Statistical modeling of higher-order interactions: recent developments and future challenges

Catherine Matias

Centre National de la Recherche Scientifique, Paris, France
Sorbonne Université, Paris, France
Université de Paris Cité, Paris, France

Data Science, Statistics, and Visualisation (DSSV) - June 2026

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

Higher-order interactions I

Motivations

- Networks or graphs focus on pairwise interactions
- These type of pairwise interactions can already be quite elaborate: undirected/directed, binary/weighted, simple/multiple, static/dynamic, multiplex or multi-layers, ...
- Nonetheless pairwise interactions are not sufficient to describe the nature of complex interactions :
  - e.g. the presence of a 3rd species may modify the interaction of 2 other species ;
  - e.g. a collaboration between 3 authors is stg different from 3 pairwise collaborations between these same authors ;
- Collective interactions or group interactions are richer than just pairwise interactions

$\hookrightarrow$ These are called higher-order interactions (HOI).

Higher-order interactions II

Where do we find HOI?

- Social networks: triadic and larger groups (as early as Simmel, 1950)
- Scientific co-authorship,
- Interactions between more than two species in ecological systems,
- HOI between neurons in brain networks,
- Metabolites in chemical reactions,
- etc

These interactions **CAN NOT** be represented by a graph.

Higher-order interactions III

A nice review (2020) to start on the topic:

Contents lists available at ScienceDirect

Physics Reports

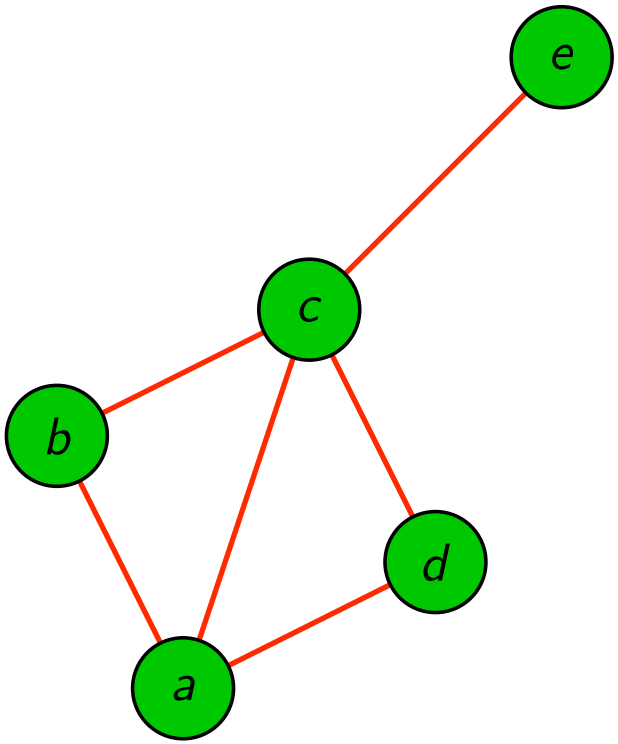
journal homepage: www.elsevier.com/locate/physrep

Networks beyond pairwise interactions: Structure and dynamics

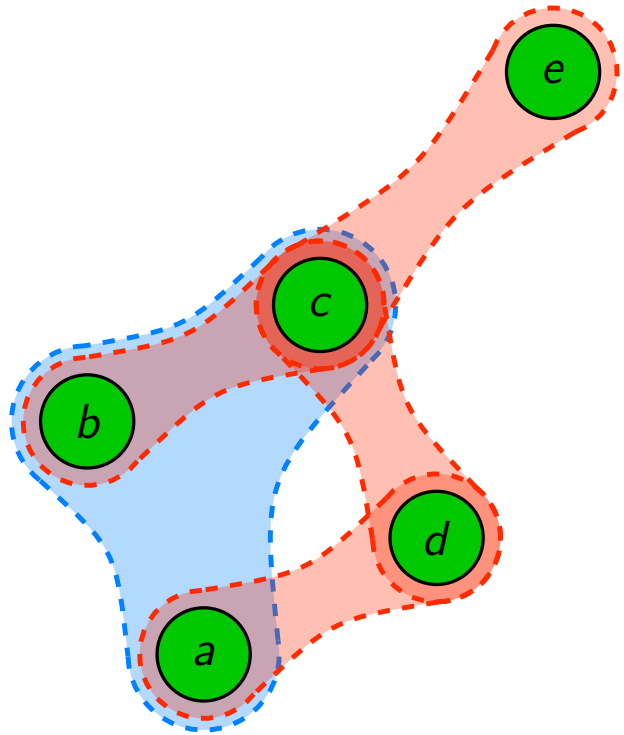
Federico Battiston [a,*], Giulia Cencetti [b], Iacopo Iacopini [c,d], Vito Latora [c,e,f,g], Maxime Lucas [h,i,j], Alice Patania [k], Jean-Gabriel Young [l], Giovanni Petri [m,n]

