increase of average values of D and C simply results from increased local connectivity, while for B the increase of average value indicates increases in the length of the contiguous paths. The decrease of the average value of S, on the other hand, indicates that the role of small cycles increases with increasing π_3 when the genealogical back reference step decreases; for small genealogical distances the connections begin to favour small cycles. Although the results allow no clear generalisation of the parameter dependencies they nevertheless support the conclusion that the individual variations in students' concept networks can result from simple and small variations in how they use the basic linking-motifs in introducing the new concepts and in how far backwards in the genealogical history the new connections reach.

5 Discussion and conclusions

We have explored students' expressions of the relational structure of concepts by analysing the concept networks they have made. The aspects of interest in these networks are the relatedness of concepts, and how certain basic linking-motifs are used in an ordered addition of new concepts to the networks. Attention has been paid to the local and global relational organisation of the concepts by measuring local clustering of nodes (concepts) around other nodes, and on how a given node participates in communicability between nodes. It was found that the networks have an appreciable local clustering but betweenness communicability and subgraph centralities of the nodes are largely comparable to corresponding randomized structures. Only a small number of nodes have higher values of these centralities than expected on the basis of the random network model. These nodes, however, are essential in creating the global order and they are structurally kinds of key nodes in conferring the overall contiguity of the conceptual connection within the network. Nodes with comparable properties are also found in real concept networks where they play a key role in regard to the content [27,32].

The results of the model and how they compare to real concept networks can be understood to be related to the design principles used by students - their use of experiments and models in connecting concept create local ordering based on triangular patterns, but with no appreciable global ordering except in the case of a few nodes. The model shows that such structures may originate from a set of simple generic rules to add new concepts to the existing network. What is needed, is a collection of basic linking-motifs to introduce new concepts: the addition of one node with one link; a node with two or three links (2-star or 3-star); or the addition of a node with 2-star and triadic closure. In addition, the links are formed on the basis of a genealogical rule where nodes are linked not to the most recently added nodes but nodes further away in the genealogical ordering. The best match is reached with back-reference where within a module steps from 6 to 9 are most probable, while between the modules the step lengths are on the average 18-20 genealogical steps. These simple rules are enough to generate networks with comparable properties as those found empirically.

The results presented here model specific types of concept networks used in learning, where students have constructed networks following specific rules to substantiate the addition of new concepts as part of the network. The construction thus requires careful considerations from students of how the inclusion of new concepts is substantiated. Such concepts networks are very different from traditional concept maps based on propositions but requiring no substantiation or epistemic justification [24–26]. At present, there are only a few examples of rule-based concept networks and more comprehensive comparisons are not yet possible. However, the results presented here interestingly parallel recent findings in learning lexicons [22,23]. Also in that case the relational structure of words is closely connected to the learning of the meaning of words. In the case of lexicons, the closeness centrality appears to be indicative of a key word [22,23]. In the case of rule-based networks, the high communicability betweenness centrality is a characteristic feature of key concepts because the key concepts of rule-based networks must be substantiated through contiguous paths of other concepts [27]. In both cases, however, key concepts can be recognised through their structural role.

The results provided here suggest that learners handle knowledge so that they process the relational aspect of conceptual knowledge in rather small pieces, finding the connections on the basis of the affiliation of concepts in the procedures (experiments and models), where they are used. Furthermore, there is a finite genealogical span of a few concepts in making these affiliations. At a more general level these results support the view that students process and order their knowledge by using simple basic linking-motifs and genealogical affiliation schemes. In short, there seems to be a preference for a certain parsimony in handling the knowledge. Of course, this finding is not very unexpected, but nicely confirmed here through structural analysis of knowledge representations. These notions indicate that simple basic elements are enough to understand at least some relevant aspects of how students conceptualise scientific knowledge and organise complex knowledge.

Acknowledgements: This work was supported by the Acedemy of Finland through Grant 311449.

