This is a preliminary version of the slides that will be used for tutorials.

The slides will be revised to reflect recent studies and recommended improvements.

The final version may differ from this version.

Mining of Real-world Hypergraphs: Concepts, Patterns, and Generators

Part II. Static Structural Patterns

KAIST
Carnegie Mellon University

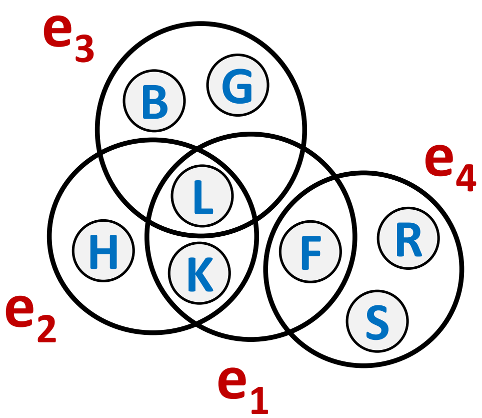
Geon Lee

Jaemin Yoo

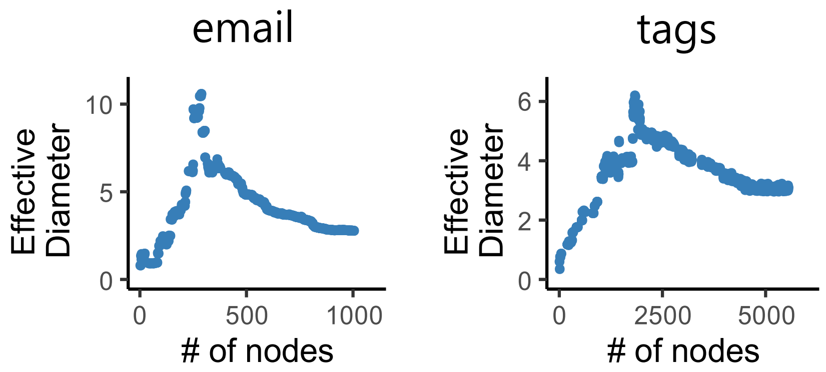
Kijung Shin

Part 1. Static Structural Patterns

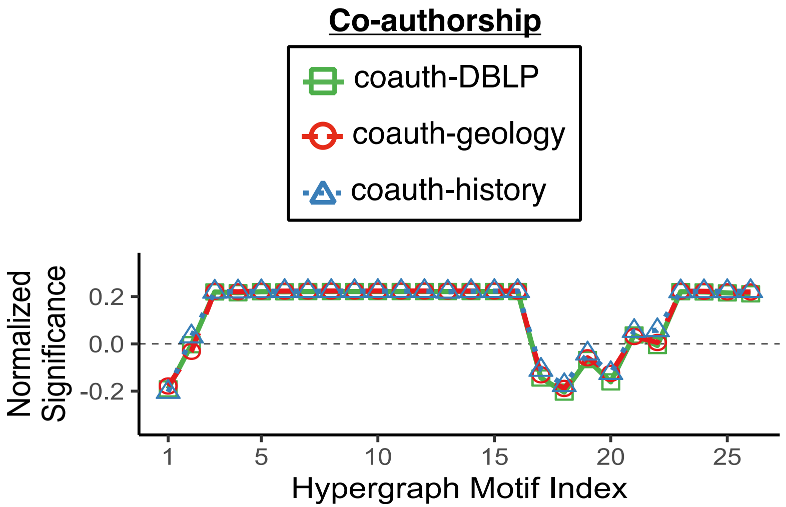
*“Given a **static** hypergraph, how can we analyze its structure?”*



Input Hypergraph



Basic Patterns (Part 1-1)



Advanced Patterns (Part 1-2)

Roadmap

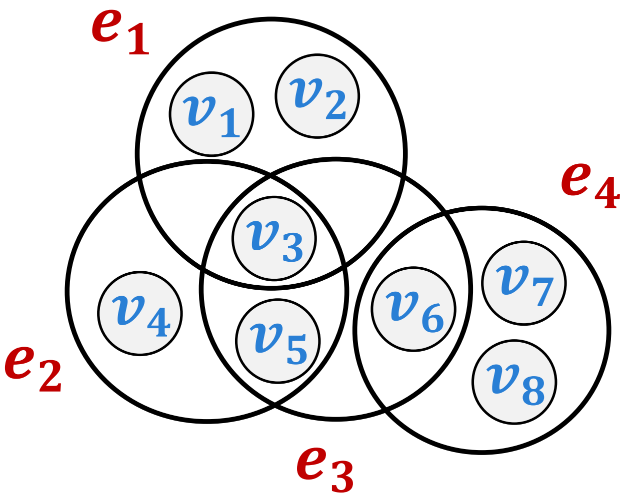
- **Part 1. Static Structural Patterns**
  - Basic Patterns **<<**
  - Advanced Patterns
- **Part 2. Dynamic Structural Patterns**
  - Basic Patterns
  - Advanced Patterns
- **Part 3. Generative Models**
  - Static Hypergraph Generator
  - Dynamic Hypergraph Generator

Part 1-1. Basic Static Structural Patterns

		Part 1. Static Patterns	Part 2. Dynamic Patterns
Basic Patterns	Node-Level	DYHS20, KKS20, LCS21	BKT18, CS22
	Hyperedge-Level	KKS20, LCS21	BKT18, CBLK21, LS21
	Hypergraph-Level	BASJK18, DYHS20, KKS20	KKS20
Advanced Patterns	Sub-hypergraph-Level	BASJK18, LMMB22, LKK20, LCS21	BASJK18, CJ21, LS21

Background

- **Degree** of a node v is the number of hyperedges containing v.
- **Size** of a hyperedge e is the number of nodes in e.



Hypergraph

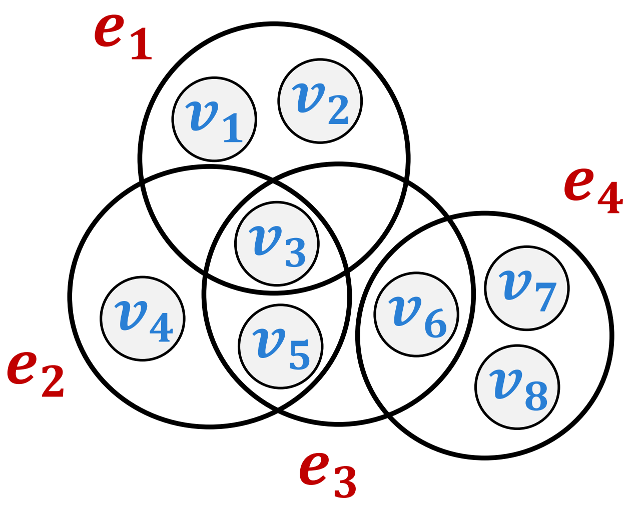
Degree of v_5 is 2.

Size of e_2 is 3.

Example

Background (cont.)

- **Incidence matrix $H = \{0,1\}^{|V|\times|E|}$** of a hypergraph $\mathcal{G} = (V, E)$ is:



Hypergraph

	e_1	e_2	e_3	e_4
v_1	1	0	0	0
v_2	1	0	0	0
v_3	1	1	1	0
v_4	0	1	0	0
v_5	0	1	1	0
v_6	0	0	1	1
v_7	0	0	0	1
v_8	0	0	0	1

Incidence matrix

$$H[i][j] = \begin{cases} \mathbf{1}, & \text{if } v_i \in e_j \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

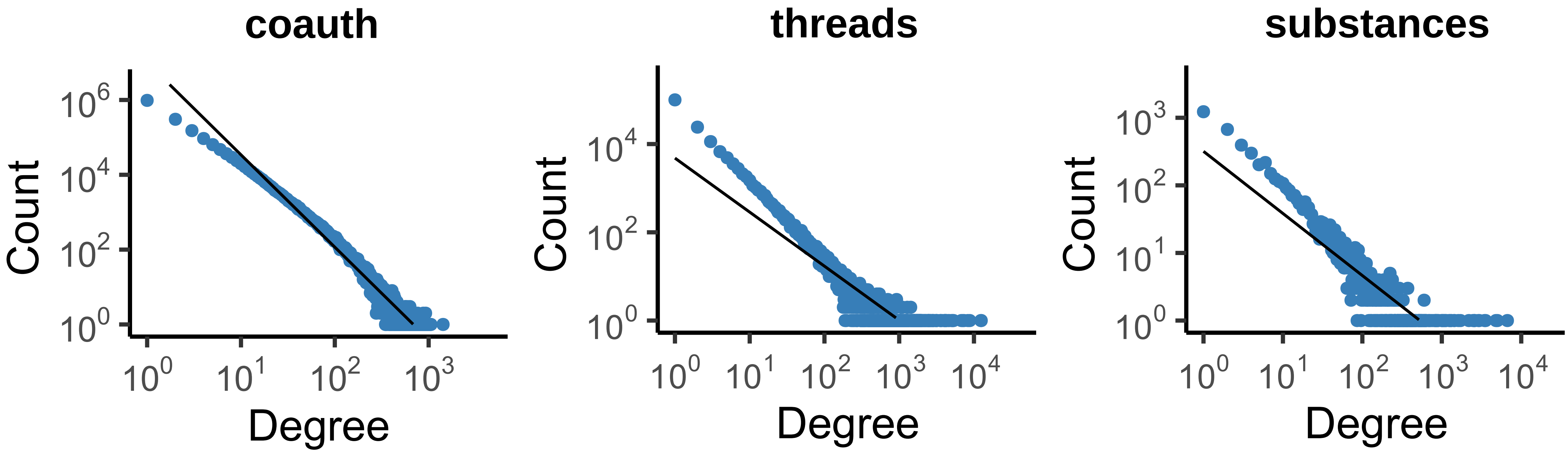
KKS20: Four Basic Static Patterns

- **P1.** Degree distribution
- **P2.** Hyperedge size distribution
- **P3.** Intersection size distribution
- **P4.** Singular value distribution

	Node	P1	
	Hyperedge	P2, P3	
	Hypergraph	P4	
	Sub-hypergraph		

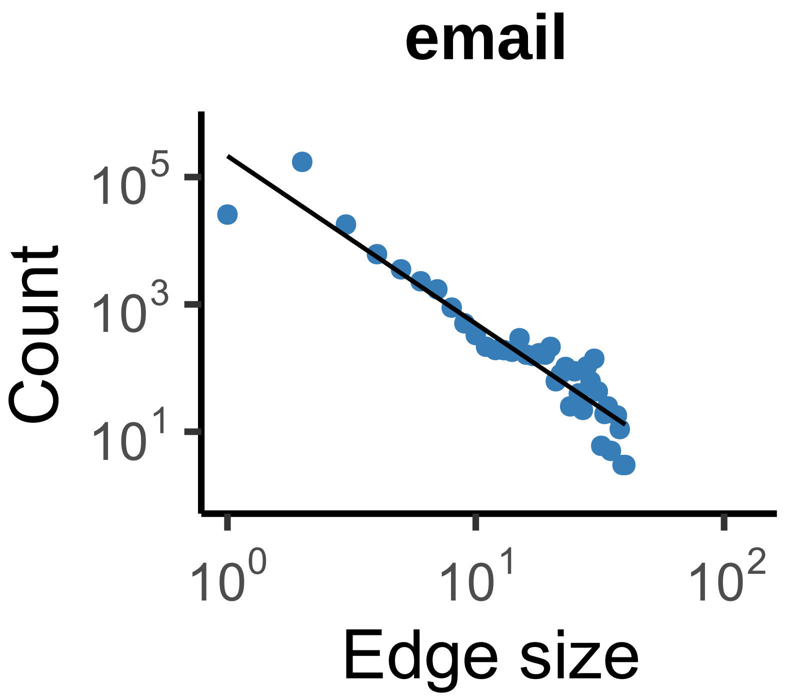
Degree Distribution

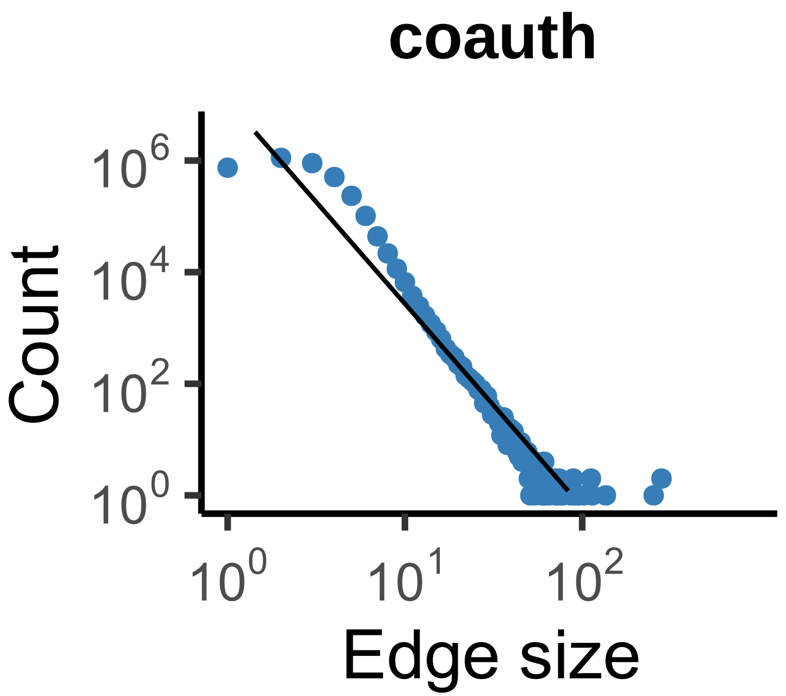
- Degree distributions of real-world hypergraphs are **heavy-tailed**.

