Hb-graphs and their application to textual datasets

Les Diablerets CUSO 22.10.2018

Xavier Ouvrard[1,2], Jean-Marie Le Goff[2], Stéphane Marchand-Maillet[1]

www.infos-informatique.net/2018-CUSO-Diablerets.html

[1]University of Geneva, [2]CERN

Research context

Research context

- PhD started in 10.2016 @ University of Geneva
  Hypergraph Modeling and Visualisation of Complex Collaboration Networks
- Done within the **Collaboration Spotting** project @ CERN
  => enhancing co-occurences in datasets

In project...

- **Datasets** modeled and stored as **labelled graphs**.
- **Co-occurences** through a **reference**.
- Multiple **facets** of dataset can be visualized.

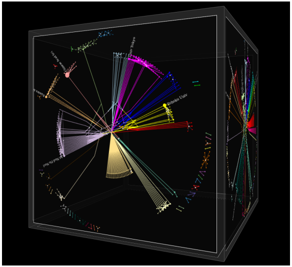


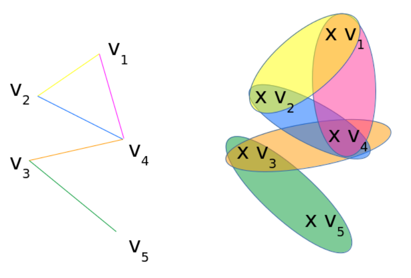
Figure 1: DataHyperCube: prototype in Ouvrard et al. [2018b]

but co-occurences are...

- Bags of elements
- **n-adic relationships**
- if bags reduced to sets:
  $\rightsquigarrow$ **hypergraphs** well fitted to model it!
- otherwise we need families of multisets
  $\rightsquigarrow$ **hb-graphs** are introduced

Hypergraphs

From graphs to hypergraphs



- **Hypergraphs** $\equiv$ generalisation of graphs to multiple nodes' links
- Hypergraphs introduced by Berge and Minieka [1973].

Definition

Bretto [2013]:

- **Hypergraph** $\mathcal{H}$: a family of subsets of a vertex set
- **Hyperedges**: elements of the family

Two visions

- **set of elements** of power set of nodes $\rightsquigarrow$ set view
- **extension of graphs** $\rightsquigarrow$ n-adic relationship view

Multisets I

Multiset and operations

- **Multiset**: a universe and a multiplicity function $A_m = (A, m)$
- **Natural multiset**: the range of the multiplicity function is a subset of $\mathbb{N}$.
- In natural multisets: two views:
  - weighted set
  - collection of objects: **bag**
- **Support** of the multiset: elements of the universe that have non zero multiplicity
- **m-cardinality** of a multiset A_m: sum of all multiplicity of elements of A.
- More in Singh et al. [2007].

"Navy blue and sky blue are blue colour names."

- A = {navy, blue, sky, color, name} (stopwords are removed, stemming done)
- $m\,(\text{navy}) = 1$
  $m(\text{blue}) = 3$
  $m(\text{sky}) = 1$
  $m\,(\text{color}) = 1$
  $m\,(\text{name}) = 1.$
- $A_m = \left\{\text{navy}^1, \text{blue}^3, \text{sky}^1, \text{color}^1, \text{name}^1\right\}$
  $\#_m A_m = 7$
- TF view: $A_{m'} =$
  $\left\{\text{navy}^{1/7}, \text{blue}^{3/7}, \text{sky}^{1/7}, \text{color}^{1/7}, \text{name}^{1/7}\right\}$
  $\#_m A_{m'} = 1$

Multisets II

Vector representation

<u>**Given:**</u> a natural multiset $A_m = (A, m)$ of universe $A = \{\alpha_i : i \in [\![n]\!]\}$ and multiplicity function m. It yields:

$$A_m = \left\{ \alpha_{i_j}^{m\left(\alpha_{i_j}\right)} : \alpha_{i_j} \in A_m^\star \right\}.$$

- **Vector representation**: $\overrightarrow{A_m} = (m(\alpha))_{\alpha \in A}^\top$.
- Sum of the elements of $\overrightarrow{A_m}$: $\sharp_m A_m$
- $|A|$ elements to be described but only $|A_m^\star|$ are non-zero
- => useful for building incidence matrix of hb-graphs

"Navy blue and sky blue are blue colour names."

