Hypergraph Modularity and Clustering

Bogumił Kamiński
Valérie Poulin
Paweł Prałat
Przemysław Szufel
François Théberge*
theberge@ieee.org

August 2019

Outline

Graph modularity

We can write the modularity of a partition $\mathbf{A}$ of graph G as:

$$
\begin{aligned}
q_G(\mathbf{A}) &= \sum_{A_i \in \mathbf{A}} \left(\frac{e_G(A_i)}{|E|} - \frac{(vol(A_i))^2}{4|E|^2} \right) \\
&= \frac{1}{|E|} \sum_{A_i \in \mathbf{A}} \left(e_G(A_i) - \mathop{\mathbb{E}}_{G \in \mathcal{G}} (e_G(A_i)) \right)
\end{aligned}
$$

$e_G(A_i) = |\{e \in E : e \subseteq A_i\}|$ is the *edge contribution* and, $\mathbb{E}_{G \in \mathcal{G}}(e_G(A_i))$ is the *degree tax*.

Chung-Lu Model

In model II, we select m edges $e = (u_1, u_2)$ where each u_i is independently sampled from V according to the multinomial distribution where $p(v_i) = deg_G(v_i)/vol(V)$.

Let $\mathcal{CL}_2(G)$, the distribution of graphs obtained with model II. For $G' = (V, E') \sim \mathcal{CL}_2(G)$:

- $\mathbb{E}_{G' \sim \mathcal{CL}_2(G)}(deg_{G'}(v_i)) = deg_G(v_i), 1 \leq i \leq n.$
- we always have $|E'| = |E| = m$,
- there can be multi-edges,
- there can be self-edges,
- complexity is $O(m)$.

Chung-Lu Model

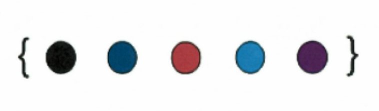
Lemma

The degree tax term in the modularity function for graph G is the expected value of the edge contribution term over the graphs $G' \sim \mathcal{CL}_2(G)$.

Can we generalize this model for hypergraphs?

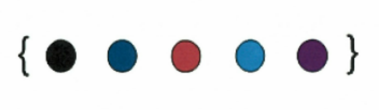
More complex relations

Relations may involve several entities ...

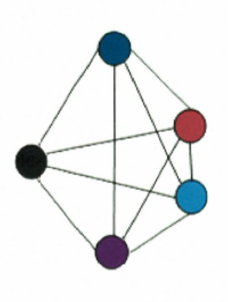


More complex relations

Relations may involve several entities ...

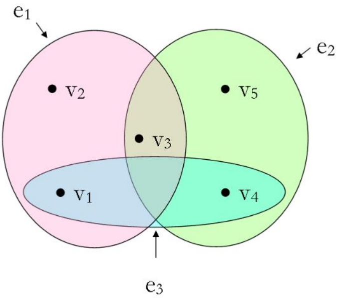


which are often represented via 2-section subgraphs (loss of information).



Hypergraphs

- $H = (V, E)$ where $|V| = n$, $|E| = m$
- $e \in E$: (hyper)-edges where $e \subseteq V$, $|e| \geq 2$
- Edges can have weights
- We consider undirected hypergraphs



Why Hypergraphs?

Some data are better modeled with hypergraphs, such as:

- email exchanges
- tracking co-locations
- categorical data modeling
- numerical linear algebra (ref: P.A. Papa and I.L. Markov)

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

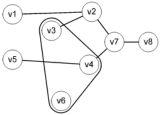


Figure 1: An example of a nearly block-diagonal matrix and corresponding hypergraph. Each row of the matrix corresponds to a hyperedge, and each column corresponds to vertices $v1$ through $v8$. Recursively bisecting the graph aligns the blocks of the matrix on the diagonal.

Hypergraphs

- Few hypergraph-based algorithms exist in data science
- They are typically slower
- Some have an equivalent graph representation

(q) Can we define a modularity function on hypergraphs?