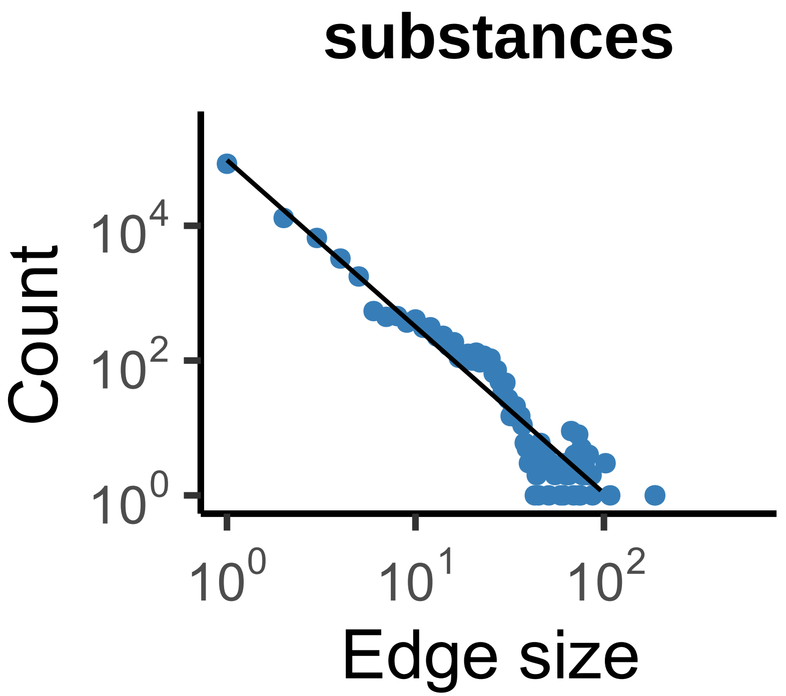


Hyperedge Size Distribution

- Size distributions of real-world hypergraphs are **heavy-tailed**.

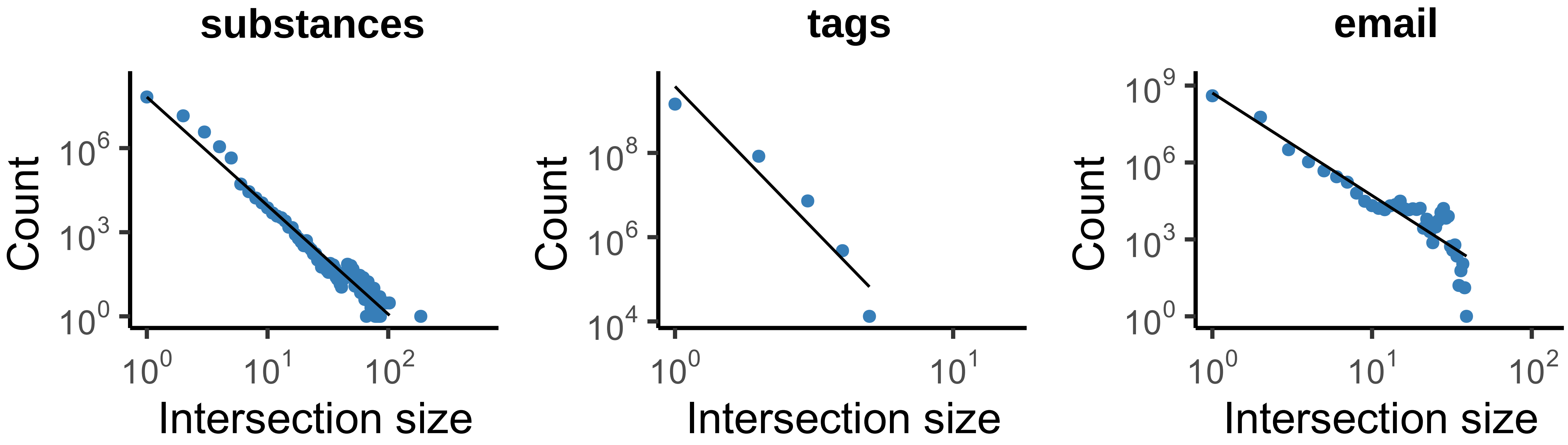






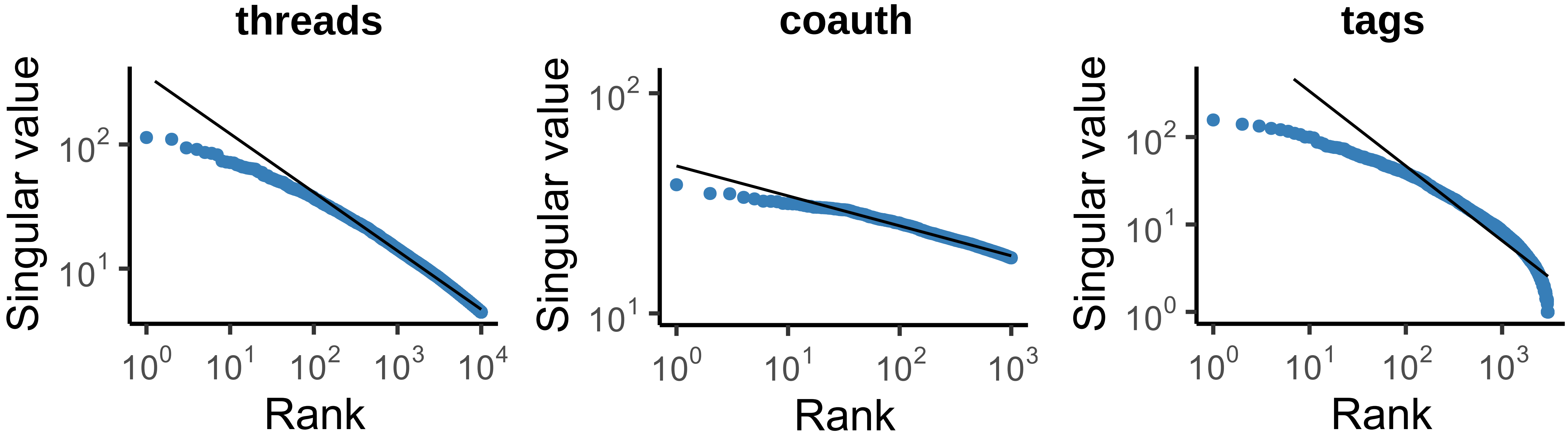
Intersection Size Distribution

- Intersection size distributions of real-world hypergraphs are **heavy-tailed**.



Singular Value Distribution

- Singular values of the incidence matrices of real-world hypergraphs are **skewed** and **heavy-tailed**.



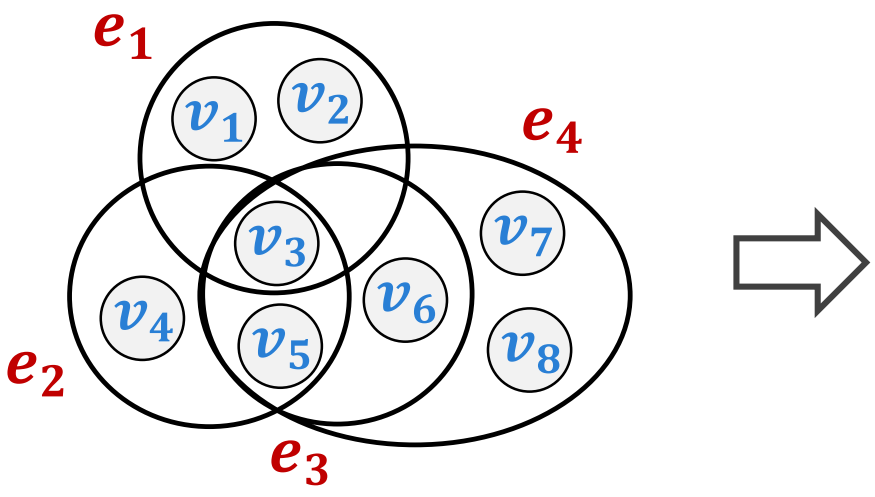
LCS21: Two Basic Static Patterns

- **P1.** Pair/triple-of-nodes degree distribution
- **P2.** Hyperedge homogeneity distribution

	Node	P1	
	Hyperedge	P2	
	Hypergraph		
	Sub-hypergraph		

Pair/Triple Degree Distribution

- **Degree of pair/triple of nodes** is the number of hyperedges overlapping the nodes.



Hypergraph

Degree of $\{v_3, v_5\}$ is 3.

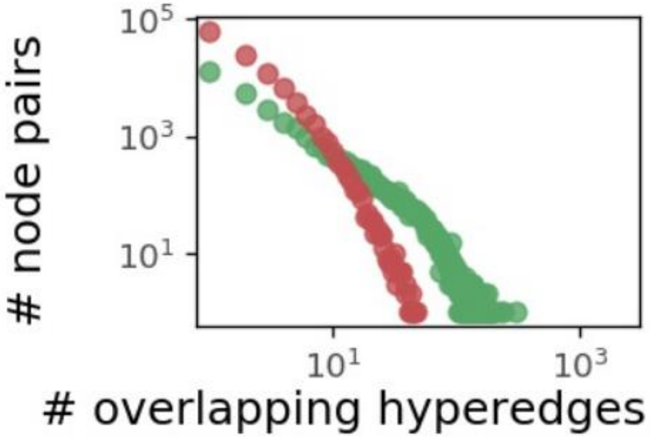
Degree of $\{v_3, v_5, v_6\}$ is 2.

Example

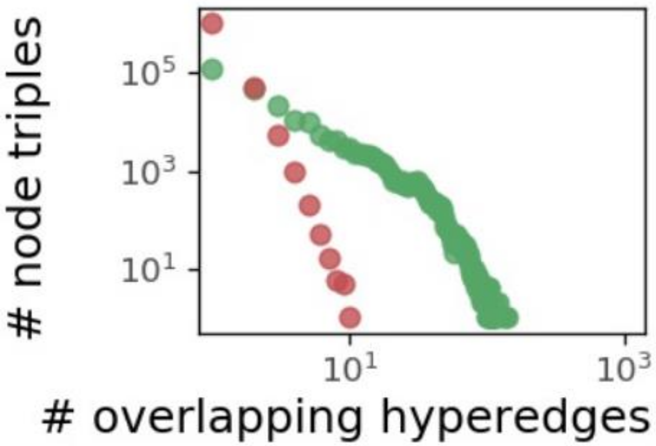
Pair/Triple Degree Distribution (cont.)

- Degree distributions of pair/triple of nodes in real-world hypergraphs are **more skewed with a heavier tail** than those in randomized ones.

Pair-of-Nodes



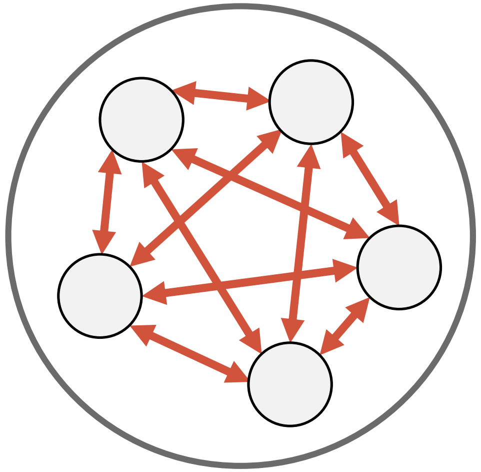
Triple-of-Nodes



Real-world hypergraph
Randomized hypergraph

Hyperedge Homogeneity

- **Homogeneity** of a hyperedge $\boldsymbol{e}$ is the average number of hyperedges overlapping at all the pairs of nodes in the hyperedge.



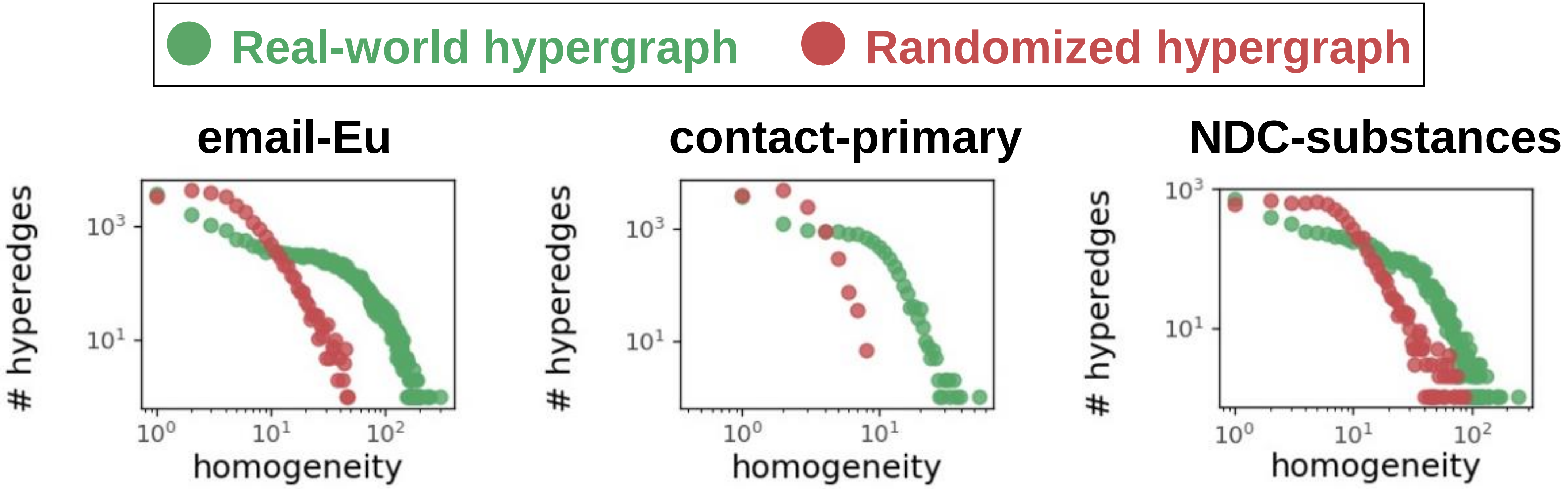
Hyperedge $\boldsymbol{e}$

Number of hyperedges overlapping nodes u and v.

$$\textbf{homogeneity}(\boldsymbol{e}) := \begin{cases} \dfrac{\sum_{\{u,v\}\in\binom{e}{2}} |E_{\{u,v\}}|}{\binom{|e|}{2}}, & \text{if } |\boldsymbol{e}| > \mathbf{1} \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

Hyperedge Homogeneity (cont.)

