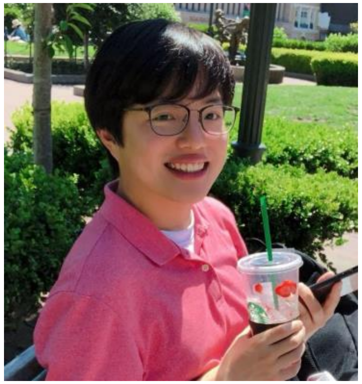
KAIST

SIAM 2022 International Conference on Data Mining

On the Persistence of Higher-Order Interactions in Real-World Hypergraphs

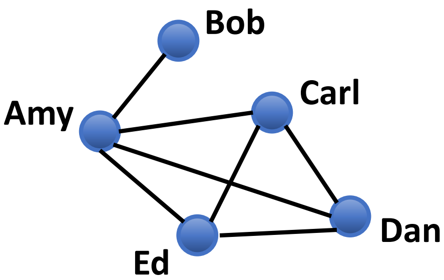
Hyunjin Choo



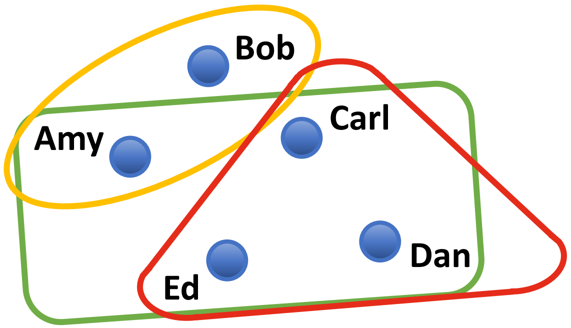
Kijung Shin

Hypergraph

- A **hypergraph** is a generalization of an ordinary graph
- A **hyperedge** joins an **arbitrary** number of nodes



(a) Ordinary Graph

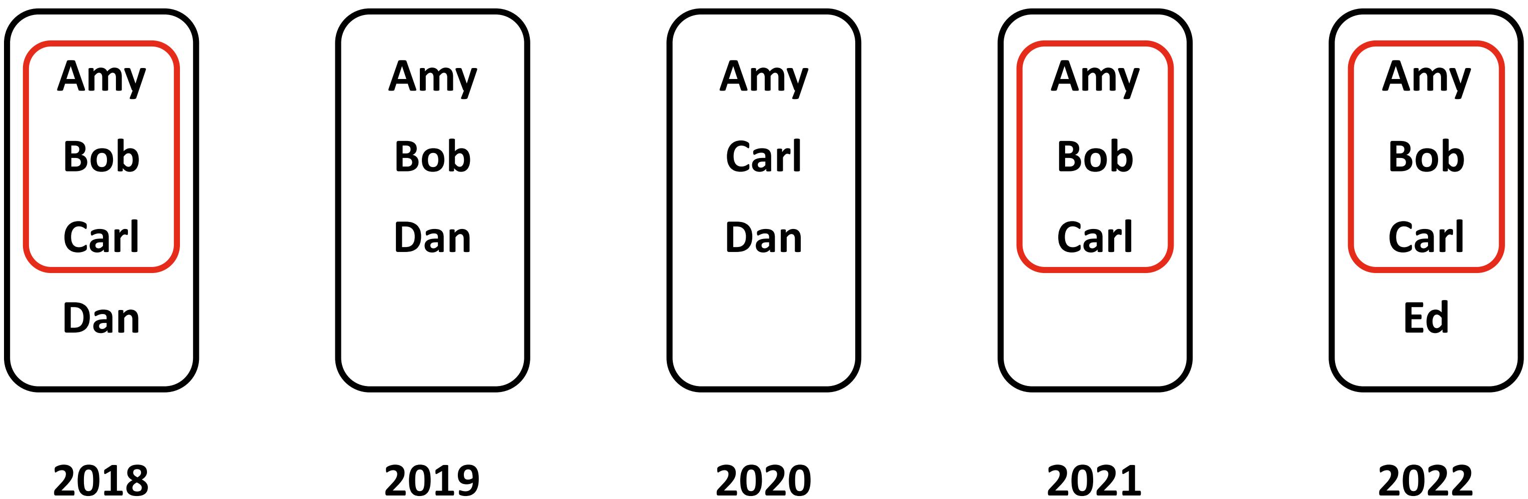


(b) Hypergraph

- ➢ Sender and receivers of an email
- ➢ Co-authors of a publication
- ➢ Items co-purchased by a customer

