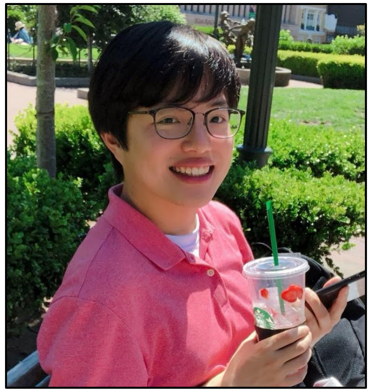
How Do Hyperedges Overlap in Real-World Hypergraphs? Patterns, Measures, and Generators

Geon Lee*

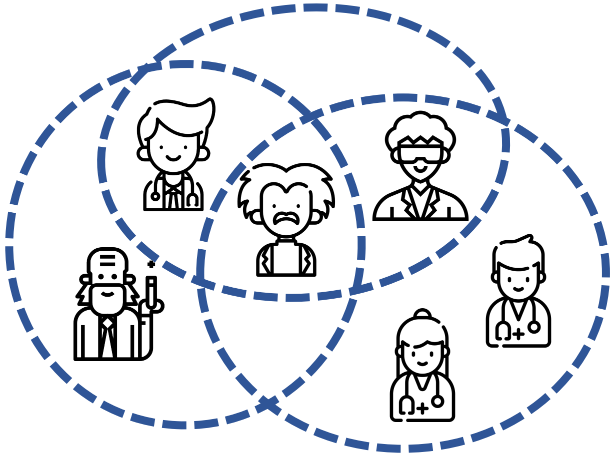
Minyoung Choe*



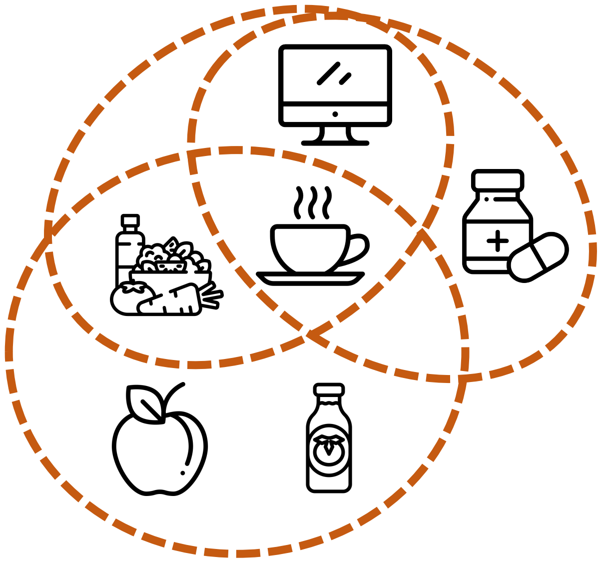
Kijung Shin

Hypergraphs are Everywhere

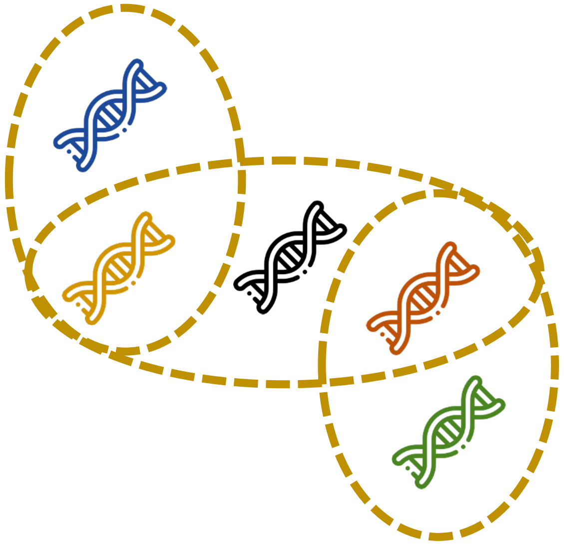
- Hypergraphs consist of nodes and hyperedges.
- Each hyperedge is a subset of any number of nodes.
- Hyperedges can overlap in infinitely many different ways.



Collaborations of Researchers



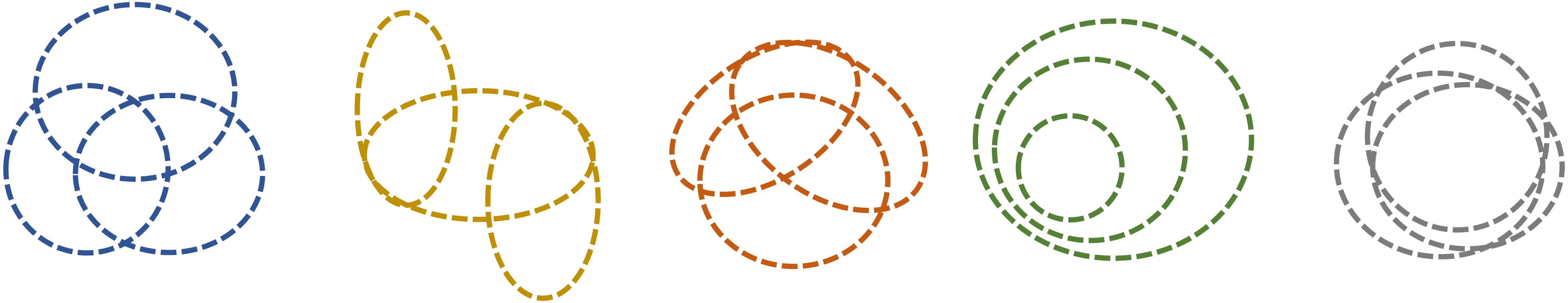
Co-purchases of Items



Joint Interactions of Proteins

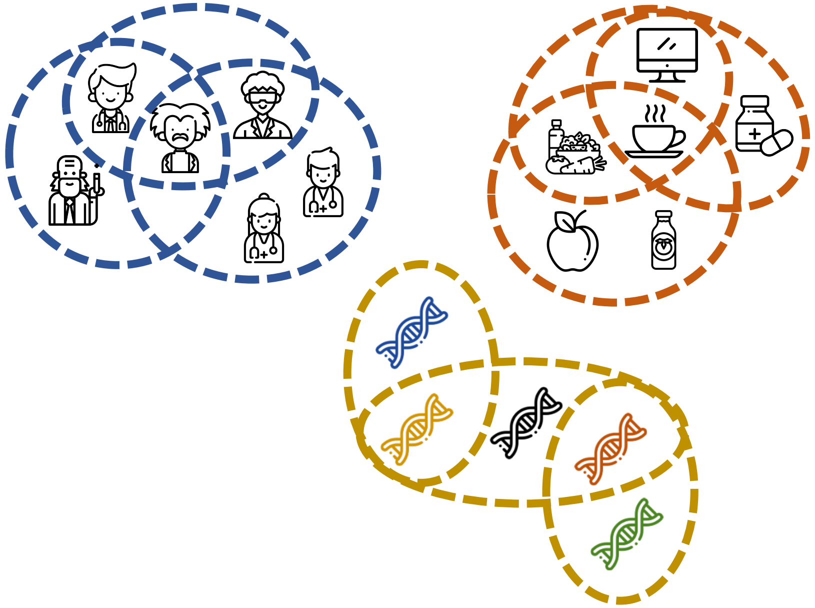
Our Questions

Q1 How do hyperedges overlap in real-world hypergraphs?



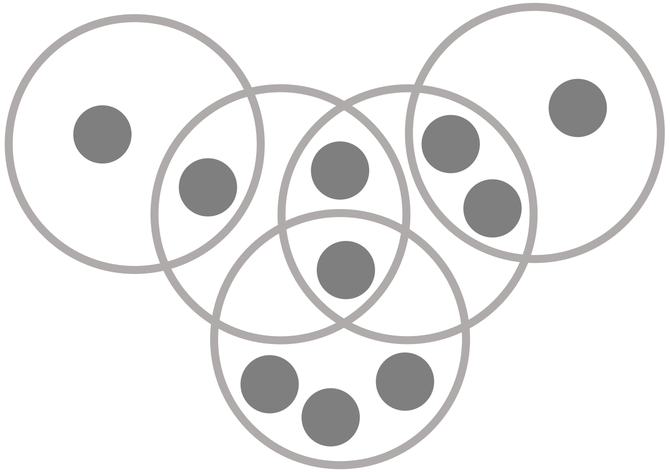
Our Questions (cont.)

Q2 Are there any non-trivial patterns that distinguish real-world hypergraphs from random hypergraphs?



Real-world Hypergraphs

VS

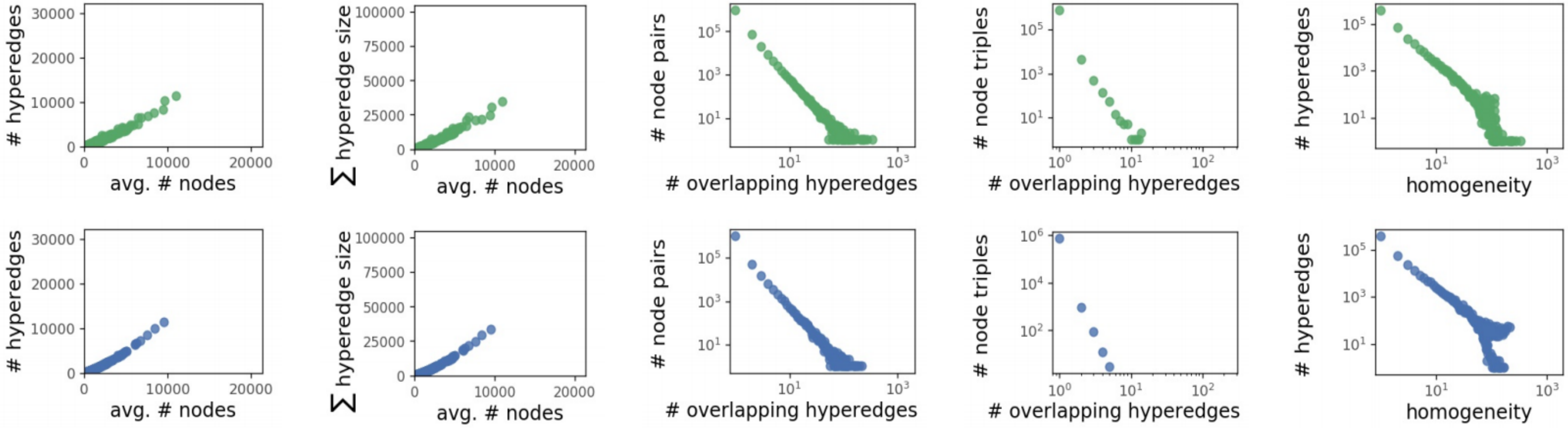


Random Hypergraph

Our Questions (cont.)