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



DYHS20: Five Basic Static Patterns

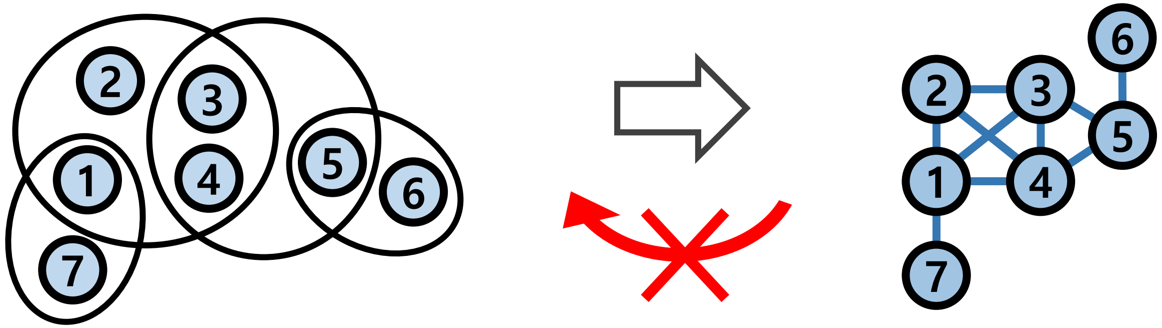
- **P1.** Heavy-tailed degree distribution
- **P2.** Skewed singular values distribution
- **P3.** Giant connected component
- **P4.** High clustering coefficient
- **P5.** Small effective diameter

	Node	P1	
	Hyperedge		
	Hypergraph	P2,P3, P4,P5	
	Sub-hypergraph		

M. T. Do, S. Yoon, B. Hooi, K. Shin, "Structural Patterns and Generative Models of Real-world Hypergraphs", **KDD 2020**

Multi-level Decomposition

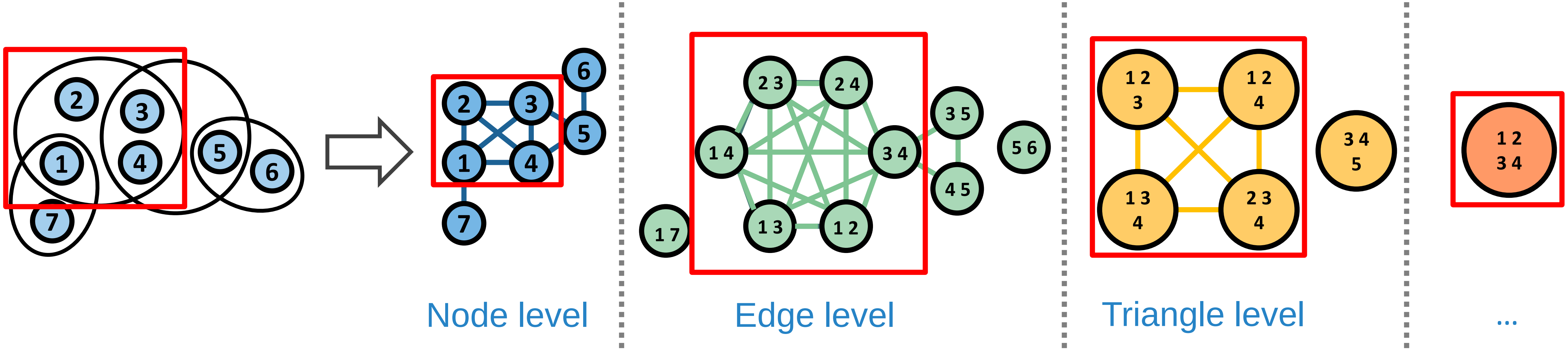
- **Hypergraphs**: not straightforward to analyze.
  - Complex representation
  - Lack of tools
- Projection (a.k.a., **clique expansion**)
  - Information loss
  - No higher-order information



Only interactions at
the level of nodes

Multi-level Decomposition (cont.)

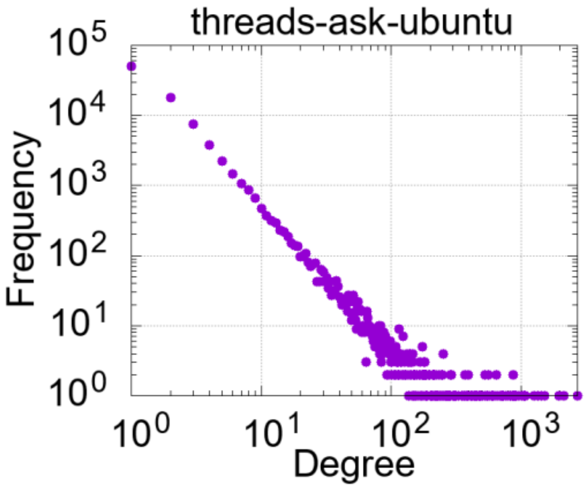
- **Multi-level decomposition**
  - Representation by pairwise unipartite graphs
  - Leveraging existing tools & measurements
  - No information loss: Original hypergraph is reconstructible



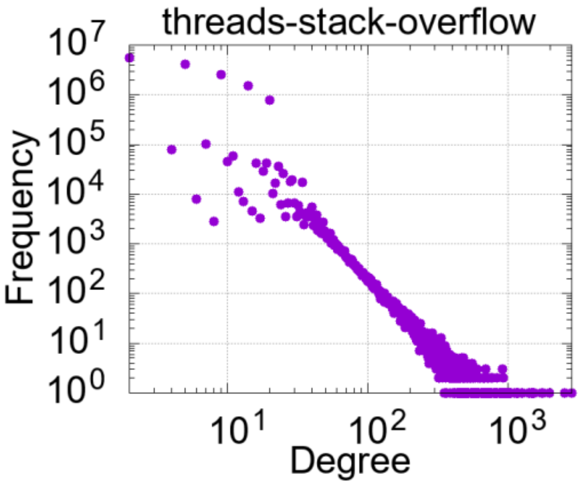
M. T. Do, S. Yoon, B. Hooi, K. Shin, "Structural Patterns and Generative Models of Real-world Hypergraphs", **KDD 2020**

Degree Distribution

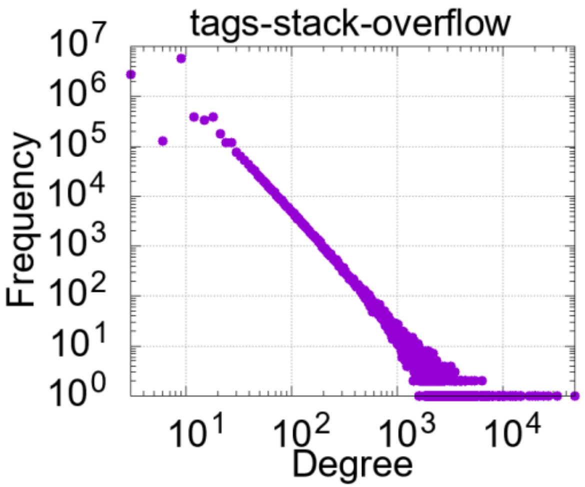
- At <u>every</u> decomposition level, degree distributions are **heavy-tailed**.



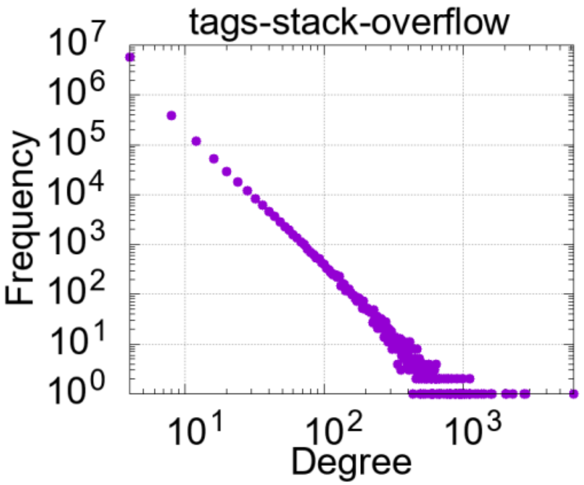
Node level



Edge level



Triangle level

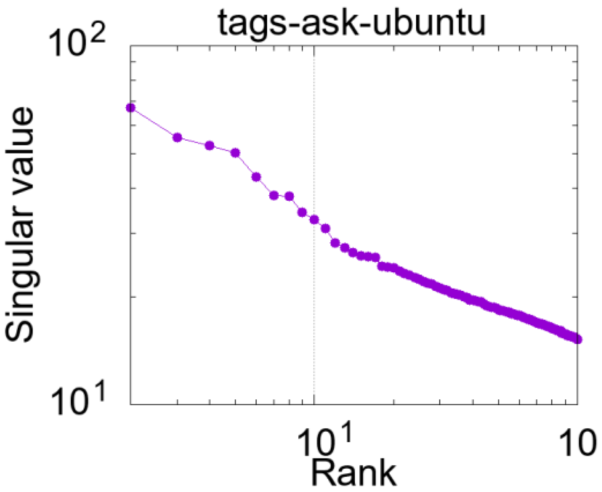


4-clique level

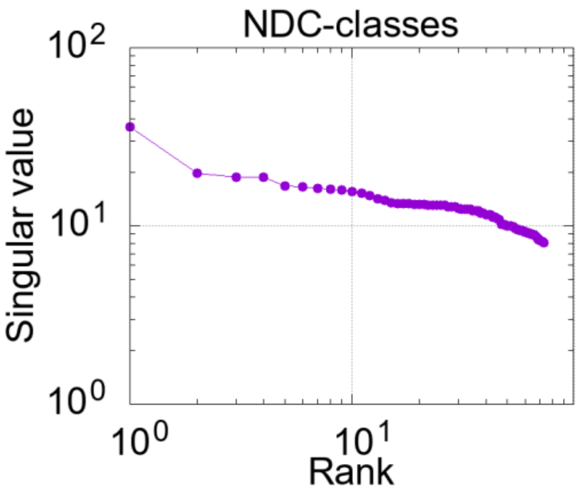
M. T. Do, S. Yoon, B. Hooi, K. Shin, "Structural Patterns and Generative Models of Real-world Hypergraphs", **KDD 2020**

Singular Value Distribution

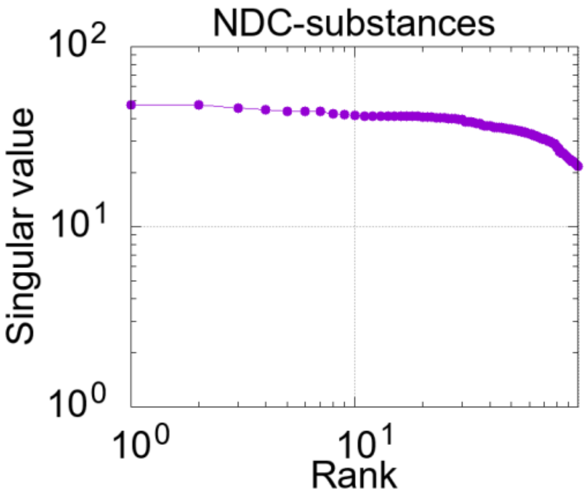
- At <u>every</u> decomposition level, singular value distributions are **heavy-tailed**.



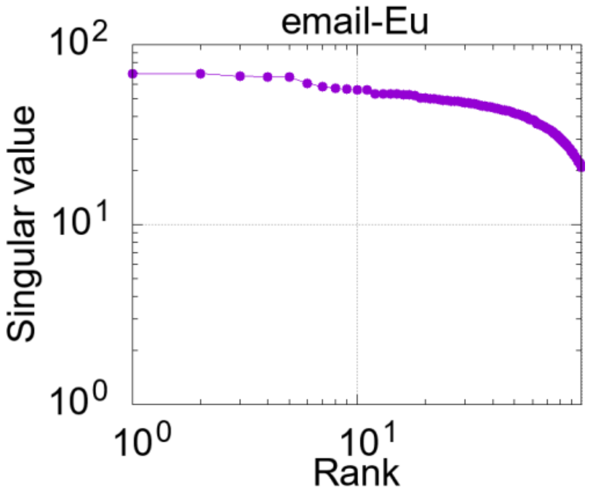
Node level



Edge level



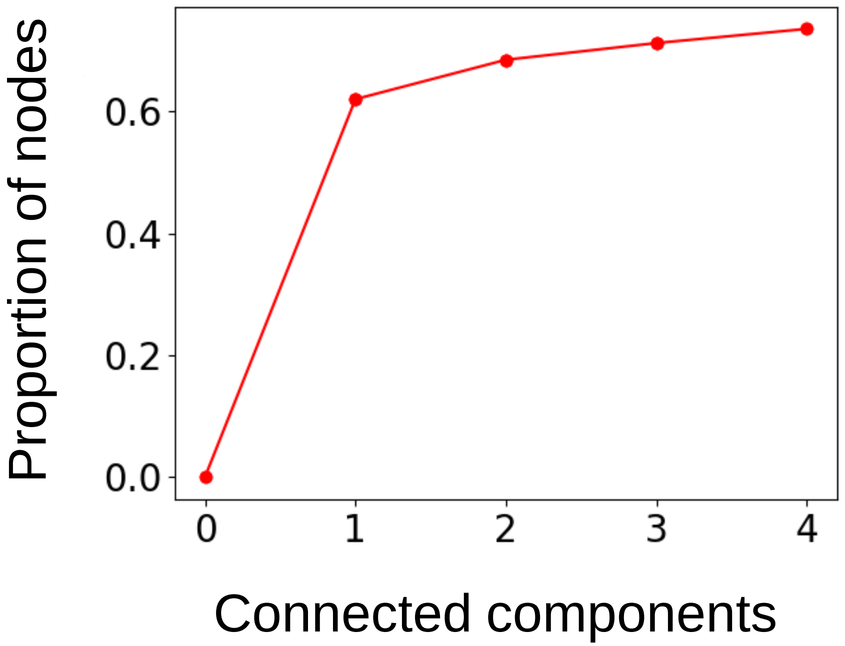
Triangle level



4-clique level

Giant Connected Component

- At <u>every</u> decomposition level, a large proportion of nodes are **connected**.

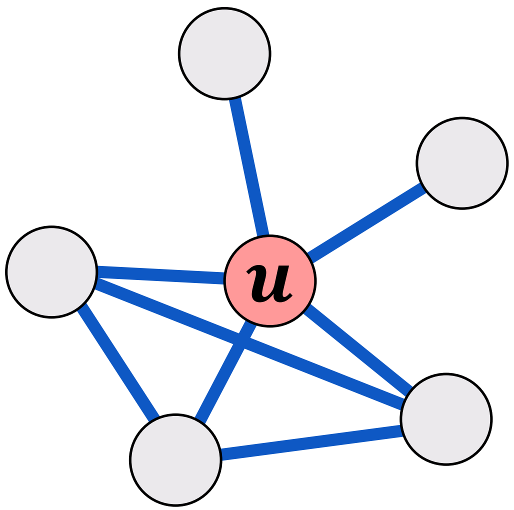


Level	Node ($k=1$)	Edge ($k=2$)	Triangle ($k=3$)	4clique ($k=4$)
coauth-DBLP	0.86	0.57	0.05	0.0006
coauth-Geology	0.72	0.5	0.06	0.0005
coauth-History	0.22	0.002	0.002	0.001
DAWN	0.89	0.98	0.91	0.52
email-Eu	0.98	0.98	0.86	0.41
NDC-classes	0.54	0.62	0.27	0.19
NDC-substances	0.58	0.82	0.36	0.02
tags-ask-ubuntu	0.99	0.99	0.79	0.21
tags-math	0.99	0.99	0.91	0.35
tags-stack-overflow	0.99	0.99	0.92	0.42
threads-ask-ubuntu	0.65	0.09	0.02	0.01
threads-math	0.86	0.61	0.03	0.0004
threads-stack-overflow	0.86	0.32	0.004	3e−5

Proportion of nodes in the largest connected component

High Clustering Coefficient

- At every decomposition level, there is high likelihood of having links between "**friends of friends**."



