A natural history of networks: Modeling higher-order interactions in geohistorical data

Alexis Rojas[1, 2*], Anton Holmgren[1], Magnus Neuman[1], Daniel Edler[1, 3, 4], Christopher Blöcker[1] and Martin Rosvall[1]

[1]Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden

[2]Department of Computer Science, University of Helsinki, 32611, Finland

[3]Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Gothenburg, Sweden

[4]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE-405 30 Gothenburg, Sweden

*** Correspondence:**
Corresponding Author
alexis.rojasbriceno@helsinki.fi

Keywords: paleobiology, geohistorical data, complex networks, higher-order, Infomap, map equation.

Abstract

Paleobiologists often employ network-based methods to analyze the inherently complex data retrieved from geohistorical records. Because they lack a common framework for designing, performing, evaluating, and communicating network-based studies, reproducibility and interdisciplinary research are hampered. The high-dimensional and spatiotemporally resolved data also raise questions about the limitations of standard network models. They risk obscuring paleontological patterns by washing out higher-order node interactions when assuming independent pairwise links. Recently introduced higher-order representations and models better suited for the complex relational structure of geohistorical data provide an opportunity to move paleobiology research beyond these challenges. Higher-order models can represent the spatiotemporal constraints on the information paths underlying geohistorical data, capturing the high-dimensional patterns more accurately. Here we describe how to use the Map Equation framework for designing higher-order models of geohistorical data, address some practical decisions involved in modeling complex dependencies, and discuss critical methodological and conceptual issues that make it difficult to compare results across studies in the growing body of network paleobiology research. We illustrate multilayer networks, hypergraphs, and varying Markov time models for higher-order networks in case studies on gradient analysis, bioregionalization, and macroevolution, and delineate future research directions for current challenges in the emerging field of network paleobiology.

1. Current challenges and opportunities for network-based paleobiology research

Network science is transforming scientific research and thinking in the twenty-first century. Many natural and social phenomena can be described as networks, where nodes represent individual components, and links indicate their interactions. Network science studies high-dimensional, heterogeneously structured, complex systems and their underlying processes (Barabási and Pósfai 2016). In recent years, standard network models based on pairwise or direct interactions between individual components have been applied to almost every area of paleontological research, including biostratigraphy (Muscente et al. 2019), biogeography (Dunhill et al. 2016; Kiel 2017; Rojas et al. 2017; Kocsis et al. 2018; Jeon et al. 2021), macroecology (Roopnarine 2010), and macroevolution (Muscente et al. 2018; Kocsis et al. 2021). However, methodological inconsistencies and conceptual issues in the emergent interdisciplinary field of network paleobiology make it challenging to reproduce experiments and compare outcomes across studies, for instance macroevolutionary patterns delineated using standard (Muscente et al. 2018) and higher-order models (Rojas et al. 2021). The complexity of the high-dimensional and spatiotemporally resolved data retrieved from geohistorical records also raises questions about the limitations of the standard network models: How accurately do they represent the complex local (outcrops, stratigraphic sections), regional (geological basins), and global scale systems examined in paleobiology?

Although network science provides methods for statistical analysis and machine learning of relational data (Brandes et al. 2013; Lambiotte et al. 2019), paleobiologists often describe network analysis as a tool for visualization and qualitative assessment (Huang et al. 2016; Penn-Clarke and Harper 2020; Ye et al. 2021). While network visualization techniques are powerful tools for exploratory data analysis (Perri and Scholtes 2020), this misrepresentation reflects a need to adapt research practices in quantitative paleobiology based on theoretical and methodological advances of network science. Recent studies on the deep-time fossil record (Eriksson et al. 2021; Rojas et al. 2021) suggest that the most critical conceptual issue in the growing body of network paleobiology is ignoring the extent to which the choice of network model impacts the results. There are also methodological inconsistencies, including inadequate descriptions of the input network, incomplete explanations of the clustering approach, and uncritical acceptances of the network partition without validation. The lack of a common framework obstructs interdisciplinarity, reproducibility, and communicability, and calls for standardized research practices in the emergent field of network paleobiology: how to design, perform, communicate, and evaluate network studies.

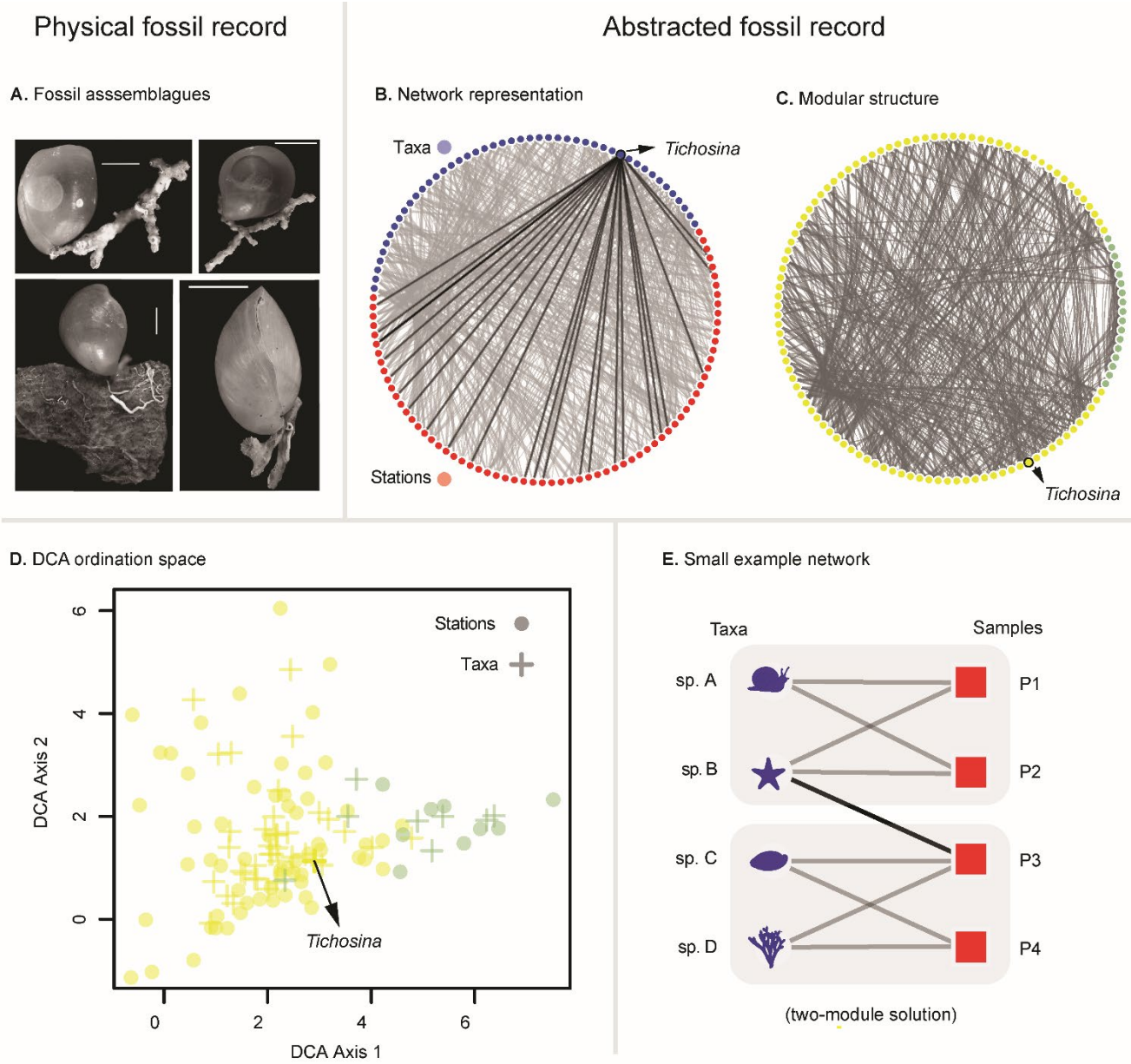
In this overview, we describe how to use the Map Equation framework (Rosvall and Bergstrom 2008; Edler et al. 2017) to identify important patterns in geohistorical data with higher-order network models. Specifically, we describe the concept of higher-order interactions in networks, address some practical decisions involved in network-based modeling of geohistorical data, and illustrate alternative network models, including multilayer networks, hypergraphs, and varying Markov time models for higher-order networks, with cases studies on depth gradient analysis, bioregionalization, and macroevolution. In addition, we delineate future research directions for current challenges in network paleobiology. We focus on the Map Equation framework and its applications in paleobiology research because it is an increasingly popular alternative to the standard statistical approaches currently used in paleobiology research (Kocsis et al. 2021; Rojas et al. 2021; Pilotto et al. 2022; Viglietti et al. 2022). The associated software called Infomap for finding community structure in standard and higher-order networks is also widely used in natural and social sciences

(Law et al. 2022; Lazaridis et al. 2022; Martins et al. 2022). The Map Equation framework for higher-order networks combines sedimentological, ecological, morphological, taxonomic, and any other data retrieved from geohistorical records, enabling an integrated investigation of the complex interactions between plate tectonics, global climate change, and evolution of life.

2. The Map Equation framework for higher-order networks

Research in network paleobiology has focused on delineating community structure that reveals, for example, biozones, bioregions, or evolutionary faunas, based on statistical regularities (Figure 1A-D). This unsupervised learning task known as community detection makes up a whole subfield of network science: network scientists have developed various methods for different purposes and research questions (Schaub et al. 2017). The Map Equation framework uses random walks on a network as a proxy for chains of interdependencies or flows of some sort in the underlying system. It consists of an objective function – the map equation – and its search algorithm Infomap. The map equation measures the quality of community structure through the modular description length of the flows (Rosvall and Bergstrom 2008). Minimizing the map equation over possible network partitions with Infomap identifies how network flows organize in communities and helps explain geohistorical systems because modular network flows capture their important components. Flow-based methods for community detection, such as the Map Equation Framework (Rosvall et al. 2019), are attractive for geoscientists also because they can explore the biosedimentary record at multiple scales (Eriksson et al. 2021) and capture movement patterns across nodes, such as species moving over their geographic ranges (Lambiotte and Rosvall 2012). Previous research has found that the Map Equation outperforms other approaches when operating on various benchmark networks (Lancichinetti and Fortunato 2009; Aldecoa and Marín 2013; Kheirkhahzadeh et al. 2016; Ghasemian et al. 2019).