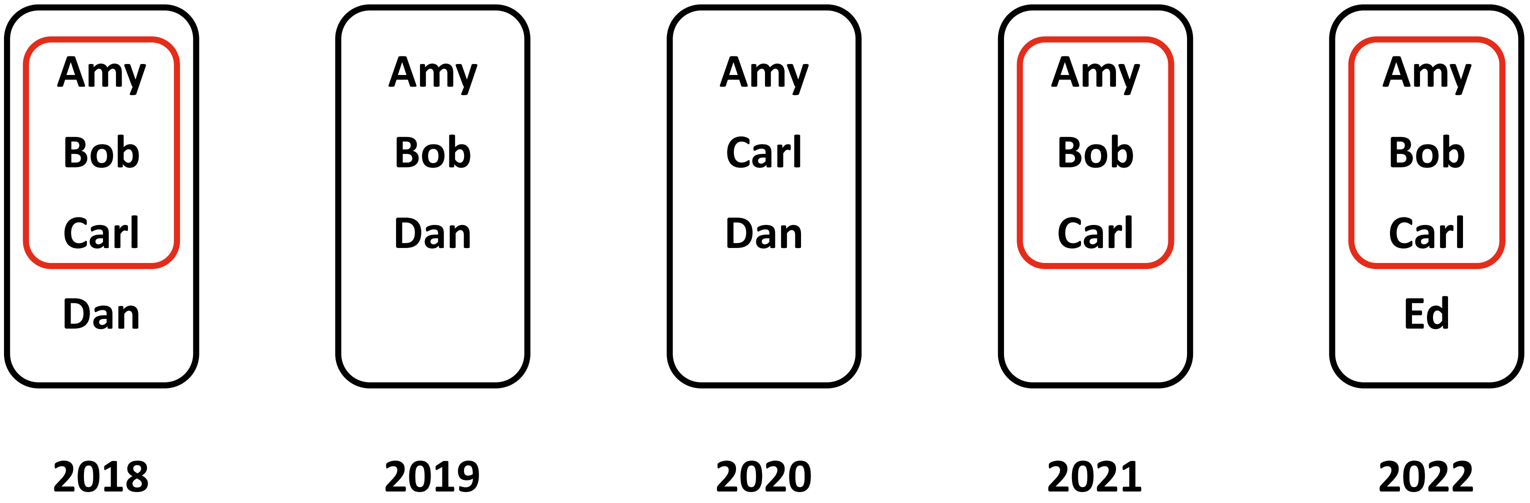
Higher-Order Interaction (HOI)

- A **higher-order interaction (HOI)** is the **co-appearance** of a set of nodes in any hyperedge
  - E.g.) If A, B, and C publish a paper together, any of $\{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}$ becomes a HOI



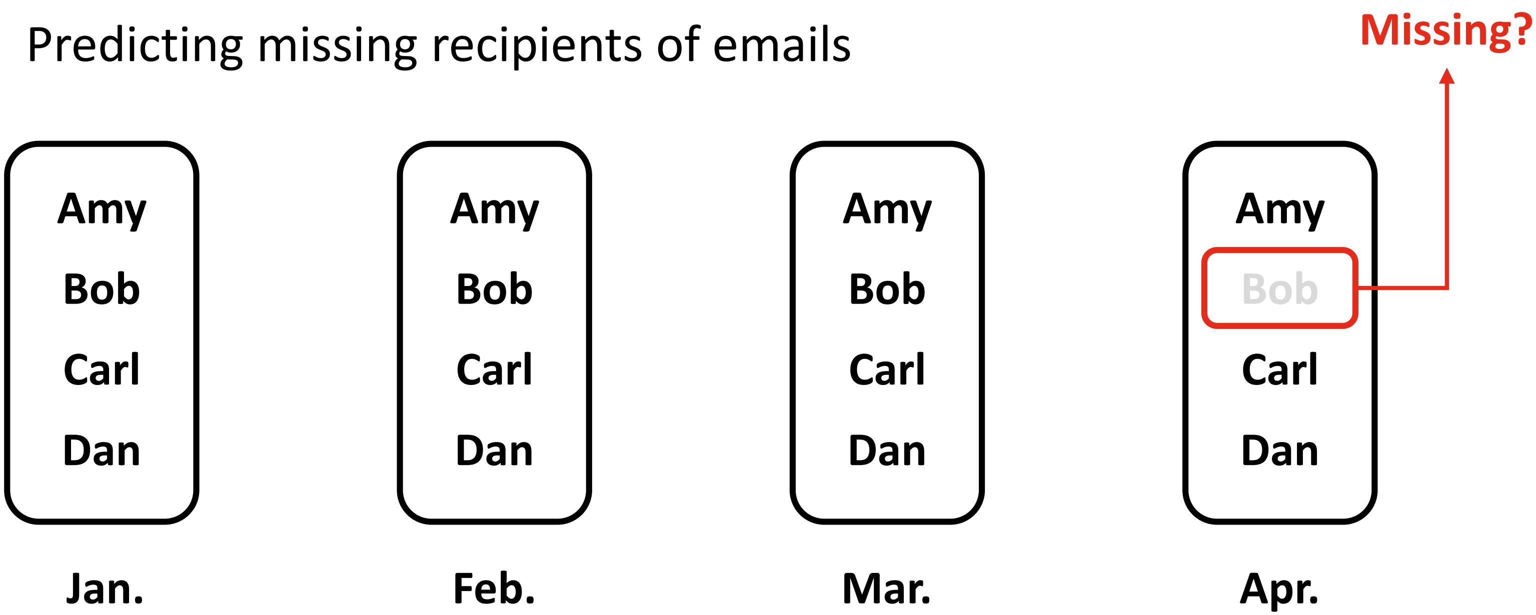
Persistence of HOIs

- HOIs can appear **repeatedly** over time
- **Persistence** of repeated HOIs can be used to measure the strength or robustness of group relations



Applications

- Predicting the persistence of HOIs has many **potential applications**
  - Recommending groups (e.g., Facebook groups) in social networks
  - Recommending multiple items together
  - Predicting missing recipients of emails



Our Questions

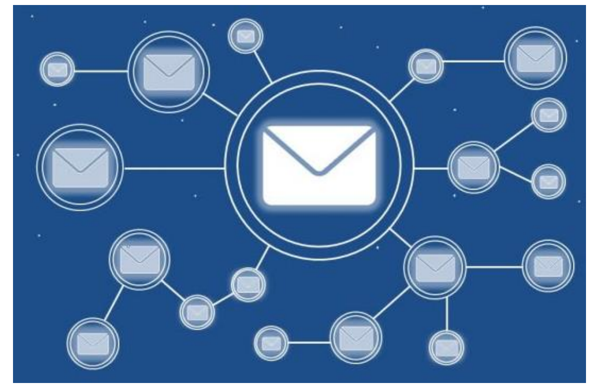
1. How do HOIs in real-world hypergraphs **persist over time**?
2. What are the **key factors** governing the persistence?
3. How accurately can we **predict** the persistence?