Local clustering coefficient

of node u is $C_u = \frac{3}{10}$.

Clustering coefficient

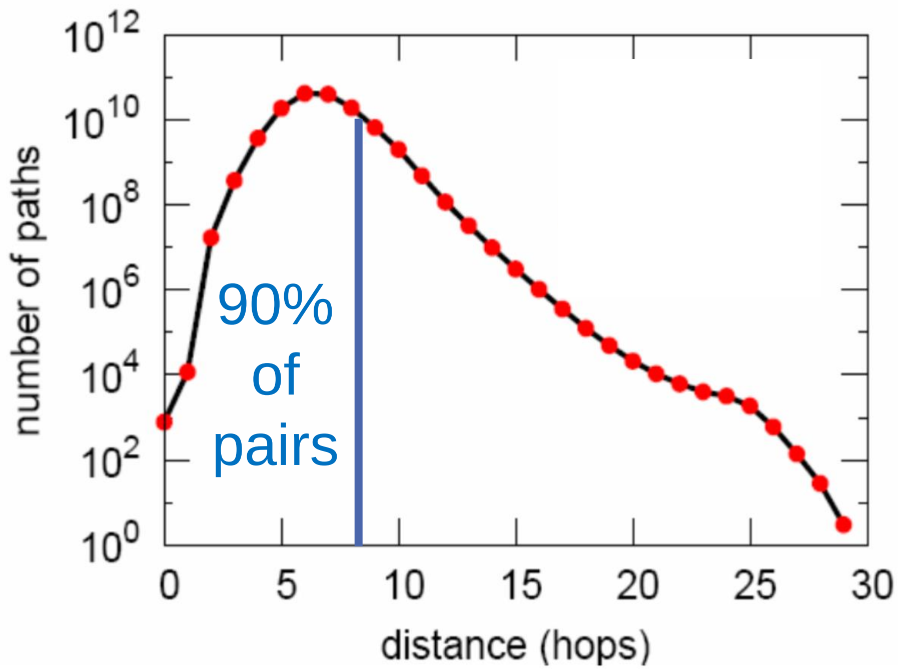
of the graph is $C = \frac{1}{|V|}\sum_{v \in V} C_v$.

Example

Dataset	Clust. coeff.	
	Real	Null.
coauth-DBLP	0.60	0.31 ±1e−4
coauth-Geology	0.57	0.42 ±2e−4
coauth-History	0.24	0.26 ±2e−4
DAWN	0.64	0.30 ±9e−5
email-Eu	0.49	0.36 ±5e−4
NDC-classes	0.61	0.32 ±2e−3
NDC-substances	0.40	0.17 ±6e−4
tags-ask-ubuntu	0.61	0.14 ±7e−5
tags-math	0.63	0.46 ±2e−4
tags-stack-overflow	0.63	0.03 ±1e−6
threads-ask-ubuntu	0.11	0.19 ±7e−4
threads-math	0.32	0.12 ±1e−4
threads-stack-overflow	0.18	0.12 ±2e−5

Small Effective Diameter

- At every decomposition level, most pairs of connected nodes are **reachable** within a small distance.



Dataset	# Nodes	Eff. diameter	
		Real	Null.
coauth-DBLP	1,924,991	6.8	6.7 ±9e−3
coauth-Geology	1,256,385	7.1	6.8 ±8e−3
coauth-History	1,014,734	11.9	17 ±0.19
DAWN	2,558	2.6	1.85 ±8e−5
email-Eu	998	2.8	1.85 ±7e−5
NDC-classes	1,161	4.6	2.6 ±6e−3
NDC-substances	5,311	3.5	2.5 ±9e−3
tags-ask-ubuntu	3,029	2.4	1.9 ±2e−5
tags-math	1,629	2.1	1.8 ±1e−4
tags-stack-overflow	49,998	2.7	1.9 ±2e−6
threads-ask-ubuntu	125,602	4.7	11.9 ±0.042
threads-math	176,445	3.7	4.9 ±4e−3
threads-stack-overflow	2,675,995	4.5	5.9 ±2e−3

Roadmap

- **Part 1. Static Structural Patterns**
  - Basic Patterns
  - Advanced Patterns **<<**
- **Part 2. Dynamic Structural Patterns**
  - Basic Patterns
  - Advanced Patterns
- **Part 3. Generative Models**
  - Static Hypergraph Generator
  - Dynamic Hypergraph Generator

Part 1-2. Advanced Static Structural Patterns

		Part 1. Static Patterns	Part 2. Dynamic Patterns
Basic Patterns	Node-Level	DYHS20, KKS20, LCS21	BKT18, CS22
	Hyperedge-Level	KKS20, LCS21	BKT18, CBLK21, LS21
	Hypergraph-Level	BASJK18, DYHS20, KKS20	KKS20
Advanced Patterns	Sub-hypergraph-Level	BASJK18, LMMB22, LKK20, LCS21	BASJK18, CJ21, LS21

BASJK18: One Advanced Static Pattern

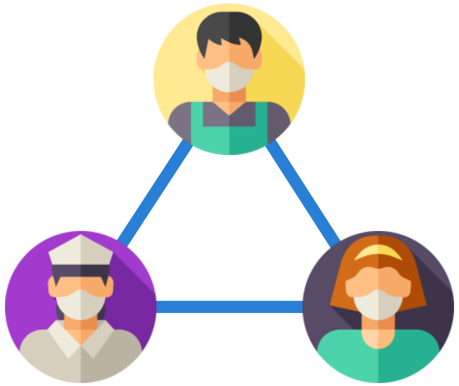
- **P1.** Open and closed triangles

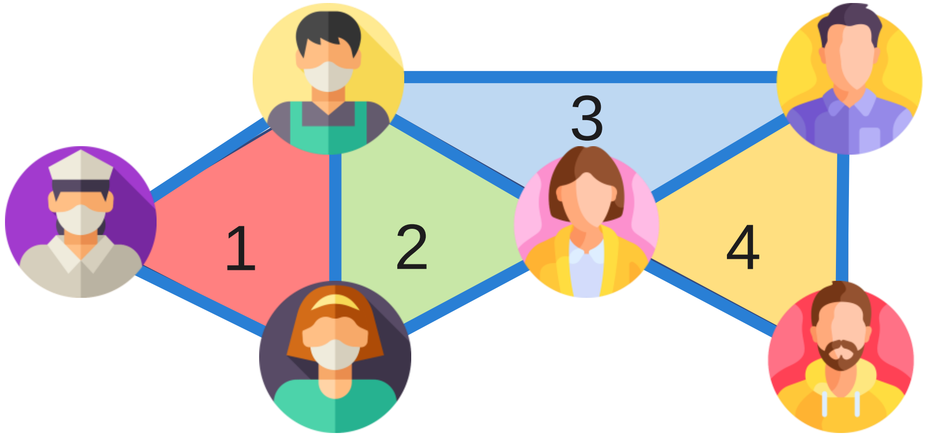
	Node		
	Hyperedge		
	Hypergraph		
	Sub-hypergraph	P1	

A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, J. Kleinberg, "Simplicial Closure and Higher-order Link Prediction", **PNAS (2018)**

Background

- A **triangle** is 3 nodes connected to each other.
- The **counts** of triangles is an important primitive.
  - E.g., Community detection, spam detection, link prediction





A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, J. Kleinberg, “Simplicial Closure and Higher-order Link Prediction”, **PNAS (2018)**

Triangles in Hypergraphs

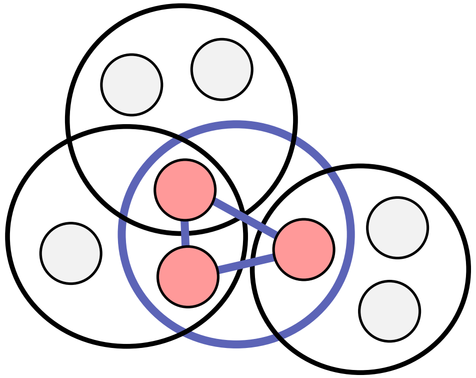
?

Question:

How can we define **triangles** in hypergraphs?

Answer:

Tri-wise relations (i.e., group interactions of three nodes) should be taken into account.



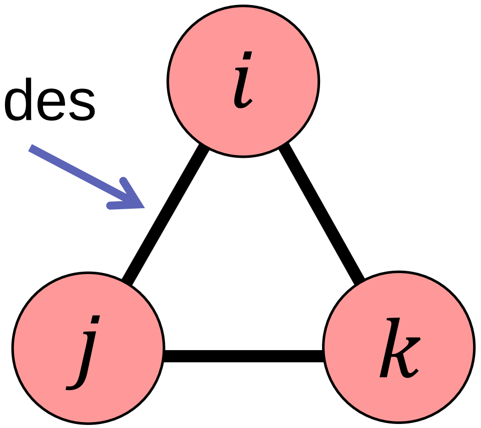
!

A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, J. Kleinberg, “Simplicial Closure and Higher-order Link Prediction”, **PNAS (2018)**

Open and Closed Triangles: Definition

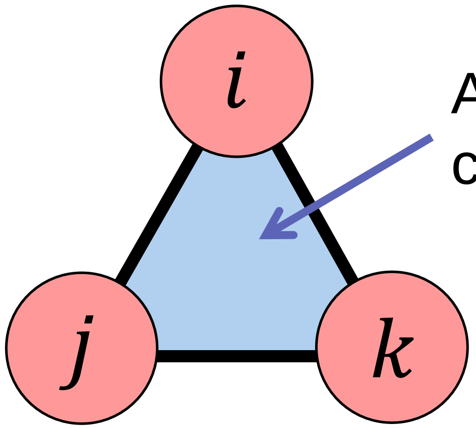
- There are **two types of triangles** in hypergraphs.
  - **Closed triangles** cannot be captured by pairwise graphs.