Pairwise vs HOI

HOI are defined as sets of interacting entities.
e.g. $V = \{a, b, c, d, e\}; \mathcal{I} = \{\{a, b, c\}, \{a, d\}, \{c, d\}, \{c, e\}\}$



(a) Pairwise interactions



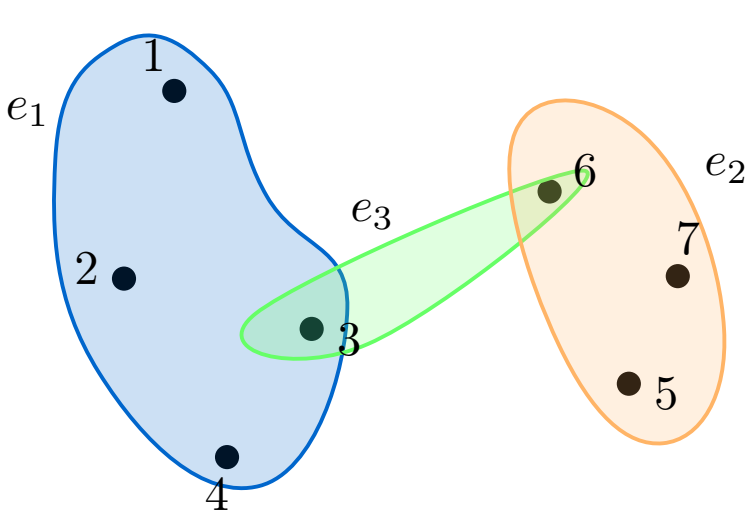
(b) A HOI in blue

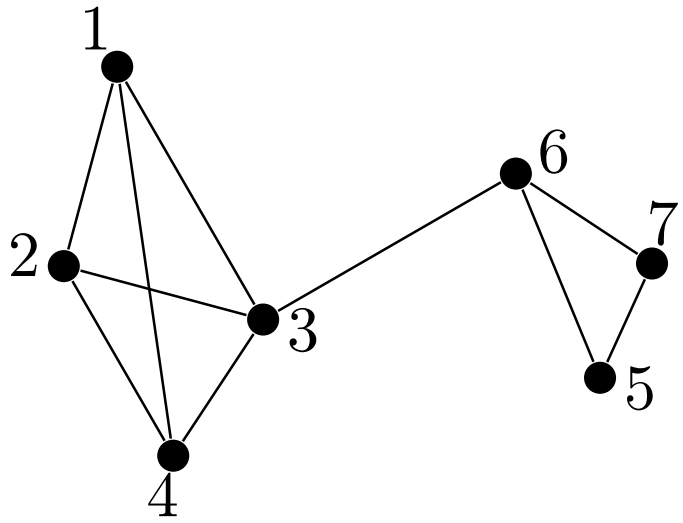
The order of an interaction is the number of entities that interact.

