Quantifying the Impacts of Expanding Social Protection on Efficiency and Equity : Evidence from a Behavioral Microsimulation Model for Ghana

Osei, Robert

2019

Network cartography of university students' knowledge landscapes about the history of science: landmarks and thematic communities

Henri Lommi and Ismo T. Koponen[*]

*Correspondence:
ismo.koponen@helsinki.fi
Department of Physics, University of
Helsinki, Gustaf Hallstromin katu 2,
P.O. Box 64, FI-00014 Helsinki,
Finland

Abstract

We examine students' representations of their conceptions of the interlinked nature of science history and general history, as well as cultural history. Such knowledge landscapes of the history of science are explored by using the knowledge cartographic, network-based method of analysis to reveal the key items, landmarks, of the landscapes. We show that Katz centrality and Katz centrality efficiency are robust and reliable measures for finding landmarks. It is shown that landmarks are most often persons but include also colligatory landmarks, which refer to broader sets of events or ideas. By using Katz centrality we study how landmarks depend on periodisation of the networks to see what kinds of changes occur by changing the time window on history. The community structure of the networks is studied by using the Louvain method, to reveal the strong thematic dependence of the communities. When landmarks are studied in relation to community structure, it is found that colligatory landmarks gain importance in relation to person-centred landmarks. Network-based cartography thus reveals many features about landmarks, how communities emerge around them and how they depend on periodisation, which traditional methods can only detect or identify with difficulty. Such knowledge has direct impact on the design and planning of education and courses which could better address the need to facilitate a deeper understanding of the related nature of science history and history in general.

Keywords: Concept networks, Learning history of science, Thematic attributes, Katz centrality, Community structure

Introduction

In science education, the history of science is seen to serve the understanding of science as culture and its role in culture and society in general. Science history is then on a par with a broader history, as a basis for understanding how our culture has evolved and how different acts in the past are connected to those we encounter and see today (Gooday et al. 2008; Russell 1981; Galili and Hazan 2001; Leite 2002; Höttecke and Silva 2011). The basic skeletal scaffolding of such knowledge consists of chronologically ordered and appropriately periodised facts of what happened, when and by whom. These simple facts form the basic fabric on which more involved interpretations and deeper understandings of their consequences are located. Such a picture is a kind of historic landscape, which is

complex because of many intertwined but chronologically ordered connections contained within it (van Boxtel and van Drie 2012; van Drie et al. 2009; van Drie and van Boxtel 2008).

The multiple historical events, multiple perspectives and their interleaving produce a complex landscape of the history of science as an embedded part of broader history. In learning, students thus need to produce a complex, network scaffolding of key concepts that provide landmarks for navigating in that complex landscape. Many of these key concepts are assumed to be colligatory concepts (van Boxtel and van Drie 2012), which refer to sets of events or ideas having importance over extended periods or having overarching impacts over long temporal durations. A learner then encounters a situation, where before a deeper understanding is possible to achieve, a vast number of details must be known and acquired. This is a well-known challenge in learning history (van Boxtel and van Drie 2012; van Drie et al. 2009; van Drie and van Boxtel 2008; Dawson 2009; 2004). Only when students have acquired a rich enough knowledge base does it become possible to construct a big picture; a landscape of history and history of science being a part of that.

The scaffolding and knowledge base associated with it that allows learners to place historical knowledge in correct periods and contexts is called a "frame of reference" or "historical overview knowledge" (van Boxtel and van Drie 2012). It has been suggested that the students' knowledge on which such scaffoldings are based should contain chronologically ordered key concepts, so that key concepts can act as landmarks for navigating in the historical landscape. The most useful key concepts which act as landmarks are supposedly related to phenomena of long duration or very important events or ideas. Examples of such key concepts are, for example, the Scientific Revolution, the Industrial Revolution, the Enlightenment and the Reformation. Such concepts colligate many facts of persons, events, ideas or inventions and group them into meaningful sub-clusters, subordinated to the key concepts. This motivates referring to these key concepts as colligatory concepts (van Boxtel and van Drie 2012), which are the most important and luminous landmarks in the historical landscape. In constructing the scaffoldings, students must pay attention to the chronology of events, to periodisation, and to their duration and extension in time.

Research on how students construct their views of historical landscapes has shown how difficult such a task is for students (van Boxtel and van Drie 2012; van Drie et al. 2009; van Drie and van Boxtel 2008; Dawson 2009; 2004). Very often students' constructions of history landscapes differ from the landscapes which are targeted in teaching and instruction: Persons as landmarks may be overemphasised (Leite 2002), important landmarks may be missing, or periodisation is conceived incorrectly (van Boxtel and van Drie 2012; van Drie et al. 2009; van Drie and van Boxtel 2008). We study here the historical landscape of a group of ten university students (pre-service teachers), as revealed through a set of preparatory tasks to explore chronologically the science history of the four centuries between 1570 and 1930 divided into six time periods and how that history of science was embedded in the culture, society and politics of the same era. The periodisation followed from using certain events of significance like the initial and terminal dates. The landscapes are thus outcomes of the students' personal judgements on the relevance of given topics such as characters, ideas, inventions and events during those centuries. The landscapes are also affected by how the students

were able to recover those topics from various easily available sources, mostly from Wikipedia.

We analyse the data produced by the preparatory task by using network cartography of knowledge (more briefly, network cartography). The network cartography combines different network metrics and explorative methods, such as community detection in characterising the relevant and salient features of different nodes in the network (Guimerà and Amaral 2005; Stella and De Domenico 2018). Here, we apply the network cartographic methods in the sense they have been applied in the context of knowledge discovery (Börner 2015; Börner and Scharnhorst 2009; Chen et al. 2009; Shi et al. 2015), in knowledge acquisition (de Arruda et al. 2017; Costa 2006), and in learning (Koponen and Nousiainen 2018; 2014; Siew 2018). Regarding learning, network cartography has recently been used successfully in different areas of learning disciplinary knowledge: In revealing key concepts in students' knowledge in physics (Koponen and Nousiainen 2018; 2014), and as a predictor for success in quiz-type tasks in psychology (Siew 2018). Here, a similar type of cartographic approach is used in the context of learning history of science.

The objective of the present study is to explore the structure of university students' (pre-service teachers) knowledge landscapes of history of science and its landmarks. The practical motivation of the study originates from attempts to find solutions for teaching university students to understand the history of science as part of cultural and general history. In order to do this, we need methods and means to explore students' holistic understanding of the "big picture of history", in other words, the cartography of their history landscapes. For this purpose, complex network methods and knowledge cartography turn out to be invaluable new tools. First, we construct network-like landscapes of history that emerge from the data; second, we perform a network-based cartography of the landscape. The analysis focuses on:

- Revealing the structural properties of the network.
- Finding landmarks (i.e. the key concepts).
- Finding thematic communities emerging around landmarks.
- Revealing the effect of periodisation on landmarks and communities.

The study reported here provides information of how students in a group level conceive that historic landscape, how the most important person, ideas and events are thematically arranged within it, and what is the overall structure of that historic landscape at the group level, as collated and collective knowledge. It should be noted that the objective here is thus to find out the features and properties of students' knowledge landscapes, not the structure of the landscape as it may appear in the sources (e.g. in Wikipedia) the students used.

Methods

Historical topics, even at the simplest level, are not isolated facts but a connected and contiguous set of facts. The importance of notions and facts needs to be approached from a more holistic perspective, by considering how given facts and notions (in what follows referred to as elements of the network) are linked as part of the network, and how they affect other elements in the network not directly connected to the given element. The research problem is approached by using network cartographic methods (Börner 2015; Börner and Scharnhorst 2009; Chen et al. 2009; Shi et al. 2015) based on degree and

Katz centralities (Estrada 2012; Borgatti 2005; Katz 1953; Sharkey 2017; Wang et al. 2017) and modularity optimisation (Newman and Girvan 2004; Newman 2004). These methods provide complementary information on the skeletal structure of the network, how different thematic traits are connected to different landmark nodes (high affinity nodes), and how those distributions of centralities evolve as the network expands when the temporal window on historic periods is changed.

Empirical sample

The context of this study is a physics history course for finnish pre-service physics teachers (third and fourth year of studies), in a department of physics where they complete part of their pre-service teacher studies. The aim of the course was to introduce physics history to students as a part of more general science history and as a part of cultural and general history. The course lasted seven weeks and consisted of weekly two three-hour contact teaching sessions. The sample of data which is examined here comes from a preparatory task for a weekly topic. In these pre-tasks, students were asked to identify persons, ideas, inventions, and events of each topic they thought were of major interest or importance for the era to be discussed in more detail during the week. The total sample consists of six pre-tasks, focusing on six different historical periods, which were: i) 1572–1704; ii) 1704–1789; iii) 1789–1848; iv) 1848–1900; v) 1900–1914; and vi) 1914–1928. The periods were selected so that their beginning and ending is some significant historical event: 1572, Tycho Brahe's observation of supernova; 1704, Newton's last significant contribution to Science (*Opticks*); 1789, the French Revolution; 1848, the Year of Revolution; 1900, the origins of quantum physics; 1914, WWI; 1928, the end of old quantum physics. These periods coincide roughly with the periods when mechanics, thermodynamics, electromagnetism and quantum physics were developed and matured. Moreover, the amount of students' knowledge contents they associate with each period are of comparable size, thus producing networks which are equally extensive. This periodisation provides the chronological scaffolding on which different themes are then located for further discussion. In the sample analysed here we have included preparatory tasks by only those 10 students who provided all the required preparatory tasks, thus covering all the periods.

In the preparatory tasks, students were asked to find information and facts by using easily available sources like textbooks (Holton and Brush 2001; Simonyi 2012) and introductions to physics history from Wikipedia. The tasks were thus simple, not very deep in nature and on the level of very general knowledge. The goal was to collect background knowledge for further discussions. Regarding Wikipedia, it should be noted that in the science history and history of 1572–1928 many Wikipedia sources are rather extensive and based on extensive use of primary sources. The availability of material thus supposedly did not limit the scope and topics to any great extent. For example, the Wikipedia sources on e.g. Galileo Galilei, Isaac Newton and the Scientific Revolution, which featured as the students' targets of interest for the period 1572–1704, is several hundred pages in total. The students thus needed to make selections and decisions on what to retrieve and what they decided was relevant.