Q3 How can we reproduce the patterns through simple mechanisms?

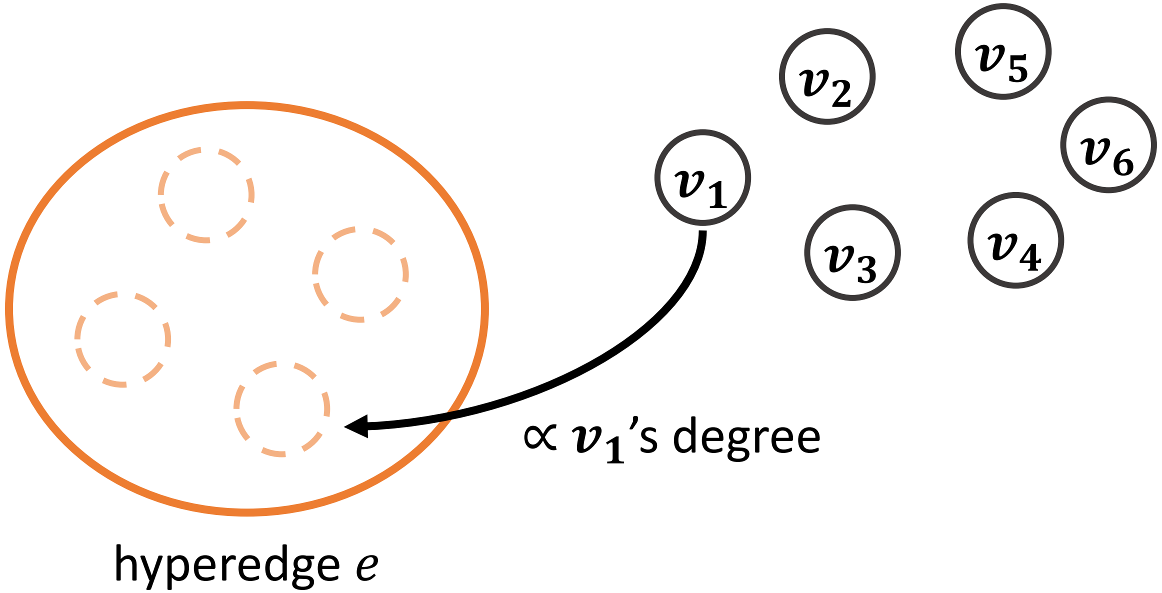
● **Real-world hypergraph** ● **Generated hypergraph**



Null Model

HyperCL: Random Hypergraph Generator (Null Model)

- Nodes are sampled with probability proportional to the degree of each node.
- The degree distribution of nodes is empirically preserved.

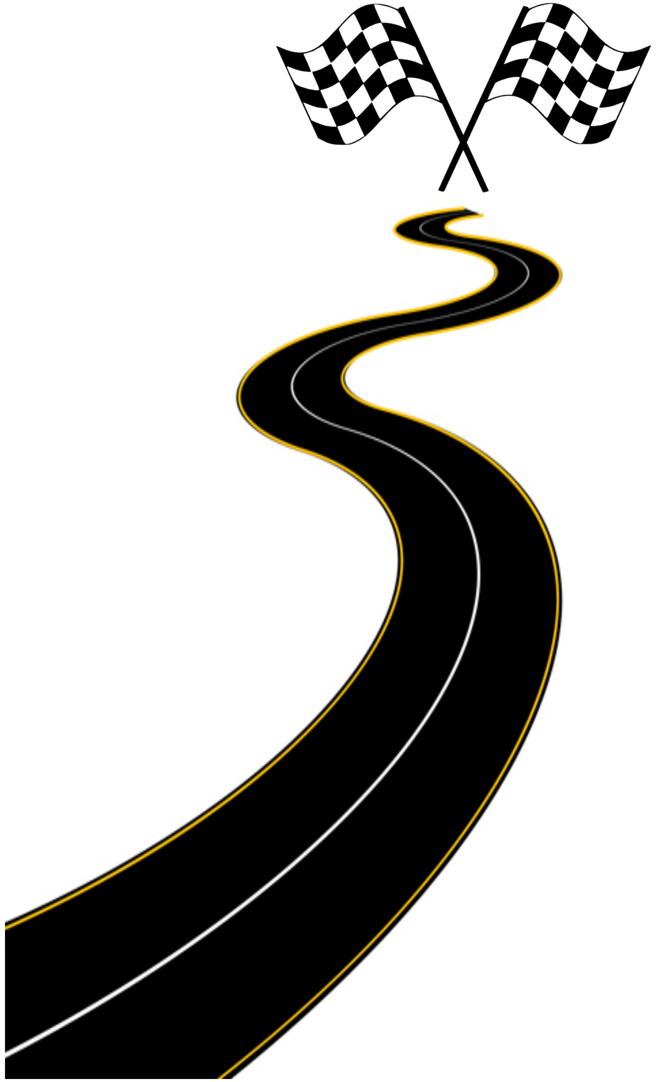


Datasets

Thirteen real-world hypergraphs from six domains

Domain	**Datasets**
Email	email-Enron, email-Eu
Contact	contact-primary, contact-high
Drugs	NDC-classes, NDC-substances
Tags	tags-ubnutu, tags-math
Threads	threads-ubuntu, threads-math
Co-authorship	coauth-DBLP, coauth-geology, coauth-history

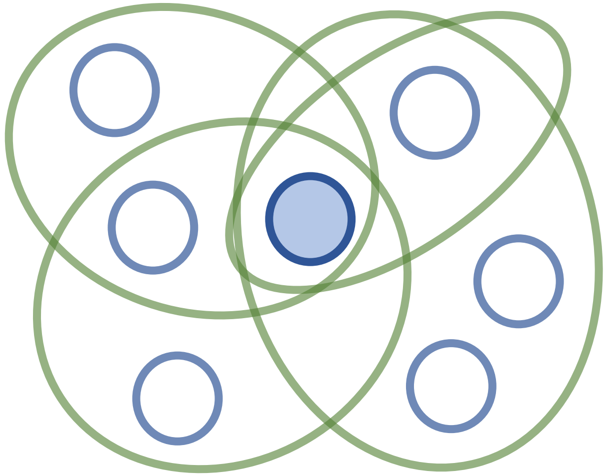
Roadmap



1. **Observation: Egonet Level**
2. Observation: Pair/Triple of Nodes Level
3. Observation: Hyperedge Level
4. Generators
5. Conclusions

Observation: Egonet Level

How substantially do the hyperedges around a node overlap with each other in the real-world hypergraphs?



We quantitatively measure this by using density and overlapness.

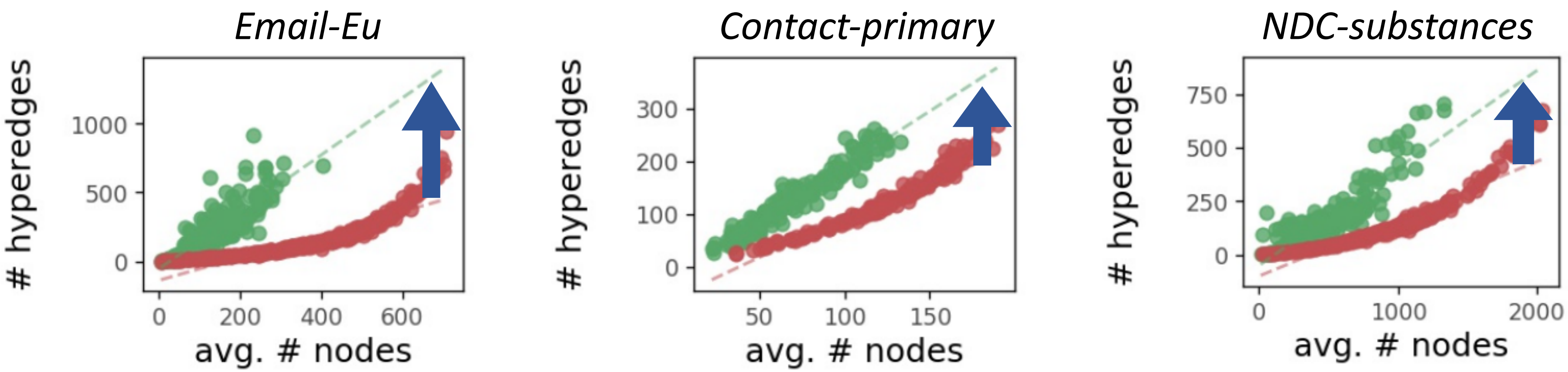
Density of Egonets

Egonet of a node ($\mathcal{E}$): set of hyperedges that contains the node

Density: $\rho(\mathcal{E}) \coloneqq \dfrac{|\mathcal{E}|}{|\bigcup_{e\in\mathcal{E}} e|}$

← number of hyperedges

← number of nodes


Email-Eu *Contact-primary* *NDC-substances*

● **Real-world hypergraph**

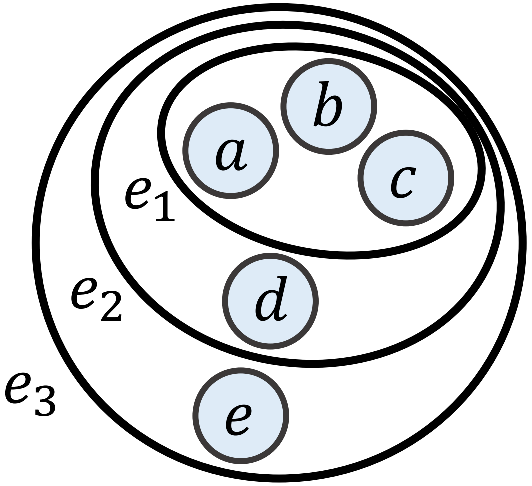
● **Randomized hypergraph**

slope
≈ average egonet density

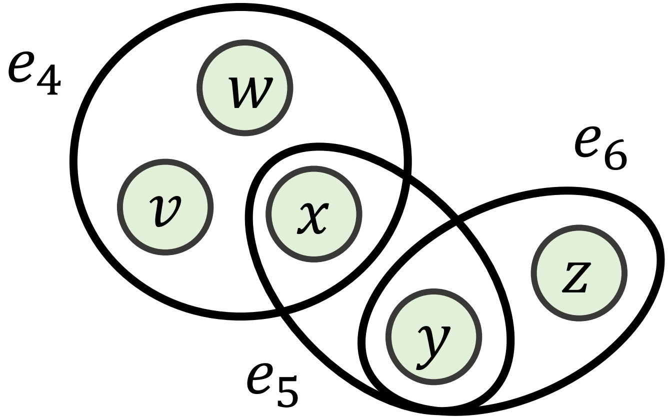
Egonets in real-world hypergraphs tend to have **higher density** than those in randomized ones.

