A guided tour on nodes clustering in hypergraphs

Catherine Matias

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

Rencontres Mathématiques de Rouen - 2024

Outline

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

This is a nice recent review (2020):

Contents lists available at ScienceDirect

Physics Reports

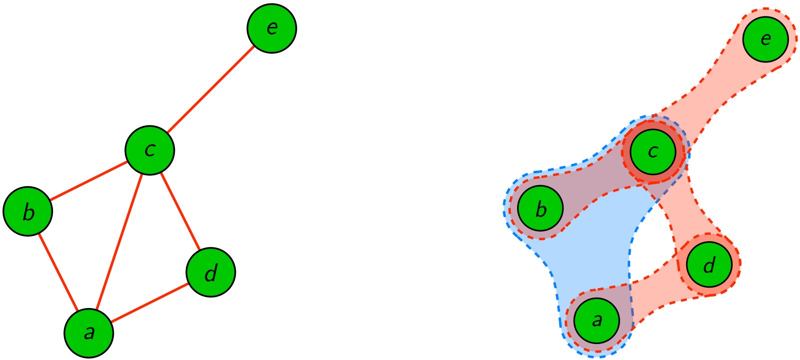
journal homepage: www.elsevier.com/locate/physrep

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. $V = \{a, b, c, d, e\}; \mathcal{I} = \{\{a, b, c\}, \{a, d\}, \{c, d\}, \{c, e\}\}$



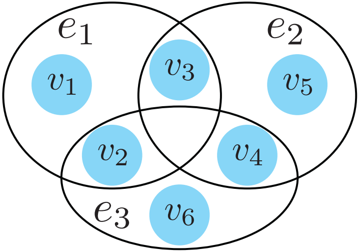
(a) Pairwise interactions

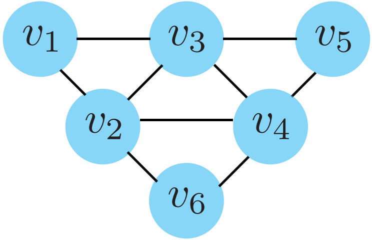
(b) A HOI in blue

The order of an interaction is the number of entities that interact - 1.

Outline

Naïve Graph representation: clique expansion graph

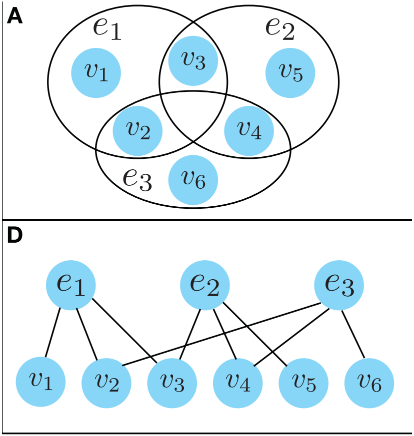




Picture from Schaub *et al.* 2021

- 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)



Picture from Schaub *et al.* 2021

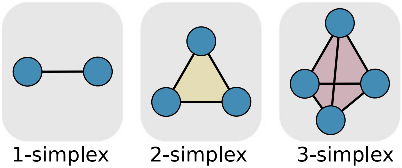
- No loss of information for hypergraphs with multiples hyperedges and self-loops;
- But "higher-order" now translates into node degrees in one part;
- 2 two parts don't play symmetric roles: statistical models on bipartite graphs are not appropriate here

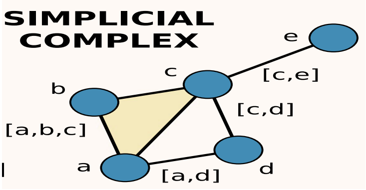
Other graph representations

- There are other graph-representations of HOIs
- But none of it may completely capture these

$\hookrightarrow$ There are 2 mathematical objects to represent HOIs : Simplicial complexes and hypergraphs.