Roadmap

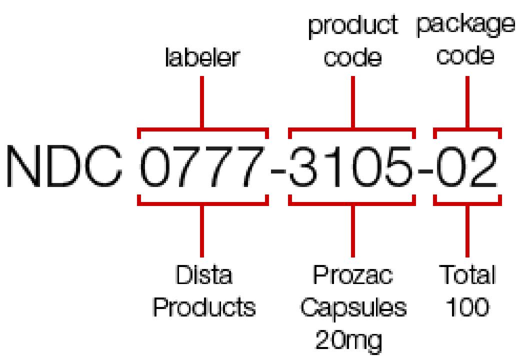
- Introduction
- **Observations <<**
  - Hypergraph-Level Analysis
  - Group-Level Analysis
  - Node-Level Analysis
- Predictions
- Conclusions

Datasets

Coauthorship



Email



NDC

#boot
#networking
#drivers
#server
#wireless

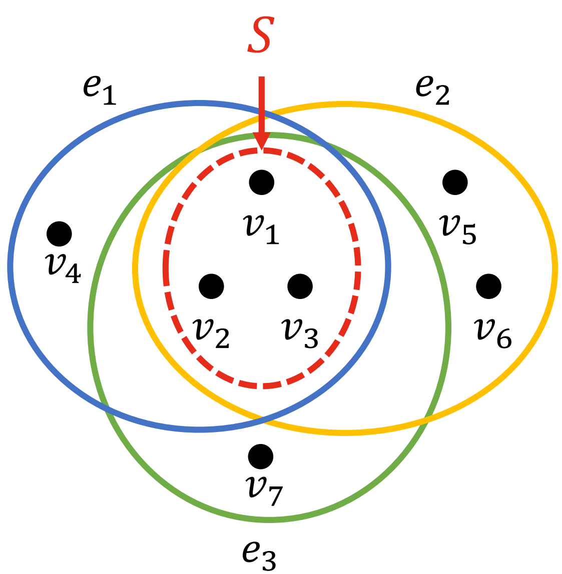
Tags

Datasets

Domain	Dataset	Node	Hyperedge	Time Unit
Coauthorship	DBLP Geology History	an author	authors	1 Year
Contact	High Primary	a person	a group interaction	1 Day 6 Hours
Email	Enron Eu	an email address	sender and all receivers	1 Month 2 Weeks
NDC	Classes	a class label	class labels applied to a drug	2 Years
	Substances	a substance	substances in a drug	
Tags	Math.sx Ubuntu	a tag	tags added to a question	1 Month
Threads	Math.sx Ubuntu	a user	users who participate in a thread	1 Month

Timestamped Hyperedges

- For each HOI S,
  - $E(S)$: Set of hyperedges containing S
  - $E(S,t)$: Set of hyperedges at time t containing S
    - ➢ Hyperedge e_i is associated with the timestamp t_i



Timestamped Hyperedges:

$e_1 = \{\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, v_4\}, \quad t_1 = 1$

$e_2 = \{\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, v_5, v_6\}, \quad t_2 = 1$

$e_3 = \{\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, v_7\}, \quad t_3 = 3$

Examples:

$S = \{v_1, v_2, v_3\}$

$E(S) = \{e_1, e_2, e_3\}$

$E(S,1) = \{e_1, e_2\}$

$E(S,2) = \emptyset$

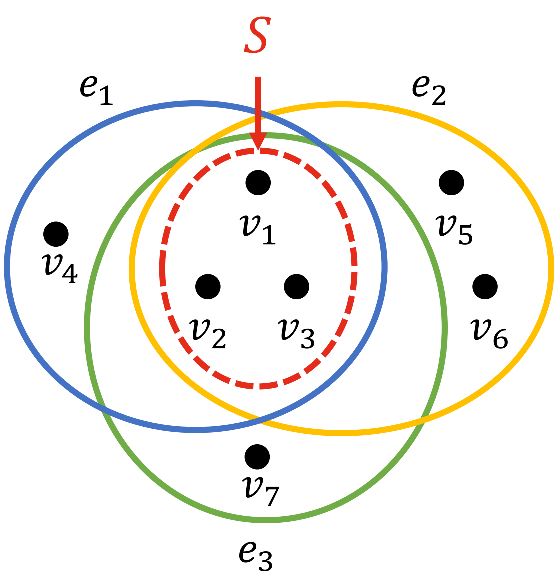
$E(S,3) = \{e_3\}$

Measure: Persistence of a HOI

- **Persistence** of a HOI S over a time range T is the number of time units in T when S co-appear in any hyperedge, i.e.,

$$P(S,T) \coloneqq \sum_{t\in T} I(S,t)$$

where $I(S,t) = \begin{cases} 1, & if\ |E(S,t)| \geq 1 \\ 0, & otherwise \end{cases}$



$E(S,1) = \{e_1, e_2\}$
$E(S,2) = \emptyset$
$E(S,3) = \{e_3\}$

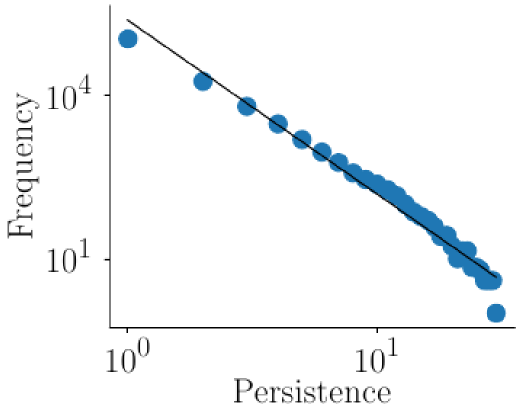
$$\boldsymbol{P(S,[1,3]) = \sum_{t=1}^{3} I(S,t) = 1+0+1 = 2}$$