- $A_m = \left\{ \text{navy}^1, \text{blue}^3, \text{sky}^1, \text{color}^1, \text{name}^1 \right\}$
  $\overrightarrow{A_m} = \begin{pmatrix} 1 & 3 & 1 & 1 & 1 \end{pmatrix}^\top$
- $A_{m'} = \left\{ \text{navy}^{1/7}, \text{blue}^{3/7}, \text{sky}^{1/7}, \text{color}^{1/7}, \text{name}^{1/7} \right\}$
  $\overrightarrow{A_{m'}} = \begin{pmatrix} 1/7 & 3/7 & 1/7 & 1/7 & 1/7 \end{pmatrix}^\top$

Motivation of the introduction of hb-graphs

- **e-adjacency tensor** of general hypergraphs (Ouvrard et al. [2017, 2018a]):
  - First proposal: built using different vertices
  - Allowing vertices duplication requires multisets
  - Hypergraphs as particular case of hb-graphs
- **Co-occurences = bags of elements**
  - Family of co-occurences retrieved
  - Natural hb-graphs well fitted
- **Individual weight on vertices per hyperedge requires multisets**
  - Diffusion by exchange (Ouvrard et al. [2018c])

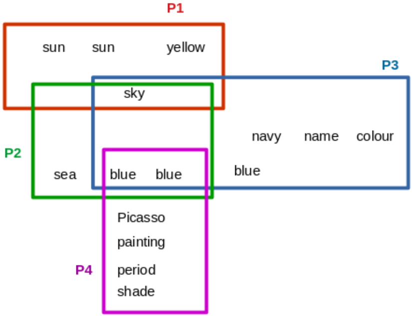
Hb-graphs: extending hypergraphs

Hyper-Bag-graph or hb-graph

- **Hb-graph**: family of multisets - called **hb-edges** - with:
  - same universe V, called **vertex set**.
  - support a subset of V.
  - **each hb-edge has its own multiplicity function**.
- **Natural hb-graph**: when all multiplicity functions have their range included in $\mathbb{N}$
- **Support hypergraph**: hypergraph of the support of the multisets
- **Star of a vertex**: multiset of all hb-edges where the vertex is, with a multiplicity the vertex multiplicity in this hyperedge
- **m-degree of a vertex**: m-cardinality of the star of this vertex
- **hypergraph**: natural hb-graph with multiplicity function ranges in $\{0, 1\}$

Four sentences:

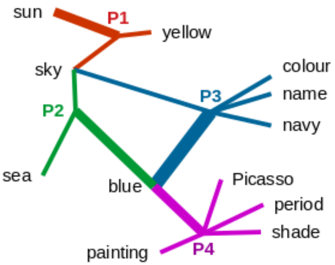
- **P1**: "The sun is in the sky and the sun is yellow."
- **P2**: "The sea is blue and the sky is also blue."
- **P3**: "Navy blue and sky blue are blue colour names."
- **P4**: "Picasso had a blue period where his paintings were in blue shade."

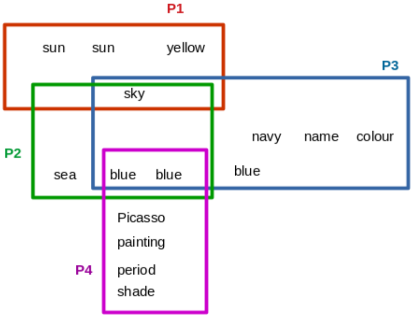


Hb-graphs: extending hypergraphs

Four sentences:

- **P1**: "The sun is in the sky and the sun is yellow."
- **P2**: "The sea is blue and the sky is also blue."
- **P3**: "Navy blue and sky blue are blue colour names."
- **P4**: "Picasso had a blue period where his paintings were in blue shade."