In the preparatory tasks, the students were asked to report the connections (at least 48 connection by each student in each week) between historical characters (scientists, philosophers, writers, artists, rulers etc.) and the ideas and inventions and historical events they found relevant and important for the given era, in science history as well as in

general history. They also reported why they considered these connections relevant. They reported the resulting connections in the form of pair-wise connections, for example [galilei ↔ heliocentricmodel], [newton ↔ gravity]. On the basis of these ordered (but not directed) pairs we constructed a network where each key word is a node in the network and the dyadic connection a link connecting the two nodes. The key words where attributed both periodically and thematically, on basis how students represented the appearance of key words in either in general history or science history. The resulting network, which represents students' knowledge, includes all the different keywords and connections reported by the student group of 10 students, consists of $N = 1613$ different nodes (keywords) and the $M = 2306$ different links between them. The number of links when all the nodes and links in all the pre-tasks are collated varies from about 50 links for the most extensively connected node to only one link for many of the nodes.

The data was curated recognising first from each dyad (as reported by students) the key words, and second, forming a list of closely synonymous words in order to reduce the irrelevant variations in the key words (the language was Finnish, which has many different flexions for same words). Eventually, all key words, which were considered synonymous, were merged to one key word. In total, about 700 synonymous words were found and resulted to 1613 different key words, which were used as basis to construct the networks.

The network representing the students' knowledge is explored by aggregating the six different time periods in chronological order, gradually shifting the time window towards modern times while keeping aggregated periods comparably extensive according to the content of the knowledge the students have attached to them. Agglomerated periodisation with long enough periods was chosen because in the case of networks (denoted by g_i-g_{vi} in what follows) representing individual, unaggregated periods i-vi, the emergence of landmarks is nearly completely determined by periodisation and persons dominate the periods. Therefore, the network analysis does not reveal interesting, deeper regularities for g_i-g_{vi}. Furthermore, in order to be a big picture instead of a piecewise and piecemeal collection of facts, there needs to be an aggregate of long enough periods of history, where the significance of characters, events and ideas change. A periodisation which is motivated by these demands and which provides robust enough networks for statistical analysis consist of networks aggregating three consecutive time periods: G_I (periods i-iii, 1572–1848), G_{II} (periods ii-iv, 1704–1900), G_{III} (periods iii-v, 1789–1914) and G_{IV} (periods iv-vi, 1848–1928). It is later shown that these networks are also heavy-tailed in the same manner and thus, comparable to each other. For comparison, also network G_{TOT} aggregating all six periods in 1572–1928 is also formed. Eventually, networks G_I-G_{IV} are treated as networks of thematically associative knowledge. Such an associative network provides a broad picture of how the different factual components that the students perceive to be of importance form the big picture of the given era.

Finding key nodes: the landmarks

We are interested in finding the nodes which are globally best connected. To detect such nodes, we have chosen to use Katz centrality (Estrada 2012; Borgatti 2005; Katz 1953; Sharkey 2017; Benzi et al. 2013), which is metric based on counting walks of different lengths. Moreover, it contains a factor which allows to tune how walks of different lengths are taken into account, i.e. how extensive portion of the network is explored. Katz centrality is therefore suitable for our purposes to explore the global connectivity of the network

and which nodes are the most important ones in contributing to global connectivity. In addition to Katz centrality, we also use Katz centrality efficiency (Katz efficiency in what follows), which is defined through the effect of isolation of the node (Sharkey 2017; Wang et al. 2017).

Katz centrality was originally introduced to measure the influence of actors in social networks, where the influence of interest is the ability to communicate with other actors in the network. Consequently, Katz centrality can be defined through a weighted sum of walks of given length l, because the lth power $\mathbf{A}^l$ of adjacency matrix $\mathbf{A}$ describing the network provides the number of walks of length l. The adjacency matrix $\mathbf{A}$ has element $[\mathbf{A}]_{ij} = 1$ when nodes i and j are connected, otherwise $[\mathbf{A}]_{ij} = 0$. Katz centrality can then be expressed as a sum (Estrada 2012; Katz 1953; Sharkey 2017; Benzi et al. 2013)

$$K_k = \left[\left(\frac{\mathbf{A}^0}{\alpha^0} + \frac{\mathbf{A}^1}{\alpha^1} + \frac{\mathbf{A}^2}{\alpha^2} + \frac{\mathbf{A}^3}{\alpha^3} + \dots \right) |\mathbf{1}\rangle \right]_k = \left[\left(\sum_{i=0}^{\infty} \frac{\mathbf{A}^i}{\alpha^i} \right) |\mathbf{1}\rangle \right]_k \tag{1}$$

where $|\mathbf{1}\rangle$ is the identity vector and the damping factor α decides how walks of different lengths are weighted. To have a convergence of the sum, α must be larger than the largest eigenvalue $\alpha_{\max}$ of the adjacency matrix $\mathbf{A}$. The damping factor α can be used to tune the extension of the neighbourhood to be included in the calculation of Katz centrality and thus, whether the local ($\alpha \gg \alpha_{\max}$) or global ($\alpha \to \alpha_{\max}$) environment of walks is contributing to Katz centrality.

Katz centrality, by performing the sum in Eq. (1), can also be written in the matrix resolvent form (Sharkey 2017; Benzi et al. 2013)

$$K_k = \left\{ \left[\left(\mathbf{I} - \frac{1}{\alpha}\mathbf{A} \right)^{-1} - \mathbf{I} \right] |\mathbf{1}\rangle \right\}_k \tag{2}$$

where $\mathbf{I}$ is the identity matrix. In Eq. 2 the zeroth contribution $\mathbf{A}^0/\alpha^0$ to the sum is extracted. This has no substantial effect on the resulting Katz centralities and rankings based on it. For better comparison between the different size of networks, Katz centrality can be normalised as $K_k' = K_k/\|K\|$, because normalisation improves comparability between networks of different size (Ghosh and Lerman 2011).

The resolvent form allows us to connect Katz centrality to similar types of centrality measures describing the communicability of network (Benzi et al. 2013). The matrix resolvent form also emerges from a viewpoint where the centrality of a node is seen from the dynamic control perspective, as a steady state solution to a linear dynamical system (Sharkey 2017). Such a description yields immediately to another slightly different but related interpretation, which sees the Katz centrality as a measure of influence, where an actor is more influential when its neighbours are also influential (Borgatti 2005).

The Katz efficiency is defined through the effect of isolation of a given node in the network and how this isolation effects the overall Katz centrality of the network. Such a definition of (centrality) efficiency emphasises the role of a node for other nodes. The Katz efficiency can be also motivated from the perspective of the control analysis of dynamic systems and has been shown to be a robust measure of the influence of the node within the network (Sharkey 2017). Katz efficiency can be defined in the normalised form as (Sharkey 2017; Wang et al. 2017)

$$E_k = 1 - \frac{\sum_i K_i^*(k)}{\sum_i K_i} \tag{3}$$

where $K_i^*(k)$ refers to the network where node k is isolated by removing all the links connected to it. It must be noted that the same value of α, which is based on the largest eigenvalue of the original network, must be used in calculating $K_i^*(k)$.

Katz centrality and efficiency both measure the influence of nodes in related but complementary ways. In the most unambiguous cases for undirected networks, ranking based on Katz centrality and efficiency should provide nearly the same ranking of nodes. In what follows, this is used as criteria to select the value of α. In the present case for networks G_I – G_{IV}, the best agreement between the ranking of nodes is obtained with $\alpha = 1.058\,\alpha_{max}$ which is thus about 6% larger than the largest eigenvalue of the adjacency matrix. In practice, the ranking based on Katz centrality is identical to the ranking obtained on the basis of Katz efficiency. The correlation of Katz centrality and efficiency is very high: For networks G_I – G_{IV}, the Pearson correlation ranges from 0.94 to 0.96, while Kendall-τ_B (Corder and Foreman 2014) are from 0.97 to 0.99. Therefore, in what follows, ranking based only on Katz centrality is used.

In addition to Katz centrality, we calculate the standard degree centrality (number of links attached to node), defined as

$$D = \frac{1}{2} \sum_{i \neq k} ([\mathbf{A}]_{ik} + [\mathbf{A}]_{ki}) \tag{4}$$

where $[\mathbf{A}]_{ik}$ is the element of adjacency matrix $\mathbf{A}$.

Finally, we need to test the statistical significance of information contained in Katz centrality and whether this information is redundant to degree centrality. This is done by rewiring all connections in the networks, by using 100 different fixed degree sequence models (Zweig 2016). The rewired model thus has the same node degrees, but all the links are shuffled. All rewirings are performed with `graph-tool` software (Peixoto 2014). The Z-scores are then calculated for the standardised Katz centralities for each node, defined as (Estrada 2012; Kolaczyk 2009)

$$Z = \frac{O - \langle O \rangle}{\sigma_O} \tag{5}$$

where O is the observable value in the empirical sample. It should be noted that the damping factor α must now be based on the largest eigenvalue of the adjacency matrix corresponding to the rewired fixed degree sequence model but by using the same value of p as is used for the original network. The reliability and statistical significance requires that the absolute values $|Z|$ of Z-scores are high enough, usually the value $|Z| = 2$ is taken as a limiting case. Assuming that the variables are normally distributed, Z-scores $|Z| = 2$ and $|Z| = 3.0$ correspond to p-values 0.02 and 0.001, respectively. Here, we have chosen to use $|Z| = 2$ as a cut-off for statistically significant deviations deserving special attention. As will be seen, the nodes which have Katz centrality $K > 0.05 K_{max}$ usually also have $Z > 2$.

The high Z value means that Katz centrality of a given node is for these nodes not determined nor predictable from the degree centrality (i.e. from the local connectivity). As will be shown, rewiring a network with heavy-tailed degree distributions affects the Katz centrality distribution, as is expected (Qu et al. 2015). Therefore, the nodes with high values of Z are statistically significant ones because their global connectivity is determined

by other (yet unknown) factors than local connectivity. Katz centrality, with the damping factor α chosen to provide the smallest difference between Katz centrality and efficiency, thus provide reliable and flexible measures for finding the key nodes (landmarks) in the network and also for monitoring the changing importance of landmarks when the window of periodisation is changed.