Chung-Lu Model for Hypergraphs

Consider a hypergraph $H = (V, E)$ with $V = \{v_1, \ldots, v_n\}$.

Hyperedges $e \in E$ are subsets of V of cardinality greater than one where:

$$e = \{(v, m_e(v)) : v \in V\}$$

and $m_e(v) \in \mathbb{N} \cup \{0\}$ is the multiplicity of the vertex v in e

$|e| = \sum_v m_e(v)$ is the *size* of hyperedge e, and

$deg(v) = \sum_{e \in E} m_e(v)$, and

$vol(A) = \sum_{v \in A} deg(v)$ as for graphs.

Chung-Lu Model for Hypergraphs

Let F_d be the family of multisets of size d; that is,

$$F_d := \left\{ \{(v_i, m_i) : 1 \leq i \leq n\} : \sum_{i=1}^{n} m_i = d \right\}.$$

The hypergraphs in the random model are generated via independent random experiments. For each d such that $|E_d| > 0$, the probability of generating $e \in F_d$ is given by:

$$P_{\mathcal{H}}(e) = |E_d| \cdot \binom{d}{m_1, \ldots, m_n} \prod_{i=1}^{n} \left(\frac{deg(v_i)}{vol(V)} \right)^{m_i}.$$

where $m_i = m_e(v_i)$.

Chung-Lu Model for Hypergraphs

We can show that:

$$\mathbb{E}_{H' \sim \mathcal{H}}[deg_{H'}(v_i)] = \sum_{d \geq 2} \frac{d \cdot |E_d| \cdot deg(v_i)}{vol(V)} = deg(v_i),$$

with $vol(V) = \sum_{d \geq 2} d \cdot |E_d|$.

We use this generalization of the Chung-Lu model as our null model (*degree tax*) to define hypergraph modularity.

Hypergraph Modularities

Let $H = (V, E)$ and $\mathbf{A} = \{A_1, \ldots, A_k\}$, a partition of V. For edges of size greater than 2, several definitions can be used to quantify the *edge contribution* given $\mathbf{A}$, such as:

(a) all vertices of an edge have to belong to one of the parts to contribute; this is a *strict* definition;

(b) the *majority* of vertices of an edge belong to one of the parts;

(c) at least 2 vertices of an edge belong to the same part; this is implicitly used when we replace a hypergraph with its 2-section graph representation.

Strict Hypergraph Modularity

The edge contribution for $A_i \subseteq V$ is:

$$e(A_i) = |\{e \in E;\ e \subseteq A_i\}|.$$

The strict modularity of $\mathbf{A}$ on H is then defined as a natural extension of standard modularity in the following way:

$$q_H(\mathbf{A}) = \frac{1}{|E|} \sum_{A_i \in \mathbf{A}} \left(e(A_i) - \mathbb{E}_{H' \sim \mathcal{H}}[e_{H'}(A_i)] \right).$$

which can be written as:

$$q_H(\mathbf{A}) \quad = \quad \frac{1}{|E|} \left(\sum_{A_i \in \mathbf{A}} e(A_i) - \sum_{d \geq 2} |E_d| \sum_{A_i \in \mathbf{A}} \left(\frac{vol(A_i)}{vol(V)} \right)^d \right)$$

Link to Chung-Lu Model

We generalized model II over hypergraphs.

For each d, sample $|E_d|$ edges $e = (u_1, .., u_d)$ where each u_i is independently sampled from V with $p(v_i) \propto deg(v_i)$.

Let $\mathcal{CL}_2(H)$, the distribution of hypergraphs obtained this way; for $H' = (V, E') \sim \mathcal{CL}_2(H)$:

- $\mathbb{E}_{H' \sim \mathcal{CL}_2(H)}(deg_{H'}(v_i)) = deg_H(v_i), 1 \leq i \leq n.$
- we always have $|E'_d| = |E_d|$,
- there can be multi-edges, and
- there can be repeated vertices within an edge.