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

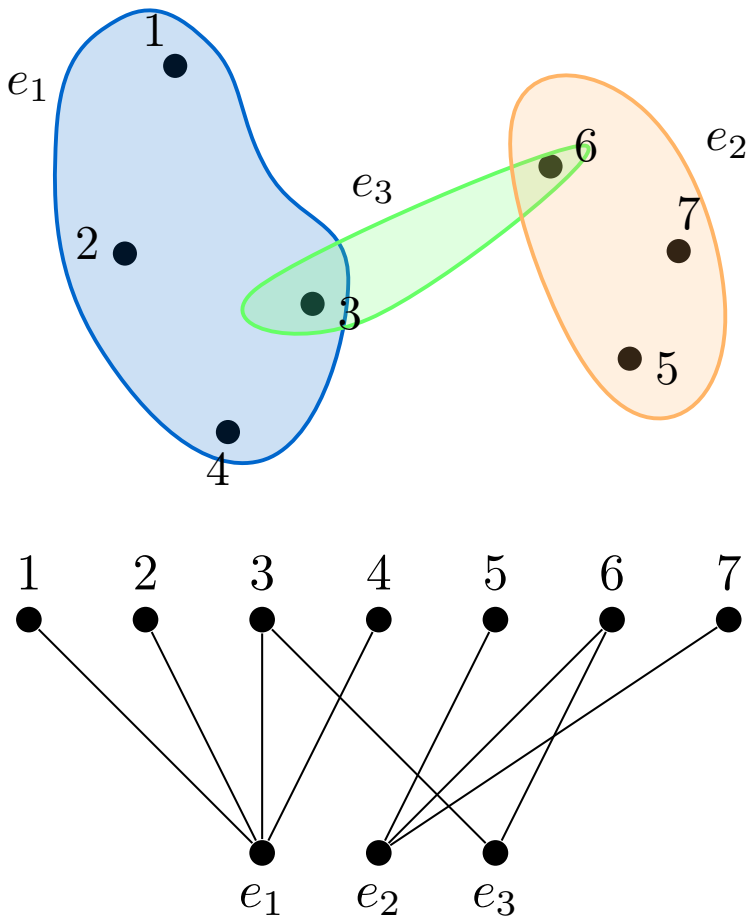
Naïve Graph representation: clique expansion graph





- Each interaction is transformed into a clique = all edges between pairs are present ;
- HOIs actually disappeared !
- Too simplistic: For e.g, in co-authorship 1 paper with 3 authors $\neq$ 3 different papers written by pairs of those authors.

Bipartite graph representation (two-modes network or star-expansion graph)



- ☺ No loss of information for hypergraphs with multiples hyperedges and self-loops;
- ☹ Statistical Modeling drawbacks:
  - Number of hyperedges is not random;
  - Configuration models will assume sizes of hyperedges are fixed;
  - Bipartite SBM impose hyperedges clusters, reducing the flexibility of the model;

HOIs representations

Relying on graphs

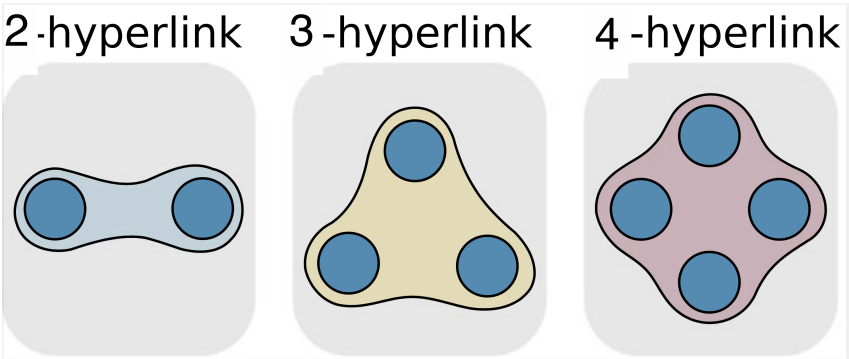
- There are other graph-representations of HOIs
- But none of these completely capture the information (at least from a statistical modeling point of view)

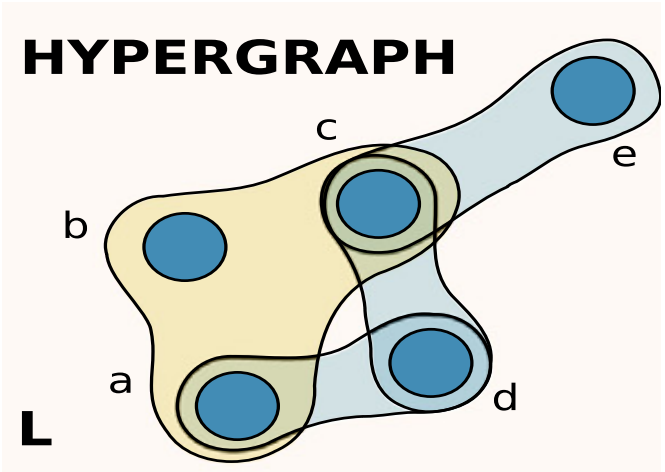
Relying on other objects

There are 2 mathematical objects to represent HOIs :

- Simplicial complexes (constrained)
- Hypergraphs

Hypergraphs I





Hypergraph definition

A hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is defined as a set of nodes $\mathcal{V} \neq \emptyset$ and a set of hyperedges $\mathcal{E}$. **Each hyperedge is a non-empty collection of k nodes** taking part in an interaction.

