


RESEARCH

Open Access



# Mapping change in higher-order networks with multilevel and overlapping communities

Anton Holmgren<sup>1\*</sup> , Daniel Edler<sup>1,2</sup>  and Martin Rosvall<sup>1</sup> 

\*Correspondence:  
anton.holmgren@umu.se

<sup>1</sup> Integrated Science Lab,  
Department of Physics, Umeå  
University, 901 87 Umeå, Sweden

<sup>2</sup> Department of Biological  
and Environmental Sciences,  
Gothenburg Global Biodiversity  
Centre, University of Gothenburg,  
405 30 Gothenburg, Sweden

## Abstract

New network models of complex systems use layers, state nodes, or hyperedges to capture higher-order interactions and dynamics. Simplifying how the higher-order networks change over time or depending on the network model would be easy with alluvial diagrams, which visualize community splits and merges between networks. However, alluvial diagrams were developed for networks with regular nodes assigned to non-overlapping flat communities. How should they be defined for nodes in layers, state nodes, or hyperedges? How can they depict multilevel, overlapping communities? Here we generalize alluvial diagrams to map change in higher-order networks and provide an interactive tool for anyone to generate alluvial diagrams. We use the alluvial diagram generator in three case studies to illustrate significant changes in the organization of science, the effect of modeling network flows with memory in a citation network and distinguishing multidisciplinary from field-specific journals, and the effects of multilayer representation of a collaboration hypergraph.

## Introduction

Complex systems are inherently dynamic. Their components influence each other through various informational and physical processes, changing interaction patterns over time. Researchers represent these interactions with networks (Edler et al. 2017; Calatayud et al. 2020; Farage et al. 2021; Calatayud et al. 2021; Neuman 2022; Edler et al. 2022a; Rojas et al. 2022) and simplify their organization with community-detection algorithms (Rosvall and Bergstrom 2008; Fortunato 2010; Schaub et al. 2017; Traag et al. 2019; Peixoto 2019). For example, community-detection algorithms that model the various processes as flows on networks assign nodes to possibly nested modules of typically densely connected nodes, among which the network flows persist relatively long (Rosvall and Bergstrom 2008). Identifying modules in multiple networks with shared nodes enables exploring organizational changes when the systems they represent change over time or between states: Modules merge and split when groups in students' social networks form and dissolve during school days, or new research fields emerge when old fields fuse or break and move apart. Various summary statistics can quantify these structural changes (Danon et al. 2005; Amelio and Pizzuti 2017; Newman et al. 2020), but they destroy essential information about how the networks change.

Alluvial diagrams with modules represented as stacks of blocks joined by stream fields were introduced to reveal network organizational changes by depicting merging and splitting modules (Rosvall and Bergstrom 2010). Researchers have successfully used them to map shifting regional tendencies in urban networks (Liu et al. 2013), study dynamics of hot topics in research fields (Ruan et al. 2017; Pal et al. 2022), track changing bitcoin user activity (Cazabet et al. 2017), and explore evolving media channel preferences across crisis phases (Petrun Sayers et al. 2021). Generating alluvial diagrams requires dedicated software to remove tedious manual work. However, current applications to generate alluvial diagrams work only for standard networks partitioned into modules.

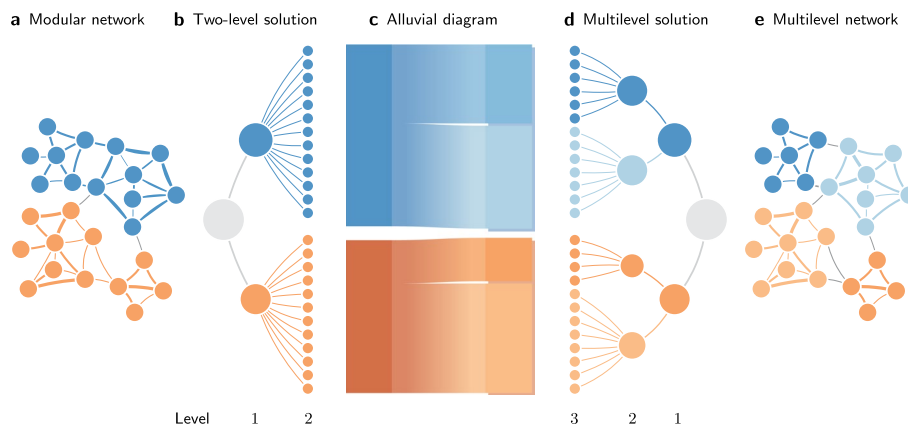
Today researchers use temporal, multilayer, and memory networks to capture interactions in complex systems with higher accuracy (Kivelä et al. 2014; Rosvall et al. 2014; De Domenico et al. 2015, 2016; Xu et al. 2016; Lambiotte et al. 2019) and multilevel modular solutions to reveal more regularities in their organization (Rosvall and Bergstrom 2011; Peixoto 2014). Multilayer networks can represent networks over time with links in time-windowed layers. Memory networks can represent higher-order network flow models where the transition rates depend on the current node and previously visited nodes. Both representations enable overlapping modules. Mapping change in these rich network representations requires generalizing alluvial diagrams and their generators to higher-order networks with multilevel and overlapping modular solutions.

Here we introduce alluvial diagrams for multilayer and memory networks with multilevel and overlapping modular solutions. We demonstrate a new alluvial generator for higher-order networks available for anyone to use at <https://www.mapequation.org/alluvial> (Holmgren et al. 2022a), and illustrate how we use it in three case studies revealing: significant changes in the multilevel organization of science over six years using parametric bootstrap resampling, multidisciplinary journals in a second-order network representation of citation flows, and the effects of multilayer representation of a collaboration hypergraph.