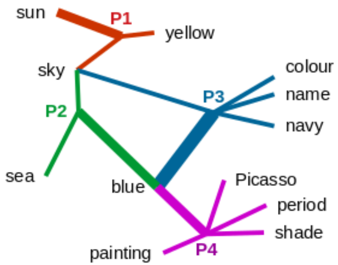




Hb-graphs: extending hypergraphs

Four sentences:

- **P1**: "The sun is in the sky and the sun is yellow."
- **P2**: "The sea is blue and the sky is also blue."
- **P3**: "Navy blue and sky blue are blue colour names."
- **P4**: "Picasso had a blue period where his paintings were in blue shade."

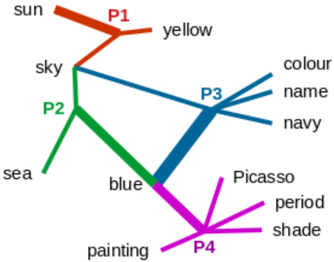


	P1	P2	P3	P4
sun	2	0	0	0
sky	1	1	1	0
yellow	1	0	0	0
sea	0	1	0	0
blue	0	1	3	2
colour	0	0	1	0
navy	0	0	1	0
name	0	0	1	0
painting	0	0	0	1
Picasso	0	0	0	1
period	0	0	0	1
shade	0	0	0	1

Incidence matrix of a hb-graph

Incidence

- hb-edges are incident if their intersection is not empty
- **Incidence matrix of the hb-graph** $\mathcal{H}$: $H = [m_j (v_i)]_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant p}}$.
- Used in: diffusion by exchange in Ouvrard et al. [2018c]
- Incidence is a pairwise concept: a vertex is incident to a hb-edge.
- The rows allow to see which hb-edges are incident: linked by rows.



$$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

How hb-graphs are useful?