Simplicial complexes vs hypergraphs I





Simplex and Simplicial complexes

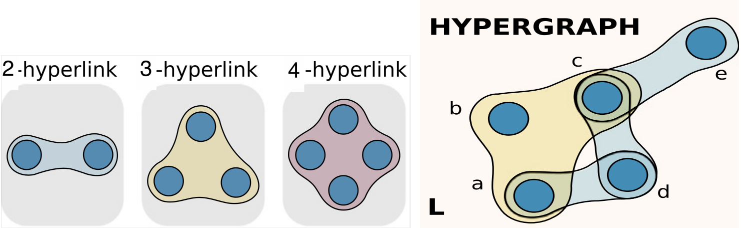
- a *k*-simplex $\sigma = \{p_0, p_1, \dots, p_k\}$ is a set of $k+1$ points (in a topological space);
- a subface of a simplex σ is any subset of points in σ;
- a simplicial complex = a collection $K = \{\sigma_1, \dots, \sigma_n\}$ of simplexes (of any size);
- a valid simplicial complex is such that $\forall \sigma \in K$, every subface of σ also belongs to K

Simplicial complexes vs hypergraphs II

(Dis)-Advantages

- ☺ strong mathematical object, very useful in many areas; e.g: statistical topological data analysis, to approximate varieties of irregular algebraic structures;
- ☹ Valid simplicial complexes impose all sub-interactions of an interaction should exist;
- ☹ points come with positions in (topological) space

Simplicial complexes vs hypergraphs III



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 distinct nodes** taking part in an interaction.

Simplicial complexes vs hypergraphs IV

Hypergraphs characteristics

- Hypergraphs naturally include the entity of 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 $\{b, c\}$.

Simplicial complexes vs hypergraphs V

Simple hypergraphs and variants

- In **simple hypergraphs**, an hyperedge appears only once and contains distinct nodes;
- May consider **nodes to appear with multiplicities** in a same hyperedge
  - Example: chemical reactions, multiplicity = stoichiometric coefficient;
  - I call these **multisets** hypergraphs;
  - generalize (in some sense) the notion of loops in graphs
- May consider **multiple** hyperedges, when a same hyperedge may appear several times (= integer-valued weight on a hyperedge);
- May introduce a **direction**: a hyperedge e is divided into 2 ordered subsets (e_1, e_2) of interacting nodes ($e = e_1 \cup e_2$);
  $\hookrightarrow$ not much used though;

NB : in the following, focus on hypergraphs.

Matrix encoding of hypergraphs

- **Incidence matrix** H, size $n \times m$ where n nb of nodes, m nb of interactions; with entry $H_{i,e} = 1$ when node i belongs to hyperedge e.
  $\hookrightarrow$ contains all the information;
  $\hookrightarrow$ enables definition of **node degrees** d_i (=rowSums of H) and **hyperedge sizes/degrees** δ_e (=colSums of H)
- **Adjacency matrix** $A = HH^\intercal - D$ has size $n \times n$, where $D = diag(d_1, \dots, d_n)$
  $\hookrightarrow$ This is the adjacency matrix of the clique expansion graph;
  $\hookrightarrow$ contains only partial information;

Outline

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; hyperedge size/degree;
- **Centrality measures**
  - relies on the notion of paths;
  - a path is a sequence $(e_1, e_2, \dots, e_t)$ 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;

Graph statistics with no natural generalization

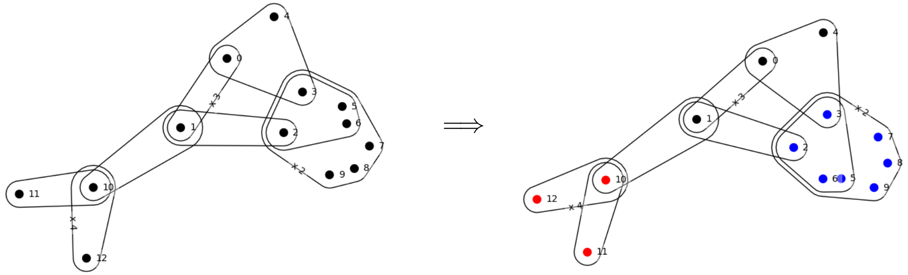
- clustering and transitivity (based on triangles);
- motifs (combinatorial complexity)

Outline

Outline

Clustering the nodes of a hypergraph I

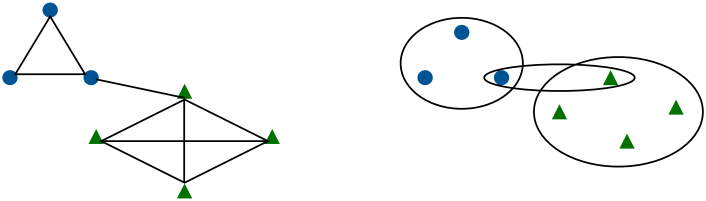
Objective



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?