## Methods

Alluvial diagrams depict changes in the modular composition between networks with stacks of blocks representing the modules (Fig. 1). Each block's height is proportional to the flow volume of the corresponding module—the total visit probability of all nodes in the module. To highlight structural change between multiple networks, a vertical stack of blocks represent each network's modular structure, and horizontal stream fields connect blocks that share nodes across neighboring networks. Like block heights, stream-field heights are proportional to the flow volume of the node overlap between corresponding modules. To reduce clutter, we order stream fields to minimize their overlap.

We use Infomap to search for multilevel modular structures with nested submodules (Edler et al. 2022b; Rosvall and Bergstrom 2011). Throughout the paper, we use multilevel to denote partitions with nested submodules as illustrated in Fig. 1d–e, and multilayer to denote the network type that stratifies connections between nodes into different layers. Infomap optimizes the map equation, the average per-step codelength on a modular description of a random walk modeling network flows (Rosvall and Bergstrom 2008). The modules are groups of nodes where the random walker spends a relatively



**Fig. 1** Schematic alluvial diagram of a multilevel network structure. **a** A weighted network with modular structure, organized into a two-level solution in **b**. **c** An alluvial diagram representation of the solutions in panels **b** and **d** using the same colors. Columns of blocks represent modules with heights proportional to the contained flow volume. The leftmost column is an ordinary two-level alluvial diagram representation. The multilevel representation to the right shows multiple levels, with the background showing the top-level organization. Stream fields connect modules in the left and right columns that share nodes. **d** Multilevel solution of the network in **e**

long time compared to exiting it and entering other modules. While we focus on modules derived from Infomap, alluvial diagrams work with output from any community detection or hierarchical data clustering method.

#### Mapping change in networks with multilevel communities

We extend alluvial diagrams to multilevel network partitions by nesting submodules in super-modules with adaptive module distances. The right multilayer stack of the schematic alluvial diagram in Fig. 1c illustrates. In the spirit of cartography, we put blocks corresponding to the top-level modules in a bottom layer to highlight the large-scale organization and provide a cleaner visualization. Optionally, we display finer-level structures in layers above the bottom layer. The right multilayer stack in Fig. 1c expands the left stack's single layer with one such extra layer corresponding to the four submodules of the multilevel modular solution. To show that deeper submodules are more closely related than their larger parent modules, we draw sibling submodules closer together than other modules. Specifically, we halve the distance between two adjacent modules for each level down in the multilevel solution.

#### Multilevel significance clustering

To separate trends from mere noise in the module assignments, we extend the significance clustering method described in ref. Rosvall and Bergstrom (2010) to multilevel partitions. The approach has three main steps: First, we search for optimal multilevel partitions for each network using Infomap. Then, to assess these partitions' robustness to slight perturbations in the data, we create a large number of independent bootstrap networks. For each bootstrap network, we search for the optimal multilevel partition using Infomap as for the original network. Finally, we summarize the variability in the bootstrap partitions by applying the significance clustering method introduced in

ref. Rosvall and Bergstrom (2010) extended to multilevel partitions. For each level in the multilevel solution of the original network, we search for the largest subset of nodes in each module or submodule that are also clustered together in at least a fraction  $p$  of solutions obtained from the parametric bootstrap procedure.

Searching for significant subsets in multilevel solutions is computationally more demanding than for ordinary two-level partitions. To improve the performance, we trivially parallelize the algorithm by running each module or submodule in separate threads.

### Mapping change in higher-order networks

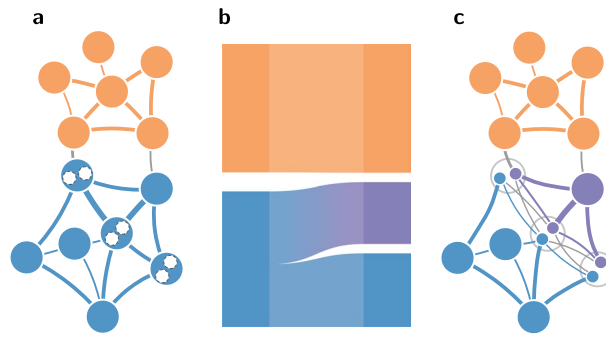
We generalize alluvial diagrams to multilayer and memory networks. Multilayer networks can model different modes of interaction or interactions that change over time in different layers. Memory networks can model dynamics that depend on from where the flows come. Infomap represents both higher-order networks with so-called *state nodes* (Edler and Bohlin 2017). In higher-order networks, we call ordinary nodes *physical nodes* to distinguish them from state nodes. In a multilayer network, one state node for each physical node and layer represents the physical node in the layer (De Domenico et al. 2015). In a second-order memory network with memory of the previous step, one state node for each physical node and incoming link represents the physical node for flows incoming along that link (Rosvall et al. 2014). In this way, the order of a memory network corresponds to the order of a Markov process, the first order being regular memoryless Markov dynamics. Physical nodes with multiple state nodes and different outgoing links can model higher-order dynamics on the network.

In theory, using alluvial diagrams for higher-order networks is no different than for ordinary networks. In practice, the many possible combinations of first- and higher-order networks, memory networks with different memory, and multilayer networks with different layers make it challenging to determine node equality in different networks because we need to match nodes across networks to draw stream fields between modules. While alluvial diagrams require networks to share a significant fraction of physical nodes, we also need their state nodes to match since they are the smallest components of higher-order networks. With no universal solution to this node-matching problem, we discuss some challenges and how we choose to solve them.

### First- and higher-order networks

Alluvial diagrams with first- and higher-order networks require matching different node types: First-order networks have only physical nodes, but higher-order networks have physical nodes and state nodes. We illustrate this schematically in Fig. 2 with a first-order network in Fig. 2a and a higher-order network with state nodes as smaller circles inside the physical nodes in Fig. 2c. We consider only hard module boundaries in the first-order network, whereas modules overlap in the higher-order network when physical nodes' state nodes are assigned to different modules. In Fig. 2c, the modules overlap in the physical nodes containing the purple and blue state nodes.