Figure 1. Network representations of geohistorical data are abstracted fossil records. A. The physical fossil record. These brachiopod shells aim to represent the benthic macrofauna from the R/V Pillsbury program in the Caribbean. Modified from Rojas et al. (2022, fig. 2). B. The abstracted fossil record. This abstracted record is a bipartite network representation of the underlying data (sampling stations × taxa matrix) (Rojas et al. 2022)(Supplementary Data 1). The brachiopod *Tichosina*, the larger component of the Cenozoic brachiopod faunas in the Caribbean, is indicated. C. Modular structure delineated via community detection with the Map Equation framework and using a Markov Time = 2 (Kheirkhahzadeh et al. 2016) (Supplementary Data 2). Modules include sampling stations and taxa. Nodes are rearranged in the circular layout by their module affiliation. Only the two larger modules are displayed; they represent 98% of the network flow. D. Modules mapped on a Detrended Correspondence Analysis (DCA) ordination space. E. Small example network showing a two-module partition.



2.1. The map equation

The map equation models a diffusion process on a network with a random walk, a succession of random steps between nodes (Rosvall and Bergstrom 2008). The random walk starts from a randomly selected sample (Figure 1E), continues to a randomly selected taxon present in the sample, and then to a randomly selected sample where the taxon is present, and so on repeatedly. The map equation uses these network flows' long-term node and link visit rates to capture the network structure in a principled way (see Rosvall and Bergstrom 2008; Rosvall et al. 2009). For example, if the network has communities of highly interconnected samples and taxa representing different biofacies, the network flows will persist for relatively long times within those communities.

The map equation employs the minimum description length principle, a central concept in information theory stating that the best hypothesis is the one that compresses the data the most by using its regularities (Rissanen 1978). The map equation specifies the theoretical lower limit of how concisely we can describe the trajectory of a random walk, given a partition of the network's nodes

into modules (Rosvall and Bergstrom 2008; Rosvall et al. 2009). Our modular description uses codewords for node visits and transitions between modules. Like in Morse code, frequently visited nodes use shorter codewords for best compression. For a modular description that can capitalize on community structure, we assign each node a unique codeword in its module and reuse short codewords between modules. A uniquely decodable modular code also requires an index codebook with codewords for entering each module and an exit codeword in each module codebook for switching to the index codebook. If the random walker steps within the module, only a single codeword is required, whereas if it steps between modules, we need three: First the exit codeword from the old module codebook, then the enter codeword for the new module from the index codebook and last the node codeword from the new module codebook. When the partition matches significant community structure in the network such that the module exit rates are low, the extra cost of describing module transition events is lower than the gain in describing steps within modules with shorter codewords. The map equation quantifies this tradeoff between efficient descriptions within many small modules and short descriptions between a few large modules. While we use codewords to explain the machinery of the map equation, the average per-step theoretical lower limit of the modular codelength given by the map equation is all we need to measure how good we are at identifying modular structure. From Shannon's source coding theorem (Shannon 1948), each Shannon entropy term of the map equation sets the lower bound on the per-step codelength of the index and module codebooks. With probability p_α for codeword α in a codebook with the set of codewords P, the Shannon entropy $H(P) = -\sum p_\alpha log_2 p_\alpha$. For a two-level partition using one index codebooks and one set of m module codebooks, the map equation is

$$L_{(M)} = qH(Q) + \sum_{i=1}^{m} p_i \, H(P^i)$$

where the first term is the average length of codewords in the index codebook and the second term is the average length of codewords in the module codebooks, both weighted by their rates of use (Rosvall and Bergstrom 2008). Q is the normalized enter rates q_i/q for each module, where $q = \sum_{i=1}^{m} q_i$ is the total rate at which the index codebook is used. For module i, P^i is the normalized visit rates for each node in the module plus the exit rate, and p_i is the rate of use for the module codebook.

In general, sample-based geohistorical data form an undirected network because the relationship between sampling units and taxa is symmetric: a random walker can move along links between samples and taxa in both directions. In an undirected network, node α's visit rate is the total weight w_α of its links divided by the total link weight of all N nodes in the network,$W = \sum_{\alpha=1}^{N} w_\alpha$. In the small example, we treat all fossil occurrences as equally important, and the network is unweighted with all link weights equal to 1. Therefore, node α's visit rate is its number of links divided by two times the total number of links in the network, since counting each node's link-end double-counts the number of undirected links. Similarly, the exit and enter rates of a module are the total number of links that crosses its boundary divided by W. In the small example networks (Figures 1E), one link with weight 1 crosses each boundary such that all enter and exit rates are $\frac{1}{18}$. With a compressed notation for the entropy, $H(w_1, w_2, \ldots, w_m) = -\sum_{i=1}^{m} \frac{w_i}{w \, log_2 \frac{w_i}{w}}$ where $w = \sum_{i=1}^{m} w_i$, the total codelength for the two-module partition is

$$L_{(M)} = \frac{2}{18} H(1,1) + \frac{10}{18} H(2,2,2,3,1) + \frac{10}{18} H(2,2,2,3,1) = 2.61 \; bits.$$

This partition minimizes the map equation (Figure 2). Overall, small modules enable short descriptions within modules because they require little information to specify the visited nodes but

may lead to long descriptions between modules from frequent and expensive module exits and entries. In contrast, large modules enable short descriptions between modules because between-module steps are rare and cheap but have long descriptions within modules from the many nodes to differentiate. Small modules where random walks persist long compress the network flows maximally and reveal the most modular regularities for the modeled network flows (Rosvall and Bergstrom 2008; Rosvall et al. 2009).

The map equation generalizes straightforwardly to multiple hierarchical levels (Rosvall and Bergstrom 2011), to higher-order network models, including so-called memory networks (Rosvall et al. 2014), multilayer networks (De Domenico et al. 2015), and hypergraphs (Eriksson et al. 2021, 2022), and to varying Markov time models (Kheirkhahzadeh et al. 2016). Recent Bayesian generalizations of the Map Equation deal with incomplete data in a theoretically founded way (Smiljanić et al. 2020, 2021).

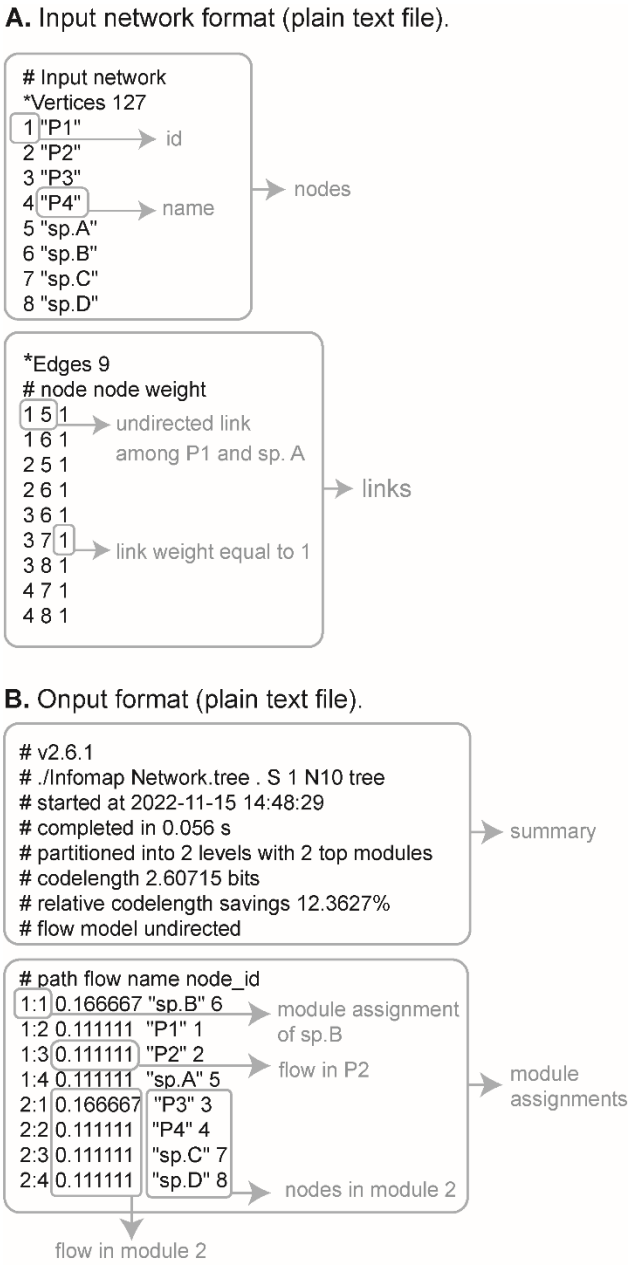


Figure 2. Input and output formats. -The Map Equation framework understands different formats. **A.** Input file in Pajek format. **B.** Output file. The resulting partition is written to a file with the extension .tree (plain text file). The formats described here correspond to the small example network in Figure 1E.

2.2. The search algorithm Infomap

For a network with n nodes assigned to m modules, there are on the order of n^m possible partitions. Even for moderately-sized networks, it is impractical to test all possible partitions to guarantee that we have found the best one. Because small changes in the network can slightly improve a partition, for practitioners it is more informative to investigate good partitions rather than focusing on finding the very best. The Map Equation framework uses a fast stochastic search algorithm called Infomap to minimize its objective function over possible node assignments (Rosvall and Bergstrom 2008). The algorithm consists of multiple search procedures (Edler et al. 2017). The core algorithm starts by assigning each node to its own module. Then it repeatedly loops through each node in random order and moves it to the module that reduces the codelength the most. Infomap repeats this procedure until no move decreases the codelength, rebuilds the network with the modules forming nodes at a coarser

level, moves these nodes into even coarser modules, and so on until no move reduces the codelength further.

To improve this two-level solution, Infomap alternates between a fine-tuning and a coarse-tuning procedure by moving individual nodes or sub-modules between modules. To find a hierarchical solution, Infomap starts from the two-level solution and iteratively builds super-module levels that compress the description of movements between modules. Then it clears the structure under each of the coarsest modules and recursively and in parallel builds sub-modules within each module until it cannot find a finer structure that decreases the hierarchical codelength. In this way, the resulting hierarchical structure of the network may have branches of different depths.