Appendix A: Network metrics and statistical analysis

Connections between nodes in concept networks are described in terms of the adjacency matrix $\mathbf{a}$, where variables a_{ij} indicate the connections between nodes i and j. If the nodes are connected, then $a_{ij} = 1$, and if there is no connection, then $a_{ij} = 0$. Here we analyse the networks as undirected ones with $a_{ij} = a_{ji}$, although the generation process is directed. For a network of N nodes these elements form a $N \times N$ dimensional adjacency matrix $\mathbf{a}$. In analysis of network properties, we pay attention to the distribution of node degrees D, local clustering coefficients C and communicability betweenness and subgraph centralities B and S, respectively.

A1. The degree D_k of the given node k simply counts the number of links attached to the node and it is given in terms of the adjacency matrix in the form [38,39]

$$D_k = \sum_i \left(a_{ik} + a_{ki}\right)/2\,. \tag{A.1}$$

A2. The local clustering coefficient C_k measures the local interconnectedness of nodes through three-cycles or triangles. This quantity is of particular importance here, because the design principle of the networks favours the formation of triangles. The local clustering coefficient C_k of node k is defined as the ratio of number of triangles to the 2-stars (connected triples) as [38,39]

$$C_k = \frac{\sum_{i>j} a_{kj} a_{ki} a_{ji}}{\sum_{i>j} a_{kj} a_{ki}}\,. \tag{A.2}$$

The local clustering coefficient takes values between 0 and 1.

A3. The subgraph centrality S_k operationalises the role of closed walks, i.e. cycles. The kth diagonal element of the nth power of the adjacency matrix gives the number of cycles starting from a given node k and involving n nodes (or links). In defining the subgraph centrality each closed walk is divided by the factorial of the length of the walk to compensate for the rapidly (though not strictly factorially) increasing number of walks when the length of the walk increases. The subgraph centrality of node is then obtained from the matrix exponential of adjacency matrix in the form [33–35]

$$S_k = \frac{[\,\exp(\mathbf{a})\,]_{kk}}{\sum_k [\,\exp(\mathbf{a})\,]_{kk}} \tag{A.3}$$

A4. The communicability betweenness B_k measures the relative ease to pass from one node to another node so that a given node k is included. For the communicability we use here a similar definition as for the subgraph centrality in that the longer paths are assigned lesser weights, in proportion of the factorial $l!$ of the length l. The communicability betweenness is then conveniently expressed in terms of matrix exponentials [34,33,35]

$$B_k = \frac{1}{A} \sum_{i,j} \frac{[\exp(\mathbf{a})]_{ij} - [\exp(\mathbf{a}+\mathbf{b})]_{ij}}{[\exp(\mathbf{a})]_{ij}}, \; i \neq j \neq k. \tag{A.4}$$

Matrix $\mathbf{b} \equiv \mathbf{b}(k)$ has nonzeros only in row and column k, so that these row and column has -1 where the adjacency matrix $\mathbf{a}$ has +1. Then the denominator in Eq. (A.4) counts all the paths between nodes i and j where node k is included. The normalization factor is given by $A = (N-1)^2 - (N-1)$ for a network of N nodes [33–35].

A5. The statistical significance of the observable values $O \in \{D, C, B, S\}$ as measured from the networks is estimated by comparing the results of the analysis to results obtained from a configuration-model (null-model), where the degree sequence is identical to the original network, but the links are established randomly [39,33]. The statistical significance of the observable O is assessed by calculating the so-called Z-scores defined as [19,33].

$$Z = \frac{O - \langle O \rangle_R}{\sigma_R} \tag{A.5}$$

where O is the observable value in the simulated ensemble and $\langle O \rangle_R$ the corresponding value in the configuration model and $< \sigma >_R$ the corresponding standard deviation. Statistical significance requires that Z-values are high enough. In what follows a value $1.65 < Z < 2.33$ (corresponding p-values 0.01< p<0.05) is consider as significant (marked by *), $2.33 < Z < 3.093$ (corresponding 0.01<p<0.001) highly significant (**) and $Z > 3.09$ (corresponding p<0.001) extremely significant (***).