As we need a one-to-one match across networks to draw stream fields, we cannot match all state nodes in the higher-order network to one first-order node. To overcome this problem, we first split the first-order nodes into pseudo-state nodes, which we depict with small dashed circles in Fig. 2a. We create as many pseudo-state nodes as there are state nodes in



**Fig. 2** Schematic first-order and higher-order networks and alluvial diagram representation. **a** A first-order network with pseudo-states (white dashed circles) matching the state nodes in **c**. **b** An alluvial diagram representation. **c** A higher-order network with state nodes (small blue and purple circles) in the physical nodes. We only show the state nodes whose physical nodes are present in two modules

the matching physical node in the higher-order network. Then, we divide first-order node  $i$ 's flow volume  $\pi_i^{(1)}$  among its pseudo-states  $\alpha$  proportionally to their matching state nodes' fraction of the flow  $\pi_{i\alpha}^{(2)}$  as

$$\pi_{i\alpha}^{(1)} = \pi_i^{(1)} \frac{\pi_{i\alpha}^{(2)}}{\pi_i^{(2)}}. \tag{1}$$

This procedure gives a one-to-one match between nodes in first- and higher-order networks, and we can draw multiple stream fields from a single first-order node (Fig. 2b).

**Memory networks**

Drawing alluvial diagrams for memory networks requires matching state nodes representing corresponding memory in different networks. We match state nodes across networks by encoding their memory in their ids such that state nodes representing the same memory share the same id in different networks. As long as the networks are not too large, we can encode memory of order  $n$  into a single binary number by dividing the binary number into  $n$  parts: We divide the number into two parts in a second-order memory network with memory of the previous step. With  $N$  physical nodes, we use the  $b = \lceil \log_2 N \rceil$  most significant bits of the state id to encode the previously visited node  $i$  and the  $b$  least significant bits to encode the currently visited node  $j$ , resulting in the state id

$$\alpha_{i \rightarrow j} = i \ll b + 1 \vee j, \tag{2}$$

where  $\ll$  is the arithmetic left-shift operator and  $\vee$  is the logical or. For example, we encode the link from physical node 2 to physical node 3 along the path represented by the trigram  $1 \rightarrow 2 \rightarrow 3$  as

$$\begin{aligned} \alpha_{1 \rightarrow 2} &= 1 \ll 2 + 1 \vee 2 = 1000_2 \vee 10_2 = 1010_2 = 10, \\ \alpha_{2 \rightarrow 3} &= 2 \ll 2 + 1 \vee 3 = 10000_2 \vee 11_2 = 10011_2 = 19, \end{aligned}$$

resulting in the directed link  $10 \rightarrow 19$  between state nodes 10 and 19. This encoding scheme works for up to  $N = 2^{16} = 65,536$  physical nodes with 32-bit ids and second-order memory.

### **Multilayer networks**

When comparing multiple multilayer networks with  $N$  layers, we encode the physical node  $i$  in layer  $l$  with id

$$\alpha_{i,l} = i \ll b + 1 \vee l, \quad (3)$$

where  $N$  is the largest layer id represented with  $b = \lceil \log_2 N \rceil$  bits. For multilayer networks, this encoding scheme is available in Infomap using the flag `--matchable-multilayer-ids N`.

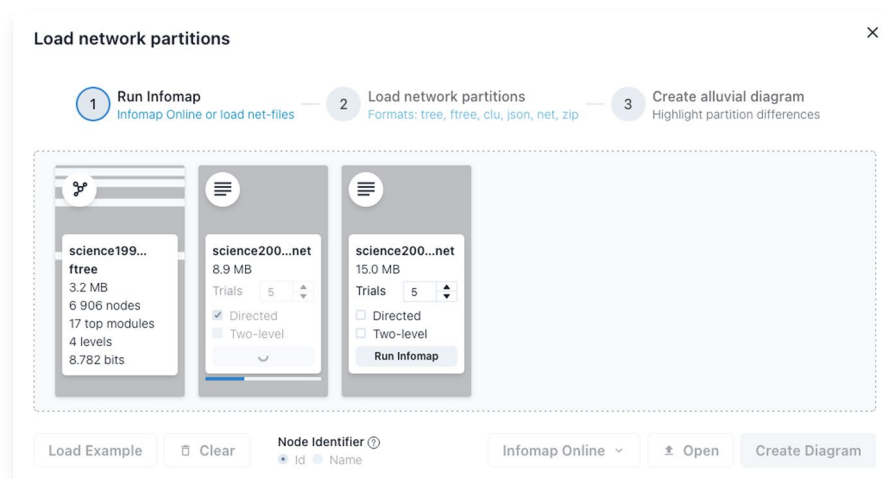
Alluvial diagrams can also visualize the layers of multilayer networks, each as a separate network. In this case, node matching is trivial as physical nodes are unique in each layer. The stream fields then connect modules that span layers.

### **Alluvial diagram generator**

We have implemented an interactive web application that generates alluvial diagrams, available for anyone to use at <https://www.mapequation.org/alluvial>. We implemented it as a client-side web application to enable researchers to use our application without programming experience or those working with sensitive data. All code runs locally in the user's web browser, and the web application does not store or upload network data to any server. We implemented it using TypeScript and React, and we display the diagrams using scalable vector graphics (SVG) (see Additional file 1: Fig. S2 in the SI for how we model the data structures).

While the most efficient community detection pipeline is to run the stand-alone C++ version of Infomap and load the resulting partitions, we have embedded a version of Infomap compiled to JavaScript with Emscripten (Zakai 2011). This embedded Infomap version supports the same network inputs as C++ Infomap, but only a subset of Infomap's features, including reading directed or undirected input, choosing the number of optimization trials, and searching for multilevel or two-level solutions (Fig. 3). We defer the specification of input formats to Additional file 1: section SI.1. We also support loading solutions from Infomap Online (Holmgren et al. 2022b), a fully featured web-based version of Infomap.