When modeling network dynamics on standard networks through random walks, a single node type usually simultaneously represents a physical component of the system and describes the flows with the nodes' links (Figure 3A-B). In contrast, the Map Equation framework for higher-order networks distinguishes physical nodes, representing the system's components, from state nodes, describing the system's internal flows (Edler et al. 2017)(Figure 3C). State nodes in the Map Equation framework can represent, for instance, memory of previous steps, layers in multilayer network representations (Rojas et al. 2021), lumped states, or any other complex relationship in the underlying geohistorical data. To capture the higher-order nature of flows on networks with state nodes, Infomap applies the same procedures to the state nodes with aggregated visit rates for all state nodes of the same physical node assigned to the same community.

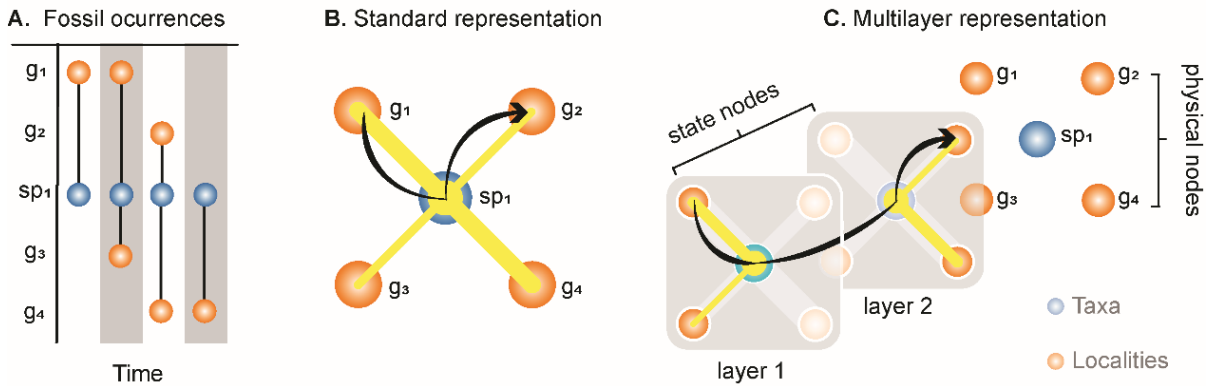


Figure 3. Network representations of geohistorical data. **A.** Temporal occurrence data. **B.** Standard bipartite network representation created by aggregating fossil occurrences (sp. 1) into arbitrary spatiotemporally explicit units (g1 to g4). This standard network uses the same nodes to represent the physical components of the system as well as to describe system's flows. A random walk as the simplest information flow model forms paths across independent pairwise links on this network washing out higher-order node interactions. **C.** Multilayer network representation. The Map Equation framework for higher-order networks distinguishes physical nodes that represent the system's components from state nodes describing the system's internal flows. This higher-order network better captures the constraints on the information paths and thus flows tend to stay among units from each layer.

1. Higher-order network models in paleobiology

1.1. The complex relational structure of the geohistorical data

Geohistorical records, either stratigraphic sections, boreholes, ice cores, or archaeological sites, are inherently complex. Despite their limitations (Kidwell and Holland 2002; Kidwell and Tomasovych 2013), the high-dimensional and spatiotemporally resolved data retrieved from individual geohistorical records allow for evaluation of past biotic responses to natural and human-induced environmental changes at local to regional scales (Council 2005; Scarponi and Kowalewski 2007; Dietl and Flessa 2011; Durham and Dietl 2015). Although high-precision chronological studies have improved our understanding of past biotic crises (Smith et al. 2018), compilations of individual geohistorical records are central for studies at larger spatiotemporal scales. Fossil occurrences of the benthic marine invertebrates in the Paleobiology Database (PaleoDB)(Peters and McClennen 2016) have become the benchmark data for network-based research on macroevolution, macroecology, and biogeography (Rojas et al. 2017, 2021; Kocsis et al. 2018; Muscente et al. 2018). In most cases, PaleoDB collections have geographic information and are assigned to a geological stage, enabling the modeling of temporal constraints. PaleoDB collections also have lithostratigraphic and sedimentological information and sometimes include taphonomy and body-size data. Each fossil occurrence in PaleoDB belongs to one collection, has a name with a specific taxonomic resolution, and is linked to an independent taxonomic classification with associated ecological information. These complex relational data describe the structure of the Phanerozoic life at multiple taxonomic levels and spatiotemporal scales.

There are also numerous databases covering specific taxonomic groups, time intervals, or geographic regions (Williams et al. 2018; Community 2020). For instance, The Strategic Environmental Archaeology Database compiles high-resolution archaeological and paleontological data (Buckland and Eriksson 2014). In most cases, the sample's age is a value taken from the original literature sources and obtained from a range of dating methods varying in precision and accuracy (Buckland 2014). Because samples may have an age range larger than the length of the preferred bin interval, modeling temporal constraints of high-resolution data using multilayer networks is challenging. This question has been approached using the Map Equation framework to investigate the recent fossil record of European beetles (Pilotto et al. 2022). Their multilayer network analysis relaxes the temporal constraints, allowing a random walker to move toward neighboring layers without exceeding the age limits of the samples in the data. In practice, with ordered layers representing 500-year time intervals and an accepted age range of 2000 years in the filtered samples, authors allowed a random walker to relax toward the first two layers in each direction. This approach accounts for the age uncertainty inherent to the samples, making it possible to explore high-resolution geohistorical data using multilayer representations.

1.1. Higher-order networks capture the complexity of the geohistorical data

One of the major conceptual changes in modern paleobiology research has been the distinction between the physical fossil record, consisting of in-situ or ex-situ specimens, and abstracted representations based on data retrieved from this physical record (Sepkoski 2013; Allmon et al. 2018). We capitalize on this idea by explicitly considering improved network representations of geohistorical data. Intuitively, a good model of the physical fossil record should be maximally parsimonious yet sufficiently complex to capture the complex interactions of spatiotemporally resolved and high-dimensional geohistorical data. Researchers designing network representations of geohistorical data must recognize that their choices make various assumptions about the spatiotemporal structure and dynamics of the biosedimentary record, such as whether or not to

describe temporal constraints. These assumptions impact the outcome, such as whether or not the solutions capture larger-scale patterns. In paleobiology, researchers tacitly assume that using benchmark data guarantees reproducibility, ignoring that different network representations and modeling decisions may impact the observed patterns. Here we argue that setting benchmarks is required to improve reproducibility and communicability in network paleobiology. This challenge is also an opportunity to move beyond standard network representations, toward higher-order models (Benson et al. 2016) that better capture the complexity of the geohistorical data.

Network representations of geohistorical data are usually standard models based on pairwise connections between taxa (e.g., species, genus) and sampling units (e.g., collections, localities, grid cells) (Figure 3A). This standard approach consists of modeling dynamical processes on these networks with first-order flows, using a memory-less random walker whose movements are determined only by its current node. This approach oversimplifies the dynamics of the system because it ignores higher-order node interactions, or dependencies, in the underlying geohistorical data (Figure 3B). Such dependencies can be modeled by equipping the random walker with memory, letting it choose its steps based on its current node but also on the previously visited node or nodes. Whereas standard models use a single node type to represent the physical components of the system, including taxa and sampling units, and model flows through one-step dynamics on their links, the Map Equation framework for higher-order networks introduces abstract nodes to describe the different states in which the physical nodes can be in the system, so-called state nodes (De Domenico et al. 2015)(Figure 3C). For instance, the multilayer network representation of the benchmark data on the Phanerozoic fossil record is a higher-order model in which a physical node representing a given taxon contains several state nodes carrying information on the geological stage where this taxon occurs (Rojas et al. 2021). Therefore, this higher-order representation created through the Map Equation framework is a form of memory network (Edler et al. 2017). First-order models of the Phanerozoic benthic marine faunas without memory ignore the temporal constraints inherent to the underlying system and obscure the macroevolutionary pattern (Figure 4). When comparing unipartite, bipartite, and multilayer representations of this benchmark data, the multilayer network achieves the shortest codelength and the best compression (Table 1).

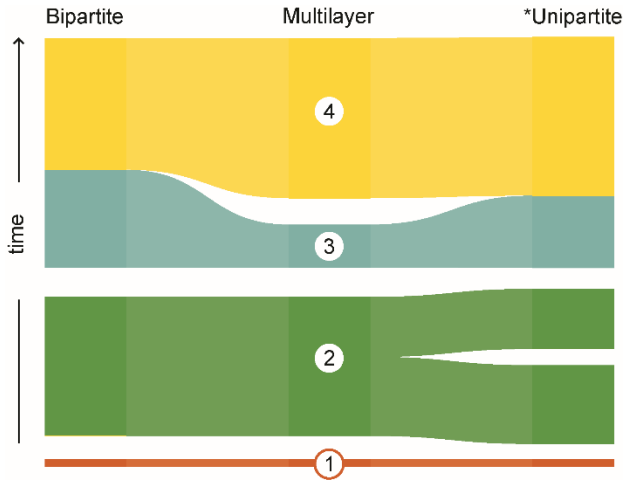


Figure 4. Different network representations capture different aspects of the benchmark data on the Phanerozoic benthic marine faunas (Peters and McClennen 2016). Bipartite and unipartite representations ignore the temporal constraints inherent to the biosedimentary record. *Unipartite projection obtained from the bipartite network by rescaling the Markov time (Kheirkhahzadeh et al. 2016). Alluvial diagram representing 98% of the network flow in each case (Supplementary Data 3-7).

Model	Physical nodes	State nodes	Levels	Top modules (99%)	Codelength (bits)	Compression (%)
Unipartite*	23203	—	2	4	11.0646	8.6009
Bipartite	23203	—	3	3	10.5169	13.1253
Multilayer	23203	64508	7	4	**6.2826**	**55.2804**

Table 1. Comparison of the modular structure of different network representations of the benchmark data on the Phanerozoic marine faunas. The multilayer representation achieves the shortest codelength and the best compression. For each partition, we measure the compression by using the corresponding one-level partition as a baseline. *Unipartite projection obtained from the bipartite network by rescaling the Markov time (Kheirkhahzadeh et al. 2016).