Methods for node clustering I

3 types of methods

1. **Modularity-based** approaches
   - Different hypergraph modularity definitions: what kind of communities do they favour?
   - Note that for computational reasons, these focus on ***multisets-hypergraphs*** where nodes may be repeated in a same hyperedge;
   - This is not always appropriate, e.g. co-authorship dataset;
   - In the context of graphs, absence of self-loops and multiple edges are known to generate pbms in modularity approaches
2. **Spectral clustering** has been generalized to hypergraphs but
   - it tends to favour groups of the same size;
3. **Stochastic Blockmodels**

Methods for node clustering II

Challenges

- Look for general clusters and not only *communities*
- Methods should come with a procedure to select the number of groups K

Modularity-based approaches I

Newman-Girvan modularity for graphs

For a clustering $\mathcal{C} = (C_1, \dots, C_K)$ 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;
- automatically selects a number of clusters

Modularity-based approaches II

Hypergraph case

- Many different generalizations exist for hypergraphs, based on different notions of communities
- We have compared methods in Poda & Matias (2024) 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.

Spectral hypergraph partitioning I

Graphs case - intuition

- When there are communities, adjacency matrix is structured as almost block diagonal;
- A Laplacian of the graph is a normalised version of the adjacency matrix
- The eigendecomposition of the adjacency matrix or of a Laplacian should reveal the communities
- This is linked to embedding: the nodes are sent to a new vector space (corresponding to the principal eigenvector), where proximity is correlated with connection in the graph

Spectral hypergraph partitioning II

Hypergraphs case

See for e.g. Ghoshdastidar & Dukkipati (2014,2017)

- Hypergraph Laplacian $L = I - D^{-1/2} H \Delta^{-1} H^\intercal D^{-1/2}$
- Compute leading eigenvectors and run k-means on rows
- No proposal to select for the number of groups (is there an eigengap?)

Why should you prefer stochastic blockmodels?

Apart from the fact that statistics are always the best option ;)

Critics

- Both methods look for *communities* and not general clusters (e.g. hubs or peripherical nodes);
- Both tend to favour groups of the same size;
- For computational reasons, modularity approaches have focused on ***multisets-hypergraphs*** (where nodes may be repeated in a same hyperedge);
  $\hookrightarrow$ assumption not always appropriate, e.g. co-authorship dataset;
  $\hookrightarrow$ with which impact?
- Modularity maximization is difficult; only local maximum is found;
- None of these methods comes with a statistical criterion to select the number of groups.

Hypergraphs Stochastic Blockmodels

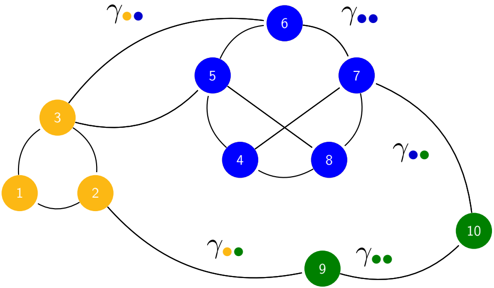
Our SBM proposal (joint work with Luca Brusa)

- We focus on **simple** graphs (instead of multisets-hypergraphs);
- We define **a stochastic blockmodel** to cluster the nodes of a hypergraph
  - We establish **parameter identifiability** results;
  - We propose a **variational expectation-maximisation** 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`).

Outline

1. The need for higher-order interactions
2. Capturing higher-order interactions
3. Statistics on hypergraphs
4. Clustering entities in hypergraphs
   - Different approaches
   - Stochastic blockmodel for hypergraphs
5. Experiments
6. Conclusions

Stochastic block model (binary graphs)



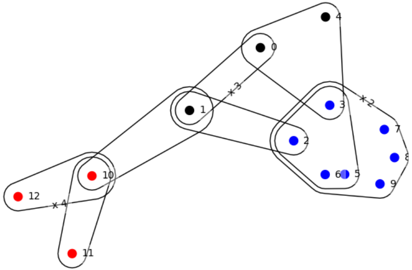
$n = 10, Z_{5\bullet} = 1$

$A_{12} = 1, A_{15} = 0$

Binary case (parametric model with $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$)

- K groups (=colors ●●●).
- $\{Z_i\}_{1 \le i \le n}$ i.i.d. vectors $Z_i = (Z_{i1}, \dots, Z_{iK}) \sim \mathcal{M}(1, \boldsymbol{\pi})$, with $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ groups proportions. Z_i not observed (latent).
- Observations: presence/absence of an edge $\{A_{ij}\}_{1 \le i < j \le n}$,
- Conditional on $\{Z_i\}$'s, the r.v. A_{ij} are independent $\mathcal{B}(\gamma_{Z_i Z_j})$.