Any hyperedge that contains the pair of nodes



Open Triangle

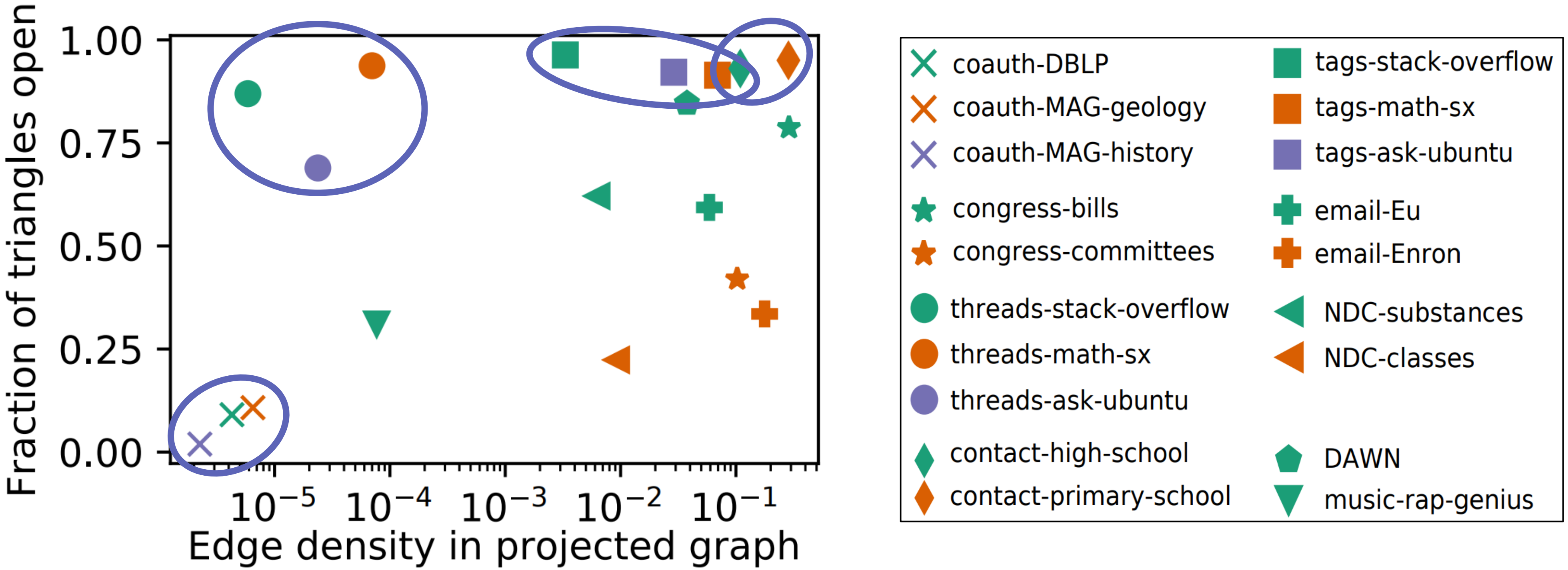
Any hyperedge that contains all 3 nodes



Closed Triangle

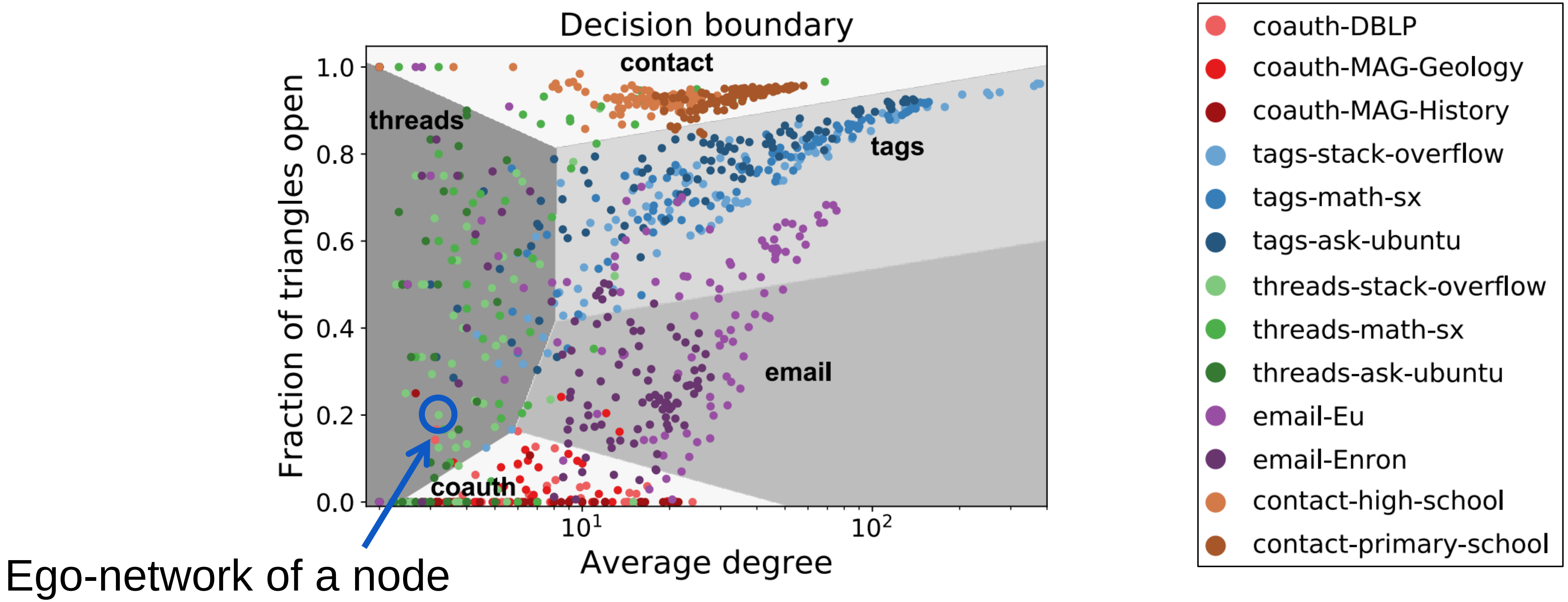
Triangles across Domains

- **Fractions of open triangles** are similar within domains.



Triangles across Domains (cont.)

- **Fractions of open triangles** are similar within domains.



LMMB20: One Advanced Static Pattern

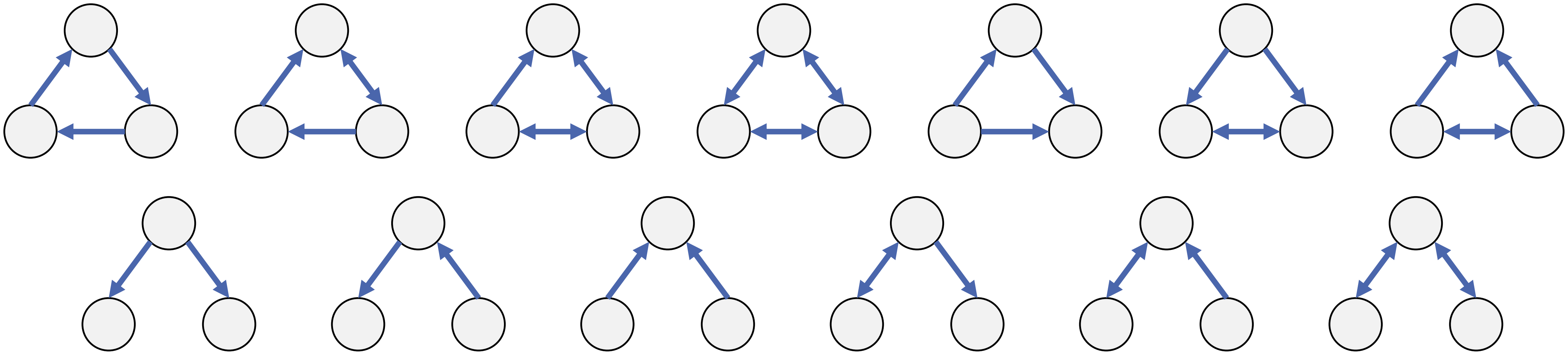
- **P1.** Higher-order network motifs

	Node		
	Hyperedge		
	Hypergraph		
	Sub-hypergraph	P1	

Q. F. Lotito, F. Musciotto, A. Montresor, F. Battiston, "Higher-order Motif Analysis in Hypergraphs", **Communication Physics (2022)**

Background

- **Network motifs** are fundamental building blocks of complex networks.
  - They appear in real-world hypergraphs at a frequency that is **significantly higher** than randomized hypergraphs.



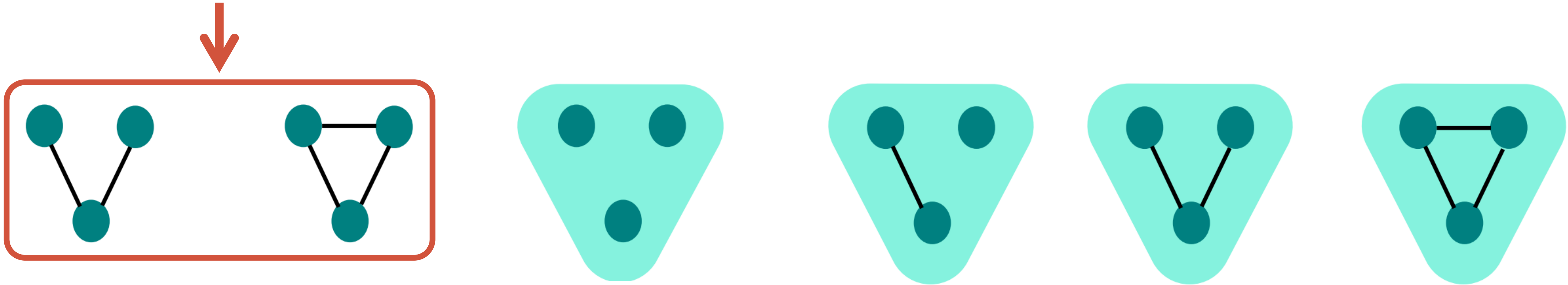
13 different types of 3-node network motifs

Q. F. Lotito, F. Musciotto, A. Montresor, F. Battiston, "Higher-order Motif Analysis in Hypergraphs", **Communication Physics (2022)**

Higher-order Network Motifs: Definition

- **Higher-order network motifs** are a generalization of network motifs.
- They additionally consider **group interactions** between the nodes.

Network motifs can only describe 2 patterns.



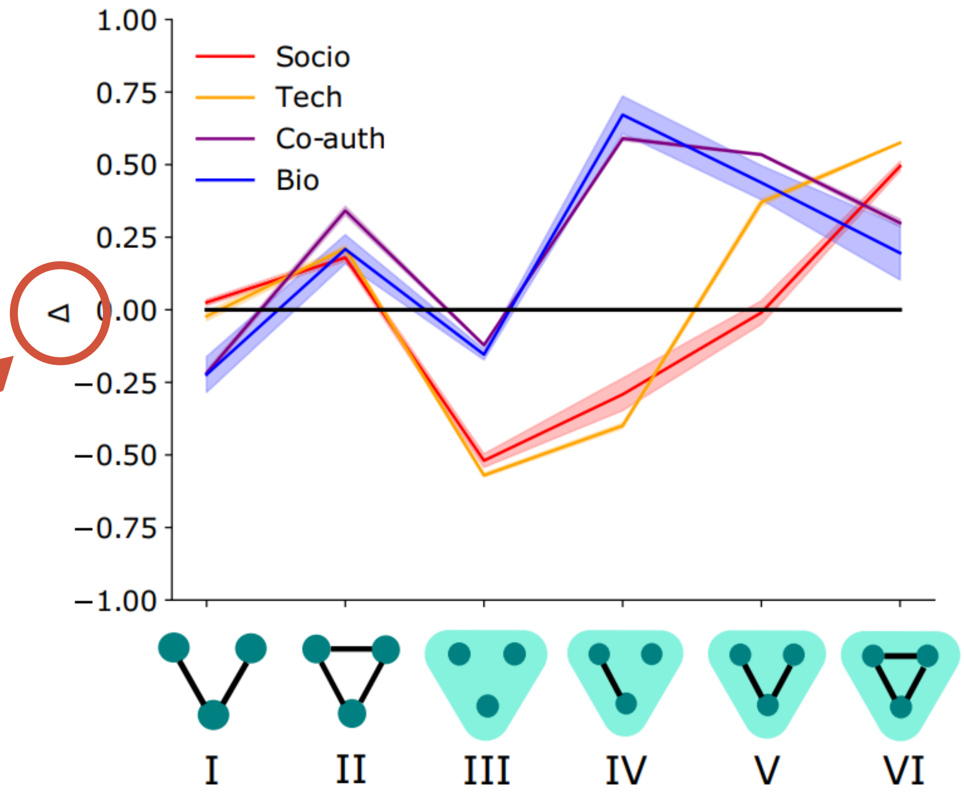
6 different types of 3-node higher-order motifs

Q. F. Lotito, F. Musciotto, A. Montresor, F. Battiston, "Higher-order Motif Analysis in Hypergraphs", **Communication Physics (2022)**

Comparison across Domains

- Different **higher-order motifs** are highlighted in each domain.

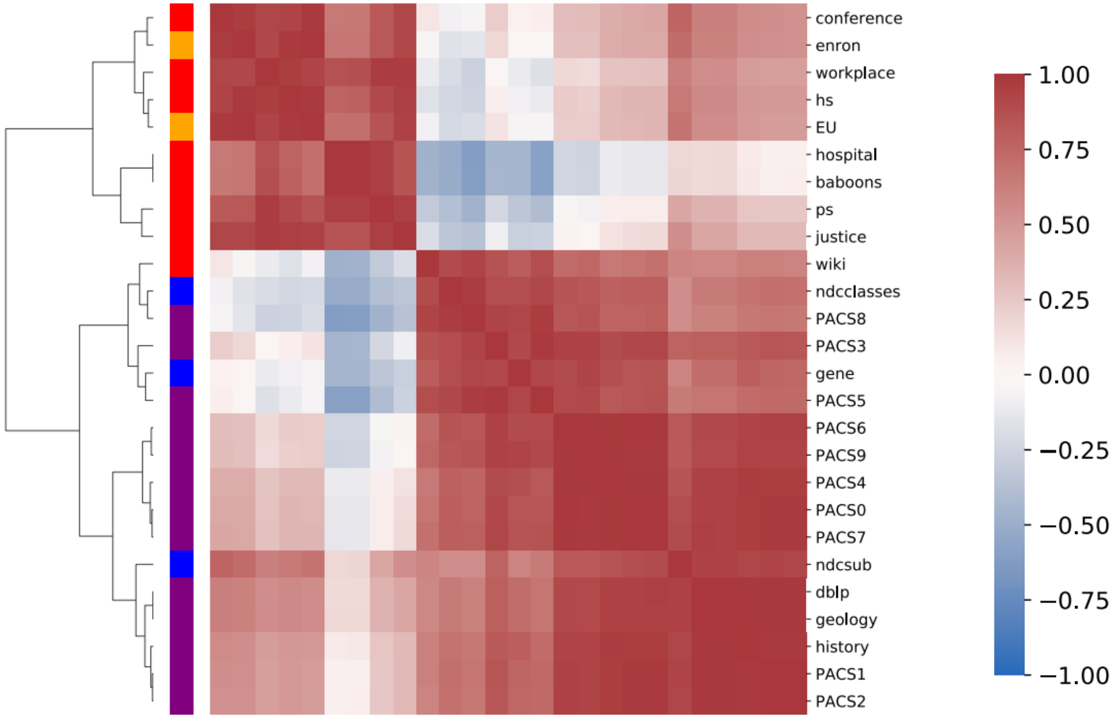
Abundance of higher-order motifs relative to random networks



Comparison across Domains (cont.)

- Distributions of **higher-order motifs** are similar within domains and different across domains.

Social domain
E-mail domain
Co-authorship domain
Biological domain

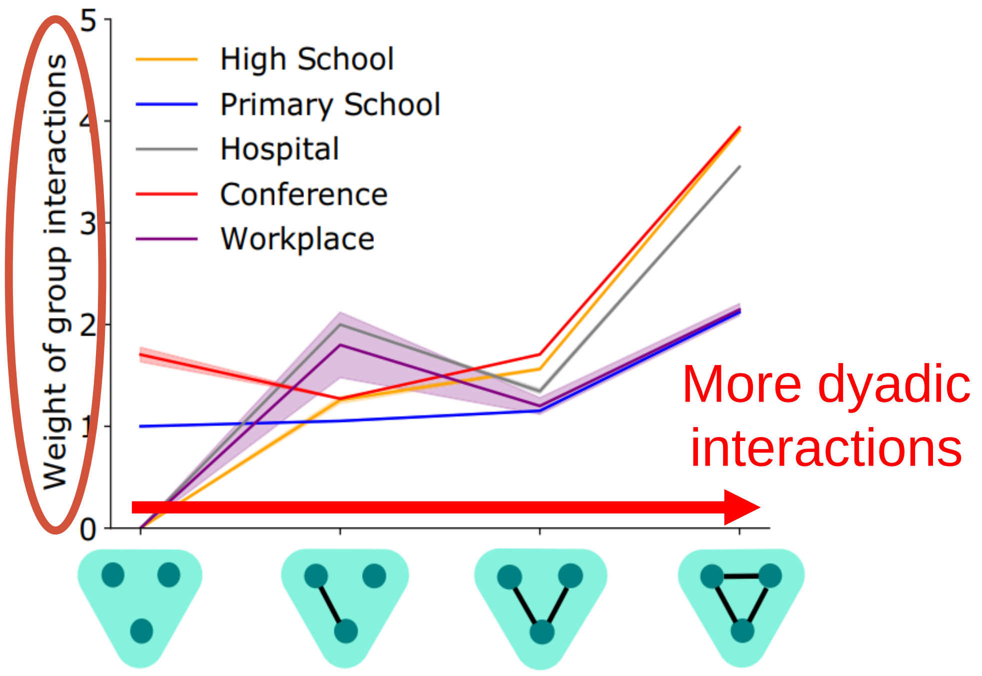


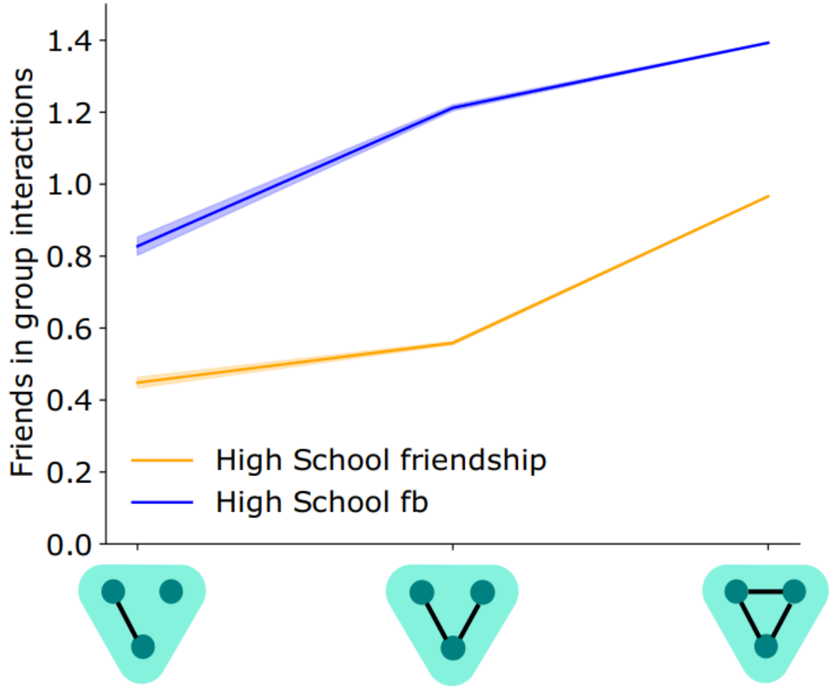
Q. F. Lotito, F. Musciotto, A. Montresor, F. Battiston, "Higher-order Motif Analysis in Hypergraphs", **Communication Physics (2022)**

Structural Reinforcement

- **Weight of each hyperedge** (i.e., the number of times each group interaction occurs) is correlated with **the number of pairwise links**.