1.1. Alternative higher-order network models for geohistorical data

1.1.1. Multilayer networks

Multilayer networks are models used to represent complex systems with multitype interactions. In the Map Equation framework, multilayer networks can be used to model temporal and non-temporal data. Temporal constraints in real geohistorical systems include ordered geological stages (Rojas et al. 2021)(Figure 3), arbitrary temporal bins (Pilotto et al. 2022), and biostratigraphic frameworks (Viglietti et al. 2022). Interlayer dynamics in multilayer networks are often modeled based on the intralayer information (Eriksson et al. 2022). For instance, when layers represent geological stages, intralayer link structure describes the constraints on the network dynamics at a given stage, whereas interlayer link structure is created through neighborhood flow coupling between state nodes of the same physical node (Rojas et al. 2021). In practice, the Map Equation framework can generate interlayer links using a relax rate (r), with a random walker moving between nodes within a given layer guided by intralayer links with probability $1-r$ and relaxing to adjacent layers guided by links between state nodes of the same physical node with a probability r (Edler et al. 2017). Previous studies show that a relax rate equal to 0.25 is large enough to capture temporal structures but small enough to preserve intralayer information (Aslak et al. 2018). Limiting the relaxing to the nearest layers gives multilayer networks a bidirectional geohistorical time arrow.

Multilayer networks can also be used to model geohistorical data lacking temporal constraints. In general, a geohistorical record with its multiple biotic and abiotic components can be conceptualized as a complex system with a multilayered structure, where each layer describes a particular interaction between sampling units, for instance, those resulting from the fossil composition and sedimentological features, independently. Physical nodes in this two-layered network can represent sampling units of any scale (e.g., samples, beds, members, formations), taxa of any resolution (e.g., species, genus), and sedimentologically defined groups (e.g., textural classes). Whereas each sampling unit in this multilayer network would be represented by a number of state nodes equal to the number of multitype interactions considered in the study, other physical components would be represented by a single state node because they occur only in one layer. In this network, movements between state nodes within each layer represent pairwise interactions, whereas movements between nodes across layers represent higher-order interactions (De Domenico et al. 2015). Our first example illustrates this approach in a case study based on the marine invertebrate fossil record (Holland and Patzkowsky 2004).

1.1.1. Varying Markov time models for higher-order networks

The fossil record does not have a unique and optimal level of description but instead multiple levels representing different scales in the organization of life. Flow-based community detection approaches usually assume one-step dynamics on the links, corresponding to Markov time 1. Modeling network dynamics at shorter or longer Markov times captures the modular structure at different resolutions (Delvenne et al. 2010). These Markov time models can be used to explore structure and dynamics in both first- and higher-order networks (Kheirkhahzadeh et al. 2016). They are especially useful when the modular structure of an empirical network does not show a hierarchical organization (e.g., Penn-Clarke and Harper 2020) or for two-level solutions in the Map Equation framework. Exploring various time scales to reveal finer or coarser partitions allows connecting time scales of the dynamics to the structural scales present in the network (Lambiotte et al. 2014). Recently, varying Markov time models were used to explore larger-scale modular patterns of the Holocene succession and present-day nearshore seabed of the Adriatic Sea (Scarponi et al. 2022). In networks representing the deep-time fossil record, the specific relationship between the Markov time and time scales of the evolution of the Earth-Life system has not been explored.

In the Map Equation framework, the discrete (Kheirkhahzadeh et al. 2016) and continuous (Schaub et al. 2012) time evolution of a Markov process on the network is defined by the parameter Markov time. Intuitively, when using shorter Markov times than 1, the average transition rate of a random walk is lower than the encoding rate (see Map Equation section), the same node is encoded multiple times in its trajectory, and smaller modules are delineated. In contrast, when using longer Markov times than1, the average transition rate is higher than the encoding rate, only some nodes on its trajectory are encoded, and larger modules are delineated. In practice, using Markov time 2, we explicitly explore the two-step dynamics on the network. In the second example, we illustrate varying Markov time models using a case study on the mid-Devonian biogeography of the brachiopods where one-step dynamics on the links (Markov time 1) in a relatively small bipartite network does not capture a hierarchical nested organization.

1.1.1. Hypergraphs

Hypergraphs are network models used to study complex systems in which an arbitrary number of its components can interact. These so-called multigroup interactions, represented through hyperedges connecting all the nodes involved in the interaction, differ from binary contacts represented through links in conventional network models (Carletti and Fanelli 2022). A recent study provides the first hypergraph representation of data derived from the fossil record (Eriksson et al. 2021). This network analysis employs the Map Equation framework to model global occurrences of the benthic marine animals from Cambrian (541 MY) to Cretaceous (66 MY), sourced from the PaleoDB, as a hypergraph where physical nodes are fossil taxa linked through weighted hyperedges that connect all taxa occurring at each stage. Because this network explicitly models temporal constraints in the underlying paleontological data as hyperedges, it captures large-scale temporal structure and dynamics of system, especially when spatial data are lacking (Figure 5).

A recent study derived unipartite, bipartite, and multilayer network representations of hypergraph flows and evaluated how the choice of both network representation and specific random-walk model impacts the number, size, depth, and overlap of multilevel communities in geohistorical data (Eriksson et al. 2022). To create each different network model, this study represents hyperedges as nodes in the bipartite network, projected the bipartite flow to create a unipartite flow representation (i.e., two-step dynamics obtained with a Markov time = 2)(Kheirkhahzadeh et al. 2016), and created state nodes for each hyperedge to which a node belongs to construct the multilayer network, showing

that the Map Equation framework can be used to model multigroup interactions. Overall, the results illustrate the advantages of using multilayer network representations of data derived from the fossil record over bipartite and unipartite representations to quantify macroevolutionary patterns, with different random walk models, including and excluding self-links, providing similar solutions.

Figure 5. Schematic diagram illustrating a hypergraph representing temporal occurrence data. In this diagram, physical nodes are fossil taxa connected through hyperlinks. Hyperlinks are depicted with different colors and connect taxa from the same time interval. Taxa recorded in different time intervals are represented by state nodes, depicted with different shades of grey.

3. Visualizing higher-order multiscale structure in networks

Network visualizations are essential tools for understanding the modular structure and dynamics of higher-order networks describing geohistorical data. However, standard graphic tools developed to represent first-order interactions among network components fail to capture higher-order and multiscale community structure and depend on the arbitrary scale of analysis (Peixoto and Rosvall 2017; Perri and Scholtes 2020). To overcome these limitations, the Map Equation framework provides graphic tools for mapping higher-order and multiscale community structure in network partitions. They are freely available as a client-side web application at https://www.mapequation.org.

3.1. Infomap Network Navigator

Higher-order networks representing geohistorical systems are usually complex with hierarchical modular structures. The Network Navigator tool was developed to explore such complex structures in real networks. The tool creates interactive maps of hierarchical network partitions with aggregated inter-module links. These, possibly directed, links are drawn with lengths inversely proportional -- and width and color saturation proportional to the flow volume between modules. Weakly connected modules are placed further apart with narrower links between them than strongly connected modules. Circles represent the modules, with areas proportional to the contained flow volume and border thicknesses proportional to the exiting flow (Figure 6). Like Google Maps, the modules can be explored by zooming in and highlighting more detail until the lowest-level leaf nodes are shown. The Network Navigator was recently used to visualize the multiscale organization of the Phanerozoic benthic marine faunas, highlighting how the large-scale evolutionary faunas are built up from lower-scale biogeographic entities (Rojas et al. 2021).

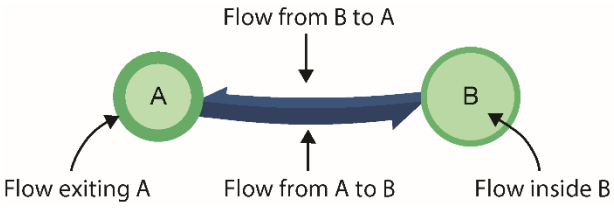


Figure 6. Aggregated inter-module links used by the Network Navigator. Modules and nodes as circles with areas proportional to flow and border widths proportional to exiting flow. Bidirectional links with width proportional to inter-module flow

3.2. Alluvial diagrams

When comparing network partitions, researchers are interested in comparing changes in the overall modular structure rather than the link structure. Alluvial diagrams are visualization tools that highlight changes in modular structure across different network partitions, including, for instance, optimized, bootstrapped, suboptimal, and planted solutions. In the paleobiology literature, alluvial diagrams have been used primarily to compare alternative partitions obtained from bootstrapping (Rojas et al. 2021) and sensitivity analysis (Pilotto et al. 2022). Recently, alluvial diagrams have been used to compare solutions representing alternative biostratigraphy models for the late Permian-mid Triassic Beaufort Group in South Africa (Viglietti et al. 2022).Because it is impractical and typically unnecessary to represent all individual nodes, an alluvial diagram highlights changes in the most important nodes' module assignment across partitions. In the Map Equation framework, the flow volume determines node importance and is calculated when Infomap searches for the optimal partition (Rosvall and Bergstrom 2008)(see Figure 2).

The alluvial diagram is constructed by grouping nodes with the same module assignment to simplify and highlight changes between network partitions (Figure 7A). Shown side-by-side, each partition is represented by a vertical stack of modules in the order we choose. To highlight the nodes that change module assignment between partitions, we draw so-called streamlines between all modules in adjacent partitions that contain the same node. The modules' flow volume determines their heights, and the streamlines' heights are proportional to the flow of the nodes that the connected modules have in common (Figure 7B-C). The alluvial diagram showing the macroevolutionary patterns obtained from different network representations of a large paleontological dataset (Figure 4) illustrates how this graphic tool helps to better understand changes in the modular structure between partitions.

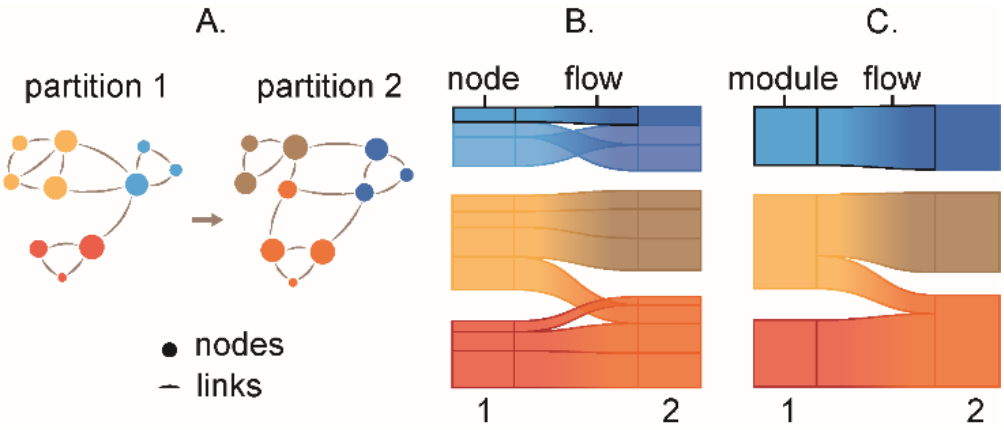


Figure 7. Visualizing change in network partitions using alluvial diagrams. To compare networks with the same sets of nodes, we assemble at least two network partitions (a). Then, we group nodes

in the same module in stacked bars with height proportional to the node's flow volume and connect corresponding nodes with streamlines (b). Finally, to highlight how the partitions change, we aggregate the nodes into modules (c). To compare additional network partitions, we add more stacks of bars to the right and repeat the procedure b-c.