HyperSBM formulation



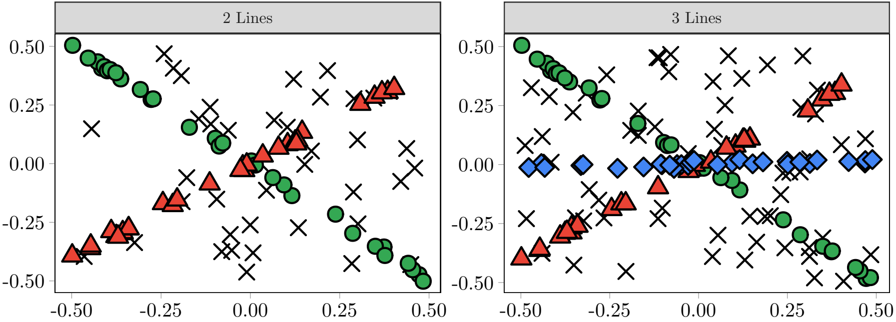
- $\mathcal{H} = (\mathcal{V}, \mathcal{E})$,
- For each $2 \leq m \leq M$, let $\mathcal{V}^{(m)} = \{\{i_1, \dots, i_m\} : i_1, \dots, i_m \in \mathcal{V} \text{ and } i_1 \neq \dots \neq i_m\}$, set of unordered node tuples of size m;

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

Outline

Line clustering through hypergraphs I

2 experiments: 2 lines (3 groups) and 3 lines (4 groups)



Line clustering through hypergraphs II

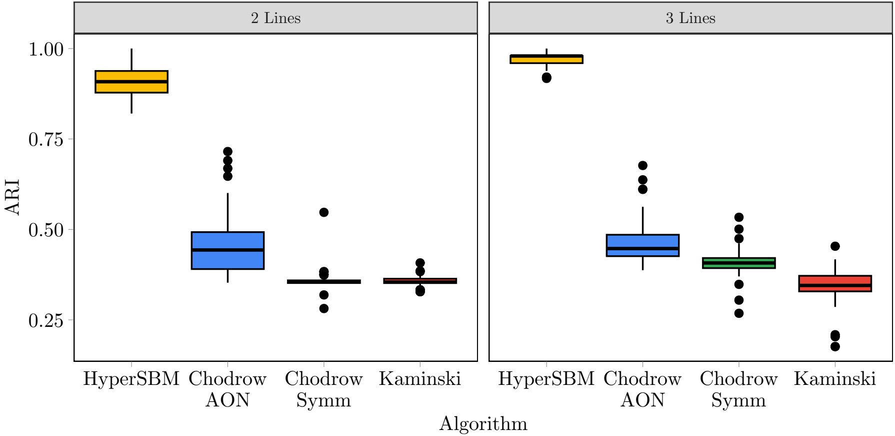
Hypergraph construction

- Select 3 points at random and fit a line
- If residual distance is less than a threshold, draw a hyperedge between those 3 points
- Globally set signal:noise hyperedge ratio = 2
- Repeat to obtain 100 3-uniform hypergraphs

Data characteristics

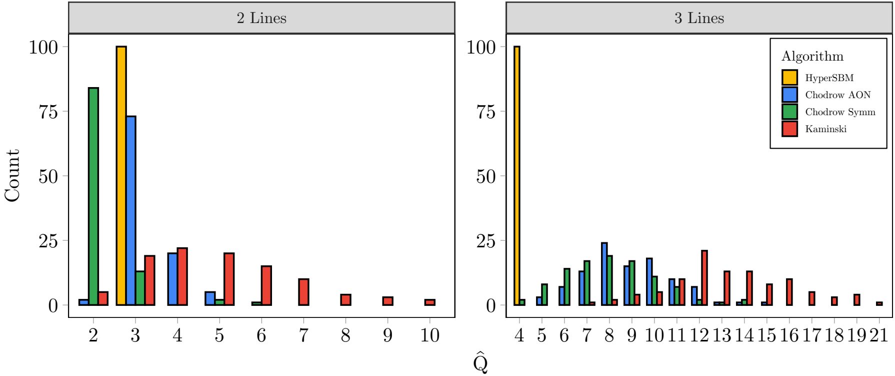
	Pts/line	Noisy pts	Total nb pts	mean nb of hyperedges
2 lines	30	40	100	1070.84
3 lines	30	60	150	587.7