Finding thematic communities

Communities of interest here are parts of the network, which are groups of nodes connected to each other more densely than to nodes, which belong to other densely connected groups; i.e. connections within the group are dense while between the groups the connections are sparse. To find such communities one can define modularity Q in the form (Estrada 2012; Newman and Girvan 2004)
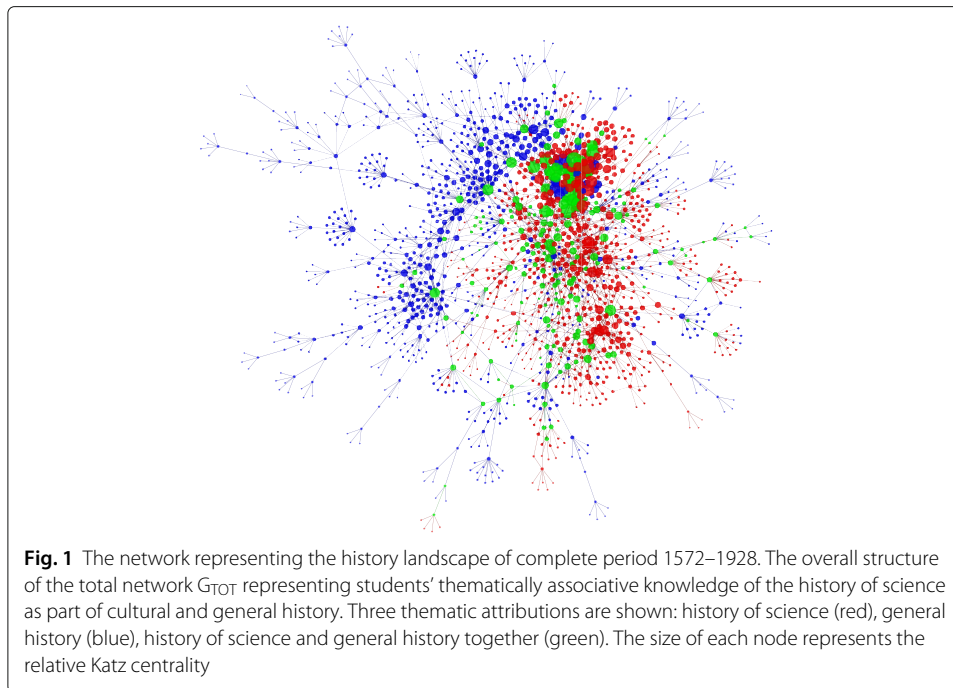
$$Q = \frac{1}{2M} \sum_{ij} \left([\mathbf{A}]_{ij} - \frac{1}{2M} D_i D_j \right) \delta(c_i, c_j) \qquad (6)$$

where $[\mathbf{A}]_{ij}$ represents the link (edge) between nodes i and j, $D_i =$ is the degree of node i (i.e the sum of the edges attached to node), M is the total number of edges, and c_i is the community to which node i is assigned. The term $D_i D_j / (2M)$ describes the expected number of edges between nodes i and j if the network is rewired. The δ-function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and describes whether two nodes belongs to same community or not. The modularity Q thus defined obtains values from -1 (fully non-modular) to 1 (fully modular), while the corresponding modularity for a random network is $Q = 0$.

The community structure of the networks can be found by optimisation of modularity, by division of the nodes into groups so that the largest possible value of modularity is achieved (Newman and Girvan 2004; Blondel et al. 2008; Fortunato 2010). To optimise the modularity, we use here the Louvain method (Blondel et al. 2008), implemented as `Python-Igraph` routine `community-multilevel` (Csardi and Nepusz 2006). The communities thus resolved turn out to reveal a very clear thematic uniformity contained in the networks representing the students' history landscape; themes apparently act as a consolidating property for community formation. Such communities are thus very context related, thematic communities. Each thematic community, however, emerges around a few landmarks.

Results

The complete network G_{TOT} that consists of all nodes and links corresponding to the period 1572–1928 is shown in Fig. 1, revealing a clear modular structure. The thematisation is shown by different colours: History of science by red, general history by blue, and if both themes occur, by green. The value of the Katz centrality (and Katz efficiency, since they are correlated) of the nodes is shown by the size of the node. Landmarks (key nodes) are recognised as nodes with exceptionally high Katz centrality (the large nodes in Fig. 1). The overall appearance of the network immediately suggests that it is heavy-tailed. To calculate Katz centrality the values of α was selected to be $p \times \alpha_{max}$, where α_{max} is the largest eigenvalue, so that p provides the smallest difference between the values of Katz centrality K and efficiency E. Parameter p was solved for each periodised subgraph individually and was found to be from 1.042 to 1.124, depending on the size and connectivity of the subgraph. For those graphs where three periods were aggregated, the p value

Fig. 1 The network representing the history landscape of complete period 1572–1928. The overall structure of the total network G_{TOT} representing students' thematically associative knowledge of the history of science as part of cultural and general history. Three thematic attributions are shown: history of science (red), general history (blue), history of science and general history together (green). The size of each node represents the relative Katz centrality

was 1.058. In that case, the correlation between K and E is optimal and both methods to rank the importance of nodes provide nearly the same rankings. The fact that it is possible to find a damping factor which yields nearly identical rankings based on Katz centrality and efficiency ensures the robustness of Katz centrality as a measure of the node's global importance; the higher the ranking of a given node on basis of its Katz centrality the more luminous it is in the global scale. Abbreviations and symbols used recurrently in the text and figures are summarized in Table 1.
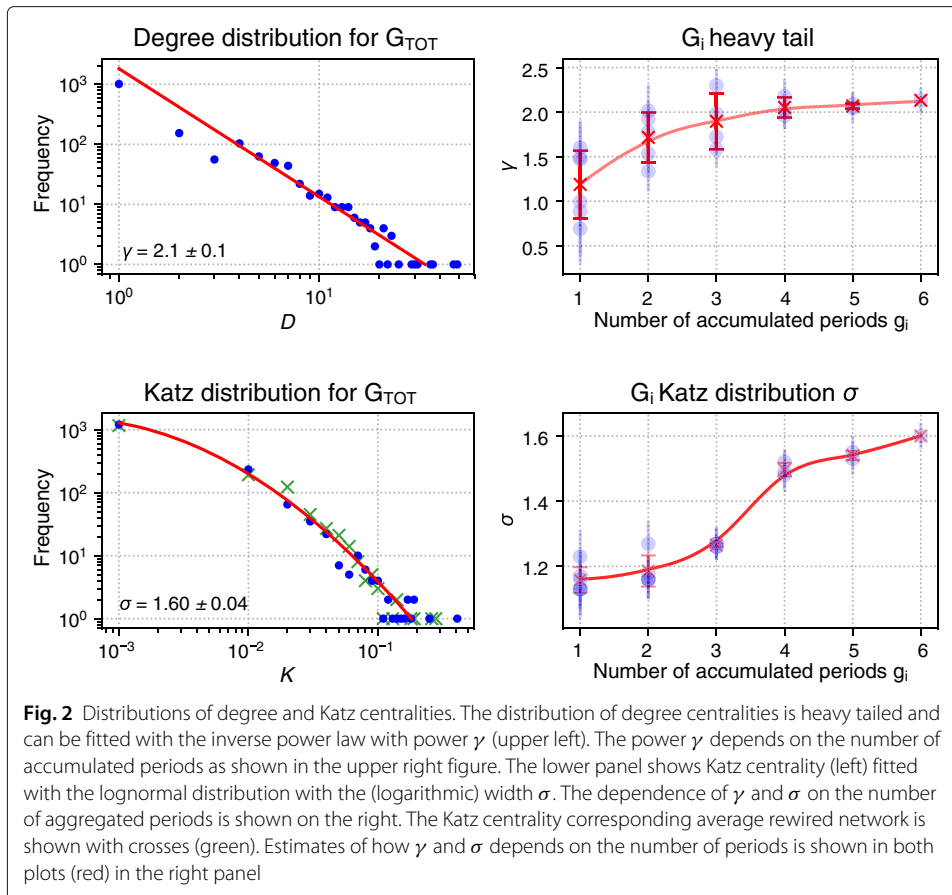
Heavy tails

The distribution of values of degree centrality D and Katz centrality K for the total network G_{TOT} is shown in Fig. 2. Heavy tails are also apparent for each six separate networks $g_i - g_{vi}$ corresponding to unaggregated periods i-vi, as well as for networks $G_I - G_{IV}$ for aggregated periods (and furthermore, all different combinations of periods i-vi). Although the degree distribution of networks do not strictly conform to the inverse power laws, fitting them with inverse power law distribution of the form

$$P(x) \propto x^{-\gamma}, \quad x \in \{D\} \tag{7}$$

provides an appropriate measure to monitor the changes in the heavy-tail of the degree distribution by using a single parameter γ.

Table 1 Summary of symbols and abbreviations used recurrently in the text and figures

Symbol	Abbreviation	Symbol	Abbreviation
A	Adjacency matrix	D	Degree centrality
$[\mathbf{A}]_{ij}$	Element *ij* of matrix **A**	K	Katz centrality
α	Damping factor	E	Katz efficiency
Z	Z-scores	Q	Modularity
N	Number of nodes	A	Assortativity
M	Number of links	C_L	Closeness centrality
γ	Inverse power	C_C	Local Clustering
σ	Width of lognormal distr.	Φ	Fragility

Fig. 2 Distributions of degree and Katz centralities. The distribution of degree centralities is heavy tailed and can be fitted with the inverse power law with power γ (upper left). The power γ depends on the number of accumulated periods as shown in the upper right figure. The lower panel shows Katz centrality (left) fitted with the lognormal distribution with the (logarithmic) width σ. The dependence of γ and σ on the number of aggregated periods is shown on the right. The Katz centrality corresponding average rewired network is shown with crosses (green). Estimates of how γ and σ depends on the number of periods is shown in both plots (red) in the right panel

The Katz centrality distributions, as shown by Fig. 2, are for present purposes better fitted with the lognormal distribution of the form

$$P(x) \propto \frac{1}{x} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x \in \{K\} \tag{8}$$

where μ is the (logarithmic) mean and σ is the (logarithmic) width as a standard deviation of the lognormal distribution. When lognormal distribution is used as a fit for heavy-tailed distributions, an appropriate characteristic is the width σ for monitoring Katz centrality distributions in different networks.