Hypergraphs II

Hypergraphs characteristics

- Hypergraphs naturally include graphs, by simply considering hyperedges of size $k = 2$;
- A hypergraph may contain a size-3 hyperedge $\{a, b, c\}$ without any requirement on the existence of the size-2 hyperedges $\{a, b\}$, $\{a, c\}$, and $\{b, c\}$.

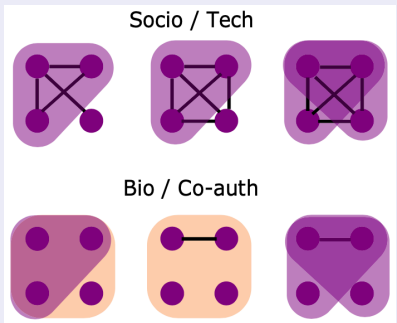
Statistical measures on hypergraphs

Graph statistics generalized to hypergraphs

- For any size $k \geq 2$, **size-k density** is = nb of size-k hyperedges $/\binom{n}{k}$
- Node degree (rowSum of H); hyperedge size/degree (colSum of H);
- **Centrality measures**
  - relies on the notion of paths;
  - a path is a sequence $(e_1, e_2, \ldots, e_t)$ of hyperedges such that 2 successive hyperedges have at least one common node ($e_i \cap e_{i+1} \neq \emptyset$);
  - concept of k-path: any 2 successive hyperedges share at least $k \geq 1$ nodes;

Other graph statistics

- Clustering and transitivity: no natural generalization! (based on triangles);
- Motifs abundance: ok but huge combinatorial complexity;



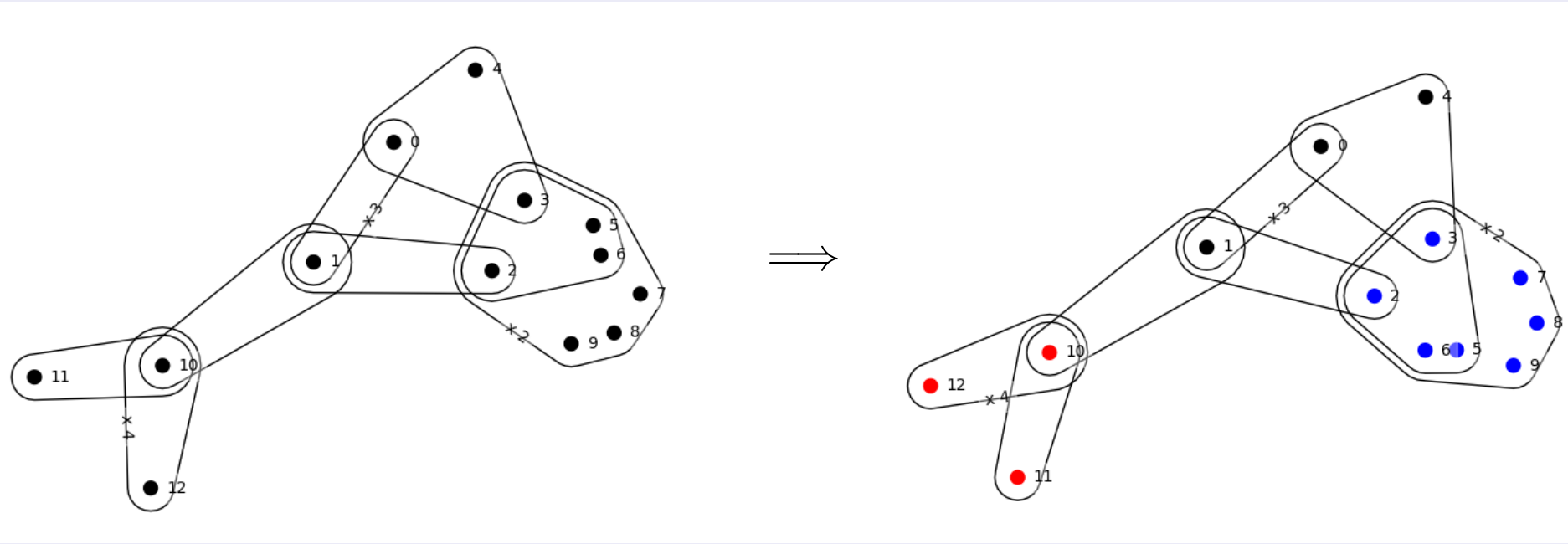
(from Lotito et al., Comm. Phys., 2022)