Density of Egonets (cont.)

Does **density** fully capture the degree of overlaps of a set of hyperedges?



$$\mathcal{E}_1 = \{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$$



$$\mathcal{E}_2 = \{\boldsymbol{e_4}, \boldsymbol{e_5}, \boldsymbol{e_6}\}$$

Our intuition: $\mathcal{E}_1$ is more overlapped than $\mathcal{E}_2$.

Density: $\rho(\mathcal{E}_1) = \rho(\mathcal{E}_2) = \frac{3}{5}$

What is the principled measure for evaluating the degree of overlaps of a set of hyperedges?

Degree of Hyperedge Overlaps

Any reasonable measure f of the hyperedge overlaps should satisfy the following axioms.

Axiom 1: Number of Hyperedges

Consider two sets of hyperedges $\mathcal{E}$ and $\mathcal{E}'$.

If $\mathcal{E}$ and $\mathcal{E}'$ have the same (1) hyperedge sizes and (2) number of distinct nodes, but $\mathcal{E}$ have more hyperedges than $\mathcal{E}'$, then $f(\mathcal{E}) > f(\mathcal{E}')$.

Degree of Hyperedge Overlaps (cont.)

Any reasonable measure f of the hyperedge overlaps should satisfy the following axioms.

Axiom 2: Number of Distinct Nodes

Consider two sets of hyperedges $\mathcal{E}$ and $\mathcal{E}'$.

If $\mathcal{E}$ and $\mathcal{E}'$ have the same (1) number of hyperedges and (2) size distribution of hyperedges, but $\mathcal{E}$ have less distinct nodes than $\mathcal{E}'$, then $f(\mathcal{E}) > f(\mathcal{E}')$.

more overlap

Degree of Hyperedge Overlaps (cont.)

Any reasonable measure f of the hyperedge overlaps should satisfy the following axioms.

Axiom 3: Sizes of Hyperedges

Consider two sets of hyperedges $\mathcal{E}$ and $\mathcal{E}'$.

If $\mathcal{E}$ and $\mathcal{E}'$ have the same (1) number of distinct nodes and (2) number of hyperedges, but $\mathcal{E}$ have larger hyperedges than $\mathcal{E}'$, then $f(\mathcal{E}) > f(\mathcal{E}')$.

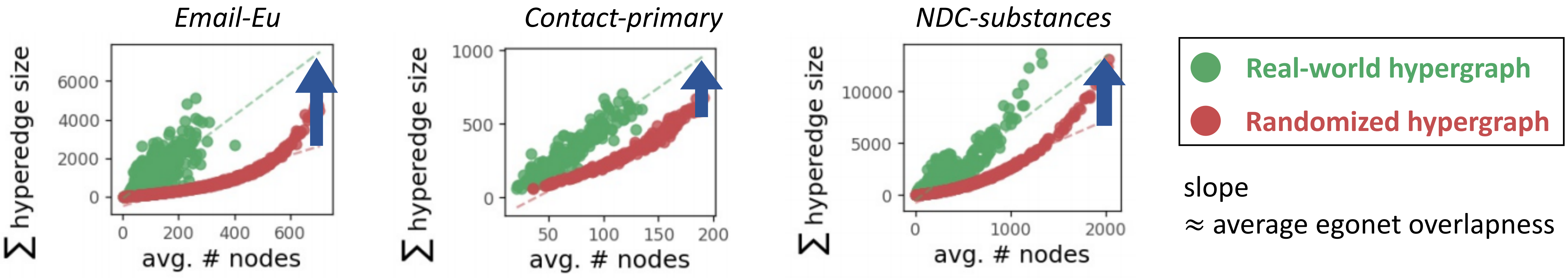
Overlapness of Egonets

Egonet of a node ($\mathcal{E}$): set of hyperedges that contains the node

Overlapness: $o(\mathcal{E}) := \frac{\sum_{e\in\mathcal{E}} |e|}{|\bigcup_{e\in\mathcal{E}} e|}$

← sum of the hyperedge sizes

← number of nodes

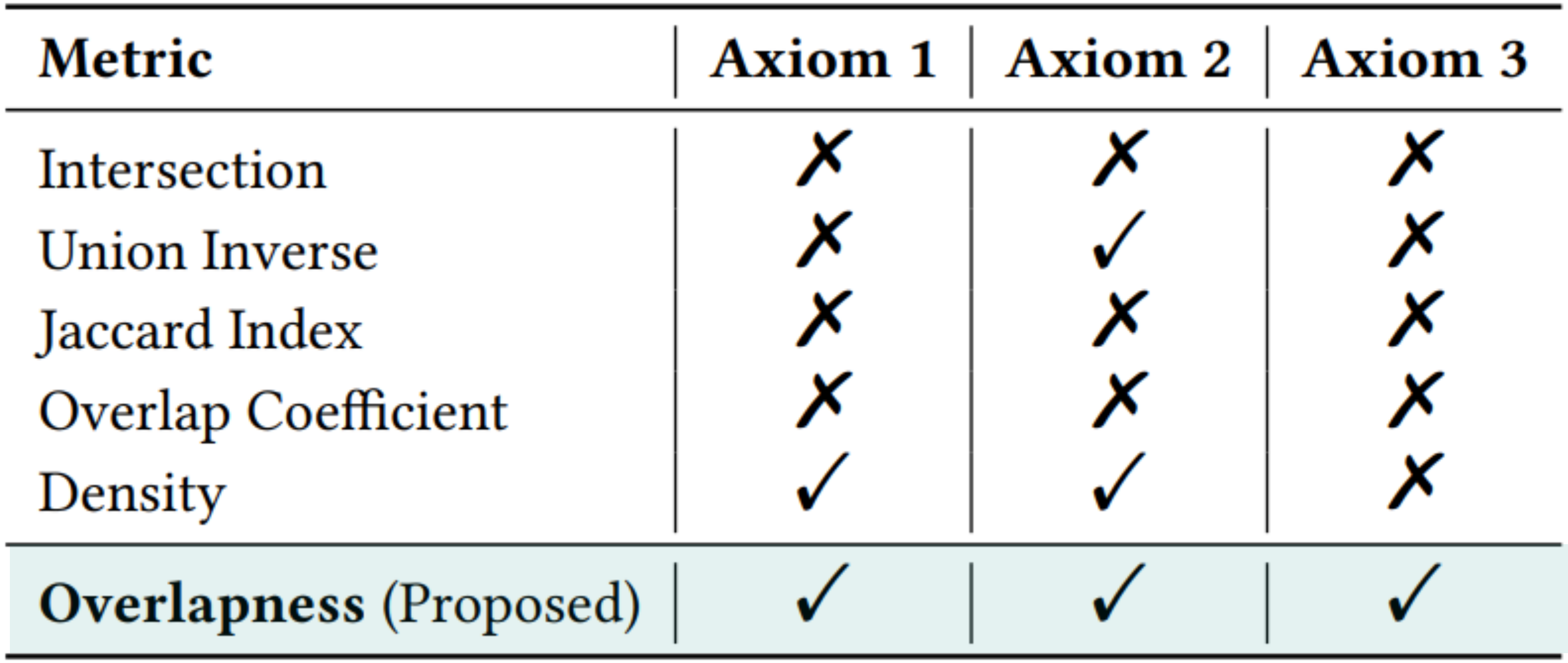


slope
≈ average egonet overlapness

Egonets in real-world hypergraphs tend to have **higher overlapness** than those in randomized ones.

Overlapness of Egonets (cont.)

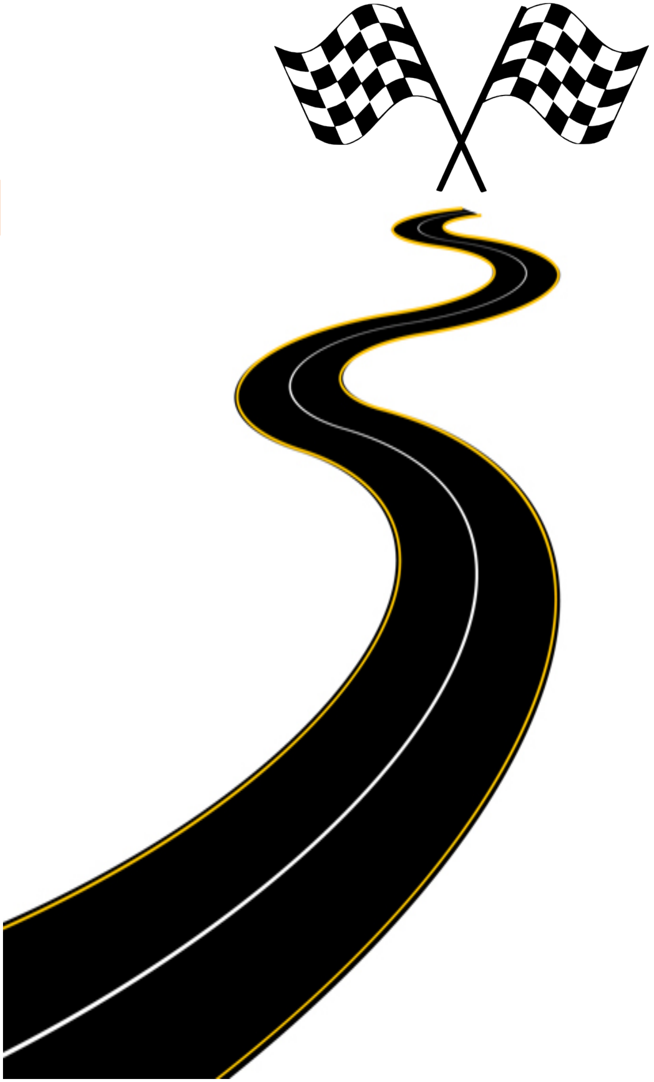
Overlapness satisfies all the axioms while others does not.