Roadmap

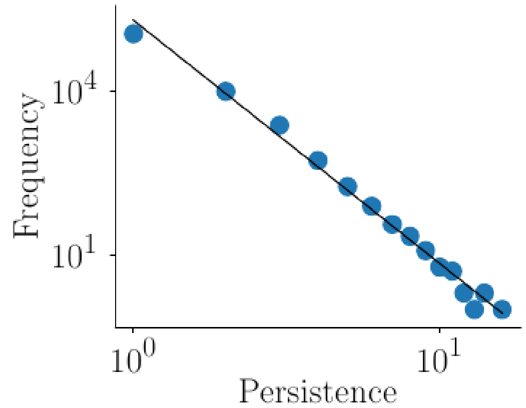
- Introduction
- Observations
  - **Hypergraph-Level Analysis <<**
  - Group-Level Analysis
  - Node-Level Analysis
- Predictions
- Conclusions

Persistence vs. Frequency

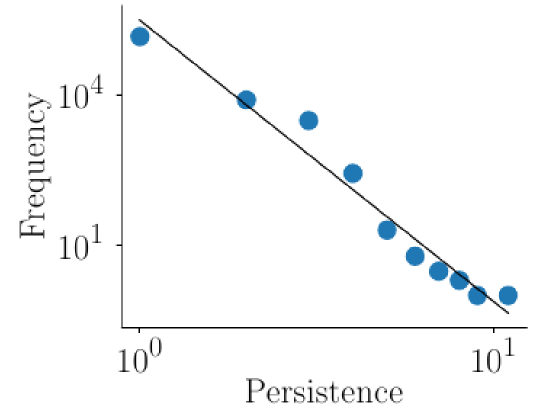
Obs. 1: Persistence of HOIs tends to follow a **power-law**.



DBLP ($|S| = 2$)



DBLP ($|S| = 3$)



DBLP ($|S| = 4$)

	R^2 **of Fitted Line**		
Size of HOIs	2	3	4
Average over all 13 datasets	**0.90**	**0.90**	**0.90**

Persistence vs. Size of HOIs

Obs. 2: As HOIs grow in size, their **average persistence** and the **power-law exponents** of fitted power-law distributions tend to decrease.

Dataset	Average Persistence (Relative)			Power-Law Exponent (Relative)		
Size of HOIs	2	3	4	2	3	4
Average over all 13 datasets	**1.00**	**0.72**	**0.63**	**1.00**	**0.71**	**0.59**

Roadmap

- Introduction
- Observations
  - Hypergraph-Level Analysis
  - **Group-Level Analysis <<**
  - Node-Level Analysis
- Predictions
- Conclusions

Group Features vs. Group Persistence

- We examined the relations between the structural group features and the persistence of HOIs (i.e., group persistence)
- We measured the **Pearson correlation coefficient (CC)** and **normalized mutual information (MI)** between the persistence and each structural feature to examine the relation between them
  - Normalized mutual information scales from 0 (no mutual information) to 1 (perfect correlation)

Group Features: Definition

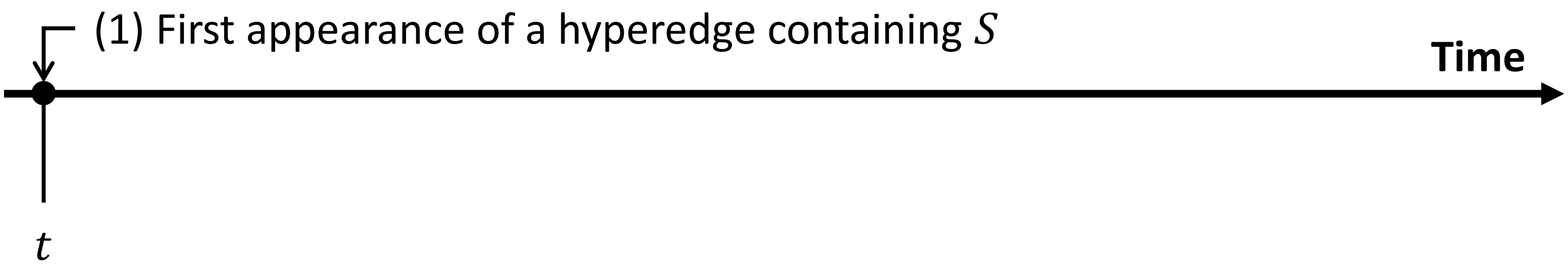
- Basic structural features of each HOI S:
  - #: number of hyperedges including S
  - Σ: sum of sizes of hyperedges containing S
  - $\cup$: number of hyperedges overlapping S
  - $\Sigma \cup$: sum of sizes of hyperedges overlapping S
  - $\cap$: number of common neighbors of S
  - $\mathcal{H}$: entropy in sizes of hyperedges containing S
- Group structural features of each HOI S:
  - ➢ (1) #, (2) #/∪, (3) Σ/(Σ ∪), (4) ∩, (5) #/∩, (6) Σ/∩, (7) Σ/#, (8) $\mathcal{H}$