4. Robustness evaluation

4.1. Identifying significant assignments in network partitions

To assess the support of network modules in the underlying data, it is convenient to construct a set of bootstrapped networks through resampling of the data in a standard manner (Efron and Tibshirani 1993). This approach enables calculating summary statistics such as mean values and standard deviations and identifying features that occur in a large enough proportion of bootstrapped networks to be considered robust. The Map Equation framework includes a method (significance clustering (Rosvall and Bergstrom 2010), that identifies sets of nodes that are significantly assigned to a module in a reference partition. For a given module in the reference partition, the largest subset of nodes clustered together in at least 95% of all bootstrap partitions represents its significant core. The significance clustering of the Map Equation framework has been used to distinguish gradual from abrupt biotic transitions in the fossil record, with abrupt events interpreted when temporally adjacent modules are significant and standalone (Pilotto et al. 2022). In addition, different aspects of a network partition can be evaluated across a set of bootstrap networks using set-theoretic measures such as the Jaccard index and measures built upon concepts from information theory (Vinh et al. 2009). The resampling procedure can also consider the underlying data by, for example, considering a discrete distribution if the data represent counts, and also a truncated distribution to avoid false negatives (Rojas et al. 2021).

4.2. Exploring alternative solutions

Finding the best network partition is generally a non-convex optimization problem and the practitioner therefore needs to consider the possibility of multiple solutions. To this end, methods exploring the solutions and their quality have been developed. One such method estimates the minimum number of searches required by the search algorithm Infomap to map the complete solution landscape, ensuring that the best solution is obtained and that alternative solutions of lesser quality can be explored (Calatayud et al. 2019). We illustrate this approach with the multilayer network representing the Phanerozoic fossil record of the benthic marine faunas, where alternative partitions are embedded using a dimension reduction technique (McInnes et al. 2018)(Figure 8), and the distance between partitions is calculated using a weighted version of the Jaccard distance. In this example, the codelength varies between the partition clusters, with the best partition in the middle cluster. In practice, these variations have minimal impact on the large-scale patterns, with only a few nodes alternating between modules.

Although the approach to explore the solution landscape was developed for practitioners to find an optimal modular description when using the Map Equation framework, it highlights the importance of exploring sub-optimal partitions when dealing with real geohistorical systems, which can provide a better understanding of complex patterns. For instance, the extinction at the end of the Cretaceous, despite being a significant and abrupt event, is not captured at the resolution of the landscape illustrated in Figure 7. To find a solution showing a Cretaceous-Neogene global transition at the highest hierarchical level, we have to study less optimal solutions (Rojas et al. 2021). Overall, by

looking at a range of solutions ordered from highest to lowest quality, we can get some insights about the relative importance of the different events shaping the Phanerozoic life history.

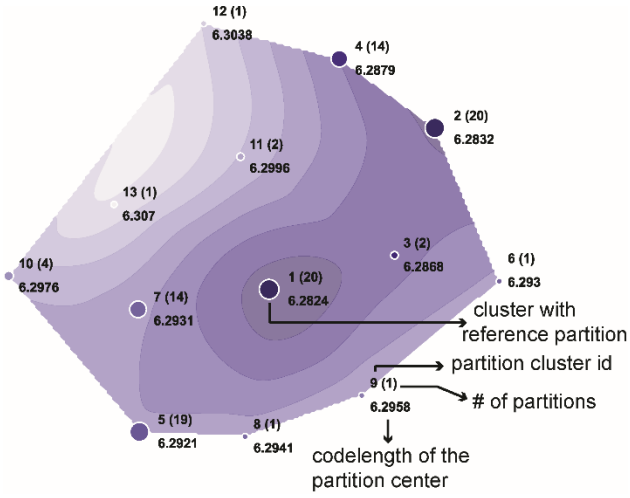


Figure 8. Quality of alternative partitions of the multilayer network representing the Phanerozoic fossil record of the benthic marine faunas mapped in a two-dimensional space. Circles represent clusters of network partitions, located at the cluster center, with size proportional to the number of partitions grouped into the cluster. Map isolines are constructed using the Jaccard distance between partitions. Despite differences in their quality (codelength), all partition clusters identified at the selected scale show a modular structure with four modules representing the Phanerozoic evolutionary faunas. At the selected scale, there is not a cluster of solutions representing a three-tier model (Sepkoski 1981).

5. Case studies on the fossil record

5.1. Delineating litho-biofacies by using multilayer networks

Understanding how the distribution of organisms along environmental gradients changes through time is a primary research area in paleobiology (Patzkowsky and Holland 2012). Indirect ordination techniques applied to species occurrence data have been successfully employed to recover environmental gradients in the sedimentary record. In general, environmental gradients are interpreted from mapping taxa or samples, coded by external factors (e.g., life habit for taxa and depositional environment for samples), into the reduced ordination space. Here we provide a multilayer network analysis that combines taxon abundance and sample attributes into the modeling. Specifically, we conceptualize the biosedimentary record as a complex system with a multilayered structure by creating a network representation with two layers, one describing taxonomic composition and the other sedimentologically-defined relationships between sampling units (Figure 9A). The underlying data were obtained from a basin-scale study carried out on Late Ordovician outcrops in central Kentucky (Holland and Patzkowsky 2004).

Physical nodes in this two-layered network represent sampling units, taxa, and sedimentological groups. Layers in this representation independently capture fossil assemblages (biofacies) and sedimentological groups (lithofacies) in the geohistorical record. However, the multilayer network reveals the so-called litho-bio facies (Figure 9B). This higher-order approach directly interprets the gradients obtained via ordination analysis in two ways, delineating modules that comprise strongly connected taxa, samples, and sedimentological descriptors, and partitioning the gradient into discrete regions (Figure 9C). Although the sedimentological information underlying this case study is relatively simple, our network model can be easily extended to represent geochemical, taphonomic, and any other complex relationship between sampling units, as well as relationships between taxa

(i.e., ecology, body size). Although beds are the fundamental units of both stratigraphy and paleontology (Patzkowsky and Holland 2012), sampling units in this multilayer framework can represent stratigraphic units of any scale. In practice, depending on the stratigraphic resolution of the underlying geohistorical data, this multilayer network model can be used to capture multitype relationships between samples, beds, members, or broader units.