Metric	**Axiom 1**	**Axiom 2**	**Axiom 3**
Intersection	✗	✗	✗
Union Inverse	✗	✓	✗
Jaccard Index	✗	✗	✗
Overlap Coefficient	✗	✗	✗
Density	✓	✓	✗
Overlapness (Proposed)	✓	✓	✓

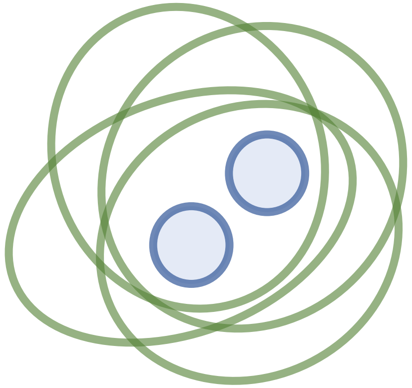
Roadmap

1. Observation: Egonet Level
2. **Observation: Pair/Triple of Nodes Level**
3. Observation: Hyperedge Level
4. Generators
5. Conclusions

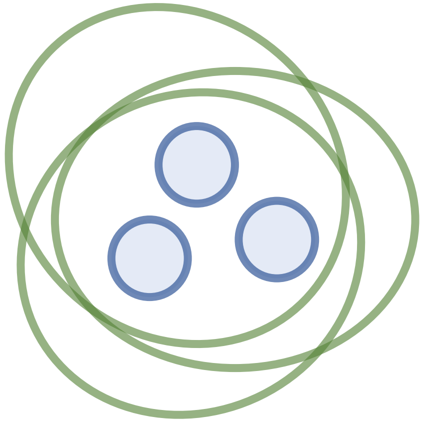


Observation: Pair/Triple of Nodes Level

How many hyperedges overlap at a pair or triple of nodes in the real-world hypergraphs?



Pair of nodes



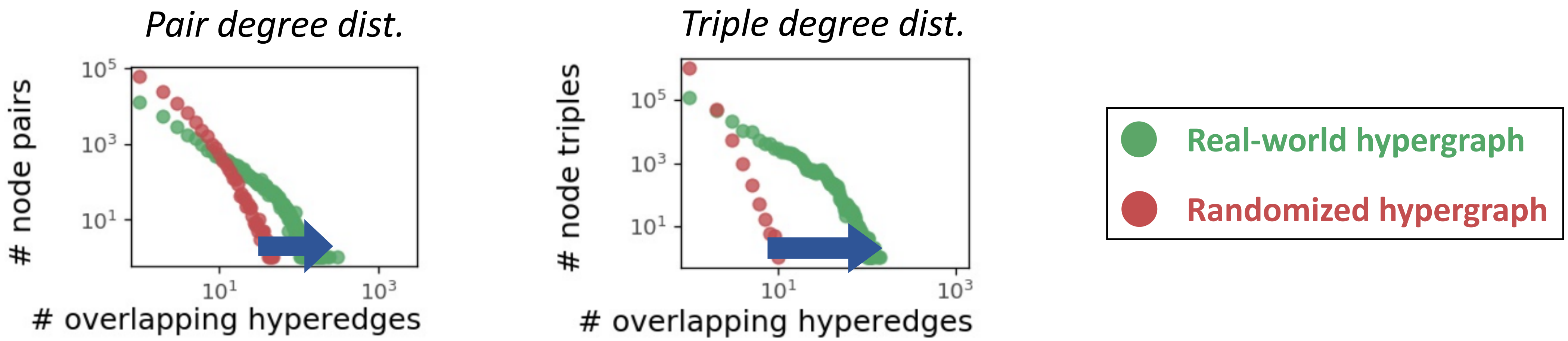
Triple of nodes

We extend the concept of degree to pairs and triples of nodes.

Degree of Node Pair/Triple

E_S: set of hyperedges overlapping at subset S of nodes

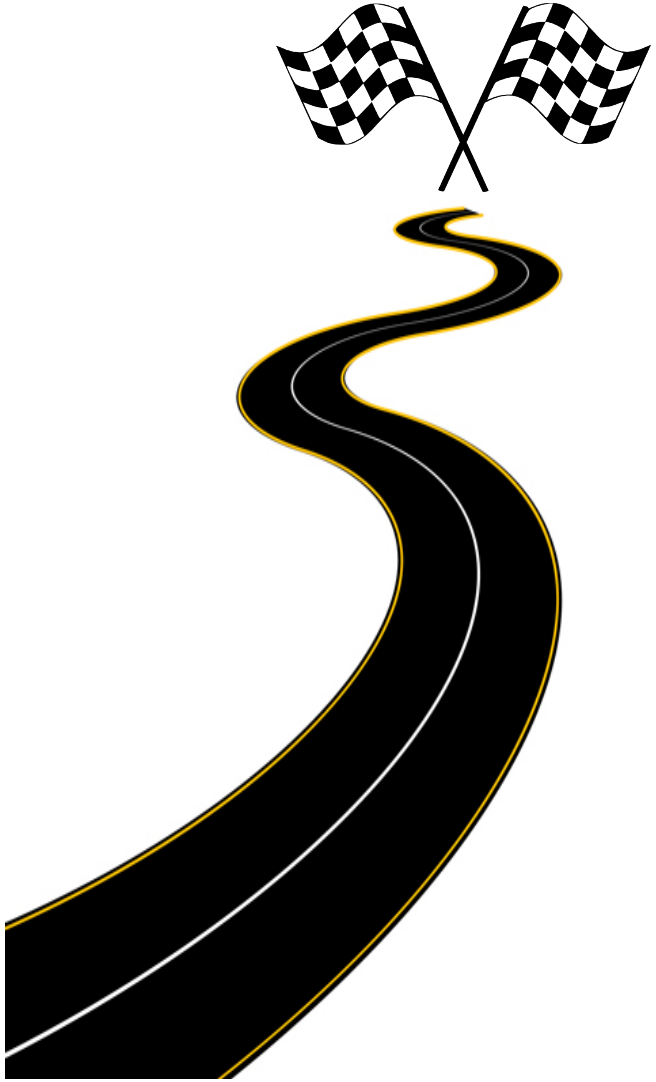
Consider the number of hyperedges overlapping at each **pair or triple of nodes**: $|E_{\{i,j\}}|$ and $|E_{\{i,j,k\}}|$.



The distribution of the number of hyperedges overlapping at each node pair & triple is **more skewed with a heavier tail** in real-world hypergraphs than in randomized ones.

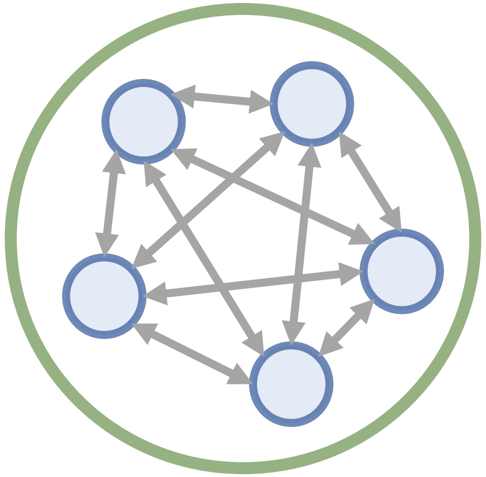
Roadmap

1. Observation: Egonet Level
2. Observation: Pair/Triple of Nodes Level
3. **Observation: Hyperedge Level**
4. Generators
5. Conclusions



Observation: Hyperedge Level

How structurally similar are nodes that form hyperedges together related to each other in the real-world hypergraphs?



We define a new measure to investigate the similarity.

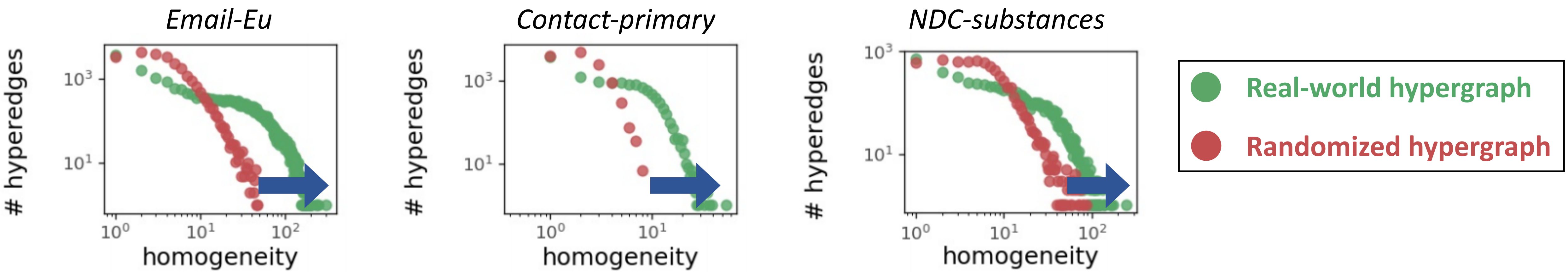
Homogeneity of a Hyperedge

How to measure the similarity among the nodes forming a hyperedge?

Homogeneity of a hyperedge:

$$homogeneity(e) \coloneqq \begin{cases} \frac{\sum_{\{u,v\}\in\binom{e}{2}} |E_{\{u,v\}}|}{\binom{|e|}{2}}, & if\ |e| > 1 \\ 0\,, & otherwise \end{cases}$$

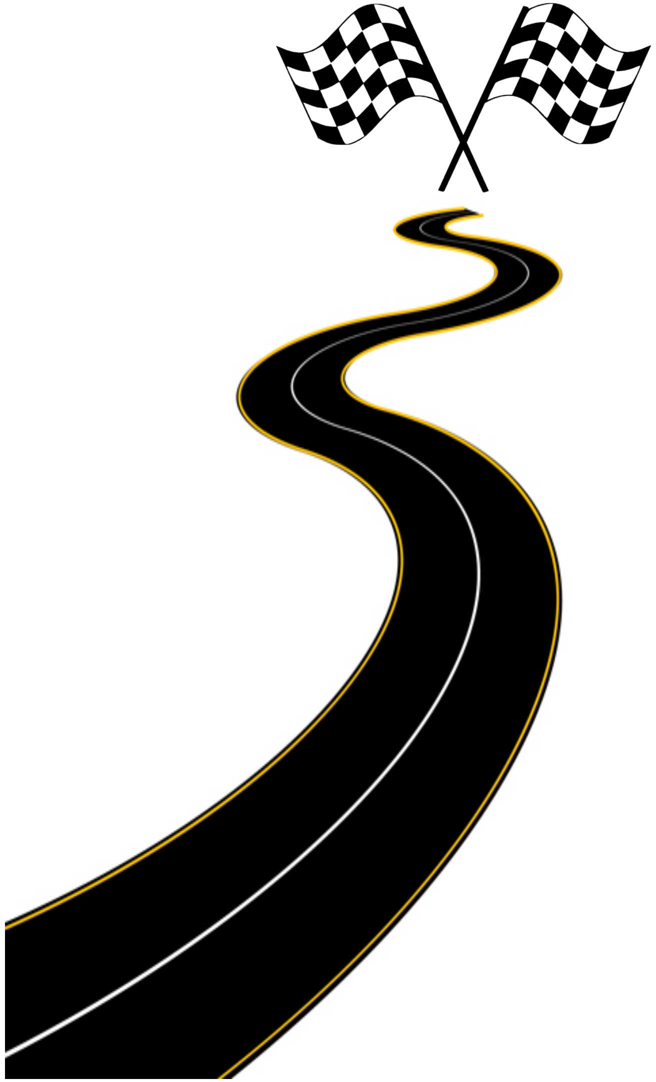
← Average number of hyperedges overlapping at all the pair of nodes in the hyperedge.