density of hyperedges containing S

avg. sizes of hyperedges containing S

Measure: Structural Features & Persistence

1) HOI S appears in a hyperedge for the first time at time t

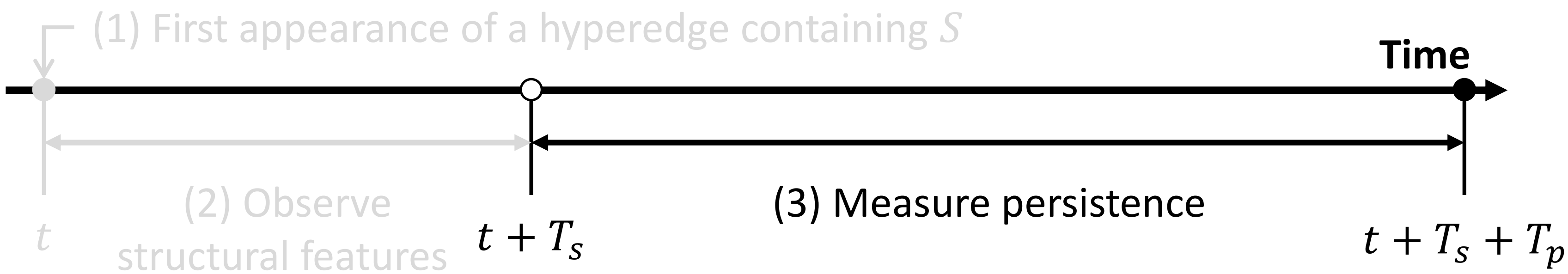


Measure: Structural Features & Persistence

1) HOI S appears in a hyperedge for the first time at time t
2) Compute its structural features using only the hyperedges appearing between time $t+1$ and $t+T_s$

Measure: Structural Features & Persistence

1) HOI S appears in a hyperedge for the first time at time t
2) Compute its structural features using only the hyperedges appearing between time $t+1$ and $t+T_s$
3) Measure its persistence between time $t+T_s+1$ and $t+T_s+T_p$



- We set $T_s = 5$ and $T_p = 10$

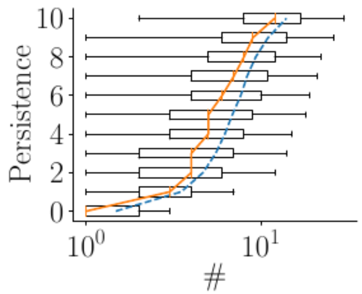
Group Features vs. Group Persistence

Obs. 3: Persistence of each HOI S is positively correlated with (a) the **number of hyperedges containing S** and (b) the **entropy in the sizes of hyperedges containing S**.

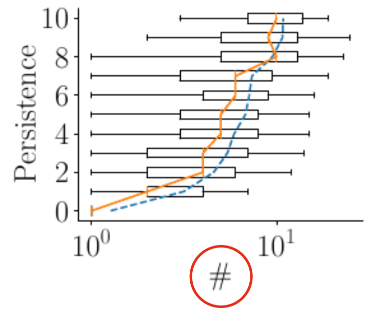
	Size of HOIs	$\#$	$\frac{\#}{\cup}$	$\frac{\Sigma}{\Sigma \cup}$	$\cap$	$\frac{\#}{\cap}$	$\frac{\Sigma}{\cap}$	$\frac{\Sigma}{\#}$	$\mathcal{H}$
MI	2	0.13	0.11	0.14	0.05	0.10	0.12	0.10	**0.15**
	3	0.11	0.06	0.08	0.05	0.08	0.09	0.08	**0.12**
	4	0.11	0.05	0.07	0.06	0.07	0.10	0.07	**0.12**
	Avg.	0.12	0.08	0.10	0.05	0.08	0.11	0.08	**0.13**
CC	2	**0.36**	0.09	0.09	0.17	0.19	0.26	-0.08	0.32
	3	**0.31**	0.10	0.10	0.05	0.16	0.20	-0.09	0.25
	4	**0.30**	0.13	0.13	-0.01	0.17	0.20	-0.10	0.24
	Avg.	**0.32**	0.10	0.11	0.07	0.17	0.22	-0.09	0.27

Group Features vs. Group Persistence

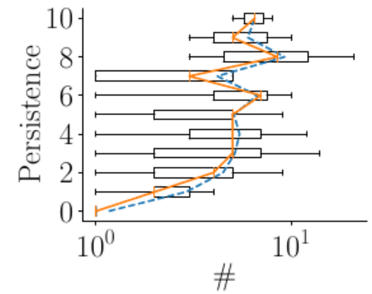
Obs. 3: Persistence of each HOI S is positively correlated with (a) the **number of hyperedges containing S** and (b) the **entropy in the sizes of hyperedges containing S**.



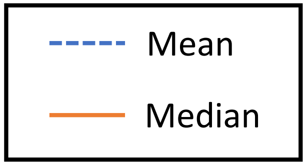
DBLP ($|S| = 2$)

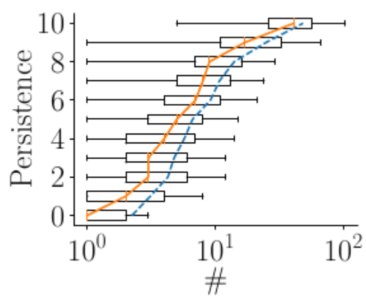


DBLP ($|S| = 3$)

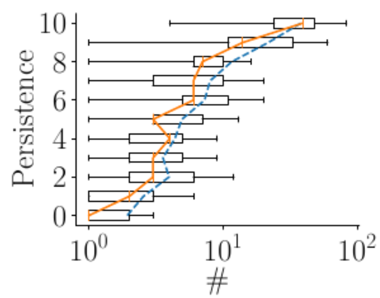


DBLP ($|S| = 4$)

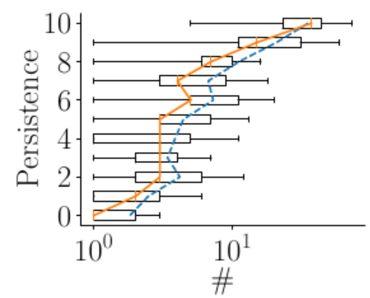




Eu ($|S| = 2$)