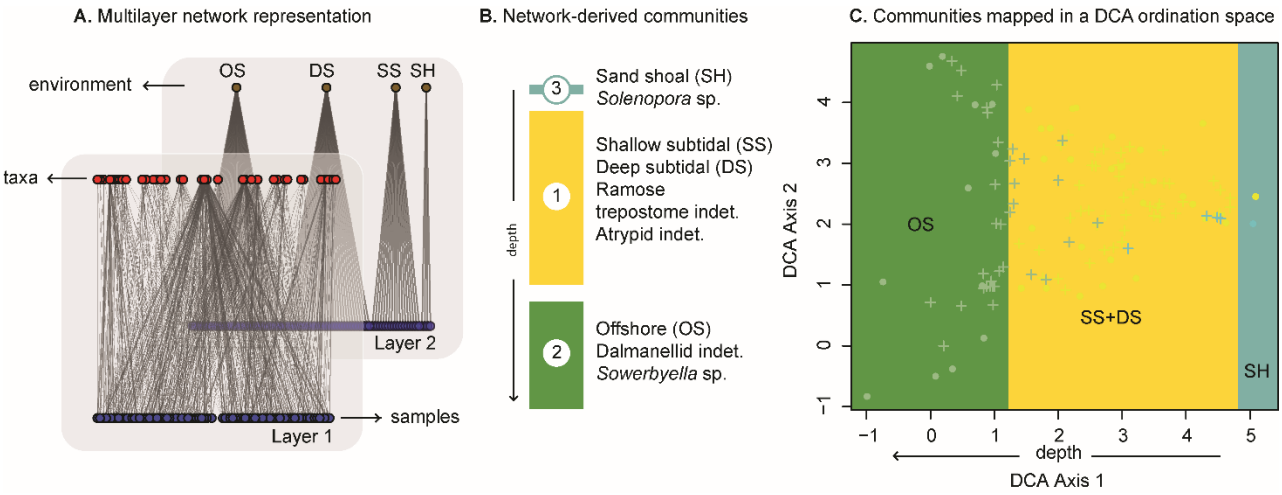


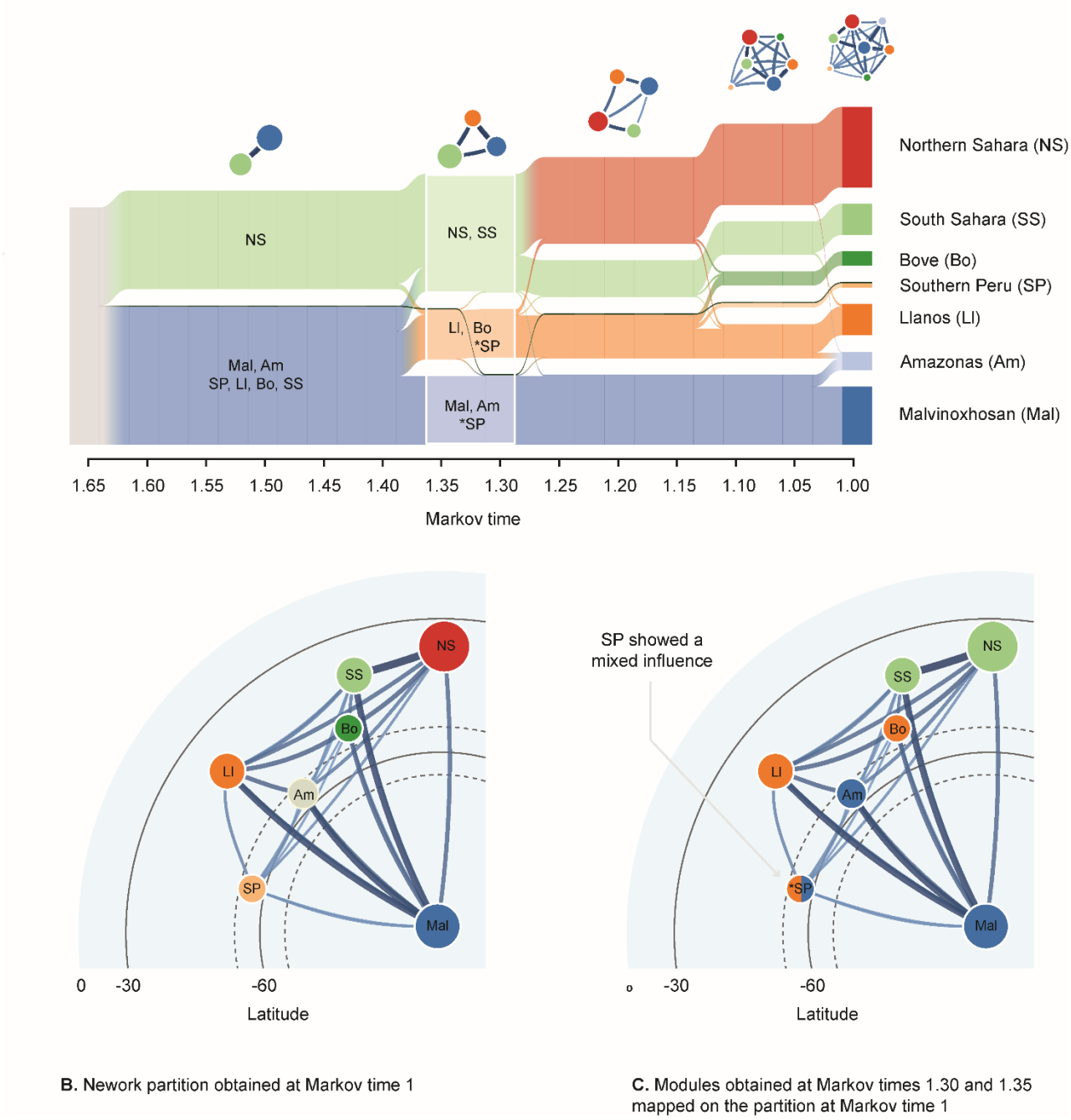
Figure 9. A. A multilayer network representing litho-biofacies from Late Ordovician outcrops in central Kentucky. In this higher order network, one layer describes the biotic component (samples × taxa matrix) and the other describes the abiotic component (samples × environments matrix) of the geohistorical record (Supplementary Data 8). **B**. Litho-biofacies delineated via community detection using the Map Equation framework. Modules in the multilayer solution include taxa, samples, and environments, and can be directly interpreted as litho-biofacies (Supplementary Data 9). **C**. Detrended Component Analysis (DCA) on the samples × taxa matrix. Network modules mapped in a DCA ordination space indicating their distribution along the depth water gradient. Background colored based on the module affiliation. Data from Holland and Patzkowsky (2004).

5.2. Validating marine bioregionalization using varying Markov time models

Research employing network-based approaches to describe biogeography in the fossil record is overwhelmingly focused on describing one-step dynamics (Markov time 1) in relatively small bipartite networks derived from the relatively limited fossil data (Penn-Clarke and Harper 2020; Ye et al. 2021), to reveal continental to global scale marine bioregions. Although first-order network representations of fossil occurrences have been shown to capture a biogeographic signal at some geological stages in the Phanerozoic (Rojas et al. 2017; Kocsis et al. 2018), these studies are unable to identify transition zones, provide a single-scale description of the bioregions and obscure larger-scale patterns. Overall, conventional approaches ignore that bioregions do not have a unique level of description but multiple levels reflecting the complex spatial structuring of marine biodiversity.

Here we describe how to use varying Markov time models through the Map Equation framework to overcome some of the limitations of the standard models currently used in paleobiology research. This case study is based on a standard bipartite network relevant for Devonian biogeography, but the approach can be applied directly to higher-order networks (Calatayud et al. 2021). We use varying Markov time models to re-examine the bioregionalization of the Middle Devonian Brachiopods from

the Old-World Realm (Penn-Clarke and Harper 2020). This approach reveals the larger-scale biogeographic structure at different resolutions. Results provide new insights into the biogeographic affinities of the brachiopod faunas from Southern Peru, which remains an open question (Figure 10). However, this case study shows that we can reveal the spatially nested hierarchical organization of marine biodiversity through varying Markov time models.



B. Nework partition obtained at Markov time 1

C. Modules obtained at Markov times 1.30 and 1.35 mapped on the partition at Markov time 1

Figure 10. Bioregionalization of the Middle Devonian Old World Realm. **A**. Varying Markov time models on the biogeographic network constructed from brachiopod occurrence data (Penn-Clarke and Harper 2020). Network partitions at different Markov times reveal the larger-scale biogeographic structures obtained at different resolutions. **B**. Network partition obtained at the Markov time 1 (7 modules). Circles represent the seven modules delineated when exploring the one-step dynamics on the assembled network. **C**. Modules obtained from partitions at Markov times 1.30 and 1.35 (3 modules) mapped on those obtained at the Markov time 1. These two partitions differ in the affiliation of the Southern Peru locality, placed alternatively into the modules representing higher (white) and lower (orange) latitudes. Overall, coarser partitions contain fewer clusters as the Markov time increases.

6. Recommendations for future research directions

Higher-order network modeling of geohistorical data holds considerable promise for paleobiology research because it provides a framework for revealing the complex interactions between biotic and abiotic components of the sedimentary record at multiple scales. Higher-order network representations better capture the spatiotemporal constraints on the information paths underlying geohistorical data, providing more accurate descriptions of paleontological patterns: Employing the Map Equation framework for higher-order networks also improves reproducibility and communicability. Our analysis has focused on regional and global-scale examples, but the Map Equation framework can be applied to geohistorical data at any scale, ranging from individual beds up to the global sedimentary record. Establishing higher-order benchmark networks from widely used compendia of paleontological data (e.g., The Paleobiology Database, NOW database) enables researchers to compare results from different studies and methods.

The standard formulation of the map equation implicitly assumes complete data. It reveals the large-scale structure of a system given the observed data (Ghasemian et al. 2019). This general approach dominates network-based research in paleobiology. However, geohistorical records are incomplete – not every species that ever lived is preserved, and not all environmental conditions of every region in the globe are recorded in the sediments (Kidwell and Flessa 1996; Council 2005). Sampling also varies due to variations in exposure and collection effort (Holland 2016). The overall effect is an abstracted fossil record partially observed, a network involving links that exist but are not represented in the model because they have not been observed (false negatives). Distinguishing missing links (true positives) from non-edges (true negatives) within the unobserved connections, a task known as link prediction, allows to improve the abstracted fossil record. Such missing links could alter conclusions when delineating network community structure and modeling network dynamics (Ghasemian et al. 2019; Blöcker et al. 2022)

In cases where we expect data to be incomplete, we should use a community-detection approach that can handle this: the Bayesian map equation models missing data with an empirical Bayes approach (Smiljanić et al. 2021). It helps resolve the important question about which of the links not present in a network are true and which are false negatives. Using the minimum description length principle (Rissanen 1978) underlying the Map Equation Framework, it is possible to use a modular partition to measure the description length of links (Blöcker et al. 2022). Assuming that links with a shorter description are more likely, we can rank the non-existing links in a network and select the best candidates for false negatives, those links that are most consistent with the network dynamics and thus most likely to exist. In practice, it is up to the researchers to examine and interpret those

predicted links in the face of the specific question at hand. Geologists and paleontologists can use this approach for stratigraphic placement of isolated samples, refining biostratigraphic models, and improving the overall description of bioregions. We believe these methodological efforts lay the ground for a fertile research direction in the emergent field of network paleobiology.

7. Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8. Author Contributions

A.B.: conceptualization, data curation, formal analysis, methodology, visualization, writing—original draft, writing, editing; A.E.: visualization, writing; M.N.: visualization, writing, editing; D.E.: writing; C.B.: writing, editing; M.R.: writing, editing, funding.

9. Funding

A.R. was partially supported by the Kone Foundation funded project *Comparing evolutionary processes in nature and society* (project number 202007064). A.E., M.N. and M.R. were supported by the Swedish Foundation for Strategic Research, Grant No.\ SB16-0089. D.E. and M.R. were supported by the Swedish Research Council, Grant No.\ 2016-00796.

Acknowledgments

We thank the contributors to the Paleobiology Database who collected data. We thank Alex Dunhill and Andrej Spiridonov for helpful comments on an early version of the manuscript. AR thanks Steve Holland for providing data to delineate Middle Upper Ordovician biofacies.

10. References

Aldecoa, R., and Marín, I. (2013). Exploring the limits of community detection strategies in complex networks. Sci Rep 3, 2216. doi: 10.1038/srep02216.

Allmon, W. D., Dietl, G. P., Hendricks, J. R., and Ross, R. M. (2018). “Bridging the two fossil records: Paleontology’s ‘big data’ future resides in museum collections,” in Museums at the Forefront of the History and Philosophy of Geology: History Made, History in the Making (Geological Society of America). doi: 10.1130/2018.2535(03).

Aslak, U., Rosvall, M., and Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. Phys. Rev. E 97, 062312. doi: 10.1103/PhysRevE.97.062312.

Barabási, A.-L., and Pósfai, M. (2016). Network science. Cambridge, United Kingdom: Cambridge University Press.

Benson, A. R., Gleich, D. F., and Leskovec, J. (2016). Higher-order organization of complex networks. Science 353, 163–166. doi: 10.1126/science.aad9029.

Blöcker, C., Smiljanić, J., Scholtes, I., and Rosvall, M. (2022). Similarity-based Link Prediction from Modular Compression of Network Flows. doi: 10.48550/ARXIV.2208.14220.