Comparison with modularity based methods I



Adjusted Rand Index

Comparison with modularity based methods II



Estimated number of groups

Co-authorship dataset I

Dataset description

- Available at http://vlado.fmf.uni-lj.si/pub/networks/data/2mode/Sandi/Sandi.htm
- Bipartite author/article graph transformed into hypergraph of authors where hyperedges link the authors of a same paper;
- We choose $M = 4$ and consider the induced largest connected component: 79 authors and 76 hyperedges (68.5% of which have size 2, while 29% have size 3 and 2.5% have size 4).

Co-authorship dataset II

Analysis through HyperSBM

- ICL selects $Q = 2$ groups, the first has only 8 authors;
- Our first group is made of authors (among) the most collaborative ones, which are also (among) the most prolific ones.
- None of these groups is a community (the first co-publishes with all, the second has low intra-group connectivity).

Comparison with hypergraph spectral clustering (HSC)

- HSC with $Q = 2$ gives a group of size 24 and one of size 55
- These groups are neither characterized by the number of co-authors nor their degrees in the bipartite graph
- Very different from our results because: spectral clustering tends to: i) extract communities ; ii) favor groups of similar size.

Outline

Conclusions

- Higher-order interactions is the new trend;
- There are already some available tools that you can test on your datasets;
  - $\hookrightarrow$ do you have such datasets?
- New progresses can only be obtained if you first **formulate new ecological questions that can be analyzed with HOIs data**

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 (2022). Model-based clustering in simple hypergraphs through a stochastic blockmodel. Technical report, arXiv:2210.05983.

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 (2014). Consistency of spectral partitioning of uniform hypergraphs under planted partition model. In *Advances in Neural Information Processing Systems*, Volume 27.

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

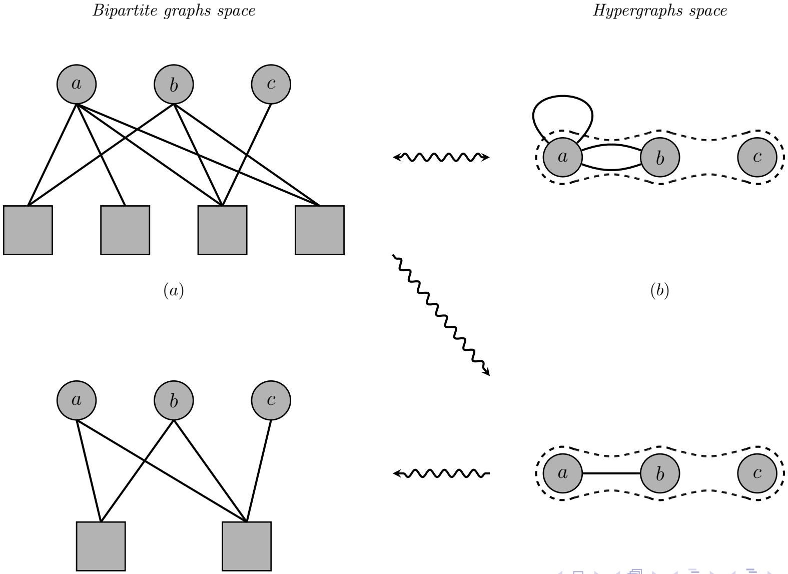
Kamiński, B., V. Poulin, P. Prałat, P. Szufel, and F. Théberge (2019). Clustering via hypergraph modularity. *PLoS One 14*(11), e0224307.

References II

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

M. T. Schaub *et al.* (2021). Signal processing on higher-order networks: Livin' on the edge... and beyond. /em Signal processing, Volume 187, 108149

Non equivalence between simple binary hypergraphs and bipartite graphs



Temporary page!

LaTeX was unable to guess the total number of pages correctly. A
was some unprocessed data that should have been added to the
this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus page
away, because LaTeX now knows how many pages to expect for t
document.