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

Clustering the nodes of a hypergraph I

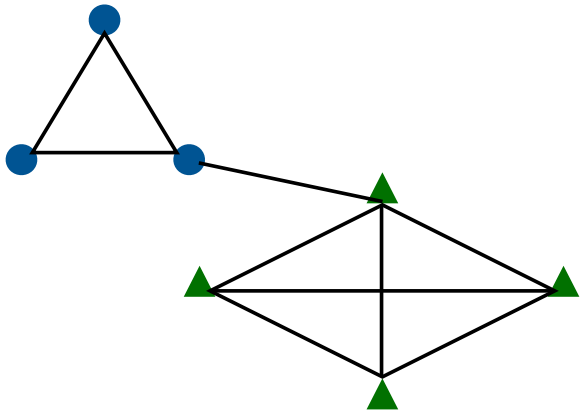
Objective

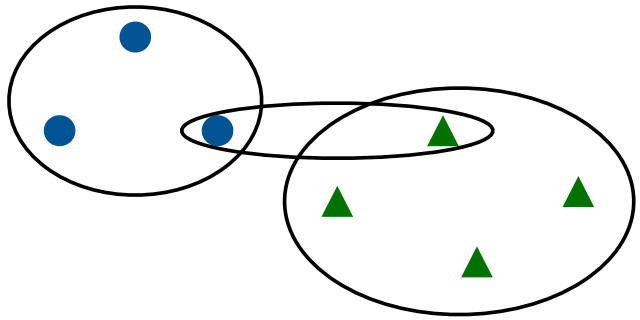


Questions: What are we looking for? Can we define *communities*?

Clustering the nodes of a hypergraph II

In a graph, a community is a set of nodes with large within-group connections and small between-groups connections.





In a hypergraph, should we weight the hyperedges wrt their sizes? If yes, how?

$\rightarrow$ This question is currently not uniquely answered.

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

Modularity-based approaches I

Graph case: Newman-Girvan modularity

For a clustering $\mathcal{C} = (C_1, \ldots, C_K)$ of the nodes of a graph $G = (V, E)$, we let

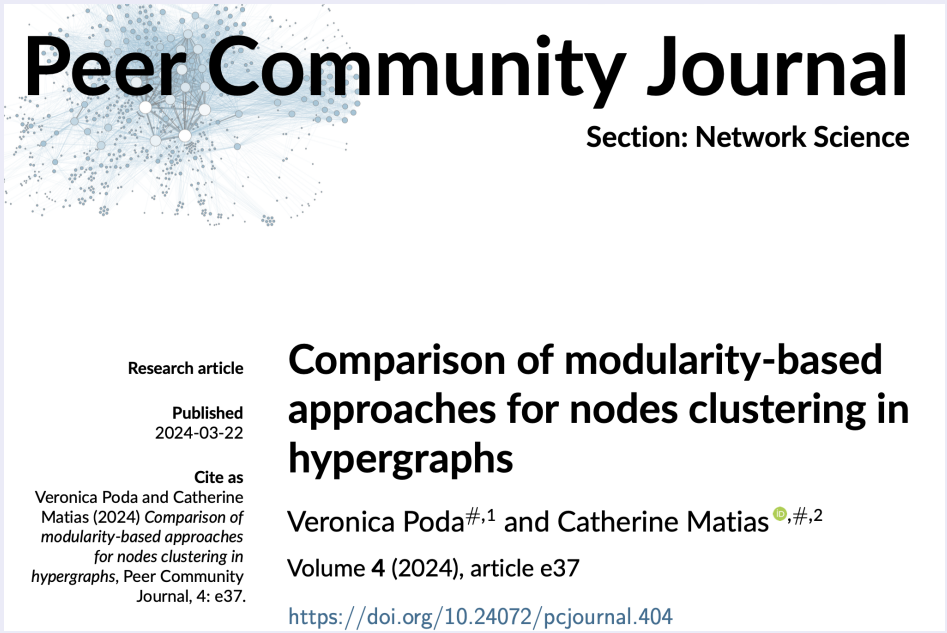
$$Q(G, \mathcal{C}) = \frac{1}{2|E|} \sum_{k=1}^{K} \sum_{u,v \in C_k} \left(A_{uv} - \frac{d_u d_v}{2|E|} \right).$$