Link to Chung-Lu Model

Lemma

The degree tax term in the modularity function for hypergraph $H = (G, V)$ and partition $\mathbf{A} = \{A_1, ..., A_k\}$ of V is the expected value of the edge contribution term over hypergraphs $H' \sim \mathcal{CL}_2(H)$.

Other Hypergraph Modularity

We can adjust the degree tax to many natural definitions of edge contribution, for example the majority definition.

In this case $(vol(A)/vol(V))^d$ (that is equivalent to $\mathbb{P}\left(\mathrm{Bin}(d, vol(A)/vol(V)) = d\right)$ becomes $\mathbb{P}\left(\mathrm{Bin}(d, vol(A)/vol(V)) > d/2\right)$.

The *majority* modularity function of a hypergraph partition is then:

$$\frac{1}{|E|}\left(\sum_{A_i \in \mathbf{A}} e(A_i) - \sum_{d \geq 2} |E_d| \sum_{A_i \in \mathbf{A}} \mathbb{P}\left(\mathrm{Bin}\left(d, \frac{vol(A_i)}{vol(V)}\right) > d/2\right)\right).$$

Other Hypergraph Modularity

Decomposing H into d-uniform hypergraphs H_d, we get the following degree-independent modularity function:

$$q_H^{DI}(\mathbf{A}) = \sum_{d \geq 2} \frac{|E_d|}{|E|} q_{H_d}(\mathbf{A}).$$

This is as before, but replacing the volumes computed over H with volumes computed over H_d for each d where $|E_d| > 0$.

Finally, we can generalize the modularity function to allow for weighted hyperedges.

Hypergraph Clustering

We seek $\mathbf{A} = \{A_1, ..., A_k\} \in \mathcal{P}(V)$, which maximize the **strict** hypergraph modularity $q_H()$.

Set $\mathcal{P}(V)$ of all partitions of V is huge.

Let: $\mathcal{S}(H) = \{H' = (V, E') \mid E' \subseteq E\}$ and define:

$$p : \mathcal{S}(H) \to \mathcal{P}(V)$$

the function that sends a sub-hypergraph of H to the partition its connected components induce on V.

We define an equivalence relation:

$$H_1 \equiv_p H_2 \iff p(H_1) = p(H_2)$$

and the quotient set $\mathcal{S}(H)/_{\equiv_p}$.

Hypergraph Clustering

Define the *canonical representative mapping*

$$f : \mathcal{S}(H)/_{\equiv_p} \to \mathcal{S}(H)$$

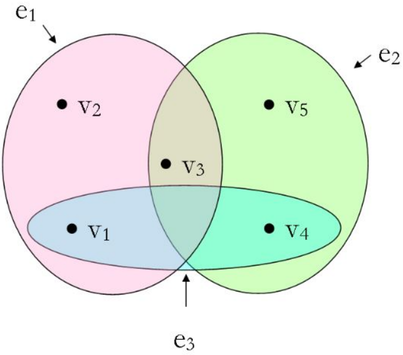
which maps an equivalence class to the largest member of the class: $f([H']) = H^*$.

Let $\mathcal{P}^*(V)$ be the image of p applied to the canonical representatives H^*.

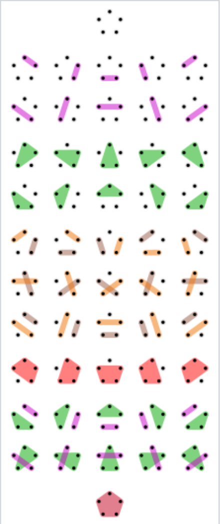
We'll show the optimal solution lies in $\mathcal{P}^*(V)$, a subset of size at most $2^{|E|}$.

Hypergraph Clustering

Consider the toy graph:

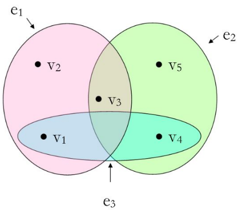


Here, $|\mathcal{P}(V)| = B_5 = 52$.