Visualisation of exchange-based diffusion

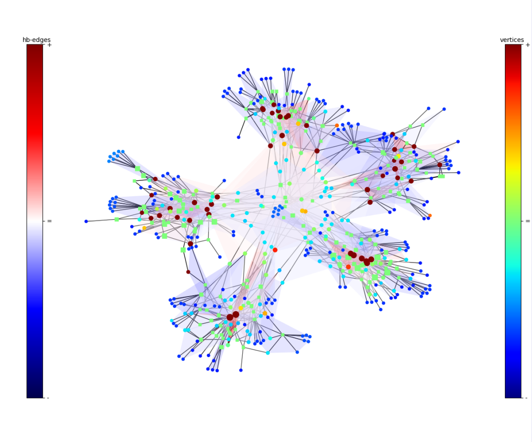


Figure 2: From Ouvrard et al. [2018c] © IEEE 2018

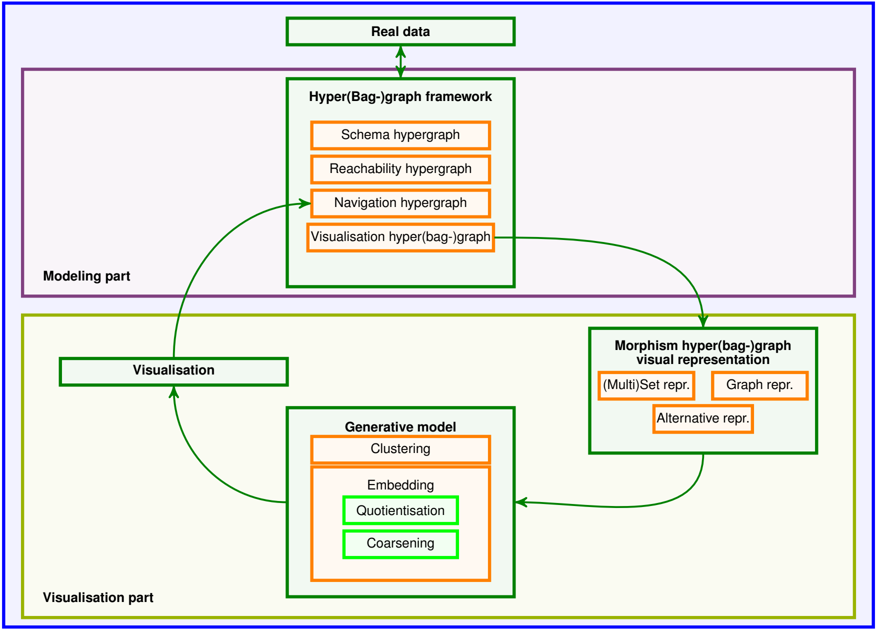
Applications

- Diffusion in hb-graphs and RW => see Ouvrard et al. [2018c]
- e-adjacency hypermatrix of hypergraphs => see Ouvrard et al. [2018a]
- Hyper(Bag-)graph modeling of datasets for information space visualisation => see next slides

On example

- 548 vertices
- 300 hb-edges
- 5 groups

Hyper(Bag-)graph modeling I



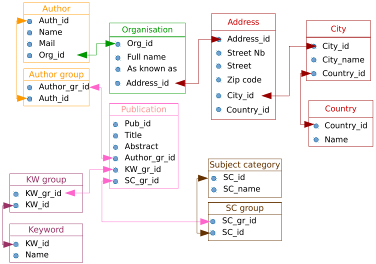
Hyper(Bag-)graph modeling II

Schema hypergraph

- In relational databases:
  - metadata instances => vertices.
  - tables => hyperedges.
  - foreign keys allow connection between hyperedges.
- In graph databases:
  - schema represents link between metadata.
  - schema not compulsory.

=> can be represented by a hypergraph, called **schema hypergraph**,

$$\mathcal{H}_{\mathrm{Sch}} = \left(V_{\mathrm{Sch}}, E_{\mathrm{Sch}}, i_{\mathrm{Sch}}\right).$$



Schema hypergraph: exploded view.

Shown on publication metadata example.

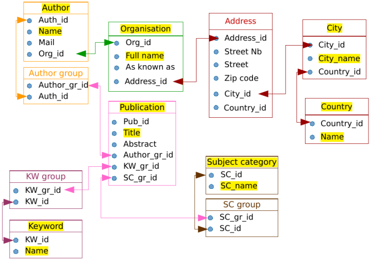
Hyper(Bag-)graph modeling III

Extracted schema hypergraph

- **Extracted schema hypergraph $\mathcal{H}_X$**: only metadata instances of interest are kept in $U \subseteq V_{\text{Sch}}$.
  The interest can be for:
  - visualisation
  - reference
  - search
  - keeping connectivity in the hypergraph.

$\mathcal{H}_X = (V_X, E_X, i_X)$, with:

- $V_X = U$,
- $E_X = \{e \cap U : e \in E_{\text{Sch}}\}$
- and $i_X = i_{\text{Sch}}|_{E_X}$.



Schema hypergraph: exploded view.
Highlighted: type of visual interest.

Shown on publication metadata example.