Hyperedges in real-world hypergraphs tend to have **higher homogeneity** than those in randomized ones.

Roadmap

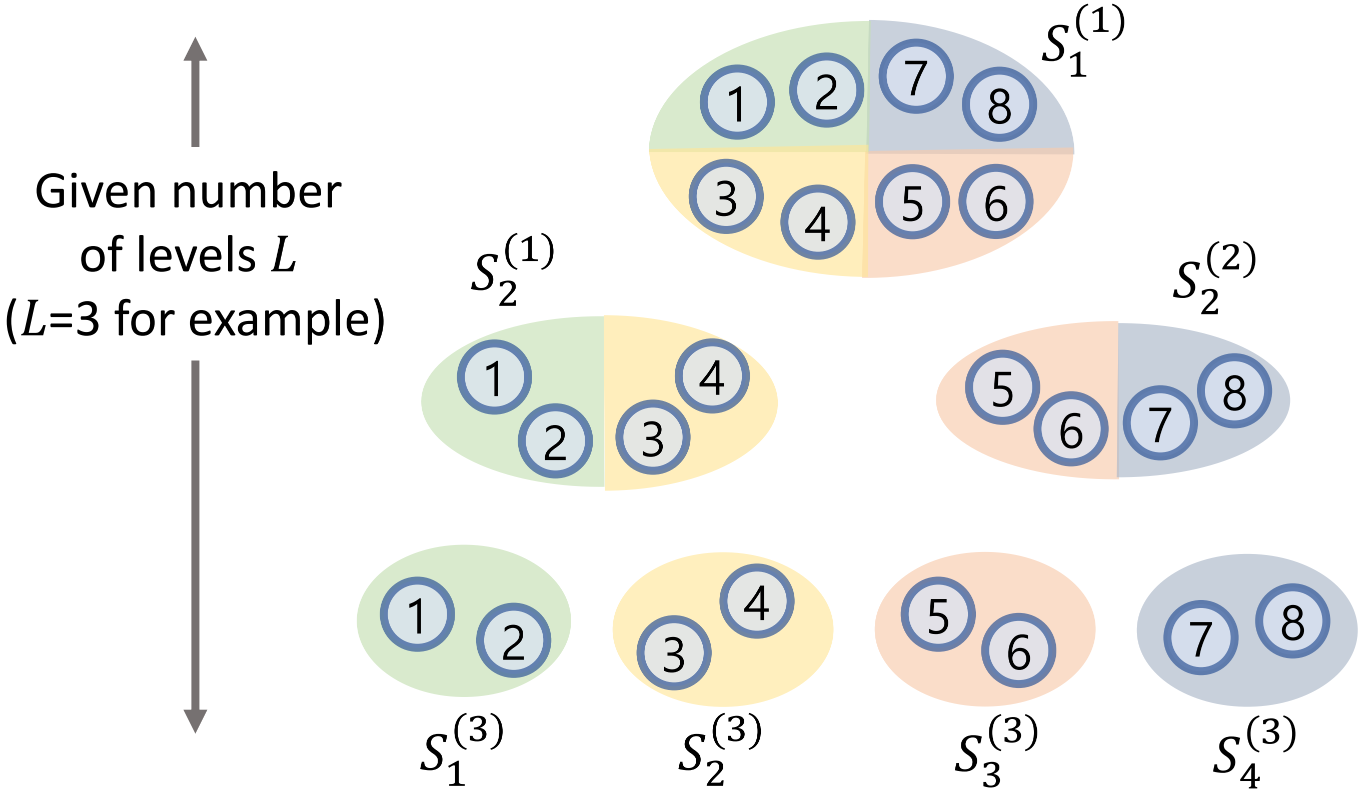
1. Observation: Egonet Level
2. Observation: Pair/Triple of Nodes Level
3. Observation: Hyperedge Level
4. **Generators**
5. Conclusions



Our Model: HyperLap

Main Idea: Extension of HyperCL

Step 1. Hierarchical Node Partitioning



Given number of levels L (L=3 for example)

At the lowest level 1, $\left|S_1^{(1)}\right| = |V|$

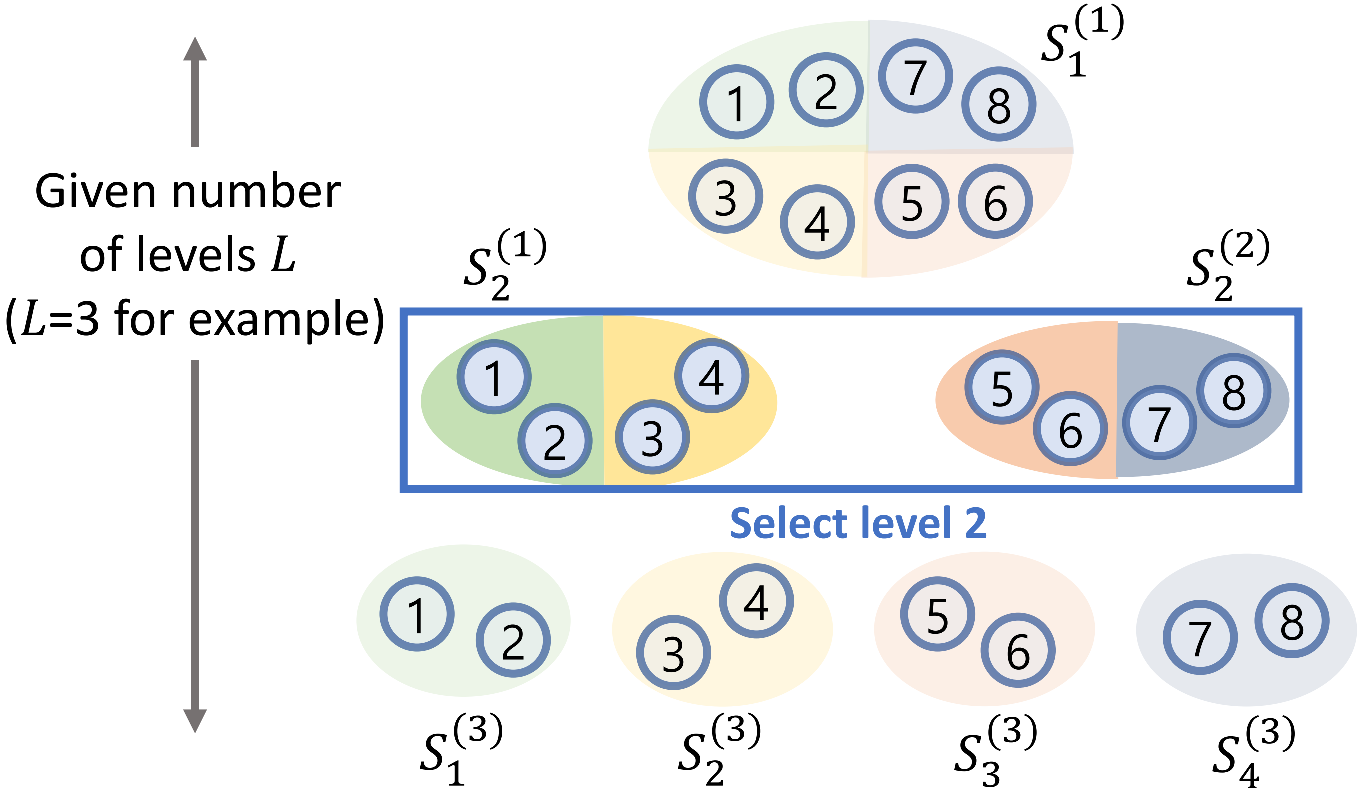
Construct lower level group $S_i^{(l)} = S_{2i-1}^{(l+1)} \cup S_{2i}^{(l+1)}$

Nodes are partitioned into 2^{L-1} disjoint groups

Our Model: HyperLap (cont.)

Main Idea: Extension of HyperCL

Step 2. Hyperedge Generation

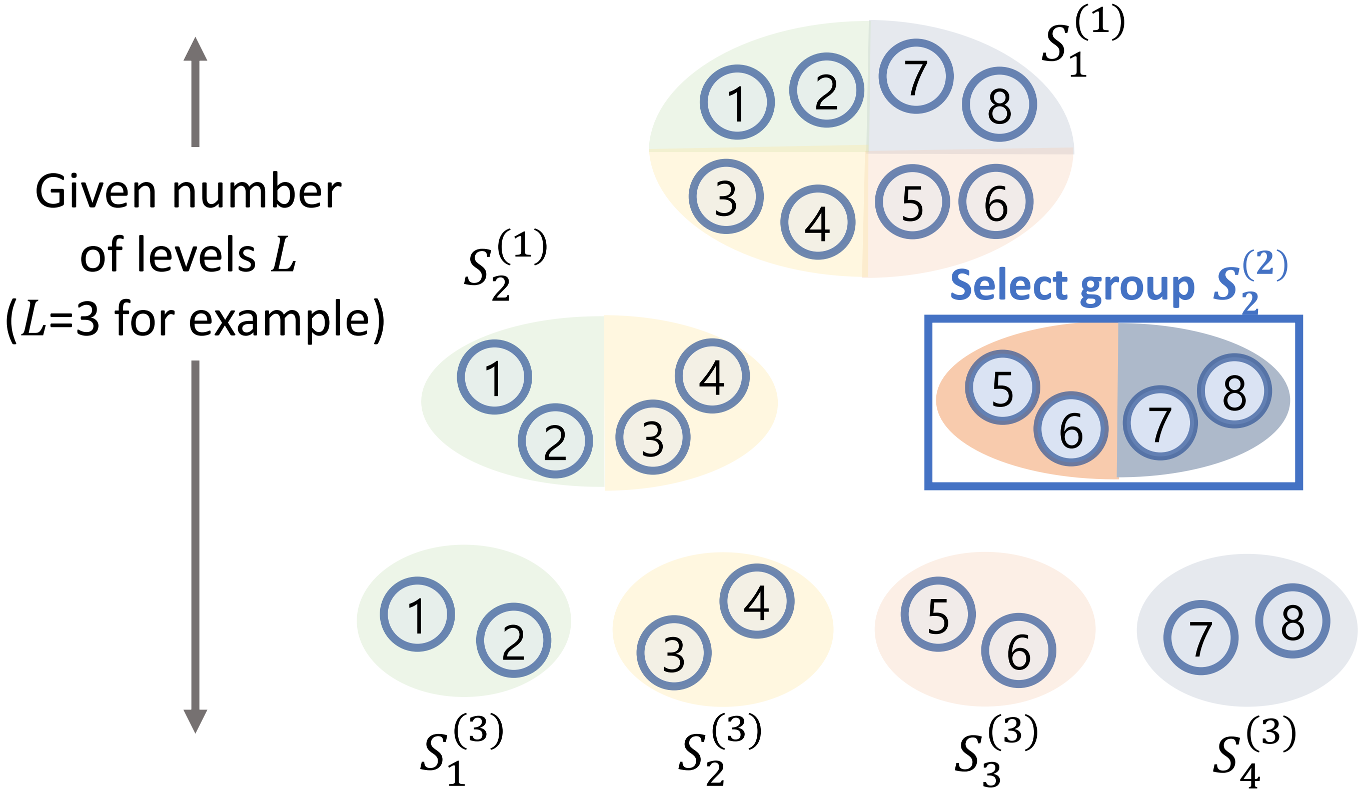


1. **Select a level with probability proportional to the given weight of each level $\{w_1, \ldots, w_L\}$.**
2. Select a group uniformly at random.
3. Sample nodes independently with probability proportional to the degree of each node to form a hyperedge.