The 52 partitions of a set with 5 elements

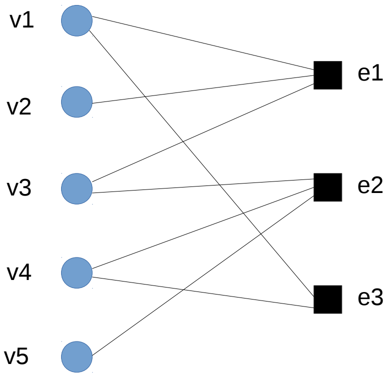
Hypergraph Clustering



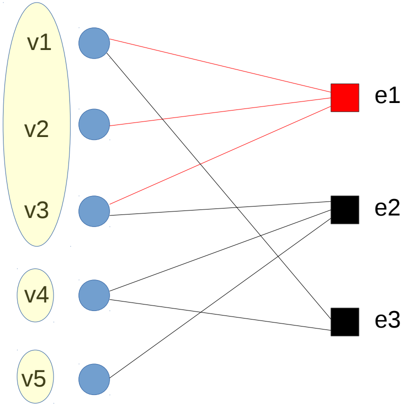
$|\mathcal{P}^*(V)| = 7$
$\leq 2^3 << B_5.$

i	$E_i \subseteq E$	$p(H_i),\ H_i = (V, E_i)$
0	$\emptyset$	$\{\{v_1\},\{v_2\},\{v_3\},\{v_4\},\{v_5\}\}$
1	$\{e_1\}$	$\{\{v_1,v_2,v_3\},\{v_4\},\{v_5\}\}$
2	$\{e_2\}$	$\{\{v_1\},\{v_2\},\{v_3,v_4,v_5\}\}$
3	$\{e_3\}$	$\{\{v_1,v_4\},\{v_2\},\{v_3\},\{v_5\}\}$
4	$\{\mathbf{e_1},\mathbf{e_2}\}$	$\{\{\mathbf{v_1},\mathbf{v_2},\mathbf{v_3},\mathbf{v_4},\mathbf{v_5}\}\}$
5	$\{e_1,e_3\}$	$\{\{v_1,v_2,v_3,v_4\},\{v_5\}\}$
6	$\{e_2,e_3\}$	$\{\{v_1,v_3,v_4,v_5\},\{v_2\}\}$
7	$\{\mathbf{e_1},\mathbf{e_2},\mathbf{e_3}\}$	$\{\{\mathbf{v_1},\mathbf{v_2},\mathbf{v_3},\mathbf{v_4},\mathbf{v_5}\}\}$

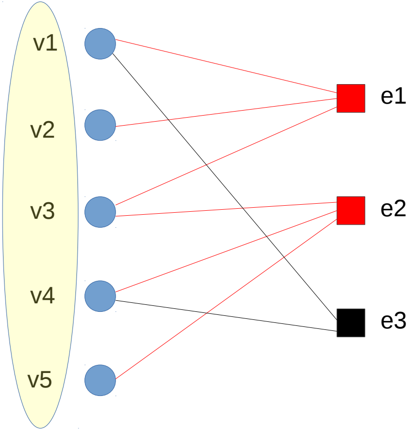
Bipartite graph view



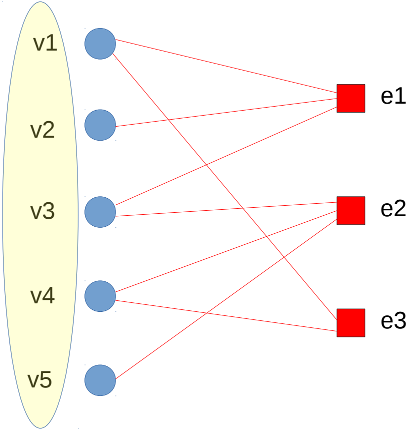
$E_1 = \{e_1\}$



$E_4 = \{e_1, e_2\}$



$E_7 = \{e_1, e_2, e_3\}$



Hypergraph Clustering

Lemma

Let $H = (V, E)$ be a hypergraph and $\mathbf{A} = \{A_1, ..., A_k\}$ a partition of V. If there exists $H' \in \mathcal{S}(H)$ such that $\mathbf{A} = p(H')$, then the edge contribution of $q_H(\mathbf{A})$ is $\frac{|E^*|}{m}$, where E^* is the edge set of the canonical representative H^* of $[H']$.

i.e. the proportion of hyperedges that are subsets of a part.