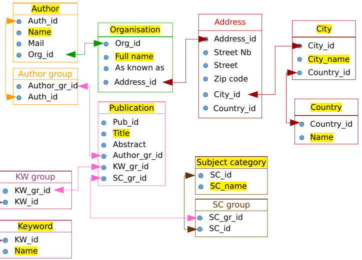
Hyper(Bag-)graph modeling IV

Reachability hypergraph

- **Reachability hypergraph**

  $\mathcal{H}_R = (V_R, E_R, i_R)$:
  obtained from $\mathcal{H}_X$ by calculating its connected components.
- Vertices of $\mathcal{H}_R$: $V_R = V_X$.
- Hyperedges of $\mathcal{H}_R$: connected components of $\mathcal{H}_X$

$$\forall e_R \in E_R : i_R(e_R) = \bigcup_{E_{\text{CC}} \in C_X} \bigcup_{e \in E_{\text{CC}}} i_X(e).$$



One hyperedge in the reachability hypergraph that contains (abusively) all the interesting tables for visualisation and reference.

e_R ={**Publication, Author, Organisation, City, Country, Subject Category, Keyword**}

Hyper(Bag-)graph modeling V

Navigation hypergraph

- **Navigation hypergraph** $\mathcal{H}_N = (V_N, E_N)$: obtained from $\mathcal{H}_R$ by choosing one hyperedge $e_R \in E_R$.
- Vertices of $\mathcal{H}_N = e_R$.
  Possible reference vertices in R_{ref}.
- Hyperedges of $\mathcal{H}_N$:

$$E_N = \{e_R \backslash R : R \subseteq R_{\text{ref}} \wedge R \neq \emptyset\}.$$

- Navigation is possible without changing reference inside a hyperedge of $\mathcal{H}_N$.

Only one $e_R \in E_R$:
e_R =**{Publication, Author, Organisation, City, Country, Subject Category, Keyword}**.

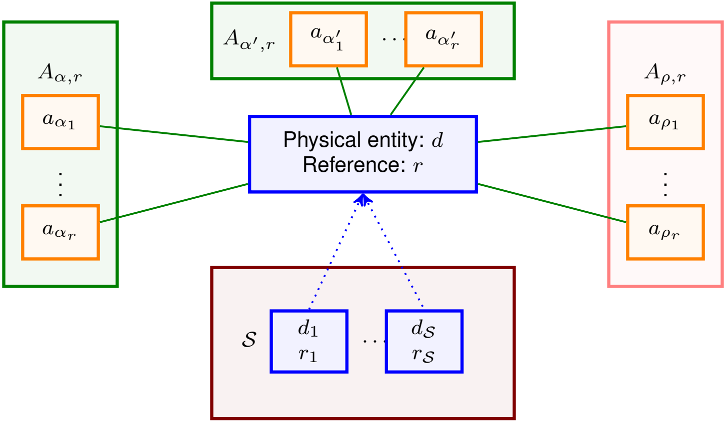
- Choosing **Publication as reference** we can have:
  - co-authors,
  - co-organisations, ...
- Choosing **Organisation as reference** we can have:
  - co-publications where a common organisation is present,
  - co-towns that have a publication where a common organisation is present...

Hyper(Bag-)graph modeling VI

- In a dataset $\mathcal{D}$, a physical entity d of reference r is fully described by:

$$(r, \{A_{\alpha,r} : \alpha \in V_{\mathsf{Sch}}\}) .$$

- $A_{\alpha,r} = \{a_1, ..., a_{\alpha_r}\}$: multiset of values of type α that are attached to d.



Hyper(Bag-)graph modeling VII

Visualisation Hyper(Bag-)graph

- For each $v \in \bigcup_{r \in \mathcal{S}} A_{\rho,r} = \Sigma_\rho$, we build a set of physical references corresponding to data d that have v in attributes of type ρ: $R_v = \{r : v \in A_{\rho,r}\}$.
- Multiset of values of type α relatively to the reference v: $\biguplus_{r \in R_v} A_{\alpha,r} = e_{\alpha,v}$.
- **Raw visualisation hb-graph** for the facet of type α/ρ attached to the search $\mathcal{S}$ is:

$$\mathcal{H}_{\alpha/\rho,\mathcal{S}} = \left(\biguplus_{r \in \mathcal{S}} A_{\alpha,r}, (e_{\alpha,v})_{v \in S_\rho} \right).$$

- By quotienting Σ_ρ and weighting => **reduced visualisation weighted hb-graph** for the search $\mathcal{S}$:

$$\mathcal{H}_{\alpha/\rho,w_\alpha,\mathcal{S}} = \left(\biguplus_{r \in \mathcal{S}} A_{\alpha,r}, \overline{E_\alpha}, w_\alpha \right).$$

- Those two hb-graphs can be reduced to their support hypergraph.