Eu ($|S| = 3$)



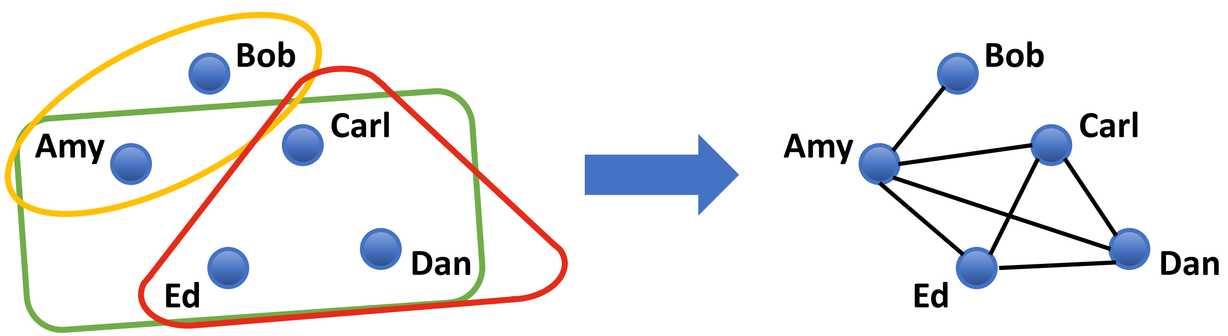
Eu ($|S| = 4$)

Node Features: Definition

- We examine the relations between the persistence of each HOI (i.e., group persistence) and the structural features of individual nodes involved in the HOI
- Structural features of each node v in the clique expansion:
  a. **degree** $d(v)$
  b. **weighted degree** $w(v)$
  c. **core number** $c(v)$
  d. **PageRank** $r(v)$
  e. **average degree of neighbors** $\bar{d}(v)$
  f. **average weighted degree of neighbors** $\overline{w}(v)$
  g. **local clustering coefficient** $l(v)$
  h. **number of occurrences of v** $o(v)$

Clique Expansion: Definition

- The **clique expansion** of a hypergraph is a pairwise graph between nodes
- It is obtained by replacing each hyperedge with the clique with the nodes in the hyperedge



Hypergraph Clique Expansion

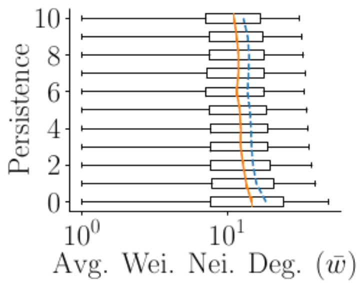
Node Features vs. Group Persistence

Obs. 4: Persistence of each HOI S is negatively correlated with the **average (weighted) degree of neighbors** of each node involved in the HOI.

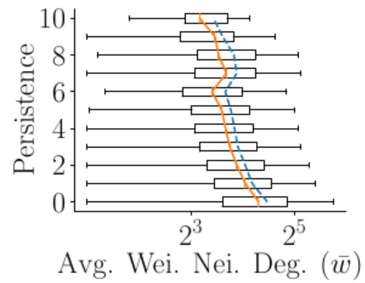
	Size of HOIs	d	w	c	r	$\bar{d}$	$\bar{w}$	l	o
MI	2	0.04	0.09	0.04	**0.17**	0.16	0.17	0.15	0.08
	3	0.03	0.06	0.04	0.09	0.09	**0.10**	0.09	0.05
	4	0.03	0.05	0.06	0.07	0.07	**0.07**	0.07	0.04
	Avg.	0.04	0.07	0.05	0.11	0.11	**0.11**	0.10	0.05
CC	2	0.05	0.09	-0.01	0.07	-0.12	**-0.14**	-0.08	0.09
	3	-0.02	0.06	-0.05	0.03	-0.11	**-0.12**	-0.02	0.05
	4	-0.07	0.03	-0.09	0.03	**-0.14**	-0.14	0.03	0.00
	Avg.	-0.01	0.06	-0.05	0.04	-0.12	**-0.13**	-0.02	0.05

Node Features vs. Group Persistence

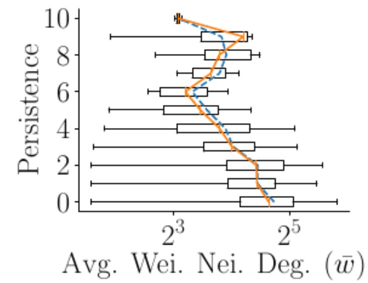
Obs. 4: Persistence of each HOI S is negatively correlated with the **average (weighted) degree of neighbors** of each node involved in the HOI.



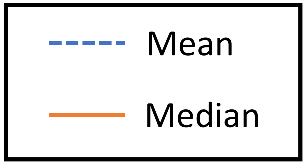
DBLP ($|S| = 2$)

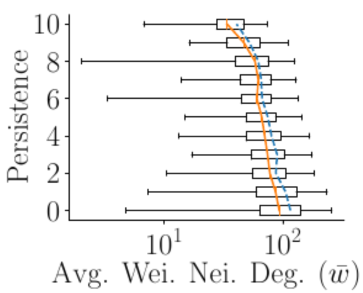


DBLP ($|S| = 3$)

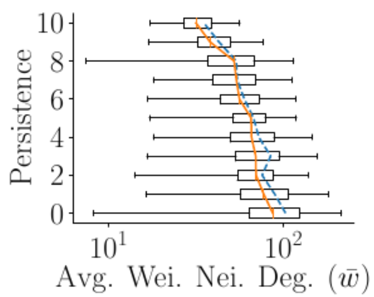


DBLP ($|S| = 4$)