Hypergraph Clustering

Lemma

Let $H = (V, E)$ be a hypergraph and $\mathbf{A}$ be any partition of V. If $\mathbf{B}$ is a refinement of $\mathbf{A}$, then the degree tax of $\mathbf{B}$ is smaller than or equal to the degree tax of $\mathbf{A}$ with equality if and only if $\mathbf{A} = \mathbf{B}$.

Hypergraph Clustering

Lemma

Let $H = (V, E)$ be a hypergraph and $\mathbf{A}$ be any partition of V. If $\mathbf{B}$ is a refinement of $\mathbf{A}$, then the degree tax of $\mathbf{B}$ is smaller than or equal to the degree tax of $\mathbf{A}$ with equality if and only if $\mathbf{A} = \mathbf{B}$.

- We prove the following by showing that for any partition, there exists some $H^* \in \mathcal{P}^*(V)$ such that $p(H^*)$ is a refinement of that partition, with the same edge contribution.

Theorem

Let $H = (V, E)$ be a hypergraph. If $\mathbf{A} \in \mathcal{P}(V)$ maximizes the modularity function $q_H(\cdot)$, then $\mathbf{A} \in \mathcal{P}^*(V)$.

Hypergraph Clustering

Previous results give the steps to define heuristic algorithms:

- for $E' \subseteq E$, let $H' = (V, E')$
- find $H^* = [H'] = (V, E^*)$ and compute *edge contribution* part of $q_H()$
- find $\mathbf{A} = p(H^*)$ and compute *degree tax* part of $q_H()$

Simple ways to search for good candidates $E' \subseteq E$:

1. **Greedy random:** shuffle the edges and add edge to E' in turn if $q_H()$ improves; repeat;
2. **CNM-like:** look for best edge to add to E' at each step;

Hypergraph-CNM

Data: hypergraph $H = (V, E)$
Result: $\mathbf{A}_{opt}$, a partition of V with modularity q_{opt}

Initialize $\mathbf{A}_{opt}$ the partition with all $v \in V$ in its own part, and q_{opt};
repeat
  foreach $e \in E$ **do**
    set $q_e = -\infty$
  end
  foreach $e \in E$ *touching two or more parts in* $\mathbf{A}_{opt}$ **do**
    compute the partition A_e obtained when merging all parts in $\mathbf{A}_{opt}$ touched by e, and compute its modularity q_e;
  end
  select edge e^* with highest q_e;
  if $q_e^* \geq q_{opt}$ **then**
    $A_{opt} = A_e^*,\ q_{opt} = q_e^*$;
  end
until $q_e^* < q_{opt}$;
output: $\mathbf{A}_{opt}$ and q_{opt}

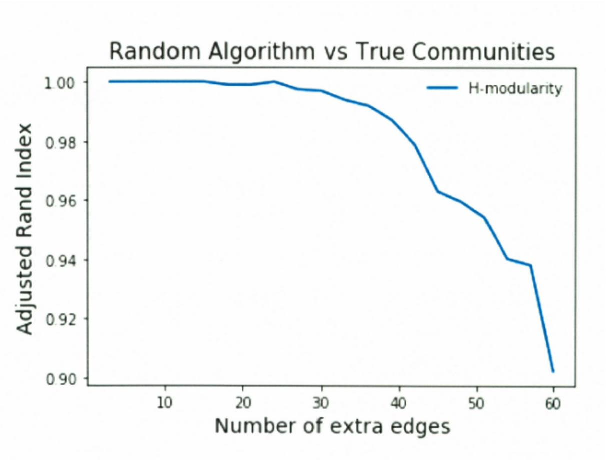
Hypergraph Clustering

Is it working?
Is $q_H()$ a ”good” objective function?