\# of group interactions





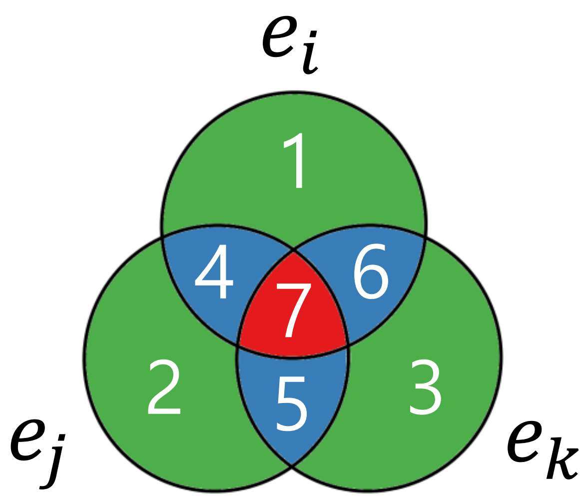
LKS20: One Advanced Static Pattern

- **P1.** Hypergraph motifs (h-motifs)

	Node		
	Hyperedge		
	Hypergraph		
	Sub-hypergraph	P1	

Hypergraph Motifs: Definition

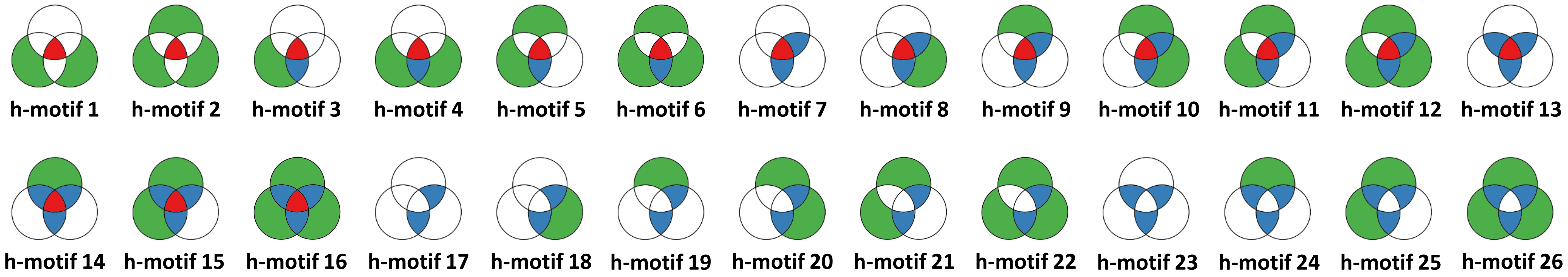
- **Hypergraph motifs (h-motifs)** describe connectivity patterns of three connected hyperedges.
- H-motifs describe the connectivity pattern of hyperedges e_i, e_j, and e_k by the emptiness of seven subsets (1) – (7).



(1) $e_i \setminus e_j \setminus e_k$
(2) $e_j \setminus e_k \setminus e_i$
(3) $e_k \setminus e_i \setminus e_j$
(4) $e_i \cap e_j \setminus e_k$
(5) $e_j \cap e_k \setminus e_i$
(6) $e_k \cap e_i \setminus e_j$
(7) $e_i \cap e_j \cap e_k$

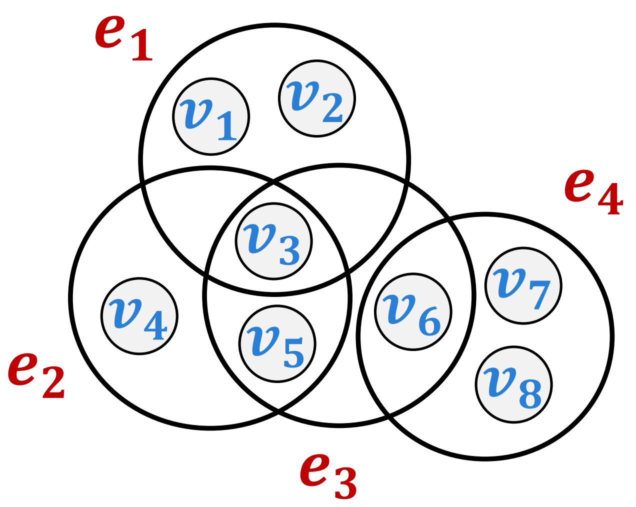
Hypergraph Motifs: Definition (cont.)

- While there can exist 2^7 h-motifs, **26 h-motifs** remain once we exclude:
  1. symmetric ones
  2. those cannot be obtained from distinct hyperedges
  3. those cannot be obtained from connected hyperedges

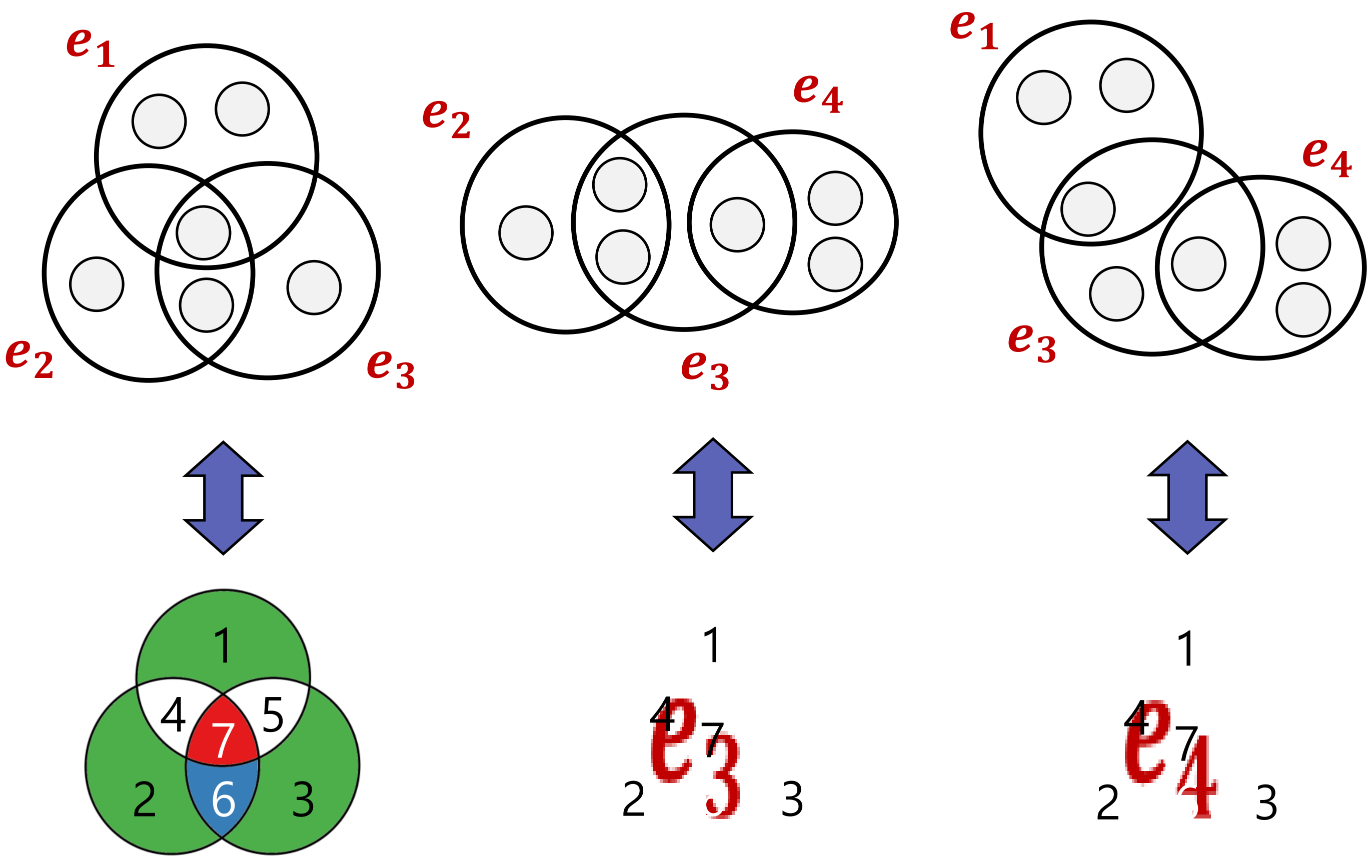


Hypergraph Motifs: Example

- Example: A hypergraph with 8 nodes and 4 hyperedges



Hypergraph



Hypergraph Motifs: Properties

Details

Exhaustive

H-motifs capture connectivity patterns of **all possible** three connected hyperedges.

Unique

Connectivity pattern of any three connected hyperedges is captured by **at most one** h-motif.

Size Independent

H-motifs capture connectivity patterns **independently of the sizes of hyperedges**.

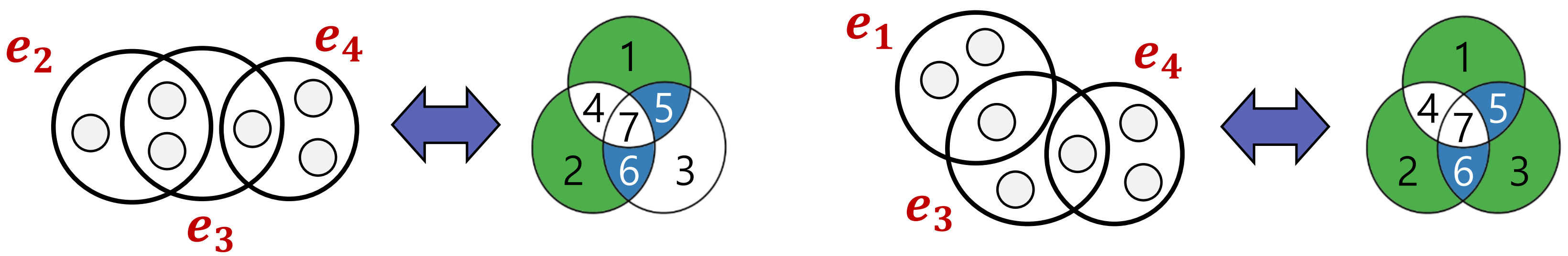
Hypergraph Motifs: Properties (cont.)

Question:

Why are **non-pairwise relations** considered?

Answer:

Non-pairwise relations play a key role in capturing the local structural patterns of real-world hypergraphs.



For example, $\{e_2, e_3, e_4\}$ and $\{e_1, e_3, e_4\}$ have same pairwise relations, while their connectivity patterns are distinguished by h-motifs.

Characteristic Profiles

?

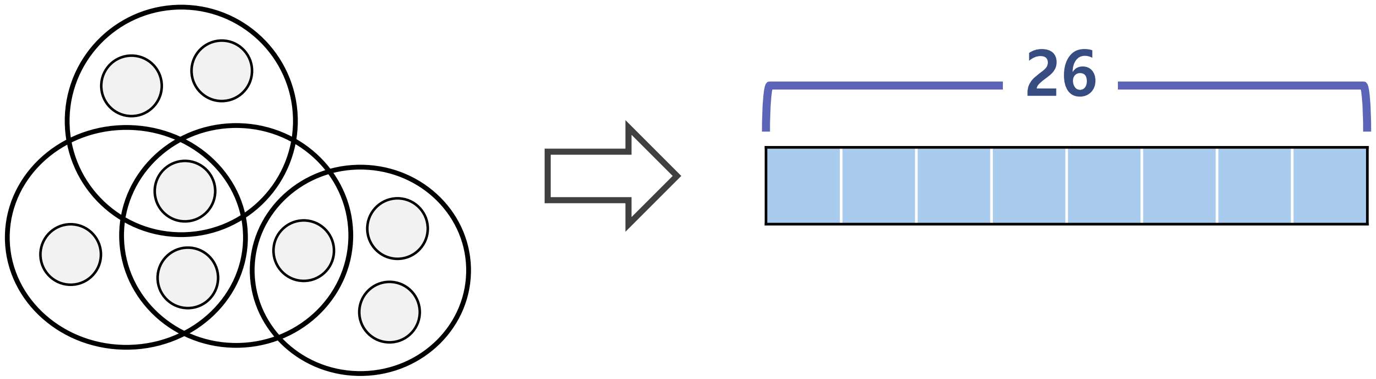
Question:

How can we summarize h-motif properties of hypergraphs?