Our Model: HyperLap (cont.)

Main Idea: Extension of HyperCL

Step 2. Hyperedge Generation

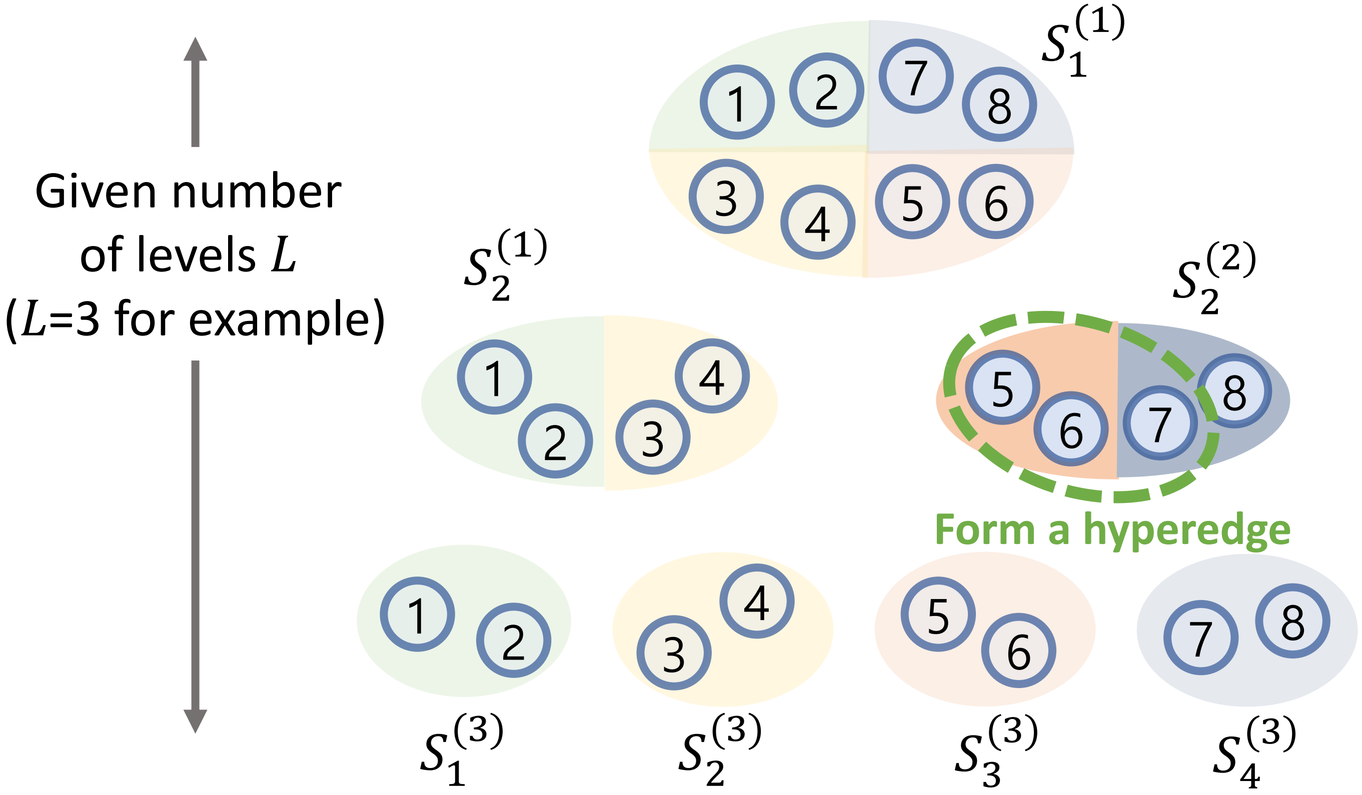


1. Select a level with probability proportional to the given weight of each level $\{w_1, \ldots, w_L\}$.
2. **Select a group uniformly at random.**
3. Sample nodes independently with probability proportional to the degree of each node to form a hyperedge.

Our Model: HyperLap (cont.)

Main Idea: Extension of HyperCL

Step 2. Hyperedge Generation



1. Select a level with probability proportional to the given weight of each level $\{w_1, \dots, w_L\}$.
2. Select a group uniformly at random.
3. **Sample nodes independently with probability proportional to the degree of each node to form a hyperedge.**

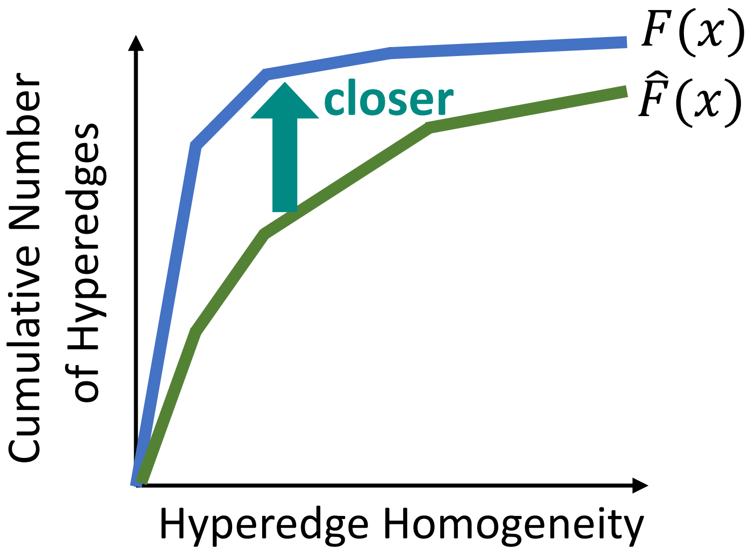
Our Model: HyperLap^{+}

HyperLap^{+} automatically tunes the parameters of HyperLap.

Objective: Minimize the hyperedge homogeneity distance $\boldsymbol{HHD(G, \hat{G})}$.

$$\boldsymbol{HHD(G, \hat{G})} = \max_{\boldsymbol{x}} \{ |\boldsymbol{F(x)} - \boldsymbol{\hat{F}(x)}| \}$$

where F and $\hat{F}$ are the cumulative hyperedge homogeneity distribution of hypergraphs G and $\hat{G}$, respectively.

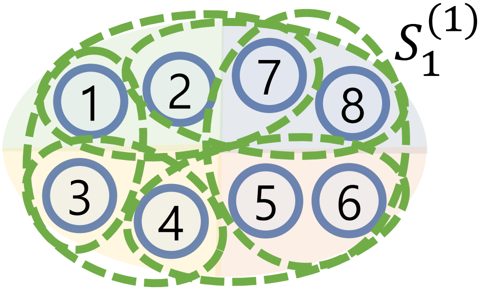


$$\min_{w_1, \ldots, w_L} HHD(G, \hat{G}) \text{ where } w_1 + \cdots + w_L = 1$$

Learnable parameters

Our Model: HyperLap^{+} (cont.)

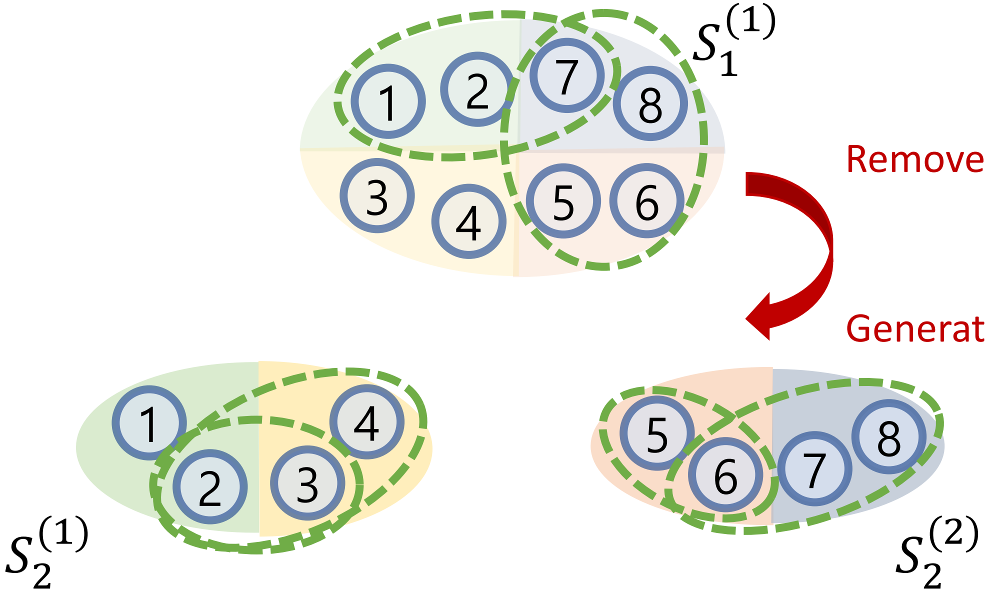
HyperLap^{+} automatically tunes the parameters of HyperLap.



1. Generate $|E|$ hyperedges at level 1.

Our Model: HyperLap$^+$ (cont.)

HyperLap$^+$ automatically tunes the parameters of HyperLap.



Search for an optimal q_1.