The degree distribution, fitted with an inverse power law, and Katz centrality distribution, fitted with a lognormal, for network G_{TOT} are shown in Fig. 2. In fitting, the lowest degrees (between 1–4) were excluded. In order to find the value of γ we used the linear regression approach implemented as `SciPy` function `linregress` (Jones et al. 2001) since the logarithms of frequencies and degrees are directly proportional. To find the values of σ of the lognormal fits, we fit the data in log-log scale by using the minimum chi-square estimation in `SciPy`. In both cases, standard deviation of residuals are used to estimate errors in fits. Note, however, that quantitatively detailed fitting is not attempted and γ and σ are used only for purposes to demonstrate the heavy-tailedness of the degree distributions.

The fitting coefficients γ and σ for six periods g_i-g_{vi}, for the aggregated periods G_I-G_{IV} and for the total period G_{TOT} are reported in Table 2. In general, the results can

be summarised by noting that networks are heavy-tailed so that networks display the presence of hubs with characteristic inverse power law fits for the degree distributions, approximately corresponding to powers from 1 to 2. The Katz centrality distributions are also heavy tailed with σ approximately in range from 1.2 to 1.6. In practice, lognormal distributions with such large values of σ are by appearance very close to inverse power law distributions. As Fig. 2 shows (at right) the powers γ and values of σ are both converging to certain fixed value respectively when at least three consecutive periods are aggregated. Note, however, that the fits are shown only for purposes to characterise the heavy tailed nature of distributions, without claiming any overarching conclusions based on the specific form of the distributions.

For completeness, we report in Table 2 average values of closeness centrality C_C, local clustering coefficient C_L, modularity Q and assortativity A (Estrada 2012). All these values are quite similar in all the networks, with values in the ranges of $0.15 < C_C < 0.24$, $0.35 < C_L < 0.44$, $0.66 < Q < 0.84$ and $-0.12 < A < -0.07$. Interestingly, clustering and modularity are relatively high for all networks. Otherwise the global values show little if any systematicity and their changes from one network to another are too small to have any greater interest.

The correlation between degree (D) and Katz (K) centrality is relatively low for G_I-G_{IV}, the Pearson and Kendall-τ_B correlation coefficients between the values of D and K are in the ranges of 0.69–0.76 (0.62 for G_{TOT}) and 0.42–0.51 (0.53 for G_{TOT}), respectively. Consequently, some, but not all, of the nodes which have a high degree centrality D also have a high Katz centrality K. Finding a relatively moderate correlation between the values of D and K indicates that despite the apparent similar generic form of distributions, local connectivity is not, after all, a very good predictor of global connectivity, which is the primary property of interest here.

In order to justify the use of Katz centrality (and efficiency) as an appropriate measure to recognise the key nodes, we compare the Katz centralities obtained for a rewired fixed degree sequence model with the Katz centralities in the original networks. The rewired

Table 2 Characteristics of networks g_X corresponding to distinct periods $X = i, \ldots, vi$ and G_X of aggregated periods $X = I, \ldots, IV$
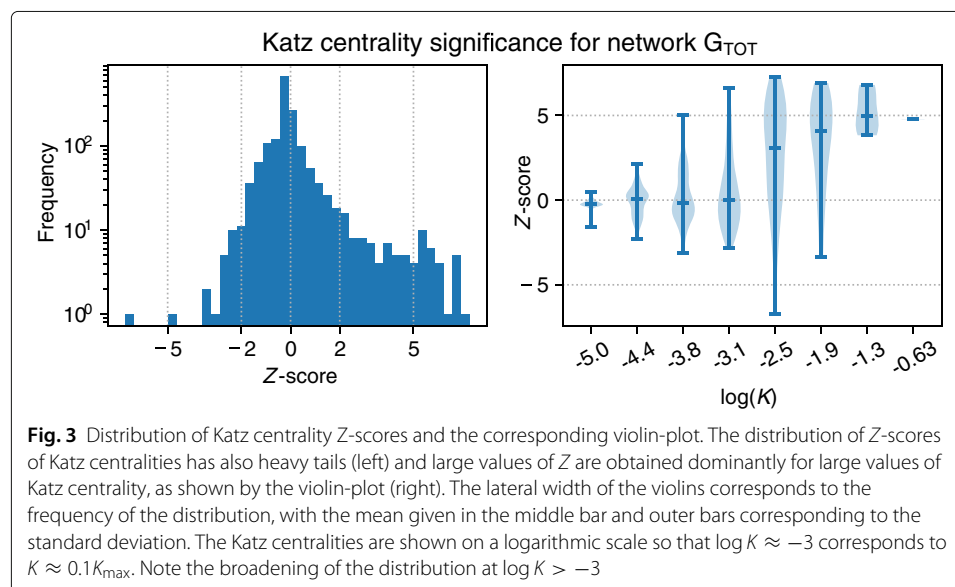
Netw.	Sizes		Fitted parameters		Correlations		Global invariants				Fragility
	N	M	γ	σ	R^2	τ_B	C_C	C_L	Q	A	Φ
g_i	239	356	1.0 ±0.3	1.23 ±0.07	0.90	0.51	0.24	0.40	0.66	-0.09	0.04
g_{ii}	311	392	1.5 ±0.4	1.17 ±0.07	0.74	0.45	0.18	0.43	0.80	-0.10	0.10
g_{iii}	326	424	1.5 ±0.2	1.13 ±0.07	0.75	0.34	0.16	0.39	0.80	-0.09	0.18
g_{iv}	158	190	0.9 ±0.6	1.13 ±0.07	0.77	0.57	0.18	0.35	0.78	-0.12	0.17
g_v	208	254	1.6 ±0.3	1.13 ±0.07	.77	0.51	0.16	0.38	0.79	-0.12	0.14
g_{vi}	308	375	0.7 ±0.3	1.16 ±0.07	0.77	0.44	0.19	0.44	0.80	-0.11	0.20
G_I	826	1212	1.7 ±0.2	1.27 ±0.05	0.71	0.51	0.18	0.37	0.78	-0.06	0.03
G_{II}	858	1149	2.0 ±0.2	1.26 ±0.04	0.75	0.42	0.16	0.39	0.83	-0.08	0.03
G_{III}	796	1053	2.3 ±0.2	1.26 ±0.03	0.69	0.46	0.15	0.37	0.84	-0.08	0.07
G_{IV}	757	992	1.6 ±0.2	1.27 ±0.05	0.76	0.48	0.17	0.39	0.83	-0.09	0.16
G_{TOT}	1613	2306	2.1±0.1	1.60 ±0.04	0.62	0.53	0.16	0.36	0.83	-0.07	0.03

Power γ is for fitted inverse power law distributions fitted to degree (D) centrality distributions. The (logarithmic) width σ is for lognormal distributions fitted to Katz (K) centrality distributions. The relative errors of fits are estimated from the standard deviation of residuals. The correlations between values D and K are for Pearson (R^2) and Kendall-τ_B ranking (τ_B) correlations. The summarised global invariants are average values of Closeness centrality (C_L), Local Clustering coefficient (C_C), Modularity (Q) and Assortativity (A). For each network, the number of nodes N and links M are provided, as well as the fragility Φ

networks are very similar to the original networks and retain the heavy-tailed appearance of Katz centrality distributions. For comparison, Katz centrality distribution averaged over 100 different rewired networks corresponding to G_{TOT} is shown in Fig. 2 along the results for original network. As is expected for heavy-tailed distributions (Qu et al. 2015), rewiring affects the centrality distribution, but it is very difficult to quantify the change directly from the distributions. To quantify the magnitude of the differences we have calculated the distribution of Z-scores for K and how they correlate with the values of K. This is shown in Fig. 3 for network G_{TOT}.

The distribution of Z-values and the violin-plot show that interesting nodes are those ones which have Katz centralities $K > 0.05$ (corresponding to $\log K \approx -3$), the average value of Z-scores exceeds a value $Z = 2$, which means that many nodes with $K > 0.05$ deviate significantly from what is expected on the basis of the rewired model (null-model) (Kolaczyk 2009). Although these nodes are rare, as seen from histogram in Fig. 3 (at the left), they are the most important nodes in the sense of global connectivity, and, furthermore, better connected than expected on basis of their local connectivity. The nodes with $K < 0.05$ are not statistically significant in the sense that their Katz centrality is predictable on basis of local connectivity (i.e. degree of node). Therefore, most of the nodes with high Katz centrality values can be considered as significant hubs in the networks which represent students' knowledge.

Because the Katz distributions are heavy-tailed, the highest-ranking nodes with a high value of Katz centrality and Katz efficiency have rapidly decreasing values when ranking decreases, thus justifying the choice to focus on nodes with high Katz centrality to find the landmarks of students' knowledge landscapes. This allows us to use the centrality values, at least to a limited degree, to assess the relative global importance (i.e. luminosity within the landscape) of landmarks (e.g. Newton has $K = 0.417$ while Galileo has $K = 0.263$, making Newton almost two times globally more luminous). Of course, attaching ordinal values to importance, through the values of K, is only for the practical purposes of making gross comparisons and should not be taken literally.