Brandes, U., Robins, G., McCranie, A., and Wasserman, S. (2013). What is network science? Netw. sci. 1, 1–15. doi: 10.1017/nws.2013.2.

Buckland, P. I. (2014). The Bugs Coleopteran Ecology Package (BugsCEP) database: 1000 sites and half a million fossils later. Quaternary International 341, 272–282. doi: 10.1016/j.quaint.2014.01.030.

Buckland, P. I., and Eriksson, E. J. (2014). "Strategic Environmental Archaeology Database (SEAD)," in Encyclopedia of Global Archaeology, ed. C. Smith (New York, NY: Springer New York), 7076–7085. doi: 10.1007/978-1-4419-0465-2_833

Calatayud, J., Bernardo-Madrid, R., Neuman, M., Rojas, A., and Rosvall, M. (2019). Exploring the solution landscape enables more reliable network community detection. Phys. Rev. E 100, 052308. doi: 10.1103/PhysRevE.100.052308.

Calatayud, J., M. Neuman, A. Rojas, A. Eriksson, and M. Rosvall. 2021: Regularities in species' niches reveal the world's climate regions. ELife 10:e58397.

Carletti, T., and Fanelli, D. (2022). "Pattern Formation on Hypergraphs," in Higher-Order Systems Understanding Complex Systems., eds. F. Battiston and G. Petri (Cham: Springer International Publishing), 163–180. doi: 10.1007/978-3-030-91374-8_5.

Council, N. R. (2005). The Geological Record of Ecological Dynamics: Understanding the Biotic Effects of Future Environmental Change. Washington, DC: The National Academies Press doi: 10.17226/11209.

De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015). Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems. Phys. Rev. X 5, 011027. doi: 10.1103/PhysRevX.5.011027.

Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2010). Stability of graph communities across time scales. Proc. Natl. Acad. Sci. U.S.A. 107, 12755–12760. doi: 10.1073/pnas.0903215107.

Dietl, G. P., and Flessa, K. W. (2011). Conservation paleobiology: putting the dead to work. Trends in Ecology & Evolution 26, 30–37. doi: 10.1016/j.tree.2010.09.010

Dunhill, A. M., Bestwick, J., Narey, H., and Sciberras, J. (2016). Dinosaur biogeographical structure and Mesozoic continental fragmentation: a network-based approach. Journal of Biogeography. doi: 10.1111/jbi.12766.

Durham, S. R., and Dietl, G. P. (2015). Perspectives on geohistorical data among oyster restoration professionals in the United States. Journal of Shellfish Research 34, 227–239. doi: 10.2983/035.034.0204.

Edler, D., Bohlin, L., and Rosvall, M. (2017). Mapping Higher-Order Network Flows in Memory and Multilayer Networks with Infomap. Algorithms 10, 112. doi: 10.3390/a10040112.

Efron, B., and Tibshirani, R. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Eriksson, A., Carletti, T., Lambiotte, R., Rojas, A., and Rosvall, M. (2022). "Flow-Based Community Detection in Hypergraphs," in Higher-Order Systems Understanding Complex Systems., eds. F.

Battiston and G. Petri (Cham: Springer International Publishing), 141–161. doi: 10.1007/978-3-030-91374-8_4.

Eriksson, A., Edler, D., Rojas, A., de Domenico, M., and Rosvall, M. (2021). How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs. Commun Phys 4, 133. doi: 10.1038/s42005-021-00634-z.

Ghasemian, A., Hosseinmardi, H., and Clauset, A. (2019). Evaluating Overfit and Underfit in Models of Network Community Structure. IEEE Trans. Knowl. Data Eng., 1–1. doi: 10.1109/TKDE.2019.2911585.

Holland, S. M. (2016). The non-uniformity of fossil preservation. Phil. Trans. R. Soc. B 371, 20150130. doi: 10.1098/rstb.2015.0130.

Holland, S. M., and Patzkowsky, M. E. (2004). Ecosystem Structure and Stability: Middle Upper Ordovician of Central Kentucky, USA. PALAIOS 19, 316–331. doi: 10.1669/0883-1351(2004)019<0316:ESASMU>2.0.CO;2.

Huang, B., Zhan, R.-B., and Wang, G.-X. (2016). Recovery brachiopod associations from the lower Silurian of South China and their paleoecological implications. Can. J. Earth Sci. 53, 674–679. doi: 10.1139/cjes-2015-0193.

Hunt, G. (2010). Evolution in Fossil Lineages: Paleontology and The Origin of Species. The American Naturalist 176, S61–S76. doi: 10.1086/657057.

Jeon, J., Liang, K., Park, J., Kershaw, S., and Zhang, Y. (2021). Diverse labechiid stromatoporoids from the Upper Ordovician Xiazhen Formation of South China and their paleobiogeographic implications. J. Paleontol.

Kheirkhahzadeh, M., Lancichinetti, A., and Rosvall, M. (2016). Efficient community detection of network flows for varying Markov times and bipartite networks. Phys. Rev. E 93, 032309. doi: 10.1103/PhysRevE.93.032309.

Kidwell, S. M., and Flessa, K. W. (1996). The Quality of the Fossil Record: Populations, Species, and Communities. Annu. Rev. Earth Planet. Sci. 24, 433–464. doi: 10.1146/annurev.earth.24.1.433.

Kidwell, S. M., and Holland, S. M. (2002). The Quality of the Fossil Record: Implications for Evolutionary Analyses. Annual Review of Ecology and Systematics 33, 561–588. doi: 10.1146/annurev.ecolsys.33.030602.152151.

Kidwell, S. M., and Tomasovych, A. (2013). “Implications of Time-Averaged Death Assemblages for Ecology and Conservation Biology,” in Annual Review of Ecology, Evolution, and Systematics, Vol 44, ed. D. J. Futuyma, 539-+.

Kiel, S. (2017). Using network analysis to trace the evolution of biogeography through geologic time: A case study. Geology, G38877.1. doi: 10.1130/G38877.1.

Kocsis, Á. T., Reddin, C. J., and Kiessling, W. (2018). The biogeographical imprint of mass extinctions. Proc. R. Soc. B. 285, 20180232. doi: 10.1098/rspb.2018.0232.

Kocsis, Á. T., Reddin, C. J., Scotese, C. R., Valdes, P. J., and Kiessling, W. (2021). Increase in marine provinciality over the last 250 million years governed more by climate change than plate tectonics. Proc. R. Soc. B. 288, 20211342. doi: 10.1098/rspb.2021.1342.

Lambiotte, R., and Rosvall, M. (2012). Ranking and clustering of nodes in networks with smart teleportation. Phys. Rev. E 85, 056107. doi: 10.1103/PhysRevE.85.056107.

Lambiotte, R., Delvenne, J.-C., and Barahona, M. (2014). Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks. IEEE Trans. Netw. Sci. Eng. 1, 76–90. doi: 10.1109/TNSE.2015.2391998.

Lambiotte, R., M. Rosvall, and I. Scholtes. 2019: From networks to optimal higher-order models of complex systems. Nature Physics 15:313–320.

Lancichinetti, A., and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. Phys. Rev. E 80, 056117. doi: 10.1103/PhysRevE.80.056117.

Law, S. R., Serrano, A. R., Daguerre, Y., Sundh, J., Schneider, A. N., Stangl, Z. R., et al. (2022). Metatranscriptomics captures dynamic shifts in mycorrhizal coordination in boreal forests. Proc. Natl. Acad. Sci. U.S.A. 119, e2118852119. doi: 10.1073/pnas.2118852119.

Lazaridis, I., Alpaslan-Roodenberg, S., Acar, A., Açıkkol, A., Agelarakis, A., Aghikyan, L., et al. (2022). The genetic history of the Southern Arc: A bridge between West Asia and Europe. Science 377, eabm4247. doi: 10.1126/science.abm4247.

Martins, A. F., da Cunha, B. R., Hanley, Q. S., Gonçalves, S., Perc, M., and Ribeiro, H. V. (2022). Universality of political corruption networks. Sci Rep 12, 6858. doi: 10.1038/s41598-022-10909-2.

Muscente, A. D., Bykova, N., Boag, T. H., Buatois, L. A., Mángano, M. G., Eleish, A., et al. (2019). Ediacaran biozones identified with network analysis provide evidence for pulsed extinctions of early complex life. Nat Commun 10, 911. doi: 10.1038/s41467-019-08837-3.

NOW Community. (2020). NOW — New and Old Worlds: Database of fossil mammals. doi: 10.5281/ZENODO.4268068.

Patzkowsky, M. E., and Holland, S. M. (2012). Stratigraphic paleobiology: understanding the distribution of fossil taxa in time and space. Chicago ; London: The University of Chicago Press.

Peixoto, T. P., and Rosvall, M. (2017). Modelling sequences and temporal networks with dynamic community structures. Nat Commun 8, 582. doi: 10.1038/s41467-017-00148-9.

Penn-Clarke, C. R., and Harper, D. A. T. (2020). Early−Middle Devonian brachiopod provincialism and bioregionalization at high latitudes: A case study from southwestern Gondwana. GSA Bulletin. doi: 10.1130/B35670.1.

Perri, V., and Scholtes, I. (2020). "HOTVis: Higher-Order Time-Aware Visualisation of Dynamic Graphs," in Graph Drawing and Network Visualization Lecture Notes in Computer Science., eds. D. Auber and P. Valtr (Cham: Springer International Publishing), 99–114. doi: 10.1007/978-3-030-68766-3_8.

Peters, S. E., and McClennen, M. (2016). The Paleobiology Database application programming interface. Paleobiology 42, 1–7. doi: 10.1017/pab.2015.39.

Pilotto, F., Rojas, A., and Buckland, P. I. (2022). Late Holocene anthropogenic landscape change in northwestern Europe impacted insect biodiversity as much as climate change did after the last Ice Age. Proc. R. Soc. B. 289, 20212734. doi: 10.1098/rspb.2021.2734.

Rissanen, J. (1978). Modeling by shortest data description. Automatica 14, 465–471. doi: 10.1016/0005-1098(78)90005-5.

Rojas, A., Calatayud, J., Kowalewski, M., Neuman, M., and Rosvall, M. (2021). A multiscale view of the Phanerozoic fossil record reveals the three major biotic transitions. Commun Biol 4, 309. doi: 10.1038/s42003-021-01805-y.