With loaded networks, the interface shows the user a top-level view of the alluvial diagram (Additional file 1: Fig. S1). The user can manipulate the diagram in several ways: expand modules to reveal their submodules, reorganize networks and modules for clarity, highlight modules or individual nodes with different colors, and change the diagram width and height. While we have implemented the features and use cases we think most researchers use, we can imagine feature requests for specific use cases. By supporting export to SVG, researchers can modify the diagrams to their needs in any vector graphics application.



**Fig. 3** Loading networks in the alluvial diagram generator. Three networks are loaded in different stages, shown as gray rectangles. The leftmost network has communities detected by Infomap. Infomap runs on the second network and has completed two out of five optimization trials. Infomap has yet to start identifying communities in the rightmost network. When all have finished, or when loading networks with communities from C++ Infomap, the user can select “Create Diagram” to create an alluvial diagram

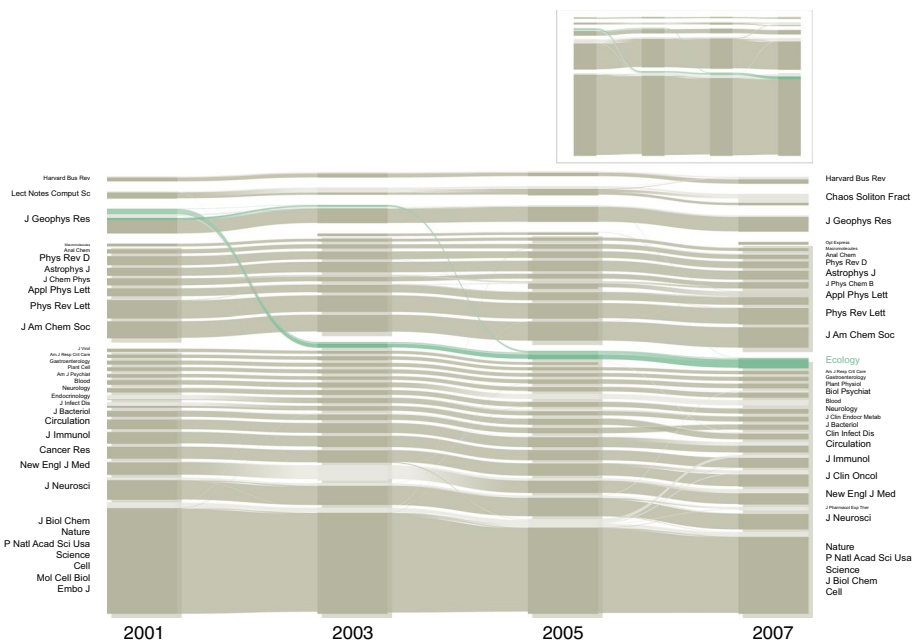
## Results

We highlight different visualization challenges in three case studies using multilevel, higher-order, and multilayer networks. In all cases, we use Infomap to identify optimal multilevel solutions using unrecorded teleportation to links with minimal impact from the teleportation rate on the results (Lambiotte and Rosvall 2012).

### Robust multilevel citation networks

First, we highlight the multilevel organization of science into research areas and fields. We use data from Thomson-Reuters Journal Citation Reports (Rosvall and Bergstrom 2010). The data include citations between journals published from 2001–2007, divided into four two-year periods. The networks have, on average, 7, 490 nodes representing journals and 586, 295 integer-weighted links representing the citation flow between them. For each year, we use Infomap with 100 optimization trials to search for the optimal multilevel solution. We use the multilevel significance clustering approach described in the “Methods” section to assess the solution’s robustness to slight perturbations in the data. First, we create 1000 independent bootstrap networks by sampling each citation weight  $w_{uv}$  from a Poisson distribution,  $\hat{w}_{uv} \sim \text{Poisson}(w_{uv})$ . Then, we use Infomap to search for the optimal multilevel solution for each bootstrap network. The bootstrap solutions have similar codelengths, with a variance of around  $10^{-5}$ . Finally, we use the significance clustering algorithm to search for the largest fraction of nodes clustered together in at least a fraction  $p = 0.95$  of the bootstrap solutions.

The resulting multilevel partitions organize science into research areas, further divided into research fields (Fig. 4). With the multilevel solution and unrecorded teleportation scheme, we do not exactly reproduce the results presented in Ref. Rosvall



**Fig. 4** Multilevel organization of science, 2001–2007. The journals organize into five large research fields, further divided into research areas. We show the top-level organization of earth sciences, economics, and computer science (the three small modules at the top), and the finer division into research fields for the physical and life sciences. The fields or areas are sorted by their citation flow and show the top-ranking journal in each research field or area. Finally, we highlight the insignificant assignment of journals clustered with Ecology in 2001 to a significantly distinct research field since 2003. At the top level shown in the inset, the same journals cluster significantly with the life sciences since 2007

and Bergstrom (2010). The life sciences show higher diversification, with more significant research fields and lower citation flows in molecular- and cell biology containing *J. Biol. Chem.*, *Nature*, *PNAS*, *Science*, *Cell*, and so on.

**First- and second-order citation networks**

In the second case study, we visualize the effects of using higher-order network models with alluvial diagrams. We organize the citation data from the Thomson-Reuters Journal Citation Reports into citation pathways (Persson et al. 2016; Wang and Waltman 2016). The data contain citations between articles published from 2007 to 2012 in the 10 000 journals with the highest impact factor, and all citation pathways contain at least one article published in 2009. When aggregated to journals, we are left with 69, 738, 205 weighted trigrams.