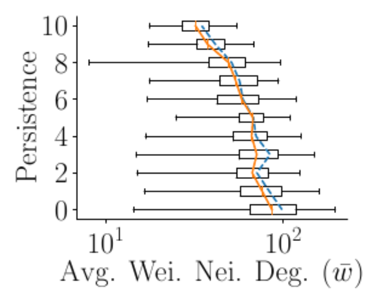




Eu ($|S| = 2$)



Eu ($|S| = 3$)



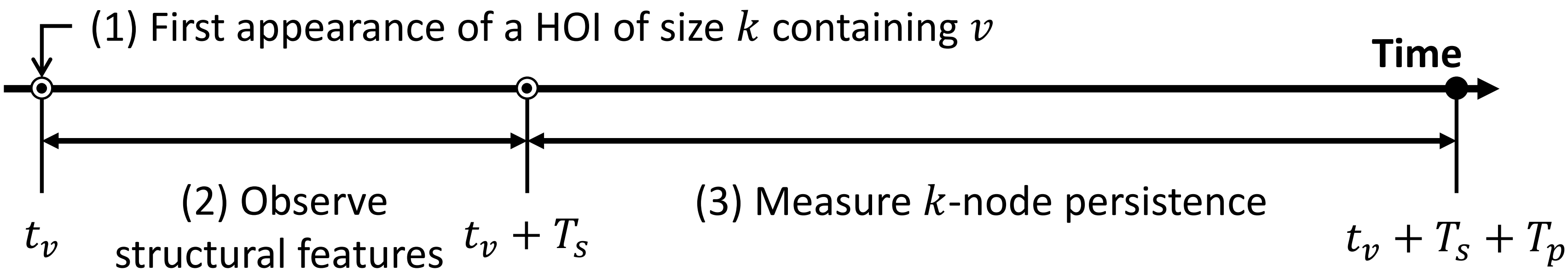
Eu ($|S| = 4$)

Roadmap

- Introduction
- Observations
  - Hypergraph-Level Analysis
  - Group-Level Analysis
  - **Node-Level Analysis <<**
- Predictions
- Conclusions

Node Features vs. Node Persistence

- We explore the relations between the structural features of each node and its k-node persistence
- **k-node persistence** of a node v: average persistence of the HOIs of size $k \in \{2,3,4\}$ that the node v is involved in
- For each node v, let t_v be the time when v is involved in any HOI of size k for the first time
  - ➢ Measure the structural node features of v using only the hyperedges appearing between time $t_v + 1$ and $t_v + T_S$



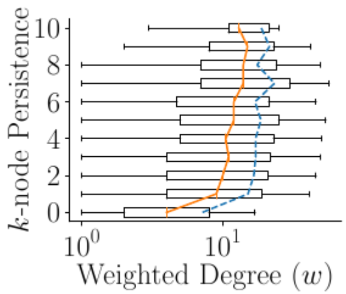
Node Features vs. Node Persistence

Obs. 5: The **weighted degree** and **number of occurrences** of each node are positively correlated with the k-node persistence of HOIs that the node is involved in.

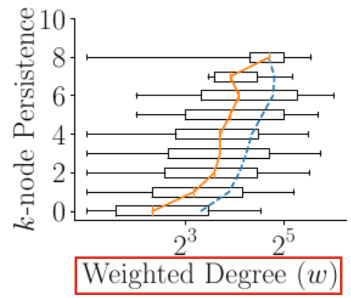
	Size of HOIs	d	w	c	r	$\bar{d}$	$\bar{w}$	l	o
MI	2	0.35	0.43	0.28	**0.53**	0.49	<u>0.51</u>	0.43	0.41
	3	0.30	0.37	0.24	**0.44**	0.42	<u>0.44</u>	0.37	0.34
	4	0.26	0.31	0.21	**0.36**	0.35	<u>0.36</u>	0.31	0.30
	Avg.	0.30	0.37	0.24	**0.44**	0.42	<u>0.43</u>	0.37	0.35
CC	2	0.15	<u>0.22</u>	0.14	0.08	0.00	-0.07	-0.02	**0.26**
	3	0.04	<u>0.16</u>	0.04	0.03	-0.04	-0.08	-0.04	**0.17**
	4	0.03	<u>0.12</u>	0.01	0.02	-0.05	-0.07	-0.04	**0.13**
	Avg.	0.07	<u>0.17</u>	0.06	0.04	-0.03	-0.07	-0.03	**0.19**

Node Features vs. Node Persistence

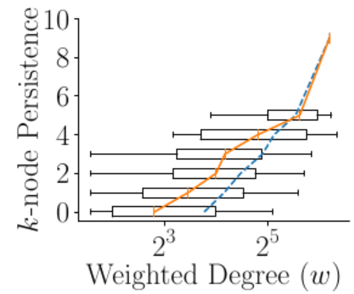
Obs. 5: The **weighted degree** and **number of occurrences** of each node are positively correlated with the k-node persistence of HOIs that the node is involved in.