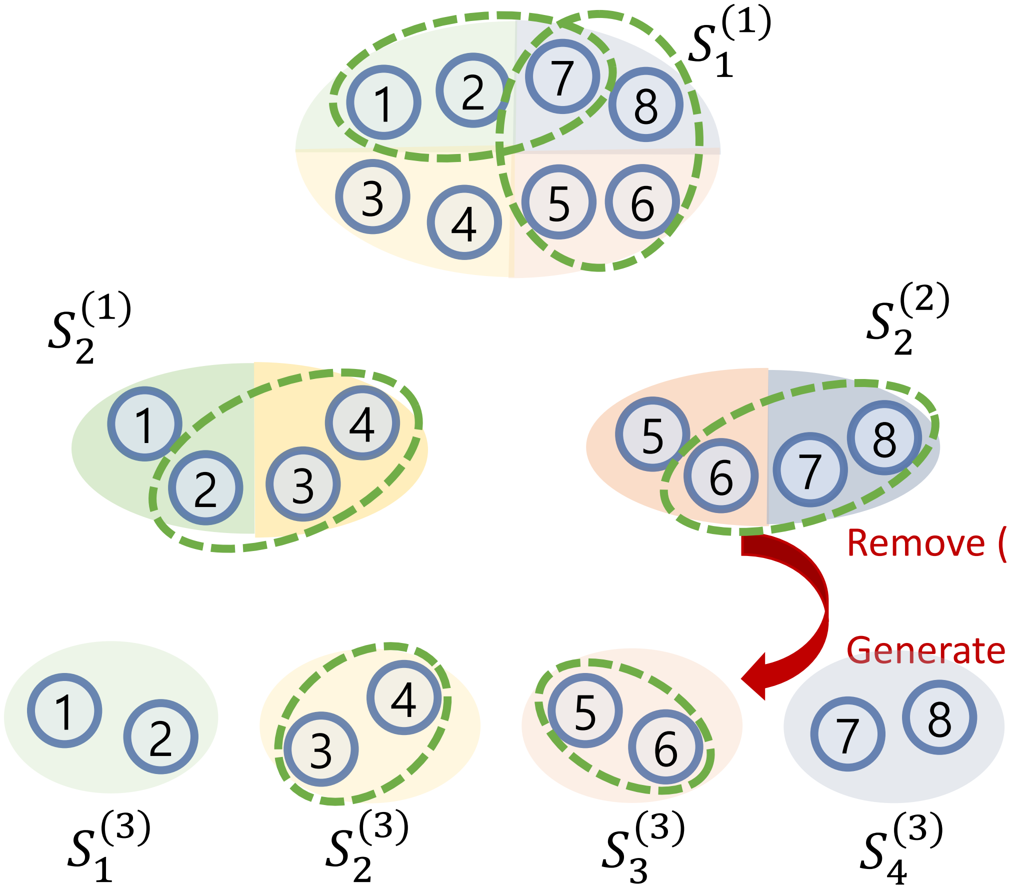
Remove $(q_1 \cdot 100)$% hyperedges at level 1

Generate $(q_1 \cdot 100)$% hyperedges at level 2

2. Replace $(q_1 \cdot 100)$% hyperedges at level 1 to level 2.

Our Model: HyperLap$^+$ (cont.)

HyperLap$^+$ automatically tunes the parameters of HyperLap.



Search for an optimal q_2.

Remove ($\boldsymbol{q_2} \cdot 100$)% hyperedges at level 2

Generate ($\boldsymbol{q_2} \cdot 100$)% hyperedges at level 3

Until Convergence

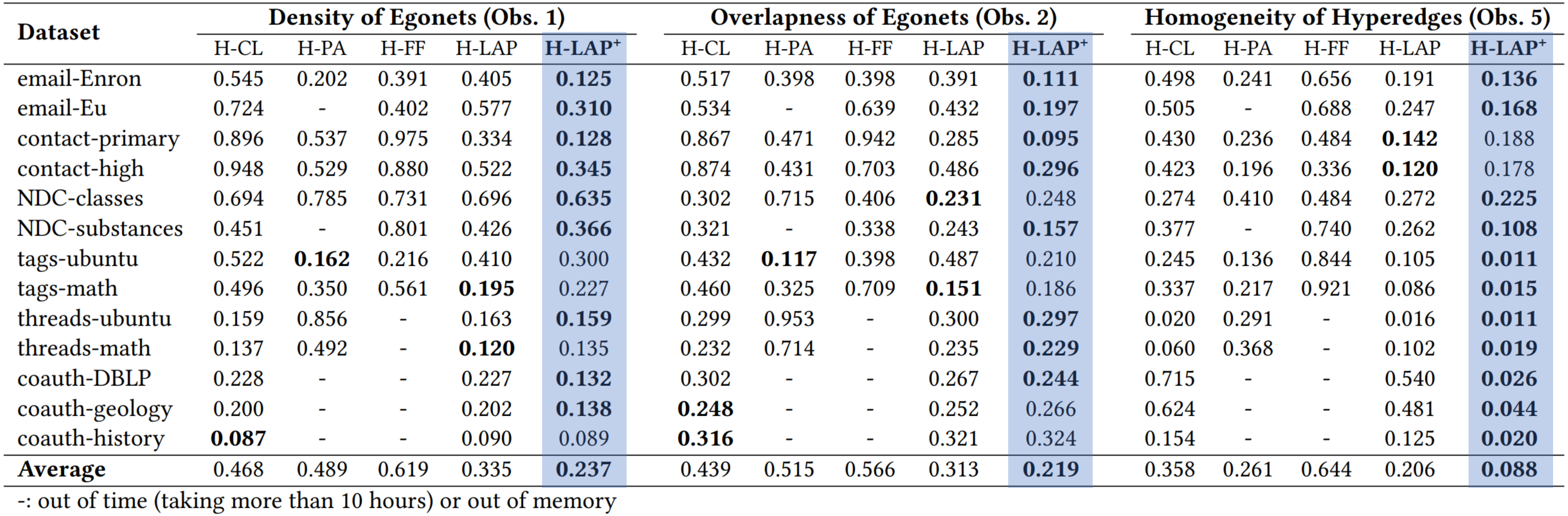
3. Replace $(q_2 \cdot 100)$% hyperedges at level 2 to level 3.

Evaluation of Our Model

HyperLap$^+$ reproduces most accurately the distributions of

(1) egonet density (2) egonet overlapness (3) hyperedge homogeneity

Measure: the similarity between the distributions derived from the real-world and the generated hypergraph by Kolmogorov-Smirnov D-statistics.

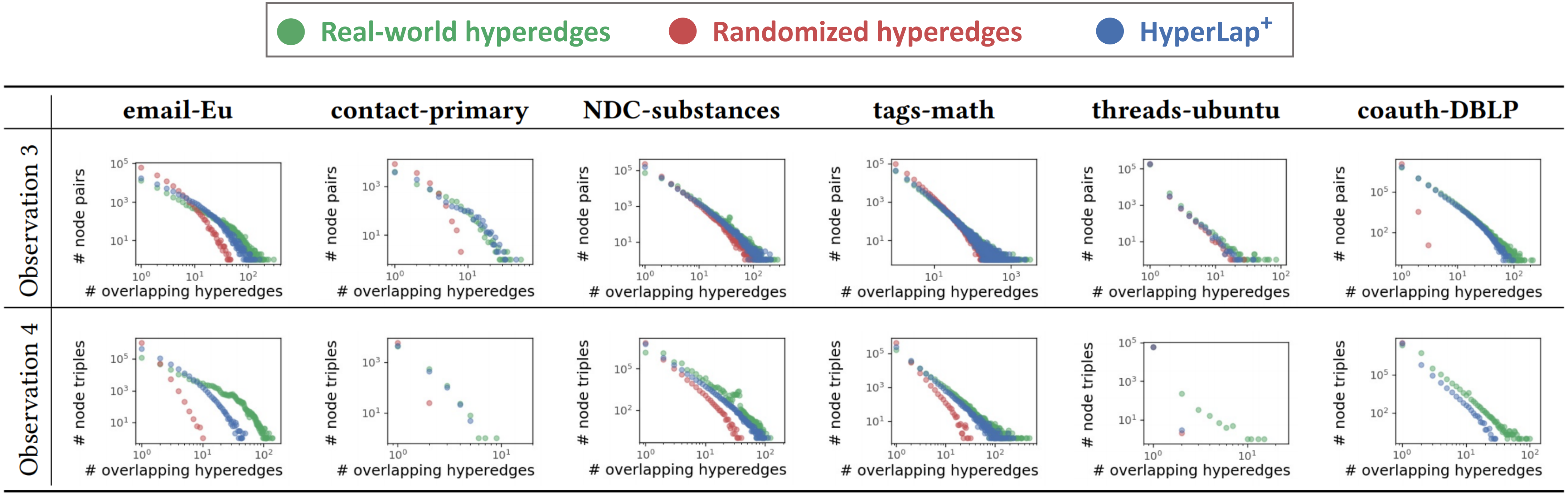


Dataset	**Density of Egonets (Obs. 1)**					**Overlapness of Egonets (Obs. 2)**					**Homogeneity of Hyperedges (Obs. 5)**				
	H-CL	H-PA	H-FF	H-LAP	**H-LAP$^+$**	H-CL	H-PA	H-FF	H-LAP	**H-LAP$^+$**	H-CL	H-PA	H-FF	H-LAP	**H-LAP$^+$**
email-Enron	0.545	0.202	0.391	0.405	**0.125**	0.517	0.398	0.398	0.391	**0.111**	0.498	0.241	0.656	0.191	**0.136**
email-Eu	0.724	-	0.402	0.577	**0.310**	0.534	-	0.639	0.432	**0.197**	0.505	-	0.688	0.247	**0.168**
contact-primary	0.896	0.537	0.975	0.334	**0.128**	0.867	0.471	0.942	0.285	**0.095**	0.430	0.236	0.484	**0.142**	0.188
contact-high	0.948	0.529	0.880	0.522	**0.345**	0.874	0.431	0.703	0.486	**0.296**	0.423	0.196	0.336	**0.120**	0.178
NDC-classes	0.694	0.785	0.731	0.696	**0.635**	0.302	0.715	0.406	**0.231**	0.248	0.274	0.410	0.484	0.272	**0.225**
NDC-substances	0.451	-	0.801	0.426	**0.366**	0.321	-	0.338	0.243	**0.157**	0.377	-	0.740	0.262	**0.108**
tags-ubuntu	0.522	**0.162**	0.216	0.410	0.300	0.432	**0.117**	0.398	0.487	0.210	0.245	0.136	0.844	0.105	**0.011**
tags-math	0.496	0.350	0.561	**0.195**	0.227	0.460	0.325	0.709	**0.151**	0.186	0.337	0.217	0.921	0.086	**0.015**
threads-ubuntu	0.159	0.856	-	0.163	**0.159**	0.299	0.953	-	0.300	**0.297**	0.020	0.291	-	0.016	**0.011**
threads-math	0.137	0.492	-	**0.120**	0.135	0.232	0.714	-	0.235	**0.229**	0.060	0.368	-	0.102	**0.019**
coauth-DBLP	0.228	-	-	0.227	**0.132**	0.302	-	-	0.267	**0.244**	0.715	-	-	0.540	**0.026**
coauth-geology	0.200	-	-	0.202	**0.138**	**0.248**	-	-	0.252	0.266	0.624	-	-	0.481	**0.044**
coauth-history	**0.087**	-	-	0.090	0.089	**0.316**	-	-	0.321	0.324	0.154	-	-	0.125	**0.020**
Average	0.468	0.489	0.619	0.335	**0.237**	0.439	0.515	0.566	0.313	**0.219**	0.358	0.261	0.644	0.206	**0.088**

-: out of time (taking more than 10 hours) or out of memory

Evaluation of Our Model (cont.)