Fig. 3 Distribution of Katz centrality Z-scores and the corresponding violin-plot. The distribution of Z-scores of Katz centralities has also heavy tails (left) and large values of Z are obtained dominantly for large values of Katz centrality, as shown by the violin-plot (right). The lateral width of the violins corresponds to the frequency of the distribution, with the mean given in the middle bar and outer bars corresponding to the standard deviation. The Katz centralities are shown on a logarithmic scale so that $\log K \approx -3$ corresponds to $K \approx 0.1 K_{max}$. Note the broadening of the distribution at $\log K > -3$

Finally, before turning on more detailed discussions concerning landmarks as based on Katz centrality, we need to examine the robustness of the networks in respect to one student's contribution on the group-level knowledge. In order to test robustness of the analysis for group level data we formed ten modified networks for each corresponding periods and aggregated periods by removing the contribution of each single student in turn. For each of these modified networks, Katz centrality was calculated using the same weight α as in calculating Katz centralities for the original network. The Pearson correlation coefficients ρ between Katz centralities of modified and original networks were calculated. On this basis we formed a measure of the fragility Φ, which is the difference of maximum (max) and minimum (min) correlation coefficient divided by the average (avr) value of correlation coefficients

$$\Phi = \frac{\rho_{\max} - \rho_{\min}}{\rho_{\text{avr}}} \tag{9}$$

The fragility is thus a measure of relative importance of a single student's contribution to the distributions. Small value of fragility $\Phi < 0.10$ means that Katz centralities are essentially unchanged by removal of one student's contribution on the group level knowledge. The largest fragility of $\Phi = 0.20$ corresponds to a situation (period g_{vi}) where one student contributed about 10% of all links. The fragility values for each network are summarised in Table 2 and those reveal that the data is more fragile in periods g_i-g_{vi} ($0.04 < \Phi < 0.20$) than in aggregated periods G_I-G_{IV} ($0.03 < \Phi < 0.16$). In general, fragility is quite modest and thus the analysis can be considered as quite robust.

The landmarks

The landmarks are recognised as globally central key nodes that have high Katz centrality K (and, correspondingly, high Katz efficiency E). In what follows, the focus of the study shifts to networks G_I-G_{IV} and to corresponding periodic windows. The network representing students' knowledge shown in Fig. 1 is studied as periodised networks G_I-G_{IV}. In each period, certain landmarks emerge as summarised in Table 3.

Table 3 Landmarks in periodic windows G_I-G_{IV}

G_I (1572–1848)			G_{II} (1704–1900)			G_{III} (1789–1914)			G_{IV} (1848–1928)		
K	D	Landmark	K	D	Landmark	K	D	Landmark	K	D	Landmark
.42	46	*Newton*	.36	37	**French revol**	.36	30	*Faraday*	.31	30	*Einstein*
.26	29	*Galileo*	.22	30	*Faraday*	.23	12	**Electromagn**	.29	49	**WW I**
.20	21	Empir meth	.18	17	*Napoleon I*	.21	9	*Ampere*	.25	22	**Quant mech**
.20	15	**Scientfc revol**	.17	31	Steam engine	.19	14	Electr curr	.23	18	*Planck*
.19	14	Heliocntr wv	.17	23	**Indstr revol**	.19	6	*Oersted*	.21	15	*de Broglie*
.18	11	*Kepler*	.17	16	Electric curr	.18	10	Volta pile	.19	17	*Bohr*
.17	14	Telescope	.16	20	*Marx*	.18	17	*Maxwell*	.17	11	*Schrodinger*
.17	14	Gravit. law	.14	10	*Ampere*	.17	10	Galvanomtr	.15	8	*Born*
.15	21	*Brahe*	.14	17	*Voltaire*	.15	26	**French revol**	.15	9	*Heisenberg*
.15	13	Gravity	.14	10	Volta pile	.13	23	Nobel prize	.13	25	Nobel prize
.14	15	*Bacon*	.13	14	*Volta*	.13	16	*Planck*	.12	16	Bohr model
.13	15	*Descartes*	.13	23	*Franklin*	.13	6	Ampere law	.12	10	*von Laue*
.13	10	Kepler laws	.13	12	**Electromagn**	.12	17	*Napoleon I*	.12	7	Energy
.12	9	Royal soc	.12	19	*Lavoisier*	.11	17	*Einstein*	.11	21	Finland indp
.11	10	*Hooke*	.12	9	Metric syst	.11	6	*Henry*	.11	7	Wave-part d

For each window, the 15 highest ranking (based on Katz centrality K) landmarks are given, with D indicating their degree centrality
The colligatory concepts of history are shown in boldface and persons in italics

The results of the analysis show that in the networks that represent the students' landscapes of history, the landmarks appear to be nearly always persons; the overall picture is thus quite person-centred. However, some of the landmarks are colligatory, referring to set of ideas or events, for example the Enlightenment and the Scientific Revolution, or the French Revolution and World War I (see Table 3). The colligatory landmarks, understood as colligatory concepts of history, are shown in boldface and persons in italics (as also in Table 4). Although not as common as landmarks related to persons, the colligatory landmarks have nevertheless a significant role in the network. This role is also evident in how

Table 4 Communities and landmarks in periods G$_I$ - G$_{IV}$

GI (1572–1848)		GII (1704–1900)		GIII (1789–1914)		GIV (1848–1928)	
%	Landmark	%	Landmark	%	Landmark	%	Landmark
Community A (6.4)		Community E (4.2)		Community I (4.4)		Community M (3.5)	
6.5	*Newton*	5.2	*Faraday*	8.3	*Faraday*	7.1	**Quant mech**
4.1	*Galileo*	4.1	Electr curr	5.3	**Electromagn**	6.8	*Planck*
3.1	Empir meth	3.4	*Ampere*	4.7	*Ampere*	5.9	*de Broglie*
3.1	**Scientfc revol**	3.2	Volta pile	4.4	Electr curr	5.6	*Bohr*
3.0	Heliocntr wv	3.2	*Volta*	4.3	*Oersted*	4.9	*Schrodinger*
2.8	*Kepler*	3.2	*Franklin*	4.2	Volta pile	4.5	*Born*
2.7	Telescope	3.1	**Electromagn**	3.8	Galvanomtr	4.3	*Heisenberg*
Community B (1.1)		Community F (3.3)		Community J (1.7)		Community N (2.3)	
9.5	**French revol**	10.9	**French revol**	10.3	*Maxwell*	12.7	**WW I**
8.3	*Voltaire*	4.8	*Marx*	6.4	*Einstein*	2.7	Arms race
3.6	*Rousseau*	4.3	*Voltaire*	5.7	*von Laue*	2.4	Versail trt
3.4	*Marx*	3.4	*Rousseau*	5.1	emg radiat	2.4	*vHindenburg*
3.1	**Liberalism**	3.2	Febr revol	4.9	S thr relat	2.1	France
2.8	Encycloped	2.9	Vienna cng	4.5	Field	2.1	Tank
2.6	Human rights	2.6	Encycloped	3.9	Maxwell eqs	2.0	Assasint FrFr
Community C (1.0)		Community G (1.9)		Community K (1.7)		Community O (1.9)	
5.1	*Faraday*	9.2	Steam engn	8.8	**French revol**	16.7	*Einstein*
3.7	Electric curr	9.1	**Industr revol**	7.1	*Napoleon I*	6.4	Energy
3.6	*Ampere*	6.3	Steam car	3.7	*Gauss*	4.7	G thr relat
3.3	Volta pile	5.0	*Watt*	3.5	Finnish war	4.6	S thr relat
3.2	*Volta*	3.6	*Newcomen*	3.5	Napolnc wars	4.1	Photoel eff
3.0	**Electromagn**	3.5	Spinning J	3.1	Vienna cng	4.1	Mass-E eqv
2.9	Battery	3.4	Steam ship	2.6	Trafalgar btl	3.3	Fission
Community D (0.9)		Community H (0.9)		Community L (1.2)		Community P (1.7)	
6.7	Steam engine	18.7	*Napoleon I*	10.8	Nobel price	8.0	Nobel prize
5.2	**Industr revol**	10.2	Napolnc wars	6.8	X-rays	7.4	*von Laue*
3.6	Steam car	7.9	Trafalgar btl	6.2	*M. Curie*	5.9	emg radiation
3.1	*Watt*	6.1	Finnish war	5.5	*Roentgen*	5.8	X-ray crystlgr
2.5	Steam ship	4.2	*Alexander I*	4.5	*Becquerel*	4.4	X-rays
2.5	Spinning J	4.0	Waterloo btl	4.0	*P. Curie*	3.3	*M. Curie*
2.3	*Newcomen*	3.3	Battles	3.6	*Nernst*	2.9	*Roentgen*

For each period, the four most extensive communities (in total 16, from A to P) are listed, with sum $\sum K$ of Katz centrality of all nodes contained in it given in parenthesis. For each community A-P, the seven highest ranking landmarks are given, with % indicating its contribution or relative Katz centrality within the community
The colligatory concepts of history are shown in boldface and persons in italics