To study the effect of a second-order model, we model the data using both first- and second-order Markov chains. We create a first-order network by discarding the first step from each trigram. For example, the trigram  $i \rightarrow j \rightarrow k$  with weight  $w$  becomes the directed link  $j \rightarrow k$  with the same weight, resulting in 69 million links between the 10, 000 nodes. Using the complete trigram data, we create a second-order network. For each trigram  $i \rightarrow j \rightarrow k$  with weight  $w$ , we create two state nodes if they do not already exist:

- $\alpha_{i \rightarrow j}$  in physical node  $j$  representing the memory of coming from  $i$ ,



- $\alpha_{j \rightarrow k}$  in physical node  $k$  representing the memory of coming from  $j$ .

We connect the state nodes with a directed link  $\alpha_{i \rightarrow j} \rightarrow \alpha_{j \rightarrow k}$  with weight  $w$ . The resulting second-order network has around 3.9 million state nodes connected by 69 million links.

Because the second-order network has two orders of magnitude more state nodes than the first-order network has physical nodes, the community detection search space is much larger, significantly impacting the computational time. The first-order network takes around two minutes for ten optimization trials, while the second-order network takes around nine hours for the same task on a 2021 MacBook Pro with the M1 Max CPU and 32 GB of RAM. The resulting first-order partition has codelength  $L^{(1)} = 8.44$  bits, five top modules, and four levels. The second-order partition has codelength  $L^{(2)} = 7.83$  bits, around 4,700 top modules, and five levels. Although the second-order partition has many top modules, most are tiny, containing only one or a few state nodes. To downplay small modules at the fringe of the citation data, we compare the partition's effective number of top modules using the perplexity  $M_{\text{eff}} = 2^{H(M)}$ , with Shannon entropy

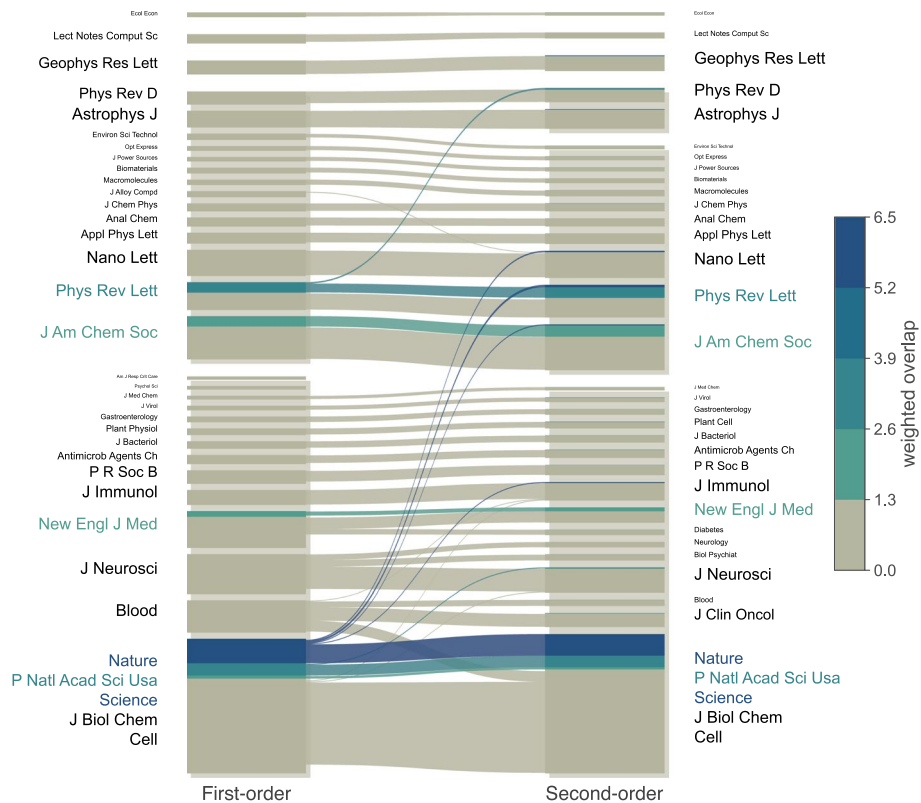
$$H(M) = - \sum_m \pi_m \log_2 \pi_m, \quad (4)$$

where  $\pi_m = \sum_{i \in m} \pi_i$  is the total flow volume of the nodes  $i$  in module  $m$ . With this metric, the first- and second-order partitions are similar with  $M_{\text{eff}}^{(1)} = 2.35$  and  $M_{\text{eff}}^{(2)} = 2.73$  effective top modules, respectively.

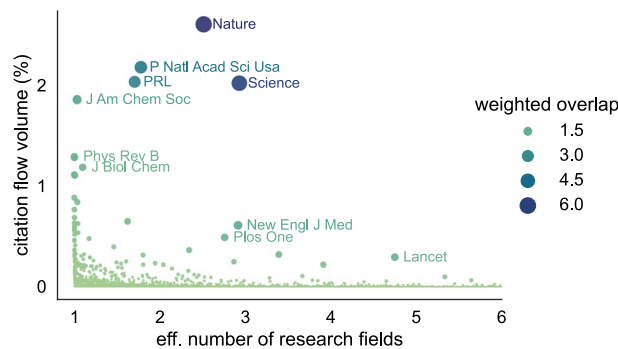
After detecting communities, we aggregate redundant state nodes in the second-order network before visualization for better performance. We lump state nodes in the same physical node and leaf module and aggregate their flows, reducing the number of states to visualize from 3.9 million to 355 thousand. After lumping, we remove any state nodes with zero flow that would not contribute to the alluvial diagram layout, further reducing the number of states to 271 thousand. Then, we create pseudo-states in the first-order network to match the higher-order state nodes. After this step, both networks contain 271 thousand state nodes. In the first-order network, all state nodes are in the same module as their physical node.

The alluvial diagram shows how the second-order model separates cosmology and astrophysics—journals clustered together with *Astrophys J.* and *Phys. Rev. D.* – from the physical sciences (Fig. 5). The cell- and molecular biology submodule containing *Nature*, *PNAS*, and *Science* grows, and the multidisciplinary journals' submodules in the life sciences divide into smaller modules.