A6. The difference between distributions p and q is estimated on the basis of the Kullback-Leibler divergence (KLD) which is an information theoretic measure of distribution difference. If distribution p is compared to distribution q KLD can be taken as a measure of the information gained when new information becomes available to update the probability distribution q to a new one p, or differently, as a measure of how much information is lost when p is approximated with q. In the case of discrete probability distributions p and q the Kullback-Leibler divergence is defined as [39,40]

$$\mathrm{KLD}(p|q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{A.6}$$

KLD is thus the expectation value of the logarithmic difference between the probabilities p_i and q_i when the probability of i is given by p_i. The KLD is always positive and defined only if for $q_i = 0$ also $p_i = 0$ holds. In addition, when p_i=0 also $p_i \log p_i = 0$. It should be noted that KLD is not a proper metrics because it is not symmetric and KLD$(p|q) \neq$ KLD$(q|p)$. In calculating the KLD we use natural logarithms, in which case KLD is given in units of nats [40]. A KLD value of 0 means complete similarity while a value of 1 means complete dissimilarity of the distributions. Here, a given value KLD < 1 can roughly be interpreted as a relative fraction of new information needed to deduce the distribution p from q.

References

[1] N. Rescher. Cognitive systematization: A systems-theoretic approach to a coherentist theory of knowledge. Rowman and Littlefield, 1979.

[2] P. Thagard. Conceptual Revolutions. Princeton University Press, 1992.

[3] L. BonJour. The Structure of Empirical Knowledge. Harvard University Press, 1985.

[4] J. Novak, and B. D. Gowin. Learning How to Learn. Cambridge University Press, 1984.

[5] I. M. Kinchin, D. B. Hay, and A. Adams. How a Qualitative Approach to Concept Map Analysis Can Be Used to Aid Learning by Illustrating Patterns of Conceptual Development. Educational Research 42 (2000) 43-57.

[6] A. M. O'Donnell, D. F. Dansereau, and R. H. Hall. Knowledge Maps as Scaffolds for Cognitive Processing. Educational Psychology Review 14 (2002) 71-86.

[7] J. C. Nesbit, and O. O. Adesope. Learning with Concept and Knowledge Maps: A Meta-Analysis. Review of Educational Research 76 (2006) 413-448.

[8] F. Amadieu, T. van Gog, F. Paas, A. Tricot, and C. Marine. Effects of Prior Knowledge and Concept-map Structure on Disorientation, Cognitive Load and Learning. Learning and Instruction 19 (2009) 376- 386.

[9] M. B. Goldwater, and L. Schalk. Relational Categories as a Bridge Between Cognitive and Educational Research. Psychological Bulletin 142 (2016) 729-757.

[10] G. S. Halford, W. H. Wilson, and S Phillips. Relational knowledge: the foundation of higher cognition. Trends in Cognitive Science 14 (2010) 497505.

[11] C. Kemp, and J. B. Tenenbaum. Structured Statistical Models of Inductive Reasoning. Psychological Review 116 (2009) 2058.

[12] C. Kemp, and J. B. Tenenbaum. The Discovery of Structural Form. PNAS 105 (2008) 10687-10692.

[13] T.H. Duong, G. S. Jo, J. J. Jung, and N. T. Nguyen. Complexity Analysis of Ontology Integration Methodologies: A Comparative Study. Journal of Universal Computer Science 15 (2009) 877-897.

[14] K. Börner. Atlas of Knowledge: Anyone Can Map. MIT Press, 2015.

[15] K. Börner, R. Klavans, M. Patek, A. M. Zoss, J. R. Biberstine, R. P. Light, V. Lariviere, and K. W. Boyack. Design and Update of a Classification System: The UCSD Map of Science. PLoS ONE 7 (2012) e39464.