certain landmarks fade in or out when the periodic window changes. Such landmarks apparently provide contiguity between different periods and act as bridges, connecting themes and events.
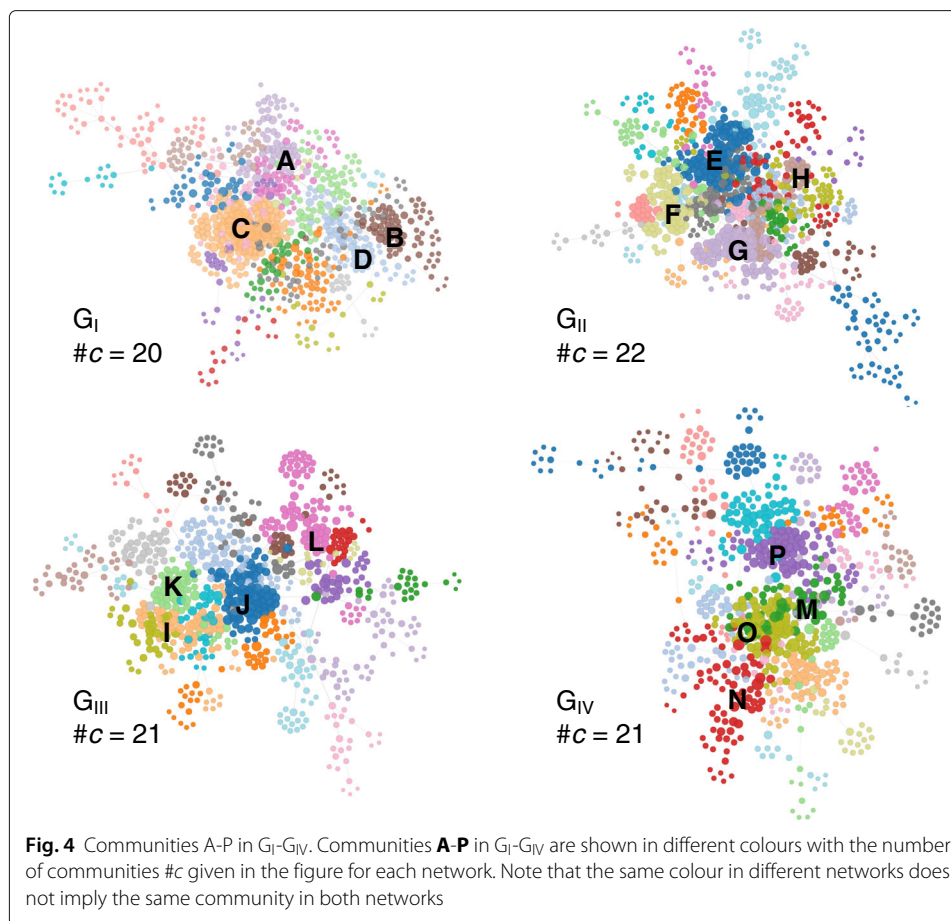
A few landmarks are astonishingly luminary revealing a very strong impact on students' conception of history of science and history in general. For example, Newton has a very central position in G_I, throughout 1572–1848, but gives eventually way to Faraday, who dominates G_{II} and G_{III}. Of colligatory luminaries, the French Revolution is the most important from G_I to G_{III}. Another question is if such a position is warranted or rather an outcome of the way history is usually treated in various sources, textbooks and popularisations. The role of landmarks in a ranking-based role as reported in Table 3, is not, however, the complete picture. As is evident from Table 3, Newton has very high Katz centrality, larger than the French Revolution, while in G_{II} the French Revolution has the largest Katz centrality and surpasses, for example, Faraday. On closer inspection, it turns out that landmarks belong to different communities and within these communities they have roles not immediately derivable from Katz centrality alone. To explore the community dependent role, we next turn to discuss the thematic communities.

Communities and landmarks

The network, as shown in Fig. 1, already reveals apparent modularity, and thus, community structure is expected. To find out how communities emerge on a structural basis, we have performed modularity optimisation. The Louvain method (Blondel et al. 2008) to find communities reveals altogether 84 different communities in graphs G_I - G_{IV}, in which different communities in different graphs corresponding to different time periods share common nodes. The community structure separated by colours and with most important communities identified by alphabetic tags is shown in Fig. 4.

The community structure is summarised in more detail in Table 4, where the top seven landmarks are listed for each of the four most important communities within each periodisation. For the communities, the most important landmarks are those nodes that have the highest values of centralities within their own community. In Table 4 the relative importance of top landmarks is seen to be several times higher than most other landmarks such as WW I or Napoleon I, which reflects the heavy-tailed distribution of the centralities of the landmarks, related to their global connectivity. On the other hand, in a few communities the relative importance is quite similar for most of the top seven members, as it is in community E and M.

The ranking of landmarks within thematic communities (Table 4) reveals a different picture than that obtained by simple landmark ranking as shown in Table 3. Now, within the communities, many colligatory concepts emerge as (colligatory) landmarks; the overall picture of the history landscape is no longer so person-centred. Among such colligatory landmarks we find landmarks related to cultural and societal changes (i.e. the Scientific Revolution, the Industrial Revolution and Liberalism), to political changes (e.g the Arms Race). In addition, several central historical events now become visible as landmarks (e.g. the French Revolution, the February Revolution and WWI). In many cases, although communities are exclusive within a given period (i.e. there is no overlap in landmarks), the same landmarks can appear in communities that emerge in different periods. This, in fact, shows a temporal and thematic contiguity between different periodical windows. In Table 4, for each of the periodical windows

Fig. 4 Communities A-P in G_I-G_{IV}. Communities **A**-**P** in G_I-G_{IV} are shown in different colours with the number of communities #c given in the figure for each network. Note that the same colour in different networks does not imply the same community in both networks

G_I - G_{IV} the four most important communities (**A**-**P**) as discussed briefly in what follows.

Community A of periodical window G_I focuses on the Scientific Revolution. In it the most luminous landmarks are Newton and Galileo Galilei. Two central ideas and inventions are also included: the empirical method and the heliocentric model. This community is the largest of all communities, and as such, rather self-contained in the sense that none of the top five landmarks it contains appear in any other community.

Community B, which is the second largest in G_I, has the theme of general cultural and political ideas, centralising on the French Revolution (and on the Enlightenment, although it does not belong to the top five landmarks) and Liberalism. Therefore, in this community Marx appears with Voltaire (and Rousseau). Landmarks which belong to this community appear again in communities F of G_{II} and K of G_{III}.

Community C of G_I is about electricity. The person-type landmarks within it are Faraday, Ampère and Volta, as one might expect. In this community, we also see the first signs of scientific phenomena and technological inventions as landmarks: the electric current and Volta's pile. Community C is again encountered in a very similar guise in periodic windows G_{II} and G_{III} but now, in both cases as a dominant community.

Community D of G_I is interesting because it is nearly entirely about technology: the steam engine, the steam engine car, steam ships and the Technological Revolution. Appropriately, the person-type landmark of this community is James Watt. This community

emerges again nearly intact as community G of G_{II} but then disappears from the top four communities.
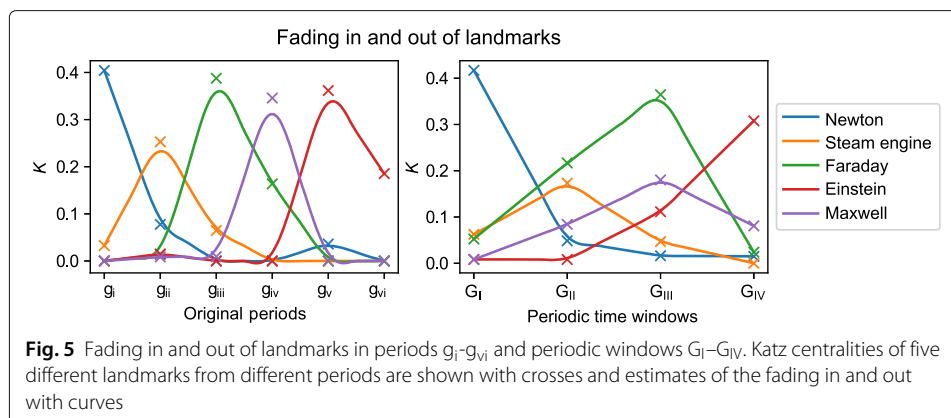
Communities E, **F** and **G** of G_{II} are very similar to communities C, B and D, respectively, contained in G_I. This means that when the periodic window is changed from 1572–1848 to 1704–1900, the community focused on the Scientific Revolution drops out and communities related to the French Revolution, electricity and (steam) technology attain the dominant role. Thus, we find in G_{II} only one genuinely new **Community H**, which is about general history, related to the Napoleonic wars (and the Continental Blockade, although it does not appear among the top five landmarks).

Communities I and **K** of G_{III} are again similar to the communities already encountered in G_I and G_{II}. Community I, which is similar to C and E, is again topmost, like E in G_{II}, while K has shifted to a lower rank, comparable to its rank in G_I. **Community J** of G_{III} is divided into the topics of electromagnetism (Maxwell and electromagnetic radiation) and the Special Theory of Relativity (Einstein and the Special Theory of Relativity), which can be both tied to the concept of the electromagnetic field.

When the time window shifts towards more modern times, to 1928, Einstein, the Special and General Theory of Relativity and other achievements of Einstein (energy as part of energy-mass equivalence) as well as photoelectric effect form a new **community O**. In G_{III} and G_{IV}, X-rays and radioactivity appear as **community L** (in G_{III}) and **community P** (in G_{IV}), both interestingly containing an institution: the Nobel prize. **Community M** of G_{IV} covers the time window 1848–1928, where quantum mechanics appears and takes the dominant position. This community, however, is very person centred. In G_{IV}, the second important **community N** is centred entirely on WW I.

The communities above show very clear and strong thematisation. They also reveal a landscape, in which the Scientific Revolution is a massive and solitary part of the landscape, the era of the Enlightenment to Liberalism is a patchwork interleaved with the Industrial Revolution and the science of electricity covering the years 1704 up to final years of 19th century and infused with a substantial portion of general history. The individuals who seem to arise as the most luminous landmarks are Newton, Galileo Galilei and Faraday. The (thematic) communities show continuity in relation to periodisation and the periodic windows. For example, community C, which is already present in G_I, gains importance in G_{II} and G_{III}, where it is the highest ranking community. Similarly, community B of G_I retains it rank in G_{II}, and community F is still present in G_{III} but as a lower rank community. There is thus fading in and fading out of communities when the periodical window changes and approaches modern times. Such fading in and out is important in maintaining the contiguity within sets of contiguous historical items and events.