Answer:

We compute a compact vector of **normalized significance of every h-motif**.

!



Characteristic Profiles (cont.)

Details

Significance of H-motif t

$$\Delta_t := \frac{M[t] - M_{rand}[t]}{M[t] + M_{rand}[t] + \epsilon}$$

# of instances of h-motif t in the **given hypergraph**

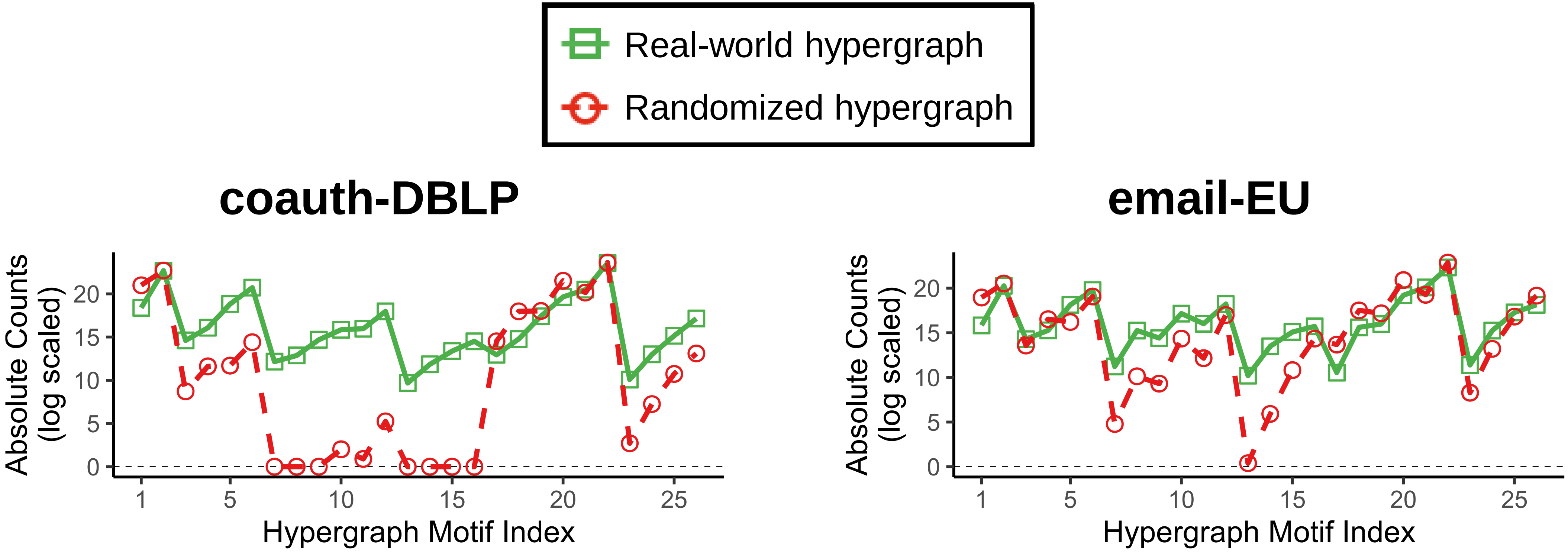
# of instances of h-motif t in the **randomized hypergraph**

Characteristic Profiles (CPs)

$$CP_t := \frac{\Delta_t}{\sqrt{\sum_{t=1}^{26} \Delta_t^2}}$$

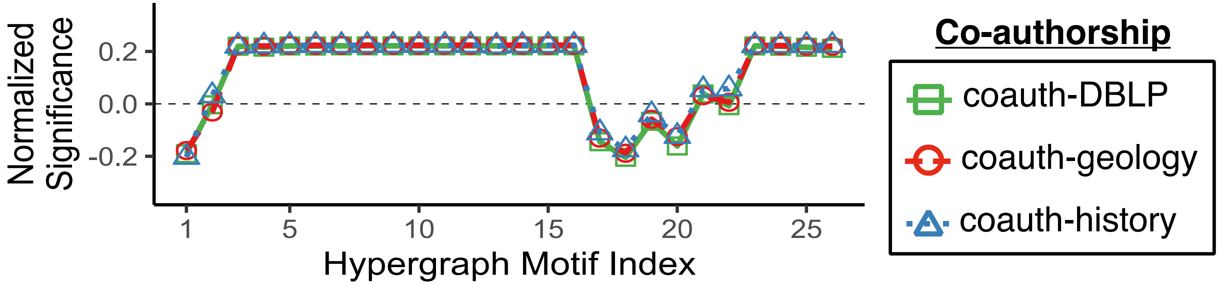
Real vs. Random

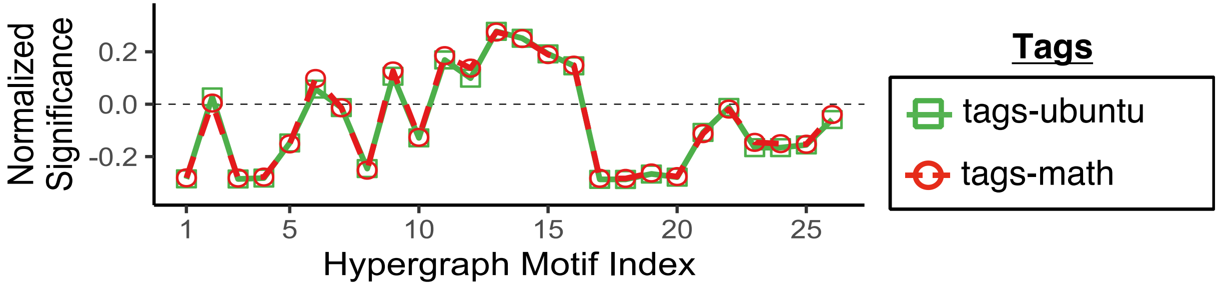
- Real-world and random hypergraphs have **distinct distributions** of h-motif instances.

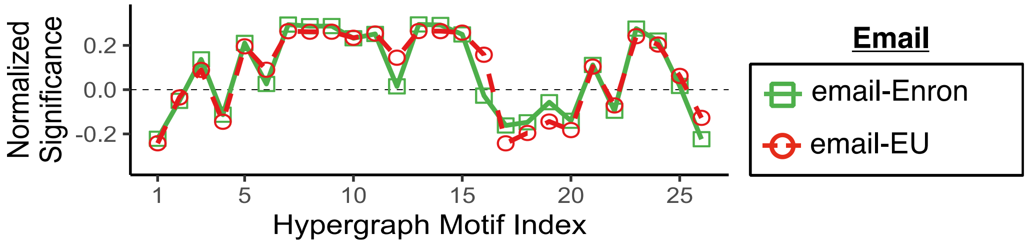


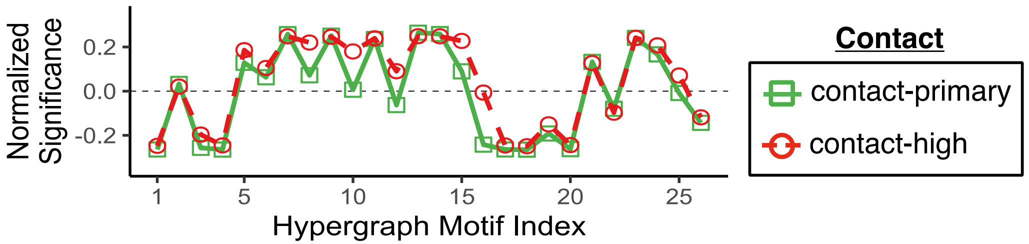
Comparison across Domains

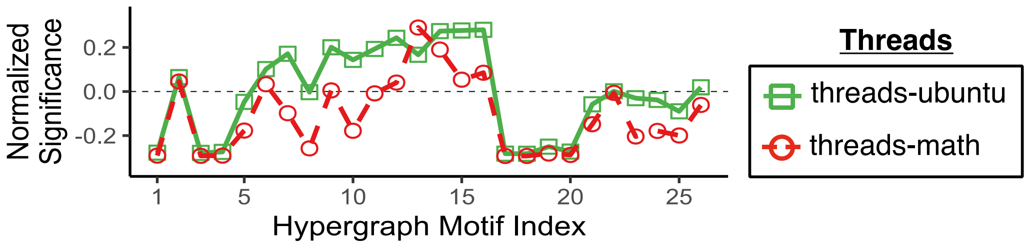
- CPs are **similar within domains** but **different across domains**.





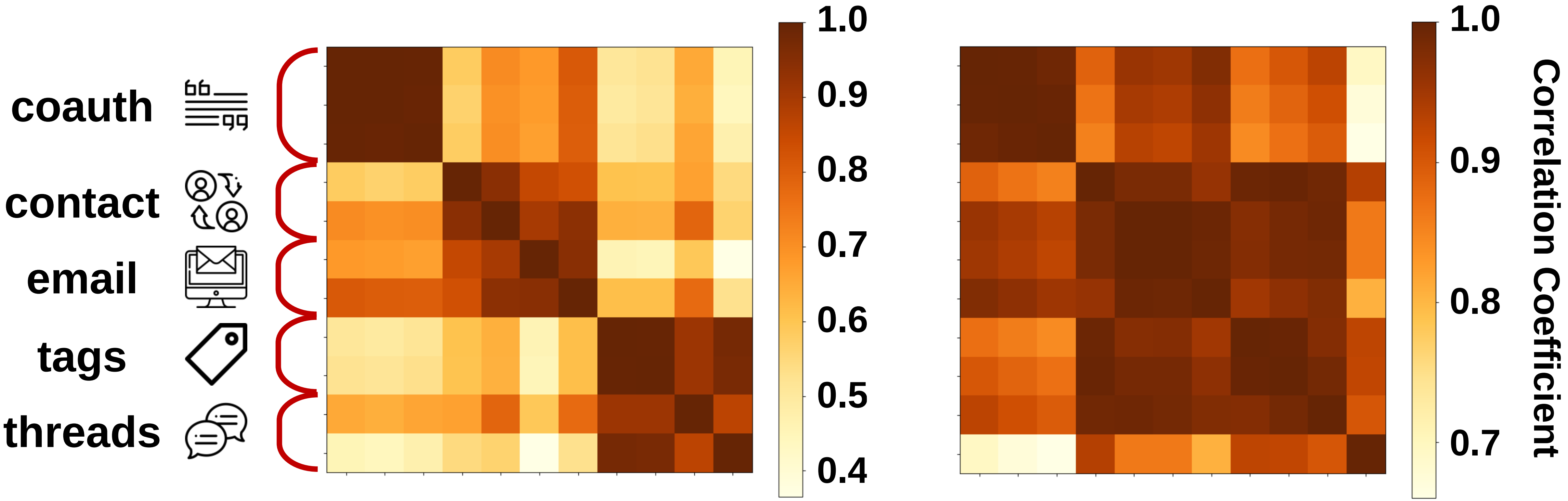






Comparison across Domains (cont.)

- **CPs based on h-motifs** capture local structural patterns more accurately than **CPs based on network motifs**.



LCS21: One Advanced Static Pattern

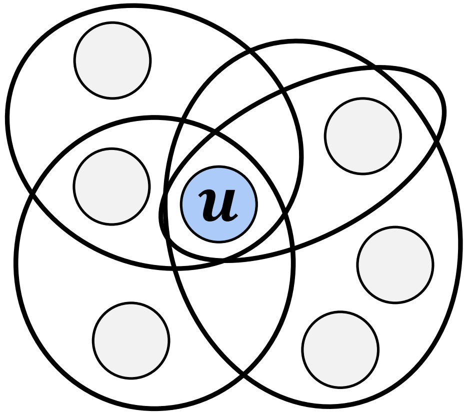
- **P1.** Density & overlapness of ego-network

	Node		
	Hyperedge		
	Hypergraph		
	Sub-hypergraph	P1	

Hypergraph Ego-network

- An **ego-network** $\mathcal{E}$ of node u is the set of hyperedges that contains u.

$$\mathcal{E}(u) \coloneqq \{e \in E: u \in e\}$$



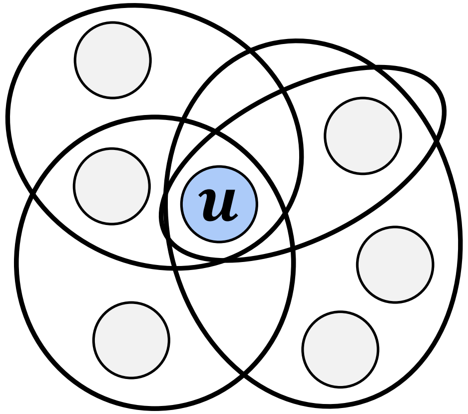
Density of Ego-networks

- **Density** measures how densely hyperedges are overlapped.

$$\rho(\mathcal{E}) := \frac{|\mathcal{E}|}{\left|\bigcup_{e \in \mathcal{E}} e\right|}$$

← # of hyperedges

← # of nodes



Hypergraph

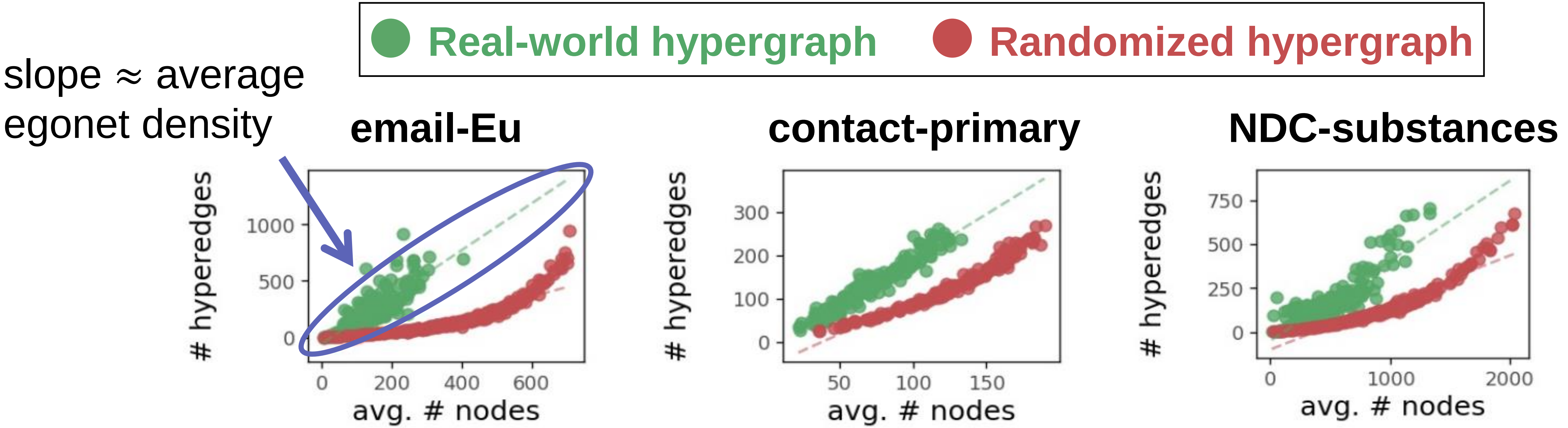
⇨

Density of egonet $\mathcal{E}(u)$ is $\frac{4}{7}$.