[16] K. Börner, and A. Scharnhorst, A. Visual conceptualizations and models of science. Journal of Informetrics 3 (2009) 191-209.

[17] F. Shi, J. G. Foster, and J. A. Evans. Weaving the fabric of science: Dynamic network models of sciences unfolding structure. Social Networks 43 (2015) 7385.

[18] C. Chen, Y. Chen, M. Horowitz, H. Hou, Z. Liu, and D. Pellegrino. Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics 3 (2009) 191-209.

[19] L. da F. Costa, Learning about knowledge: A complex network approach. Physical Review E 74 (2006) 026103.

[20] J. B. Batista, and L. da F. Costa. Knowledge acquisition by networks of interacting agents in the presence of observation errors. Physical Review E 82 (2010) 016103.

[21] J. Goñi, B. Corominas-Murtra, R. V. Solé, and C. Rodríguez-Caso. Exploring the randomness of directed acyclic networks. Physical Review E 82 (2010) 066115.

[22] M. Stella, N. M. Beckage, and M. Brede. Multiplex lexical networks reveal patterns in early word acquisition in children. Scientific Reports 7 (2017) 46730.

[23] M. S. Vitevich, and N. Castro. Using network science in the language and clinic. International Journal of Speech-Language Pathology 17 (2015) 13-25.

[24] I. T. Koponen, and M. Pehkonen. Coherent Knowledge Structures of Physics Represented as Concept Networks in Teacher Education. Science & Education 19 (2010) 259-282.

[25] I. T. Koponen, and M. Nousiainen. Pre-service physics teachers understanding of the relational structure of physics concepts: Organising subject contents for purposes of teaching. International Journal of Science and Mathematics Education 11 (2013) 325-357.

[26] M. Nousiainen. Coherence of Pre-service Physics Teachers Views of the Relatedness of Physics Concepts. Science & Education 22 (2013) 505-525.

[27] I. T. Koponen, and M. Nousiainen. Concept networks in learning: Finding key concepts in learners' representations of the interlinked structure of scientific knowledge. Journal of Complex Networks 2 (2014) 187-202.

[28] M. E. J. Newman. Random Graphs with Clustering. Physical Review Letters 103 (2009) 058701.

[29] D. Foster, J. Foster, M. Paczuski, and P. Grassberger. Communities, clustering phase transitions, and hysteresis: Pitfalls in constructing network ensembles. Physical Review E 81 (2010) 046115.

[30] B. Karrer, M. E. J. Newman. Random Acyclic Networks. Physical Review Letters 100 (2008) 118703.

[31] B. Karrer, and M. E. J. Newman. Random graph models for directed acyclic networks. Physical Review E 80 (2009) 046110.

[32] M. Nousiainen, and I. T. Koponen. Pre-service physics teachers content knowledge of electric and magnetic field concepts: Conceptual facets and their balance. European Journal of Science and Mathematics Education, 5, (2017) 74-90.

[33] E. Estrada. The structure of complex networks. Oxford University Press, 2012.

[34] E. Estrada, N. Hatano, and M. Benzi. The physics of communicability in complex networks. Physics Reports 514 (2012) 89-119.

[35] E. Estrada, D. J. Higham, and N. Hatano. Communicability betweenness in complax networks. Physica A 388 (2009) 764-774.

[36] I. S. Gradshteyn, and I. M. Ryzhik. Table of Integrals, Series and Products, 5th ed. Academic Press, 2000.

[37] A. Lipowski, and D. Lipowska. Roulette-wheel selection via stochastic acceptance. Physica A 391 (2012) 2193–2196.

[38] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of Complex Networks: A Survey of Measurements. Advances in Physics 56 (2007) 167-242.

[39] E. D. Kolaczyk. Statistical Analysis of Network Data. Springer, 2009.

[40] E. T. Jaynes. Probability Theory: The Logic of Science. Cambridge University Press: Cambridge, MA, 2003.