Hyper(Bag-)graph modeling VIII

Navigating through facets

- Reference type: ρ, current type α, target type: α'.
- Selecting vertices of type $A \subseteq A_{\alpha,S}$.
  Allows to:
  - retrieve a subset of hb-edges of $\overline{E_\alpha}$:
    $$\overline{E_\alpha}\Big|_A = \left\{e : e \in \overline{E_\alpha} \wedge (\exists x \in e : x \in A)\right\}.$$
  - retrieve the class $\overline{v}$ attached to each $e \in \overline{E_\alpha}\Big|_A$ => $\overline{V}\Big|_A$ set of class $\overline{v}$.
  - retrieve the references of type ρ: $\mathcal{V}_{\rho,A} = \left\{v : \forall \overline{v} \in \overline{V}\Big|_A : v \in \overline{v}\right\}$.
  - R_v remains the same between facets
    => group of references: $\mathcal{S}_A = \bigcup\limits_{v \in \mathcal{V}_{\rho,A}} R_v$.
  - switching to the facet of type α is then possible:
    $$\mathcal{H}_{\alpha'/\rho}\Big|_A = \left(\bigcup_{r \in \mathcal{S}_A} A_{\alpha',r}, \left(e_{\alpha',v}\right)_{v \in \mathcal{V}_{\rho,A}}\right).$$

An example

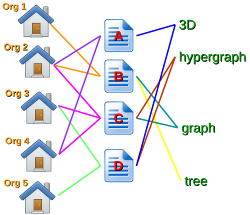


Figure 1: publication dataset

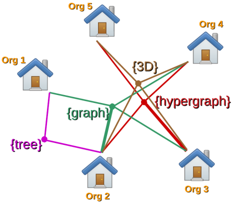


Figure 2: hb-graph view

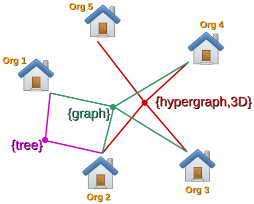
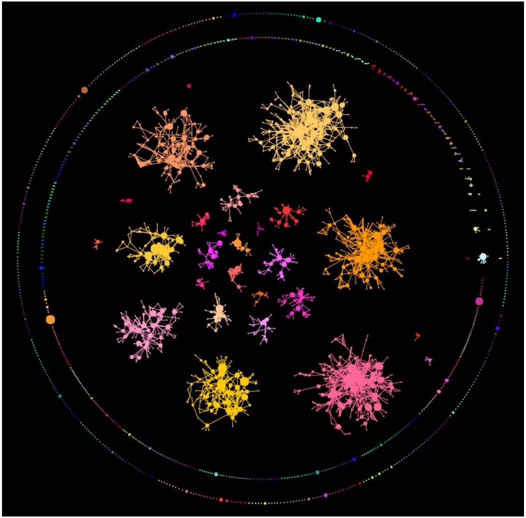