- Exact optimization is impossible; rely on Louvain algorithm (heuristic);
- Compares the nb of within-cluster edges with expected value under a null model accounting for nodes degrees;
- Louvain heuristic automatically selects a number of clusters;

Modularity-based approaches II

Hypergraph case: state of art

Many different generalizations exist for hypergraphs, based on different notions of communities;



Peer Community Journal

Section: Network Science

Research article

Published
2024-03-22

Cite as
Veronica Poda and Catherine Matias (2024) *Comparison of modularity-based approaches for nodes clustering in hypergraphs*, Peer Community Journal, 4: e37.

Comparison of modularity-based approaches for nodes clustering in hypergraphs

Veronica Poda[#,1] and Catherine Matias[#,2]

Volume 4 (2024), article e37

https://doi.org/10.24072/pcjournal.404

- We have compared methods and found that the best is Chodrow et al. (2021);
- It focuses on All-or-Nothing (AON) modularity, in which a hyperedge contributes to increase modularity only when all its nodes are in the same cluster.
- Currently the main issue is the **lack of benchmark datasets**.

Modularity-based approaches III

Intrinsic drawbacks

- Maximization is difficult;
- look for *communities* and not general clusters (e.g. hubs or peripherical nodes);
- tends to favour groups of the same size;
- Possible biases due to an underlying assumption of multisets-hypergraphs;

Open problems

- What kind of hypergraphs-communities do we want to identify with modularities?
- Design good benchmark datasets for comparing the methods;

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

Spectral hypergraph partitioning I

Graphs case - intuition

- In presence of clusters, expected adjacency matrix has low rank;
- The eigendecomposition of (the adjacency matrix or) a Laplacian matrix (normalised version of the adjacency matrix) should reveal the clusters;

Hypergraphs case Ghoshdastidar and Dukkipati (2017)

With H incidence matrix ($n \times |\mathcal{E}|$), $D = \text{diag}(d_i)$ node degrees ($n \times n$) and $\Delta = \text{diag}(|e_m|)$ hyperedges sizes ($|\mathcal{E}| \times |\mathcal{E}|$)

- Hypergraph Laplacian $L = I - D^{-1/2} H \Delta^{-1} H^\intercal D^{-1/2}$
- This is the Laplacian for a weighted clique expansion graph;
- Compute leading eigenvectors and run k-means on normalised rows;
- Weakly consistent under planted partition model under condition that the weighted clique graph identifies the clusters;

Spectral hypergraph partitioning II

Open problems

- How can we select the number of groups? (Is there an eigengap?)
- Can we design a method not based on clique expansion graph (to fully capture HOI interaction)?

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

Hypergraphs Stochastic Blockmodels

Received: 16 August 2023 | Revised: 13 May 2024 | Accepted: 1 September 2024

DOI: 10.1111/sjos.12754

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Model-based clustering in simple hypergraphs through a stochastic blockmodel

Luca Brusa[1] | **Catherine Matias**[2,3,4]

[1]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy

[2]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France

[3]Université de Paris Cité, Paris, France

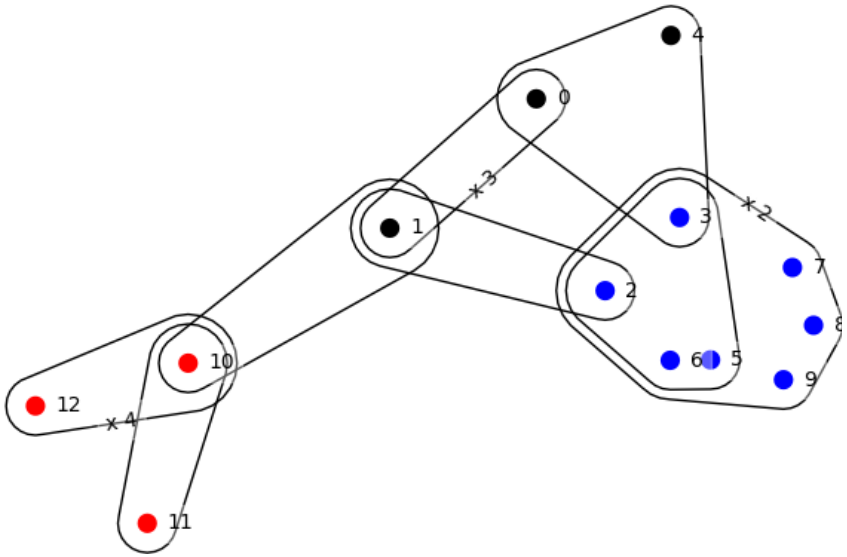
[4]Centre National de la Recherche Scientifique, Paris, France