Consider the following experiment:

- build hypergraphs with 3 communities of 20 vertices and 50 edges of size $2 \leq d \leq 5$ each;
- add $3 \leq k \leq 60$ random edges of same size(s);
- run random algorithm (with 25 repeats) several times over range of k values;
- for each k, compute mean adjusted RAND index;

Hypergraph Clustering



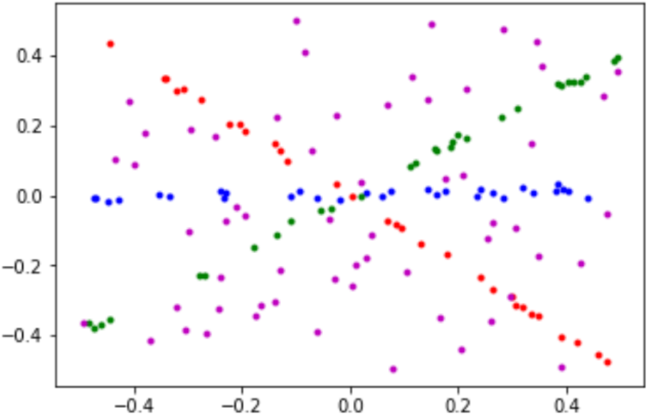
Synthetic Hypergraphs

[REF: M. Leordeanu, C. Sminchisescu, Efficient Hypergraph Clustering]

- Generate noisy points along 3 lines on the plane with different slopes
- add some random points
- select sets of 3 or 4 points (hyperedges)
  - all coming from the same line (“signal”)
  - or not (“noise”)
- Sample hyperedges for which the points are well aligned, and so that the expected proportion of signal vs. noise is 2:1.

We consider 3 different regimes: (i) mostly 3-edges, (ii) mostly 4-edges and (iii) balanced between 3 and 4-edges.

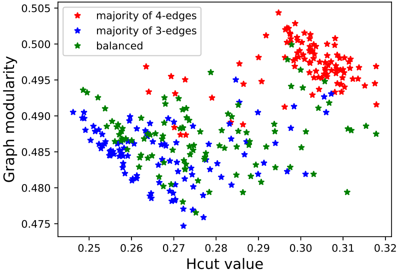
Synthetic Hypergraphs

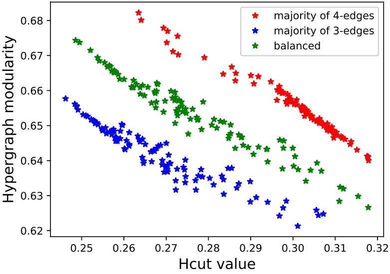


Synthetic Hypergraphs

Cluster vertices via Louvain on (weighted) 2-section graph.

Modularity vs Hcut. We observe a higher correlation with Hcut (number of splitted hyperedges) with the H-modularity.





DBLP Hypergraph

Small co-author hypergraph with 1637 nodes and 865 hyperedges of sizes 2 to 7.

We compare Louvain (over 2-section) and hypergraph-CNM (with strict modularity).

Partitioning the DBLP dataset.

algorithm	$q_H()$	$q_G()$	Hcut	#parts
Louvain	0.8613	0.8805	0.1181	40
CNM	0.8671	0.8456	0.0945	92

DBLP Hypergraph

Algorithms based on $q_H()$ will tend to cut less of the larger edges, as compared to the Louvain algorithm, at expense of cutting more size-2 edges.

Proportion of edges of size 2, 3 or 4 cut by the algorithms.

Algorithm	2-edges	3-edges	4-edges
Louvain	0.0382	0.1815	0.3158
CNM	0.0590	0.1277	0.1842

Conclusion and Ongoing Work

- Done so far:
  - generalized Chung-Lu model for hypergraphs
  - generalized modularity function to hypergraphs
  - steps toward hypergraph clustering algorithms
  - two simple heuristic algorithms: random and CNM
- Ongoing:
  - better intuition behind modularity functions
  - better, scalable clustering algorithm(s)
  - experiments on real datasets