HyperLap⁺ reproduces the heavy-tailed distributions of the number of overlapping hyperedges at each *pair* and each *triple* of nodes accurately.



Evaluation of Our Model (cont.)

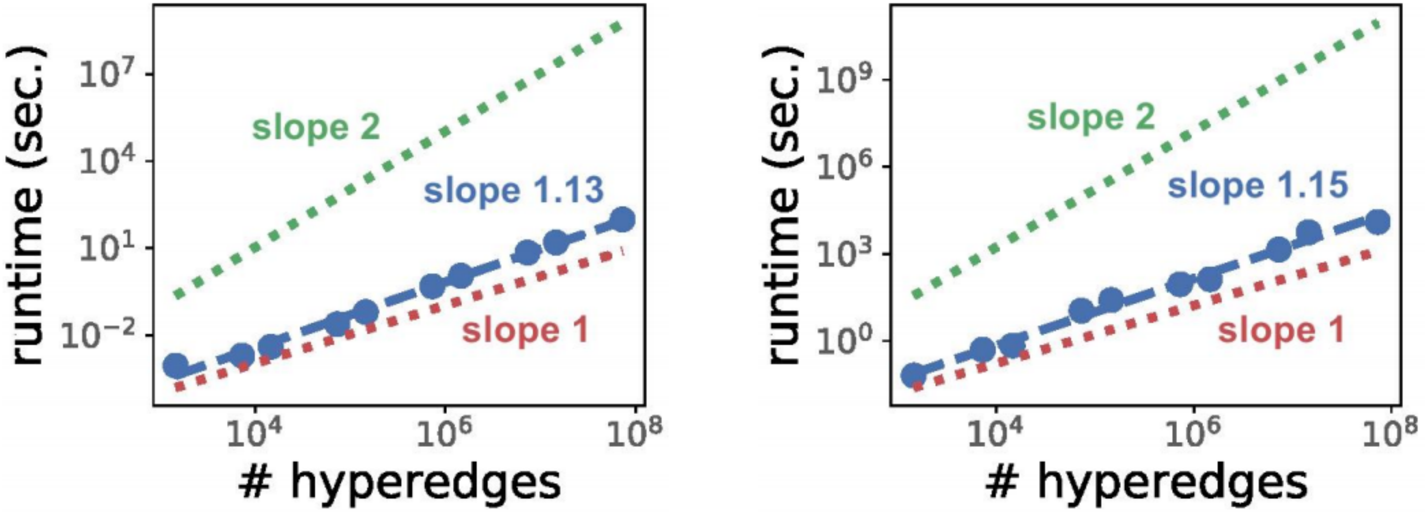
HyperLap^{+} reproduces the heavy-tailed distributions of the number of overlapping hyperedges at each *pair* and each *triple* of nodes accurately.

Dataset	**Pair of Nodes (Obs. 3)**								**Triple of Nodes (Obs. 4)**							
	Distance from Real (D-statistics)					Heavy-tail Test			Distance from Real (D-statistics)					Heavy-tail Test		
	H-CL	H-PA	H-FF	H-LAP	**H-LAP^{+}**	pw	tpw	logn	H-CL	H-PA	H-FF	H-LAP	**H-LAP^{+}**	pw	twp	logn
email-Enron	0.143	**0.056**	0.217	0.075	0.139	-2.37	-0.29	-1.53	0.089	0.295	0.136	**0.061**	0.072	-0.22	**0.38**	**0.24**
email-Eu	0.225	-	0.352	0.162	**0.066**	**0.24**	**2.75**	**2.53**	0.480	-	0.516	0.337	**0.206**	**0.41**	**2.11**	**1.96**
contact-primary	0.196	0.062	0.223	0.070	**0.051**	**9.53**	**15.74**	**13.92**	0.137	0.061	0.110	0.053	**0.031**	-1.86	-1.27	**1.23**
contact-high	0.277	**0.062**	0.141	0.127	0.067	-3.09	-0.95	-0.06	0.210	**0.131**	0.182	0.182	0.193	-3.95	-	**0.50**
NDC-classes	0.273	0.197	0.196	0.246	**0.172**	**12.15**	**14.42**	**14.04**	0.376	**0.167**	0.405	0.349	0.286	**3.22**	**7.92**	**7.34**
NDC-substances	0.272	-	0.244	0.251	**0.202**	**33.69**	**40.13**	**39.66**	0.521	-	0.591	0.492	**0.453**	**45.30**	**55.38**	**54.99**
tags-ubuntu	0.091	**0.019**	0.182	0.034	0.033	**42.33**	**43.70**	**43.55**	0.148	0.067	0.191	**0.020**	0.074	**14.25**	**15.57**	**15.43**
tags-math	0.095	0.066	0.278	0.073	**0.011**	**42.75**	**45.60**	**45.41**	0.209	**0.053**	0.286	0.113	0.079	**21.38**	**23.12**	**22.99**
threads-ubuntu	0.011	0.137	-	**0.008**	0.009	**1.28**	**1.75**	**1.75**	**0.004**	0.130	-	**0.004**	**0.004**	-1,346	-1.72	-1.72
threads-math	0.041	0.163	-	**0.014**	0.033	**15.79**	**16.66**	**16.52**	0.006	0.138	-	**0.001**	0.005	-1.49	-0.98	**0.96**
coauth-DBLP	0.224	-	-	0.191	**0.032**	**55.86**	**74.95**	**73.45**	0.215	-	-	0.214	**0.192**	**2.87**	**6.73**	**6.46**
coauth-geology	0.178	-	-	0.157	**0.040**	**31.13**	**45.08**	**44.06**	0.086	-	-	0.085	**0.069**	-0.10	**1.10**	**0.84**
coauth-history	0.033	-	-	0.030	**0.009**	**1.74**	**1.77**	**1.63**	**0.001**	-	-	**0.001**	**0.001**	-0.86	-	**0.57**
Average	0.158	0.095	0.229	0.110	**0.066**				0.193	0.130	0.302	0.147	**0.128**			

-: out of time (taking more than 10 hours) or out of memory

Scalability of Our Model

HyperLap and **HyperLap⁺** scale near linearly with the size of the considered hypergraph.

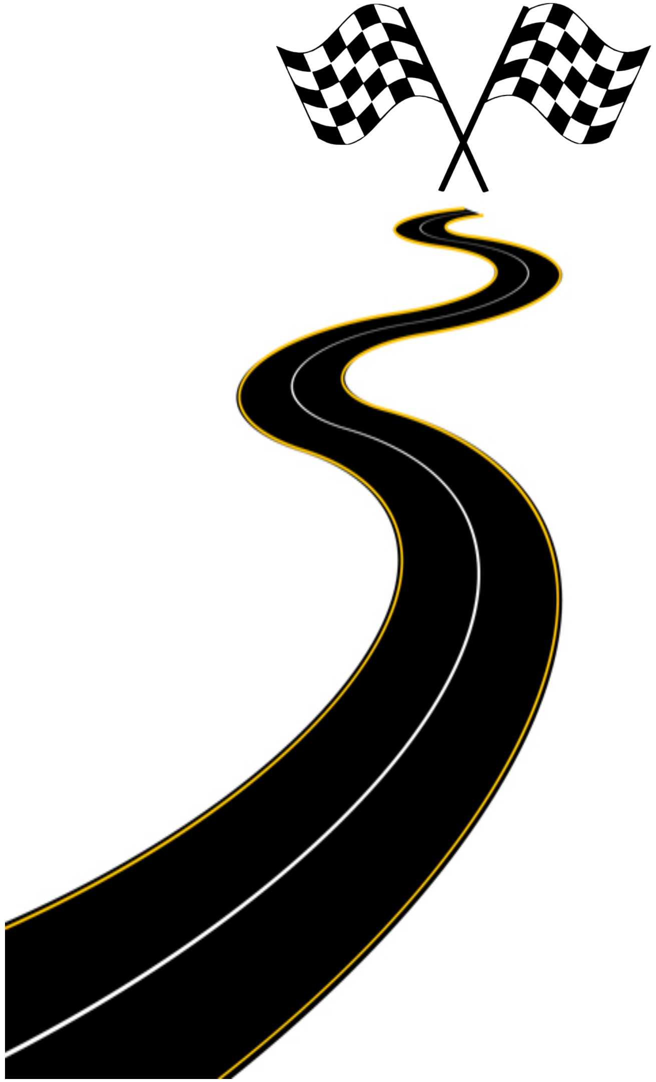


(a) HYPERLAP (generation)

(b) HYPERLAP⁺ (generation & fitting)

Roadmap

1. Observation: Egonet Level
2. Observation: Pair/Triple of Nodes Level
3. Observation: Hyperedge Level
4. Generators
5. **Conclusions**



Conclusions

Our contributions in this work:

✓ Observations in Real-world Hypergraphs

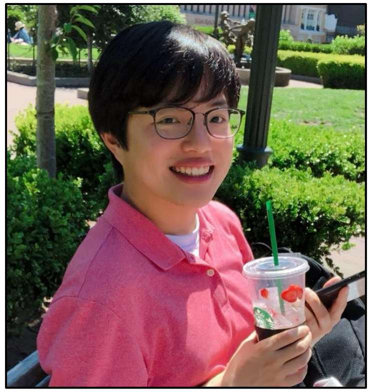
1. Egonet Level Observation
2. Pair/Triple of Nodes Level Observation
3. Hyperedge Level Observation

✓ Novel Measures

✓ Realistic Generative Models (HyperLap & HyperLap$^+$)

The code and datasets used in the paper are available at
https://github.com/young917/www21-hyperlap

How Do Hyperedges Overlap in Real-World Hypergraphs? Patterns, Measures, and Generators



Geon Lee*
Minyoung Choe*
Kijung Shin