Above all, *Nature*, *Science*, and *PNAS* are all recognized as multidisciplinary journals represented in multiple research fields. To quantify how a higher-order model captures their citation flows, we investigate in how many research fields journals are present. Since a single research field dominates most journals' citation flows, we measure the effective number of research fields. With journal  $i$ 's module-aggregated state node flow  $\boldsymbol{\pi}_i = \{\pi_{i\alpha}\}$ , we calculate its effective number of research fields  $r_i = 2^{H(\boldsymbol{\pi}_i)}$  with the entropy  $H(\boldsymbol{\pi}_i) = - \sum_{\alpha} \pi_{i\alpha} \log_2 \pi_{i\alpha}$ . With this metric, the most overlapping journals are *Bratislava Medical J.*, *Quality and Quantity*, and *Harvard Business Review* – tiny journals with only



**Fig. 5** A second-order Markov model results in overlapping multidisciplinary journals. The leftmost network is first-order, and the rightmost is second-order. Colors indicate the journal's weighted overlap in the second-order network



**Fig. 6** Influential multidisciplinary journals. The weighted overlap is the product of the journal's citation flow and its effective number of research fields. We limit the x-axis to six fields, but some small journals are in more than 30 research fields

around  $10^{-4}$  percent of the total citation flow. To highlight prominent, multidisciplinary journals and mesoscale changes in the citation flows, we weigh each journal's effective number of research fields with its total citation flow  $\pi_i = \sum_{\alpha} \pi_{i\alpha}$  for a weighted overlap

$$o_i = r_i \pi_i.$$

The journals with the highest weighted overlap are Nature, Science, and PNAS (Fig. 6).

The life sciences contain more of the multidisciplinary citation flow than the other research areas. By aggregating the weighted overlap  $o_i$  on the leaf modules  $m$ ,

$$o_m = \sum_{i \in m} o_i, \tag{6}$$

around 60 percent of the 1000 most overlapping leaf modules are in the life sciences, followed by the physical sciences with 18 percent.

### Collaboration hypergraph using different representations

Finally, we study how a hypergraph’s different first-order and multilayer network representations affect the detected communities. We use a collaboration hypergraph extracted from the 734 references in the review article “Networks beyond pairwise interactions: structure and dynamics” (Eriksson et al. 2021; Battiston et al. 2020). The referenced articles form hyperedges linking their authors. These hyperedges overlap in those authors who authored multiple papers, with the largest connected component containing 361 author nodes  $V$  in 220 hyperedges  $E$ . We illustrate a small, schematic hypergraph in Fig. 7a, where the white circles represent authors and the larger, orange circles represent papers.

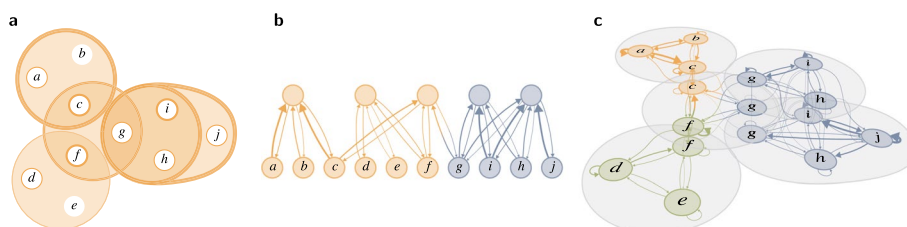
To model the flow of ideas among collaborators, we model a random walk on the hypergraph. Each hyperedge  $e$  has a weight  $\omega(e)$ , and each node  $u$  has a hyperedge-dependent weight  $\gamma_e(u)$ . We denote  $u$ ’s total incident hyperedge weight

$$d(u) = \sum_{e \in E(u)} \omega(e) \tag{7}$$

and hyperedge  $e$ ’s total node weight

$$\delta(e) = \sum_{u \in e} \gamma_e(u). \tag{8}$$

A random walker moves from node  $u$  to  $v$  with these weights in three stages by Chitra and Raphael (2019): First, choosing hyperedge  $e$  among node  $u$ ’s hyperedges  $E(u)$  with probability  $\frac{\omega(e)}{d(u)}$ . Then, choosing one of the hyperedge  $e$ ’s nodes  $v$  with probability  $\frac{\gamma_e(v)}{\delta(e)}$ . And finally, moving to  $v$ .



**Fig. 7** Schematic hypergraph with edge-dependent node weights **(a)** and flow-equivalent network representations. **b** A bipartite representation where hyperedges form hyperedge-nodes connecting all nodes in the hyperedge. **c** A multilayer representation where each hyperedge forms a layer containing the hyperedge’s nodes. The figure is adapted from Ref. Eriksson et al. (2021), licensed under CC BY 4.0

For the collaboration hypergraph, we use article hyperedge weights  $w(e) = \ln(c + 1) + 1$  where  $c$  is the number of citations for that article in December 2020 (Eriksson et al. 2021). To model the author’s unequal contributions to articles, we use hyperedge-dependent node weights (Chitra and Raphael 2019).

$$\gamma_e(i) = \begin{cases} 2 & \text{if } i \text{ is first or last author,} \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

We weigh alphabetically sorted authors uniformly because their contributions are hard to determine.