The fading in out of thematic communities is also evident when landmarks contained in the thematic communities are examined. As communities emerge around landmarks, the fading in and out of landmarks provides complementary information for tracking the importance of certain themes when periods and periodised windows on history are changed. Figure 5 shows how the Katz centrality of certain landmarks change when they are examined either in original periods g_i-g_{vi} and then in periodised windows G_I-G_{IV}. It is now seen how, for example, luminous landmarks like Newton, Faraday and Maxwell fade out as more modern times approach, but also how they are astonishingly persistent landmarks, i.e. clearly important in how students scaffold their knowledge. When, on the other hand, the fading in and out of such landmarks as Faraday, Einstein and the steam engine is

Fig. 5 Fading in and out of landmarks in periods g_i-g_{vi} and periodic windows G_I–G_{IV}. Katz centralities of five different landmarks from different periods are shown with crosses and estimates of the fading in and out with curves
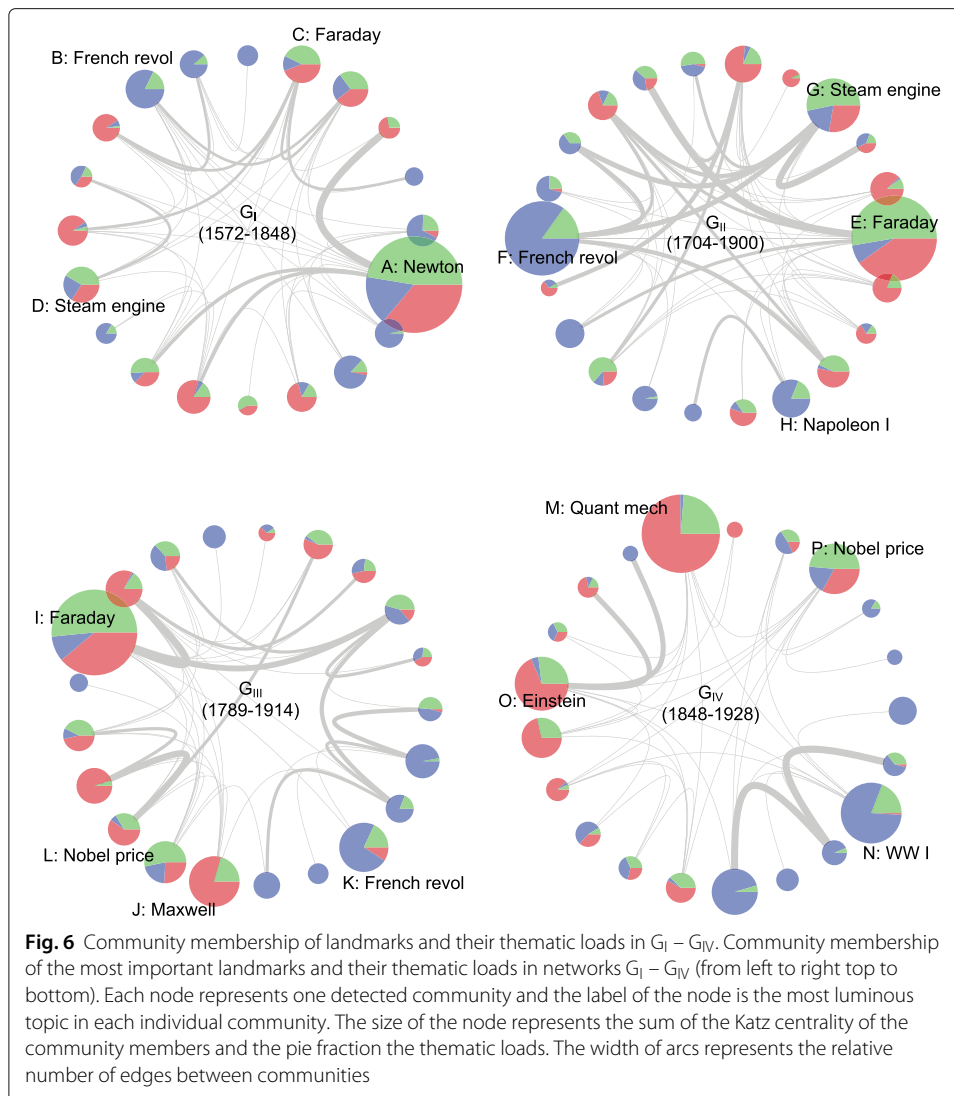
examined through periodisation, it becomes immediately clear how important such landmarks are in connecting periods and in providing contiguity. This role of landmarks is important for properly understanding the temporal duration of events.

The historic landscape and its landmarks as they appear through thematic communities and different periodic windows seem appropriate enough scaffolding for discussions where the meaning of landmarks is elaborated and inspected from different perspectives. The role of such landscape, first and foremost, is to provide a common scaffolding and common themes for discussions on how to deepen our understanding of the connections between the history of science and general history, and persons, ideas, inventions and events which are part of them both.

Changes in thematic attributions of communities also provide interesting clues to how students scaffold their knowledge of history. In the original task, the students expressed in their evaluation whether an item they provided as a dyadic, thematically associated connection, belongs to the history of science or general history. It is thus straightforward to tag the items (nodes) in three exclusive groups: history of science (red), general history (blue) or both (green). Based on the total Katz centrality of nodes in each group, we have formed the thematic loading of community as relative fractions of total Katz centralities. Now, when the network is divided into communities based on the optimisation of modularity, we can track how different communities are formed around landmarks and how the three different thematic loads are visible in different communities. This, indirectly, provides information about how the students conceive the thematic attributions and their role in the big picture, in the historic landscape in its totality. The thematic loading of communities reveals the role of landmarks changes within the community when the periodic window changes and gradually approaches modern times. The changes in thematic loadings are shown in more detail in Fig. 6, which displays how landmarks within different communities with different thematic loadings are related to each other. In that figure, a landmark in a given community is shown as having the same thematic loading as its community has, and the connection between landmarks represent the connectivity of their communities.

Figure 6 shows how the landmarks corresponding to different communities fade in and out in when periodisation moves towards more modern times, and how the thematisation loads of the landmarks change. The thematic loads of general history and history of science are more balanced for earlier periods of G_I and G_{II} than for the later period

Fig. 6 Community membership of landmarks and their thematic loads in G_I – G_{IV}. Community membership of the most important landmarks and their thematic loads in networks G_I – G_{IV} (from left to right top to bottom). Each node represents one detected community and the label of the node is the most luminous topic in each individual community. The size of the node represents the sum of the Katz centrality of the community members and the pie fraction the thematic loads. The width of arcs represents the relative number of edges between communities

G_{IV}, where the history of science becomes more separated from general history. The connections between different thematic communities are also much denser in G_I and G_{II} in comparison to G_{IV}. Both findings show that clear segregation or separation of science history from general history (and cultural history as part of it) takes place. This, of course, might reflect the objective historical evolution and professionalisation of science, but perhaps even more how modern history of science is represented in textbooks and those sources students have used. Nevertheless, the results show unambiguously that such a separation is present in students' conception of the history of science and general history. This behaviour gives clues to how students acquire, select and integrate knowledge when they have to deal with complex and information-rich situations.

Discussion and conclusions

We have here examined how the students' initial ideas of how different characters in the history of science, scientific ideas, inventions, and events in general history are connected and how that knowledge forms a kind of history landscape. The data used as the

starting point consist simply of connected pairs of these items. Despite the apparent simplicity, such pairwise associations are unexpectedly revealing when they are examined as a connected set of (thematic) associations; the separate pairwise associations form a tightly connected network. This connected network forms a kind of thematic associative network of history on which the students are able locate the events; i.e. the knowledge landscape of history. The advantages of such scaffolding for learning history has been repeatedly pointed out, thus warranting an attempt to find out how such scaffolding may appear and what it looks like.

The present study is an exercise in the cartography of students' knowledge, based on the utilisation of graph visualisation and certain centrality measures to locate the most essential elements of the history landscape. The property of interest in regard to the landscape, is the global connectivity of knowledge items within the scaffolding. When the scaffolding is represented as a network, such global connectivity can be quantified by counting all the paths (or walks), where a node is connected to other nodes in the network. For this Katz centrality provides a suitable measure. By using Katz centrality, we can find the key nodes, i.e. landmarks, of the history landscape. The key findings of the analysis, to be discussed in more detail in what follows, are:

- Degree and Katz centrality distributions are heavy-tailed.
- Landmarks are usually persons but colligatory landmarks also appear.
- Thematic communities emerge around landmarks and are historically relevant.
- Landmarks fade in and out meaningfully when the periodical window changes.

An interesting finding of the analysis is that networks, which represent students' knowledge landscapes, have clear hubs, seen as as heavy-tailed distributions of degree and Katz centrality. The heavy-tail of degree distributions, when fitted with the inverse power law, reveals exponents in the range of 1 to 2, value 1 corresponding to the usual Zipf law, which is ubiquitous in word frequency ranking distributions and many linguistic networks (Thompson and Kello 2014; Kemp and Tenenbaum 2008; Morais et al. 2013). Powers from 1.5 to 2 are encountered in information and knowledge networks like Wikipedia and in semantic networks resulting from acquiring knowledge (terms) from large knowledge networks (Thompson and Kello 2014; Masucci et al. 2011). The heavy-tailed structure of students' knowledge, as it emerges at the group level, may then reflect the structure of Wikipedia; students' knowledge landscapes has a heavy-tailed distribution of degree centrality of nodes (i.e. luminous nodes), because Wikipedia's pages have a heavy-tailed degree centrality distribution (Masucci et al. 2011). Similarly, the community structure of the students' knowledge landscape may also reflect the thematic organisation of Wikipedia, through affecting the students' landscapes owing to heuristic knowledge foraging strategy (Thompson and Kello 2014). However, it is important to note that students do not forage for knowledge from Wikipedia in a random way, but for the purpose of finding knowledge which is relevant to their task. Students make decisions about the relevance of their choices; the retention process is not random but involves consideration. This means that the outcome which students decide to express as pairwise connections, and which eventually forms the network we have analysed here, is an outcome of two independent processes, namely foraging for knowledge and making decisions about the relevance of that knowledge. Another reason, however, is connected to the fact that group-level knowledge is studied here; knowledge of individual students is aggregated or

collated. In such situations, it is very common that although an individual student's knowledge network does not show a heavy-tailed structure, the collated knowledge network is heavy tailed (Kemp and Tenenbaum 2008; Morais et al. 2013).

The landmarks (nodes that have high Katz centrality) of students' knowledge landscapes of history of science are nearly always persons, mainly scientists. The overall picture is thus quite person-centred. This is perhaps a reflection of the biography-dominated history perspective that textbooks tend to offer (Leite 2002). Such a result suggests that students' history conception is dominated by persons who are seen to be superior to most other scientists of the same era. The popularity and fame of certain heroic characters in history is certainly one reason for the emergence of the heavy tails of centralities in the network structure. One reason for strong person centred views may well be the way in which textbooks and popularisations of science tend to present the role of "great scientists" as heroic characters advancing science (Leite 2002), possibly reinforced also by the picture of scientist in popular culture. Another reason for the centrality of persons might well be related to how students make pairwise associations. The dominance of persons in pairwise associations can be also seen as a sign of ease in using persons as cues to make associative connections. In word association studies focusing on connections between cue-words and target-words, it has been found that self-beliefs about ease of memorisation affect the ability to memorise the connections (Witherby and Tauber 2017). A plausible assumption is that persons' names are easier to retrieve from the memory than colligatory concepts (like the Enlightenment) that refer to colligatory, connected set of events. This connection is only a tentative here, though it is empirically testable.