Example

Density of Ego-networks (cont.)

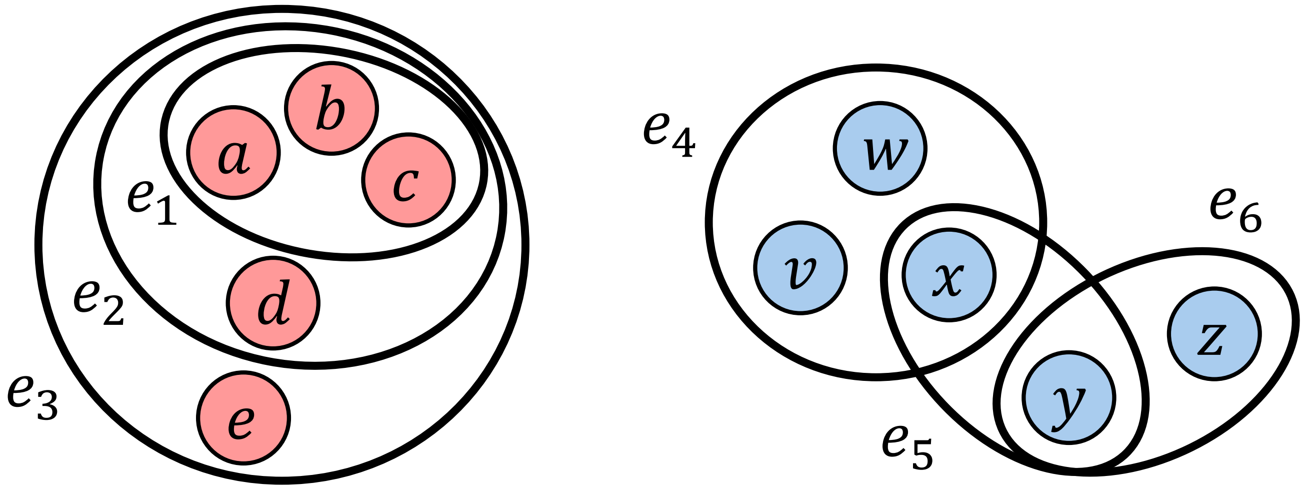
- Ego-networks in real-world hypergraphs tend to have **higher density** than those in randomized ones.



G. Lee, M. Choe, K. Shin, “How Do Hyperedges Overlap in Real-world Hypergraphs? – Patterns, Measures, and Generators”, **WWW 2021**

Density of Ego-networks (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

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

Degree of Hyperedge Overlaps

?

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

Answer:
- We present **three axioms** that any reasonable measure of the hyperedge overlaps should satisfy.
- Then, we propose **overlapness**, a new measure that satisfies all the axioms.

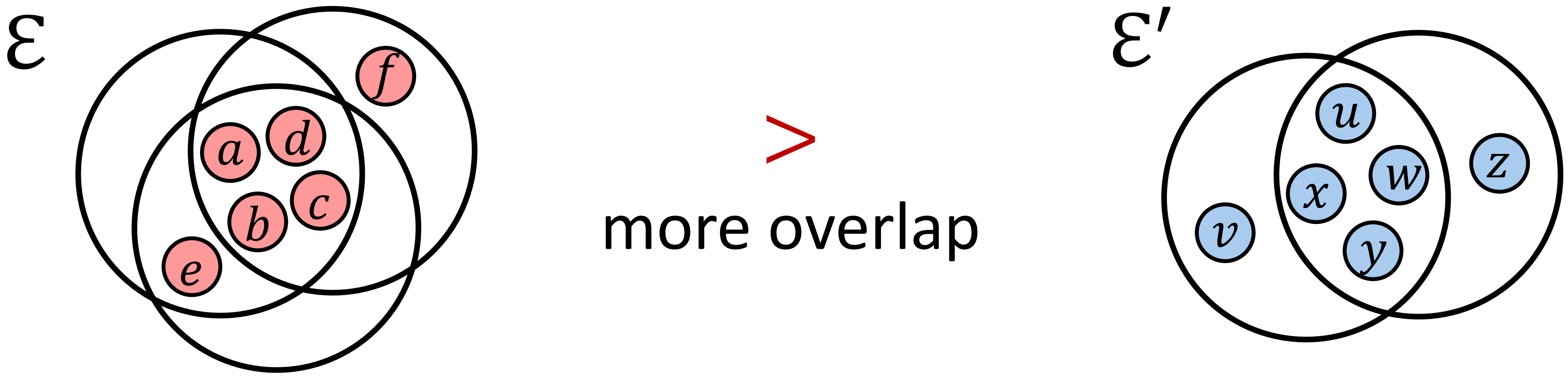
!

Degree of Hyperedge Overlaps (cont.)

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 $f(\mathcal{E}) > f(\mathcal{E}')$.

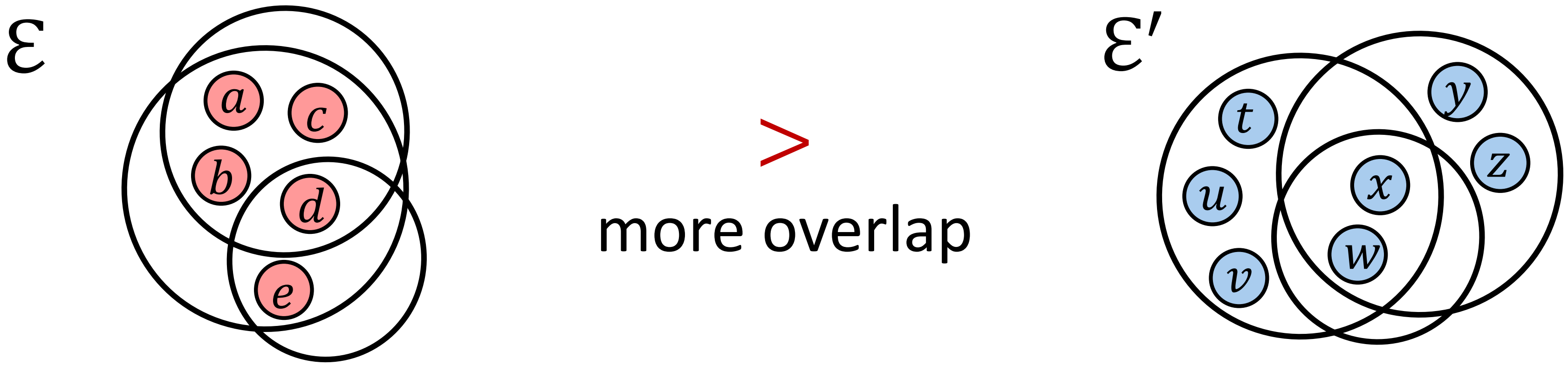


Degree of Hyperedge Overlaps (cont.)

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 $f(\mathcal{E}) > f(\mathcal{E}')$.

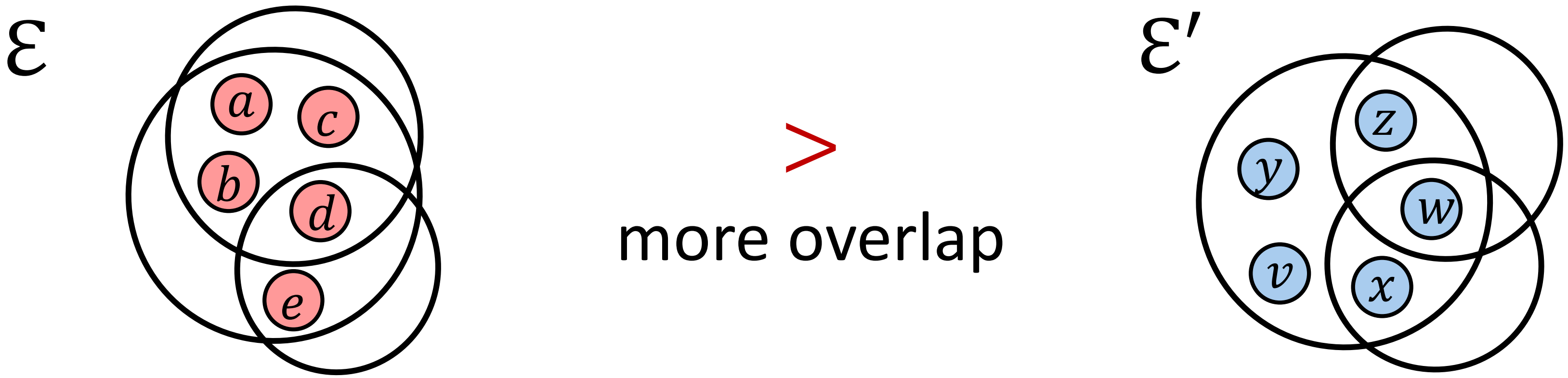


Degree of Hyperedge Overlaps (cont.)

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 distinct nodes than $\mathcal{E}'$, then $f(\mathcal{E}) > f(\mathcal{E}')$.



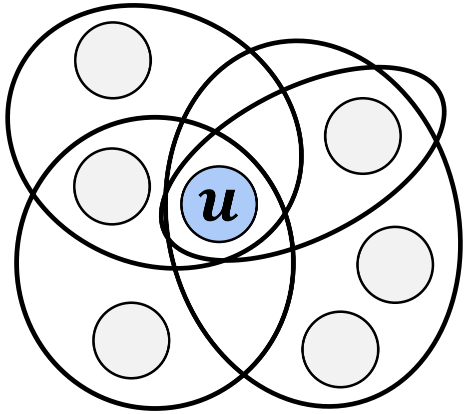
Overlapness of Ego-networks

- **Overlapness** measures how densely hyperedges are overlapped.

$$o(\mathcal{E}) \coloneqq \frac{\sum_{e \in \mathcal{E}} |e|}{\left|\bigcup_{e \in \mathcal{E}} e\right|}$$

← sum of the hyperedge sizes

← # of nodes



Hypergraph

Overlapness of egonet $\mathcal{E}(u)$ is $\frac{12}{7}$.

Example

Overlapness of Ego-networks (cont.)

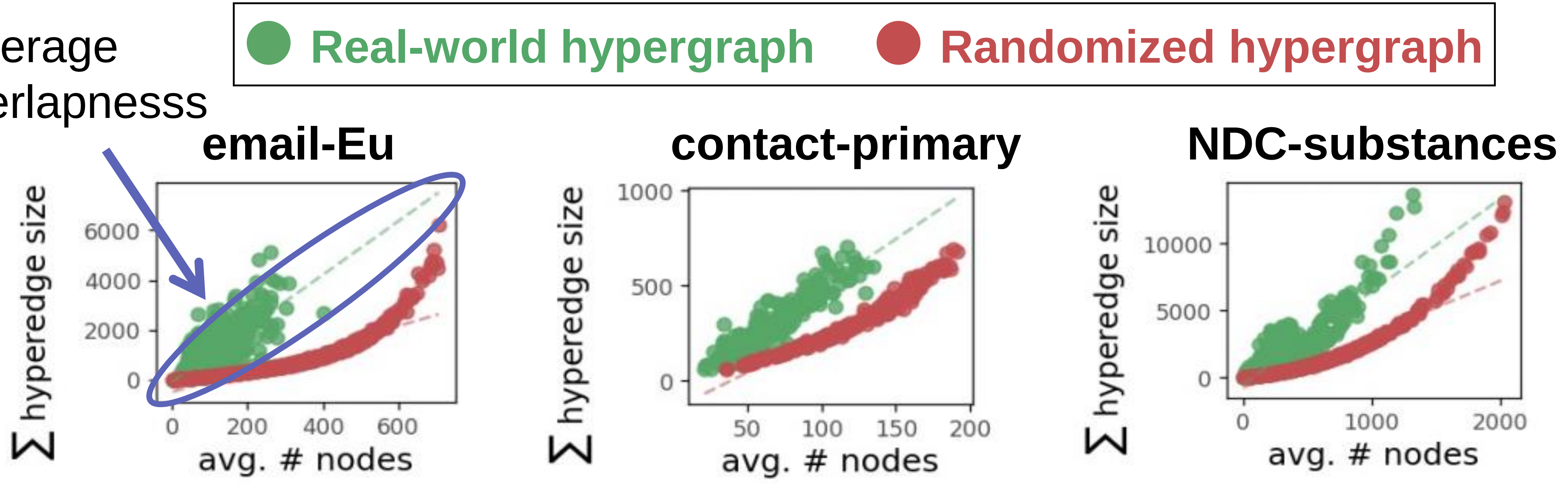
- **Overlapness** satisfies all the axioms while others do not.

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

Overlapness of Ego-networks (cont.)

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

slope ≈ average egonet overlapnesss



References

1. [KKS20] Kook, Yunbum, Jihoon Ko, and Kijung Shin. “Evolution of Real-world Hypergraphs: Patterns and Models without Oracles.” ICDM 2020.
2. [LMK21] Lee, Geon, Minyoung Choe, and Kijung Shin. “How Do Hyperedges Overlap in Real-world Hypergraphs? – Patterns, Measures, and Generators.” WWW 2021.
3. [DYHS20] Do, Manh Tuan, et al. “Structural Patterns and Generative Models of Real-world Hypergraphs.” KDD 2020.
4. [BASJK18] Benson, Austin R., et al. “Simplicial Closure and Higher-order Link Prediction.” PNAS 115(48):E11221–E11230, 2018.
5. [LMMB20] Lotito, Quintino Francesco, et al. “Higher-order Motif Analysis in Hypergraphs.” Communication Physics 5(1):1–8, 2022
6. [LKS20] Lee, Geon, Jihoon Ko, and Kijung Shin. “Hypergraph Motifs: Concepts, Algorithms, and Discoveries.” PVLDB 13(12):2256-2269, 2020.