Correspondence
Luca Brusa, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli

Abstract

We propose a model to address the overlooked problem of node clustering in simple hypergraphs. Simple hypergraphs are suitable when a node may not appear multiple times in the same hyperedge, such as in co-authorship datasets. Our model generalizes the stochastic blockmodel for graphs and assumes the existence of latent node groups and hyperedges are conditionally independent given these groups. We first establish the generic identifiability of the model

HyperSBM formulation



- $\mathcal{H} = (\mathcal{V}, \mathcal{E})$,
- For each $2 \leq m \leq M$, let $\mathcal{V}^{(m)} = \big\{\{i_1, \ldots, i_m\} : i_1, \ldots, i_m \in \mathcal{V} \text{ and } i_1 \neq \ldots \neq i_m\big\}$, set of unordered node tuples of size m;

- **Observations:** At each $\{i_1, \ldots, i_m\} \in \mathcal{V}^{(m)}$, we observe indicator variable $Y_{i_1,\ldots,i_m} = 1\{\{i_1, \ldots, i_m\} \in \mathcal{E}\}$;
- **Latent clusters:** $Z_1, \ldots, Z_n$ iid in $\{1, \ldots, Q\}$ with $\pi_q = \mathbb{P}(Z_i = q)$;
- **Conditional independence assumption**: $\{Y_{i_1,\ldots,i_m}\}_{\{i_1,\ldots,i_m\}\in\mathcal{V}^{(m)}} | \{Z_1, \ldots, Z_n\}$ are independent with $Y_{i_1,\ldots,i_m} | \{Z_{i_1} = q_1, \ldots, Z_{i_m} = q_m\} \sim \text{Bern}(B^{(m)}_{q_1,\ldots,q_m})$.

Hypergraphs Stochastic Blockmodels

HyperSBM (joint work with Luca Brusa)

- We establish **parameter identifiability** results;
- We propose a **variational expectation-maximisation (VEM)** algorithm to infer clusters and parameters;
- We propose an **ICL criterion** to select the number of clusters;
- All these tools are implemented (in C++) in a efficient `R` **package** called `HyperSBM` (`https://github.com/LB1304/HyperSBM`).

Future: FastHyperSBM

We are currently working on a faster version of the inference procedure, relying on subsampling ideas.

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

Hypergraphs Embedding (for hypergraphs clustering)

M. Cervellini et al. *Algorithms for Molecular Biology(2026)21:9*
https://doi.org/10.1186/s13015-026-00298-w

Algorithms for
Molecular Biology

RESEARCH **Open Access**

Comparing the ability of embedding methods on metabolic hypergraphs for capturing taxonomy-based features

Mattia Cervellini[1], Blerina Sinaimeri[1], Catherine Matias[2,3,4] and Alessio Martino[1*]

*Correspondence:
amartino@luiss.it
[1]Department of AI, Data and Decision Sciences, LUISS University, Viale Romania 32, 00197 Rome, Italy
[2]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France
[3]Université de Paris Cité, Paris, France
[4]Centre National de la Recherche Scientifique, Paris, France

Abstract

Background Metabolic networks are complex systems that describe the biochemical reactions within an organism through pairwise interactions between chemical compounds. While this representation is widely used to study biological function, it fails to capture the full structure of metabolic reactions, many of which involve more than two compounds. Hypergraphs offer a more natural representation, where nodes represent metabolites and hyperedges represent reactions involving multiple participants. Clustering such metabolic hypergraphs can reveal systematic differences among evolutionarily distinct organisms, providing insight into ecological constraints and evolutionary pressures.