Rojas, A., Gracia, A., Hernández-Ávila, I., Patarroyo, P., and Kowalewski, M. (2022). Occurrence of the brachiopod *Tichosina* in deep-sea coral bottoms of the Caribbean Sea and its paleoenvironmental implications. Bulletin of the Florida Museum of Natural History 59, 1–15. doi: 10.1101/2020.06.24.168658.

Rojas, A., Patarroyo, P., Mao, L., Bengtson, P., and Kowalewski, M. (2017). Global biogeography of Albian ammonoids: A network-based approach. Geology 45, 659–662. doi: 10.1130/G38944.1.

Roopnarine, P. (2010). Networks, Extinction and Paleocommunity Food Webs. Nat Prec. doi: 10.1038/npre.2010.4433.2.

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105, 1118–1123. doi: 10.1073/pnas.0706851105.

Rosvall, M., and Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PloS one, 6(4). doi: 10.1371/journal.pone.0018209.

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. The European Physical Journal Special Topics 178, 13–23. doi: 10.1140/epjst/e2010-01179-1.

Rosvall, M., Delvenne, J., Schaub, M. T., and Lambiotte, R. (2019). “Different Approaches to Community Detection,” in Advances in Network Clustering and Blockmodeling, eds. P. Doreian, V. Batagelj, and A. Ferligoj (Wiley), 105–119. doi: 10.1002/9781119483298.ch4.

Rosvall, M., Esquivel, A. V., Lancichinetti, A., West, J. D., and Lambiotte, R. (2014). Memory in network flows and its effects on spreading dynamics and community detection. Nat Commun 5, 4630. doi: 10.1038/ncomms5630.

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Scarponi, D., and Kowalewski, M. (2007). Sequence stratigraphic anatomy of diversity patterns: Late Quaternary benthic mollusks of The Po Plain, Italy. Palaios 22, 296–305. doi: 10.2110/palo.2005.p05-020r.

Scarponi, D., Rojas, A., Nawrot, R., Cheli, A., and Kowalewski, M. (2022). "Assessing biotic response to anthropogenic forcing using mollusk assemblages from the Po-Adriatic System (Italy)," in Conservation Palaeobiology of Marine Ecosystems: Concepts and Applications (The Geological Society Special Publications). In press

Schaub, M. T., Delvenne, J.-C., Rosvall, M., and Lambiotte, R. (2017). The many facets of community detection in complex networks. Appl Netw Sci 2, 4. doi: 10.1007/s41109-017-0023-6.

Schaub, M. T., Lambiotte, R., and Barahona, M. (2012). Encoding dynamics for multiscale community detection: Markov time sweeping for the map equation. Phys. Rev. E 86, 026112. doi: 10.1103/PhysRevE.86.026112.

Sepkoski, D. (2013). Towards "A Natural History of Data": Evolving Practices and Epistemologies of Data in Paleontology, 1800–2000. J Hist Biol 46, 401–444. doi: 10.1007/s10739-012-9336-6.

Sepkoski, J. J. (1981). A factor analytic description of the Phanerozoic marine fossil record. Paleobiology 7, 36–53. doi: 10.1017/S0094837300003778.

Smiljanić, J., Blöcker, C., Edler, D., and Rosvall, M. (2021). Mapping flows on weighted and directed networks with incomplete observations. Journal of Complex Networks 9, cnab044. doi: 10.1093/comnet/cnab044.

Smiljanić, J., Edler, D., and Rosvall, M. (2020). Mapping flows on sparse networks with missing links. Phys. Rev. E 102, 012302. doi: 10.1103/PhysRevE.102.012302.

Smith, S. M., Sprain, C. J., Clemens, W. A., Lofgren, D. L., Renne, P. R., and Wilson, G. P. (2018). Early mammalian recovery after the end-Cretaceous mass extinction: A high-resolution view from McGuire Creek area, Montana, USA. GSA Bulletin. doi: 10.1130/B31926.1.

Viglietti, P. A., Rojas, A., Rosvall, M., Klimes, B., and Angielczyk, K. D. (2022). Network-based biostratigraphy for the late Permian to mid- Triassic Beaufort Group (Karoo Supergroup) in South Africa enhances biozone applicability and stratigraphic correlation. Palaeontology 65. doi: 10.1111/pala.12622.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? in Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09 (Montreal, Quebec, Canada: ACM Press), 1–8. doi: 10.1145/1553374.1553511.

Williams, J. W., Grimm, E. C., Blois, J. L., et al. (2018). The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. Quat. res. 89, 156–177. doi: 10.1017/qua.2017.105.

Ye, F., Shi, G. R., and Bitner, M. A. (2021). Global biogeography of living brachiopods: Bioregionalization patterns and possible controls. PLoS ONE 16, e0259004. doi: 10.1371/journal.pone.0259004.

Data Availability Statement

Data for this study are available in the Dryad Digital Repository: https://doi.org/10.5061/dryad.sj3tx967z.

Figures captions

Figure 1. Network representations of geohistorical data are abstracted fossil records. A. The physical fossil record. These brachiopod shells aim to represent the benthic macrofauna from the R/V Pillsbury program in the Caribbean. Modified from Rojas et al. (2022, fig. 2). B. The abstracted fossil record. This abstracted record is a bipartite network representation of the underlying data (sampling stations × taxa matrix) (Rojas et al. 2022)(Supplementary Data 1). The brachiopod *Tichosina*, the larger component of the Cenozoic brachiopod faunas in the Caribbean, is indicated. C. Modular structure delineated via community detection with the Map Equation framework and using a Markov Time = 2 (Kheirkhahzadeh et al. 2016) (Supplementary Data 2). Modules include sampling stations and taxa. Nodes are rearranged in the circular layout by their module affiliation. Only the two larger modules are displayed; they represent 98% of the network flow. D. Modules mapped on a Detrended Correspondence Analysis (DCA) ordination space. E. Small example network showing a two-module partition.

Figure 2. Input and output formats. -The Map Equation framework understands different formats. **A.** Input file in Pajek format. **B.** Output file. The resulting partition is written to a file with the extension .tree (plain text file). The formats described here correspond to the small example network in Figure 1E.

Figure 3. Network representations of geohistorical data. **A.** Temporal occurrence data. **B.** Standard bipartite network representation created by aggregating fossil occurrences (sp. 1) into arbitrary spatiotemporally explicit units (g1 to g4). This standard network uses the same nodes to represent the physical components of the system as well as to describe system's flows. A random walk as the simplest information flow model forms paths across independent pairwise links on this network washing out higher-order node interactions. **C.** Multilayer network representation. The Map Equation framework for higher-order networks distinguishes physical nodes that represent the system's components from state nodes describing the system's internal flows. This higher-order network better captures the constraints on the information paths and thus flows tend to stay among units from each layer.

Figure 4. Different network representations capture different aspects of the benchmark data on the Phanerozoic benthic marine faunas (Peters and McClennen 2016). Bipartite and unipartite representations ignore the temporal constraints inherent to the biosedimentary record. *Unipartite projection obtained from the bipartite network by rescaling the Markov time (Kheirkhahzadeh et al. 2016). Alluvial diagram representing 98% of the network flow in each case (Supplementary Data 3-7).

Figure 5. Schematic diagram illustrating a hypergraph representing temporal occurrence data. In this diagram, physical nodes are fossil taxa connected through hyperlinks. Hyperlinks are depicted with different colors and connect taxa from the same time interval. Taxa recorded in different time intervals are represented by state nodes, depicted with different shades of grey.

Figure 6. Aggregated inter-module links used by the Network Navigator. Modules and nodes as circles with areas proportional to flow and border widths proportional to exiting flow. Bidirectional links with width proportional to inter-module flow

Figure 7. Visualizing change in network partitions using alluvial diagrams. To compare networks with the same sets of nodes, we assemble at least two network partitions (a). Then, we group nodes in the same module in stacked bars with height proportional to the node's flow volume and connect corresponding nodes with streamlines (b). Finally, to highlight how the partitions change, we aggregate the nodes into modules (c). To compare additional network partitions, we add more stacks of bars to the right and repeat the procedure b-c.

Figure 8. Quality of alternative partitions of the multilayer network representing the Phanerozoic fossil record of the benthic marine faunas mapped in a two-dimensional space. Circles represent clusters of network partitions, located at the cluster center, with size proportional to the number of partitions grouped into the cluster. Map isolines are constructed using the Jaccard distance between partitions. Despite differences in their quality (codelength), all partition clusters identified at the selected scale show a modular structure with four modules representing the Phanerozoic evolutionary faunas. At the selected scale, there is not a cluster of solutions representing a three-tier model (Sepkoski 1981).

Figure 9. A. A multilayer network representing litho-biofacies from Late Ordovician outcrops in central Kentucky. In this higher order network, one layer describes the biotic component (samples × taxa matrix) and the other describes the abiotic component (samples × environments matrix) of the geohistorical record (Supplementary Data 8). **B**. Litho-biofacies delineated via community detection using the Map Equation framework. Modules in the multilayer solution include taxa, samples, and environments, and can be directly interpreted as litho-biofacies (Supplementary Data 9). **C**. Detrended Component Analysis (DCA) on the samples × taxa matrix. Network modules mapped in a DCA ordination space indicating their distribution along the depth water gradient. Background colored based on the module affiliation. Data from Holland and Patzkowsky (2004).

Figure 10. Bioregionalization of the Middle Devonian Old World Realm. **A**. Varying Markov time models on the biogeographic network constructed from brachiopod occurrence data (Penn-Clarke and Harper 2020). Network partitions at different Markov times reveal the larger-scale biogeographic structures obtained at different resolutions. **B**. Network partition obtained at the Markov time 1 (7 modules). Circles represent the seven modules delineated when exploring the one-step dynamics on the assembled network. **C**. Modules obtained from partitions at Markov times 1.30 and 1.35 (3 modules) mapped on those obtained at the Markov time 1. These two partitions differ in the affiliation of the Southern Peru locality, placed alternatively into the modules representing higher (white) and lower (orange) latitudes. Overall, coarser partitions contain fewer clusters as the Markov time increases.

Table captions

Table 1. Comparison of the modular structure of different network representations of the benchmark data on the Phanerozoic marine faunas. The multilayer representation achieves the shortest codelength and the best compression. For each partition, we measure the compression by using the corresponding one-level partition as a baseline. *Unipartite projection obtained from the bipartite network by rescaling the Markov time (Kheirkhahzadeh et al. 2016).