From this hypergraph, we generate bipartite and multilayer hypergraph representations with identical node visit rates using the method described in Ref. Eriksson et al. (2021) (Fig. 7b–c). We represent walks on hypergraphs as a bipartite network by representing the hyperedges with *hyperedge nodes*, and the three stages become a two-step walk between the nodes at the bottom and the hyperedge nodes at the top in 7b. First, a step from a node  $u$  to a hyperedge node  $e$ ,

$$P_{ue} = \frac{\omega(e)}{d(u)}, \tag{10}$$

and then a step from the hyperedge node to a node  $v$ ,

$$P_{ev} = \frac{\gamma_e(v)}{\delta(e)}. \tag{11}$$

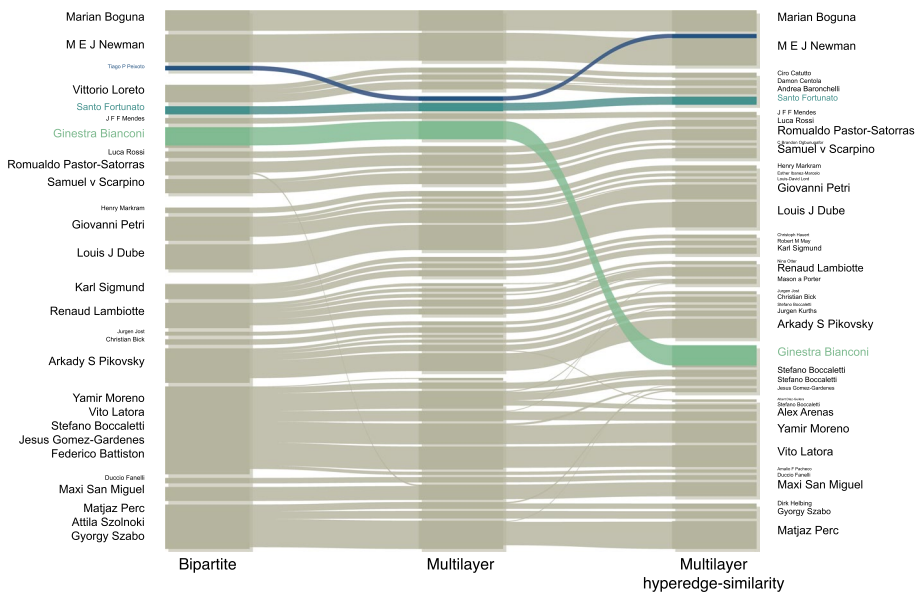
To represent the random walk on a multilayer network, we project the three-stage random-walk process down to a one-step process on state nodes in separate layers. Each hyperedge  $e$  with weight  $\omega(e)$  forms a layer  $\alpha$  with weight  $\omega(\alpha)$ . A state node  $u^\alpha$  represents  $u$  in each layer  $\alpha \in E(u)$  that contains the node (Fig. 7c). The transition rate between state node  $u^\alpha$  in layer  $\alpha$  and state node  $v^\beta$  in layer  $\beta$  is

$$P_{uv}^{\alpha\beta} = \frac{\omega(\beta)}{d(u)} \frac{\gamma_\beta(v)}{\delta(\beta)} \text{ for } \beta \in E(u, v). \tag{12}$$

With one state node per hyperedge layer that contains the node, the multilayer representation requires more nodes and links than the bipartite representation.

We also generate a multilayer network using a so-called hyperedge-similarity model that increases the probability of a random walk staying among similar hyperedges (Eriksson et al. 2021). This model reinforces community structure with modules formed by similar sets of collaborators. We let Infomap search for optimal multilevel solutions in the three network representations. As before, we create pseudo-state nodes in the bipartite network to match them with the multilayer networks’ state nodes.

The resulting partitions have effectively three or four levels. The top-level organization is most coarse-grained for the bipartite representation and most fine-grained for the hyperedge-similarity representation (Fig. 8). Only the multilayer representation assigns the submodule “Peixoto” together with the top module in which Bianconi is the highest-ranking author. It also assigns Fortunato to a different top module than the hyperedge-similarity partition. Finally, Bocalletti overlaps as the highest-ranking author in two submodules in the hyperedge-similarity partition in the same top module as Bianconi.



**Fig. 8** Network science collaboration hypergraph represented with two visit-rate-equivalent networks and a model that favors flow staying inside similar hyperedges. Modules are sorted to minimize overlap and show the names of their highest-ranking researchers. Here, we show the second-deepest level. Author groups that change top-level assignments in different networks are highlighted

### Conclusions

We have extended alluvial diagrams to higher-order networks with multilevel and overlapping communities and implemented an interactive web application available for anyone to use. In three case studies, we have used alluvial diagrams to show how the multilevel organization of science changes over time, how a second-order model compares to a first-order model, and how different hypergraph-flow equivalent networks influence the flow of ideas among network scientists.

We have focused on flow-based community detection using the map equation framework and the search algorithm Infomap. The generalized alluvial diagrams apply to any community-detection algorithm and are particularly relevant for simplifying and highlighting complex multilevel and overlapping modular descriptions of large higher-order networks.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s41109-023-00572-5>.

**Additional file 1.** Supplementary Information.

### Acknowledgements

A.H. was supported by the Swedish Foundation for Strategic Research, Grant No. SB16-0089. D.E. and M.R. were supported by the Swedish Research Council (2016-00796).

### Author contributions

AH and MR devised the study. AH performed the experiments. AH and DE implemented the interactive web application. AH and MR wrote the manuscript. All authors edited and accepted the manuscript in its final form.

### Funding

Open access funding provided by Umea University.

**Availability of data and materials**

The interactive web application is available at <https://www.mapequation.org/alluvial>, and its source code at <https://github.com/mapequation/alluvial-generator>. The multilevel significance clustering code is available at <https://github.com/mapequation/multilevel-significance-clustering>. All other relevant data and code are available at <https://github.com/mapequation/mapping-change-2>.

**Declarations****Competing interests**

The authors declare that they have no competing interests.