Methods In this study, we investigate how different graphs and hypergraphs embedding methods influence their unsupervised clustering, with the goal of capturing taxonomy-based classes. We apply 14 distinct embedding strategies to a large-scale dataset of 8467 metabolic hypergraphs. Each embedding was followed by hierarchical clustering using a fixed linkage method. To assess performance, we compared the resulting clusters against known taxonomic groupings.

Results Our findings show that the choice of hypergraph embedding has a significant effect on clustering outcomes. Among the tested methods, Bag of

Embedding metabolic pathways I

Objective

- Compare different methods for graphs/hypergraphs embedding;
- Followed by hierarchical clustering;
- Applied on metabolic pathways;
- With the aim of capturing their taxonomy (unsupervised clustering of the whole network). 4 hierarchical levels with increasing granularity:
  - C1 cellular organization. 2 classes: Eukaryotes vs Prokaryotes
  - C2 domain/kingdom. 6 classes: Animals, Archaea, Bacteria, Fungi, Plants and Protists;
  - C3 class level. 18 classes;
  - C4 genus level. 70 classes.

Embedding metabolic pathways II

Methods

- 14 embedding strategies applied on 8467 metabolic pathways;

Table 1 Summary of the 14 embedding strategies stratified by their respective family, type of input representation and size of the resulting embedding space

Family	Method	Representation	Size	References
BoW-inspired	BoH w/Jaccard distance	hypergraph	3236	[20,56]
	BoH w/Manhattan distance	hypergraph	3236	[20,56]
	BoN w/Betweenness Centrality	clique expansion	2834	Proposed
	BoN w/Degree Centrality	clique expansion	2834	Proposed
	BoN w/Node Degree	hypergraph	2834	Proposed
	BoN w/PageRank	clique expansion	2834	Proposed
Kernel-based	Histogram Kernel	hypergraph	–	[28]
	Jaccard Kernel	hypergraph	–	[28]
	Edit Kernel	hypergraph	–	[28]
	Stratified Edit Kernel	hypergraph	–	[28]
Neural Networks-based	Graph2Vec	clique expansion	[215–281]†	[43]
	HGAE (Complete)	hypergraph	400	Proposed
	HGAE (Hyperedges-Only)	hypergraph	400	Proposed
	HGAE (One-Hot+Hyperedges)	hypergraph	400	Proposed

† Depending on the problem: 215 (C1 and C2), 281 (C3), 272 (C4)

- Evaluation criteria: homogeneity $\in [0, 1]$, maximized when each cluster is composed entirely of entities belonging to a same specific taxonomic class.

Embedding metabolic pathways III

Results

- In each family, with have identified a "best performer";
- All of these use hypergraph information (instead of clique-graph): suggests a loss of information for embeddings based on cliques only;
- Overall performances of the methods quite high across all levels of taxonomy classes C1 to C4;
- C3 and C2 are the most challenging classes: at C2, Bacteria tend to split up in different (albeit pure) clusters while Eukaryotes remain in a single, separate, less pure cluster

Outline

1. Why higher-order interactions?
2. Capturing higher-order interactions
3. Clustering entities in hypergraphs
   - Modularity
   - Spectral clustering
   - Stochastic Blockmodels
4. Hypergraphs clustering via embedding
5. Overall Conclusions

Conclusions

- Higher-order interactions is a new trend;
- There are already some available tools that you can test on your datasets (see for e.g. hypergraphx, `https://github.com/HGX-Team/hypergraphx`)
- A lot of challenging questions (computational, modeling based, methodological, etc) remain open with these tools
- Those methods offer an avenue for tackling new questions on complex systems.

Any questions ?

References I

Battiston, F., G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri (2020). Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports 874*, 1–92.

Brusa, L. and C. Matias (2024). Model-based clustering in simple hypergraphs through a stochastic blockmodel. *Scand J Statist 51(4)*, 1661–1684.

Chodrow, P. S., N. Veldt, and A. R. Benson (2021). Generative hypergraph clustering: From blockmodels to modularity. *Science Advances 7*(28), eabh1303.

Ghoshdastidar, D. and A. Dukkipati (2017). Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics 45*(1), 289 – 315.

Poda, V. and C. Matias (2024). Comparison of modularity-based approaches for nodes clustering in hypergraphs. *Peer Community Journal 4*, article e37.