Thematic communities and landmarks within them provide a more detailed picture of the landmarks. The dominance of luminary persons as landmarks is diminished when we consider the history landscape as periodised and a thematised structure, by dividing it into periodic windows with different temporal durations and by finding communities within these periodic windows. The community structure arises from the fact that networks are not homogeneous but are thematically structured, consisting of clusters where connections between nodes within the clusters are denser than between nodes that belong to other clusters. The communities have very high thematic coherence. Basically, modules that are related to physics, technology and general history can be found. The emergence of thematically meaningful and coherent communities is interesting, because no such thematic ordering was part of the task; thematic ordering must arise from some internal, content-related but not explicitly visible or pre-fixed thematic family resemblance between knowledge items that belong to the communities. The obvious assumption is that this reflects the fact that the material students have used (mostly Wikipedia) is organised according to thematic organising principles, as a network of hypertext. This structure, then, guides students like an invisible hand to related knowledge items, so that most of the thematic organisation is preserved and meaningful scaffolding emerges. The thematic structure and thematic attributions thus play a crucial role in the organisation of students' knowledge and close attention to this should be paid in attempts to cartograph students' knowledge. Landmarks within the thematic communities, as they appear in students' knowledge landscape at the group level, also show historically meaningful fading in and out when the window of periodisation is changed. Some thematic communities and their landmarks (such as the French Revolution, the Industrial Revolution and Faraday) show more persistence to changes in periodisation and thus reveal their importance in

providing temporal continuity for students' knowledge landscapes. Such landmarks and thematic communities can be recognised as particularly central to students' knowledge.

Finally, the results of the study have clear implications for the teaching of history of science. The bias towards persons as landmarks contains a risk that the whole endeavour of science is attributed to a few exceptionally talented and intelligent persons; a conception which easily leads overestimation of the role of famous scientists (Leite 2002). On the other hand, when thematic communities are in focus, landmarks (and colligatory concepts) include also the Scientific Revolution, the Enlightenment, the French Revolution and the Industrial Revolution. Together with person landmarks the outcome is, after all, a satisfactory knowledge landscape, which acts as adequate if not yet optimal scaffolding for further discussions and elaborations. Some thematic groups and landmarks are, however, clearly underrepresented and missing. One underrepresented group is scientific institutions, like the Royal Society and the Academy of France, which are noted but not among the landmarks. Another underrepresented landmark is Darwinism and Evolution, which only have a small role in the group-level landscape. One would also expect to find, for example, Explorations to the Pacific and the Arctic among the landmarks. All these underrepresented landmarks would be desirable from the viewpoint of modern views about how "big picture" of history should be built not only around persons but also around events of longer duration and of more complex nature (van Boxtel and van Drie 2012; van Drie et al. 2009; van Drie and van Boxtel 2008). Clearly, in regard to instructional strategies, a step to be taken is to guide students' attention to areas which are underrepresented in their knowledge landscapes, and to make them aware of the unwarranted and strong bias there is towards a biographical and person-centred conception of history. To provide a balanced picture the teaching of science history should therefore develop approaches and scaffoldings where in addition to important persons, important events, ideas and institutions also gain better visibility (van Boxtel and van Drie 2012; van Drie et al. 2009; van Drie and van Boxtel 2008).

In summary, we propose a network-based approach to the analysis of students' representations of their conceptions of the interlinked nature of science history and general history, as well as cultural history. Many features which remain undetected, or are identified only with difficulty by traditional methods, become accessible using networks to represent how different facts are related, which of them are the most essential elements in the networks, and to see what kinds of changes occur when we change the time window on history. Such knowledge has a direct impact on the design and planning of education and courses which could better address the need to facilitate a deeper understanding of the related nature of science history and history in general.

Availability of data and material
Data is available on request from the corresponding author.

Authors' contributions
ITK has been the lead author of the manuscript, who largely conceptualised the problem and designed the network analysis. HL has made key contributions to conceptualising the problem, in collecting the data, and implementing the network analysis. Both authors read and approved the final manuscript.

References

Benzi M, Estrada E, Klymko C (2013) Ranking hubs and authorities using matrix functions. Linear Algebra Appl 438:2447–2474

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech 2008:P10008

Borgatti SP (2005) Centrality and network flow. Soc Networks 27:55–71

Börner K (2015) Atlas of knowledge: Anyone can map. MIT Press, Cambridge, MA

Börner K, Scharnhorst A (2009) Visual conceptualizations and models of science. J Informetr 3:161–172

Chen C, Chen Y, Horowitz M, Hou H, Liu Z, Pellegrino D (2009) Towards an explanatory and computational theory of scientific discovery. J Informetr 3:191-209

Corder GW, Foreman DI (2014) Nonparametric Statistics: A Step-by-Step Approach. Wiley, New York

Costa LdF (2006) Learning about knowledge: a complex network approach. Phys Rev E 74:026103

Csardi G, Nepusz T (2006) The Igraph software package for complex network research. Inter J Complex Syst:1695

Dawson, I (2004) Time for chronology. Ideas for developing chronological understanding. Teach Hist 117:14–24

Dawson I (2009) What time does that tune start? From thinking about "sense of period" to modelling history at Key Stage 3. Teach Hist 135:50–57

de Arruda HF, Silva FN, Costa LdF, Amancio DR (2017) Knowledge acquisition: A Complex networks approach. Inform Sci 421:154–166

Estrada E (2012) The structure of complex networks. Oxford University Press, Oxford

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Galili I, Hazan A (2001) Experts' Views on Using History and Philosophy of Science in the Practice of Physics Instruction. Sci & Educ 10:345–36

Ghosh R, Lerman K (2011) Parameterized centrality metric for network analysis. Phys Rev E 83:066118

Gooday G, Lynch JM, Wilson KG, Barsky CK (2008) Does Science Education Need the History of Science? Isis 99:322–330

Guimerà R, Amaral LAN (2005) Cartography of complex networks: modules and universal roles. J Stat Mech 2005:P02001

Holton G, Brush SG (2001) Physics, the Human Adventure. Rutgers University Press, New Jersey

Höttecke D, Silva CC (2011) Why Implementing History and Philosophy in School Science Education is a Challenge: An Analysis of Obstacles. Sci & Educ 20:293–316

Jones E, Oliphant T, Peterson P, et al (2001) SciPy: Open source scientific tools for Python. http://www.scipy.org. Accessed 11 Jan 2019

Katz L (1953) A New Status Index Derived from Sociometric Analysis. Psychometrika 18:39–43

Kemp C, Tenenbaum JB (2008) The Discovery of Structural Form. PNAS 105:10687–10692

Kolaczyk ED (2009) Statistical Analysis of Network Data. Springer, New York

Koponen IT, Nousiainen M (2014) Concept networks in learning: Finding key concepts in learners' representations of the interlinked structure of scientific knowledge. J Complex Networks 2:187–202

Koponen, IT, Nousiainen, M (2018) Concept networks of students' knowledge of relationships between physics concepts: Finding key concepts and their epistemic support. Appl Network Sci 3:14

Leite L (2002) History of Science in Science Education: Development and Validation of a Checklist for Analysing the Historical Content of Science Textbooks. Sci & Educ 11:333–359

Masucci AP, Kalampokis A, Equiluz VM, Hernandez-Garcia H (2011) Wikipedia Information Flow Analysis Reveals the Scale-Free Architecture of the Semantic Space. PLoS ONE 6:e17333

Morais AS, Olsson H, Schooler LJ (2013) Mapping the Structure of Semantic memory. Cog Sci 37:125–145

Newman MEJ (2004) Analysis of weighted networks. Phys Rev E 70:056131

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113

Peixoto TP (2014) The Graph-Tool Python Library. Figshare. https://doi.org/10.6084/m9.figshare.1164194

Qu J, Wang S-J, Jusup M, Wang Z (2015) Effects of random rewiring on the degree correlation of scale-free networks. Sci Rep 5:15450

Russell TL (1981) What History of Science, How Much, and Why?. Sci Educ 65:51–64

Sharkey KJ (2017) A control analysis perspective on Katz centrality. Sci Rep 7:17247

Shi F, Foster JG, Evans JA (2015) Weaving the fabric of science: Dynamic network models of science's unfolding structure. Soc Networks 43:73–85

Siew CSQ (2018) Using network science to analyze concept maps of psychology undergraduates. Appl Cognit Psychol 2018:1–7

Simonyi K (2012) A Cultural History of Physics. CRC Press, Boca Raton, FL

Stella M, De Domenico M (2018) Distance Entropy Cartography Characterises Centrality in Complex Networks. Entropy 20:268

Thompson GW, Kello CT (2014) Walking across Wikipedia: a scale-free network model of semantic memory retrieval. Front Psychol 5:86

van Boxtel C, van Drie J (2012) "That's in the Time of the Romans!" Knowledge and Strategies Students Use to Contextualize Historical Images and Documents. Cogn Instr 30:113–145

van Drie J, van Boxtel C (2008) Historical reasoning: Towards a framework for analyzing students' reasoning about the past. Educ Psychol Rev 20:87–110

van Drie J, van Riessen M, Logtenberg A, van der Meijden B (2009) "When was that date?" Building and assessing a frame of reference in the Netherlands. Teach Hist 137:14–21

Wang S, Du Y, Deng Y (2017) A new measure of identifying influential nodes: Efficiency centrality. Commun Nonlinear Sci Numer Simul 47:151–163

Witherby AE, Tauber SK (2017) The Influence of Judgments of Learning on Long-Term Learning and Short-Term Performance. J Appl Res Mem Cogn 6:496–503

Zweig KA (2016) Network Analysis Literacy: A Practical Approach to the Analysis of Networks. Springer, Wien