Received: 1 March 2023 Accepted: 3 July 2023

Published online: 11 July 2023

**References**

- Amelio A, Pizzuti C (2017) Correction for closeness: adjusting normalized mutual information measure for clustering comparison. *Comput Intel* 33:579
- Battiston F, Cencetti G, Iacopini I, Latora V, Lucas M, Patania A, Young J-G, Petri G (2020) Networks beyond pairwise interactions: structure and dynamics. *Phys Rep* 874:1
- Calatayud J, Andivia E, Escudero A, Melián CJ, Bernardo-Madrid R, Stoffel M, Aponte C, Medina NG, Molina-Venegas R, Arnan X et al (2020) Positive associations among rare species and their persistence in ecological assemblages. *Nat Ecol Evol* 4:40
- Calatayud J, Neuman M, Rojas A, Eriksson A, Rosvall M (2021) Regularities in species' niches reveal the world's climate regions. *eLife* 10
- Cazabet R, Rym B, Matthieu L (2017) Tracking bitcoin users activity using community detection on a network of weak signals. In *International conference on complex networks and their applications*. Springer, pp 166–177
- Chitra U, Raphael B (2019) Random walks on hypergraphs with edge-dependent vertex weights. In: *Proceedings of the 36th international conference on machine learning (PMLR)*, pp 1172–1181, iSSN: 2640-3498, <https://proceedings.mlr.press/v97/chitra19a.html>
- Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. *J Stat Mech Theory Exp* 2005:P09008
- De Domenico M, Lancichinetti A, Arenas A, Rosvall M (2015) Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys Rev X* 5:011027
- De Domenico M, Granell C, Porter MA, Arenas A (2016) The physics of spreading processes in multilayer networks. *Nat Phys* 12:901
- Edler D, Bohlin L et al (2017) Mapping higher-order network flows in memory and multilayer networks with Infomap. *Algorithms* 10:112
- Edler D, Guedes T, Zizka A, Rosvall M, Antonelli A (2017) Infomap bioregions: interactive mapping of biogeographical regions from species distributions. *Syst Biol* 66:197
- Edler D, Holmgren A, Rojas A, Rosvall M, Antonelli A (2022a) Infomap Bioregions 2: exploring the interplay between biogeography and evolution
- Edler D, Holmgren A, Rosvall M (2022b) The MapEquation software package. <https://mapequation.org>
- Eriksson A, Edler D, Rojas A, de Domenico M, Rosvall M (2021) How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs. *Commun Phys* 4:1
- Farage C, Edler D, Eklöf A, Rosvall M, Pilosof S (2021) Identifying flow modules in ecological networks using Infomap. *Methods Ecol Evol* 12:778
- Fortunato S (2010) Community detection in graphs. *Phys Rep* 486:75
- Holmgren A, Edler D, Rosvall M (2022a) The MapEquation alluvial diagram generator. <https://mapequation.org/alluvial>
- Holmgren A, Edler D, Rosvall M (2022b) Infomap online. <https://mapequation.org/infomap>
- Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. *J Compl Netw* 2:203
- Lambiotte R, Rosvall M (2012) Ranking and clustering of nodes in networks with smart teleportation. *Phys Rev E* 85:056107
- Lambiotte R, Rosvall M, Scholtes I (2019) From networks to optimal higher-order models of complex systems. *Nat Phys* 15:313
- Liu X, Derudder B, Csomós G, Taylor P (2013) Featured graphic. Mapping shifting hierarchical and regional tendencies in an urban network through alluvial diagrams. *Environ Plann A* 45:1005
- Neuman M (2022) PISA data clusters reveal student and school inequality that affects results. *Plos one* 17:e0267040
- Newman ME, Cantwell GT, Young J-G (2020) Improved mutual information measure for clustering, classification, and community detection. *Phys Rev E* 101:042304
- Pal R, Chopra H, Awasthi R, Bandhey H, Nagori A, Sethi T et al (2022) Predicting emerging themes in rapidly expanding COVID-19 literature with unsupervised word embeddings and machine learning: evidence-based study. *J Med Inter Res* 24:e34067
- Peixoto TP (2019) Bayesian stochastic blockmodeling. In: *Advances in network clustering and blockmodeling* pp. 289–332
- Peixoto TP (2014) Hierarchical block structures and high-resolution model selection in large networks. *Phys Rev X* 4:011047

- Persson C, Bohlin L, Edler D, Rosvall M (2016) Maps of sparse Markov chains efficiently reveal community structure in network flows with memory. arXiv preprint [arXiv:1606.08328](https://arxiv.org/abs/1606.08328)
- Petrun Sayers EL, Parker AM, Seelam R, Finucane ML (2021) How disasters drive media channel preferences: tracing news consumption before, during, and after Hurricane Harvey. *J Contingen Crisis Manag* 29:342
- Rojas A, Eriksson A, Neuman M, Edler D, Blocker C, Rosvall M (2022) A natural history of networks: higher-order network modeling for paleobiology research. *bioRxiv*
- Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci* 105:1118
- Rosvall M, Bergstrom CT (2010) Mapping change in large networks. *PloS One* 5:e8694
- Rosvall M, Bergstrom CT (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS One* 6:e18209
- Rosvall M, Esquivel AV, Lancichinetti A, West JD, Lambiotte R (2014) Memory in network flows and its effects on spreading dynamics and community detection. *Nat Commun* 5:1
- Ruan W, Hou H, Hu Z (2017) Detecting dynamics of hot topics with alluvial diagrams: a timeline visualization. *J Data Inf Sci* 2:37
- Schaub MT, Delvenne J-C, Rosvall M, Lambiotte R (2017) The many facets of community detection in complex networks. *Appl Netw Sci* 2:1
- Traag VA, Waltman L, Van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9:1
- Wang Q, Waltman L (2016) Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *J Inf* 10:347
- Xu J, Wickramaratne TL, Chawla NV (2016) Representing higher-order dependencies in networks. *Sci Adv* 2:e1600028
- Zakai A (2011) Emscripten: an LLVM-to-JavaScript compiler. In: *Proceedings of the ACM international conference companion on object oriented programming systems languages and applications companion*, pp 301–312

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---