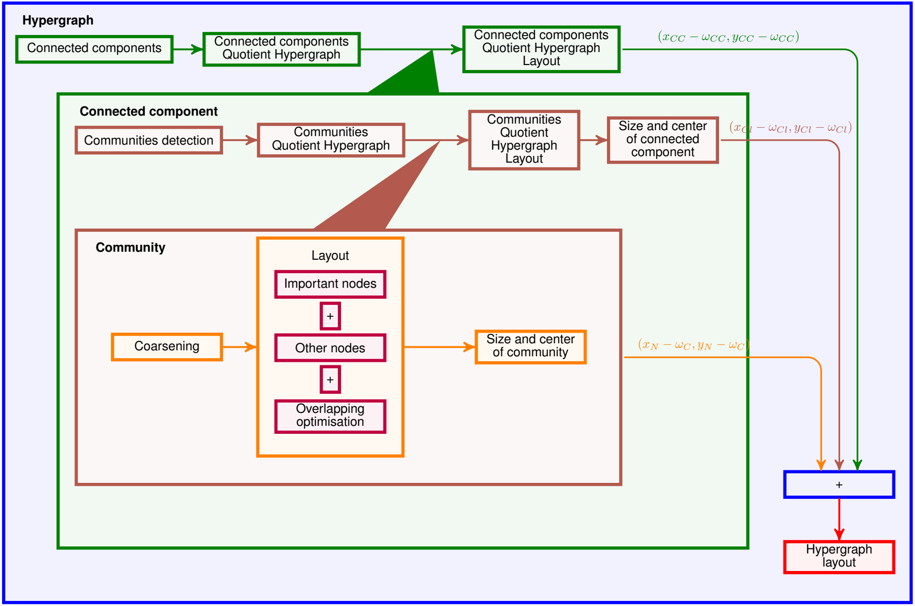
Figure 3: reduced support hypergraph

- **Aim:** Visualize **co-occurences of organisations** in reference to keywords.
- **Choose a type**:
  - α to visualize => organisations;
  - ρ to use as reference for co-occurences => keywords.
- **Figure 2:** hb-graph with:
  - vertices: organisations;
  - hb-edges: organisation co-occurences.
- **Figure 3:** support hypergraph of figure 2

Large hyper(bag-)graphs



Behind


Hypergraph
Connected components
Connected components Quotient Hypergraph
Connected components Quotient Hypergraph Layout
$(x_{CC} - \omega_{CC}, y_{CC} - \omega_{CC})$
Connected component
Communities detection
Communities Quotient Hypergraph
Communities Quotient Hypergraph Layout
Size and center of connected component
$(x_{Cl} - \omega_{Cl}, y_{Cl} - \omega_{Cl})$
Community
Coarsening
Layout
Important nodes
+
Other nodes
+
Overlapping optimisation
Size and center of community
$(x_N - \omega_C, y_N - \omega_C)$
+
Hypergraph layout

Demo

Live demo Cube

Thank you for your attention

Questions?

Bibliography I

Claude Berge and Edward Minieka. *Graphs and hypergraphs*, volume 7. North-Holland publishing company Amsterdam, 1973.

Alain Bretto. Hypergraph theory. *An introduction. Mathematical Engineering. Cham: Springer*, 2013.

Xavier Ouvrard, Jean-Marie Le Goff, and Stephane Marchand-Maillet. Adjacency and tensor representation in general hypergraphs part 1: e-adjacency tensor uniformisation using homogeneous polynomials. *arXiv preprint arXiv:1712.08189*, 2017.

Xavier Ouvrard, Jean-Marie Le Goff, and Stephane Marchand-Maillet. Adjacency and tensor representation in general hypergraphs. part 2: Multisets, hb-graphs and related e-adjacency tensors. *arXiv preprint arXiv:1805.11952*, 2018a.

Xavier Ouvrard, Jean-Marie Le Goff, and Stéphane Marchand-Maillet. A hypergraph based framework for modelisation and visualisation of high dimension multi-facetted data. *Soon on Arxiv*, 2018b.

Xavier Ouvrard, Jean-Marie Le Goff, and Stephane Marchand-Maillet. Diffusion by exchanges in hb-graphs: Highlighting complex relationships. *CBMI Proceedings*, 2018c.

D Singh, A Ibrahim, T Yohanna, and J Singh. An overview of the applications of multisets. *Novi Sad Journal of Mathematics*, 37(3):73–92, 2007.