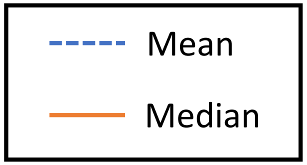
DBLP ($|S| = 2$)

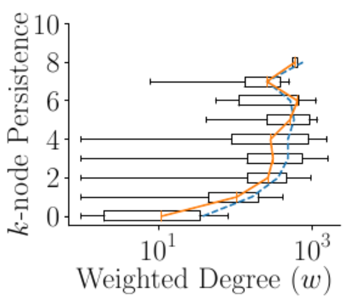


DBLP ($|S| = 3$)

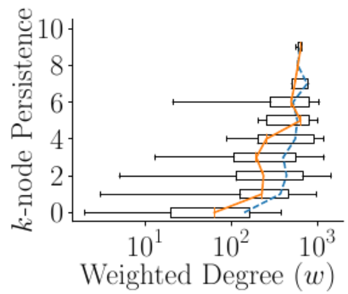


DBLP ($|S| = 4$)

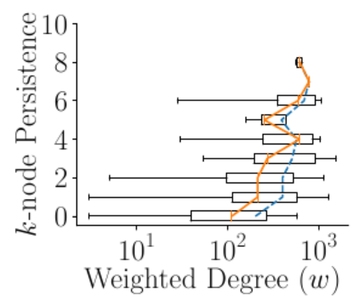




Eu ($|S| = 2$)



Eu ($|S| = 3$)



Eu ($|S| = 4$)

Roadmap

- Introduction
- Observations
    - Hypergraph-Level Analysis
    - Group-Level Analysis
    - Node-Level Analysis
- **Predictions <<**
- Conclusions

Prediction Experiments

- **Exp. 1: Predictability.** How accurately can we predict the persistence of HOIs using the structural features?
- **Exp. 2: Feature Importance.** Which structural features are important in predicting the persistence?
- **Exp. 3: Effect of Observation Periods.** How does the period of observation for measuring the structural features affect the prediction accuracy?

Problem 1: Persistence Prediction

- **Given**:
  - a **HOI S** that appears for the first time at time t,
  - all **hyperedges appearing in the past**
    - between time $t + 1$ and $t + T_s$
- **Predict**:
  - **persistence of S in the near future**
    - between $t + T_s + 1$ and $t + T_s + T_p$

Problem 2: k-Node Persistence Prediction

- **Given**:
  - a **node v** involved in a HOI of size k for the first time at time t,
  - all **hyperedges appearing in the past**
    - between time $t + 1$ and $t + T_s$
- **Predict**:
  - **k-node persistence of v in the near future**
    - between $t + T_s + 1$ and $t + T_s + T_p$

Prediction Methods

- We use all 16 structural features (8 group and 8 node features) as input features into **four regression models**:
  1) multiple linear regression (**LR**)
  2) random forest regression (**RF**)
  3) linear support vector regression (**SVR**)
  4) multi-layer perceptron regressor (**MLP**)

  ✓ **Baseline:** mean (k-node) persistence in the training set
- Training set: **2/3** of the HOIs and their persistence and **4/5** of the nodes and their k-node persistence
- Test set: the remaining ones

Evaluation Methods

- We evaluate the predictive performance of the models using two metrics:
  - **Coefficients of determination (R^2)**: measures how well the predictions approximate the real data
  - **Root mean squared error ($RMSE$)**: between predicted and real (k-node) persistence
- A **higher R^2** and **lower $RMSE$** indicate better performance

Exp. 1: Predictability

Obs. 6: The **structural features are useful** for predicting the persistence, especially when the size of the HOI is large.

Target	Prediction of Persistence of HOIs						Prediction of k-Node Persistence of Nodes					
Measure	R^{2*}			RMSE**			R^{2*}			RMSE**		
Size of HOIs	2	3	4	2	3	4	2	3	4	2	3	4
Mean	0.00	0.00	0.00	1.29	0.73	0.60	-0.01	-0.01	-0.03	0.75	0.56	0.54
SVR	0.17	0.13	0.10	1.12	0.63	0.48	0.03	0.01	0.00	0.73	0.56	0.54
LR	0.28	0.22	0.23	1.05	0.58	0.45	<u>0.17</u>	<u>0.15</u>	<u>0.09</u>	<u>0.75</u>	<u>0.71</u>	<u>0.67</u>
MLP	<u>0.34</u>	<u>0.31</u>	<u>0.37</u>	<u>0.95</u>	<u>0.53</u>	<u>0.42</u>	0.14	0.06	0.02	0.77	0.75	0.72
RF	**0.61**	**0.62**	**0.68**	**0.83**	**0.38**	**0.24**	**0.61**	**0.66**	**0.71**	**0.54**	**0.41**	**0.39**

*The higher, the better. **The lower, the better.

Measure: Feature Importance

- We use the **Gini importance** to measure the importance of each structural feature for random forest
- We compute the **rankings** of the features based on the importance

Exp. 2: Feature Importance

Obs. 7: In predicting the persistence, the **number of hyperedges containing S (i.e., $\#$)**, and the **average (weighted) degree of the neighbors of each node in S (i.e., $\bar{w}$ and $\bar{d}$)** are most useful.

Size of HOIs	$\#$	$\frac{\#}{\cup}$	$\frac{\Sigma}{\Sigma\cup}$	$\cap$	$\frac{\#}{\cap}$	$\frac{\Sigma}{\cap}$	$\frac{\Sigma}{\#}$	$\mathcal{H}$	d	w	c	r	$\bar{d}$	$\bar{w}$	l	o
2	**2.8**	10.7	8.6	13.1	13.3	9.0	9.2	8.7	9.9	8.6	8.8	5.9	*4.9*	4.3	6.4	11.9
3	*5.4*	9.2	9.2	11.8	11.2	9.6	9.8	7.9	11.2	9.1	8.4	5.7	5.1	**4.3**	6.4	12.0
4	**5.3**	9.3	9.9	10.3	10.6	8.3	8.7	*7.0*	9.5	7.3	9.2	7.7	7.7	6.3	8.0	11.0
Avg.	**4.5**	9.7	9.2	11.7	11.7	9.0	9.2	7.9	10.2	8.3	8.8	6.4	*5.9*	5.0	6.9	11.6

Feature Importance Ranking

Exp. 2: Feature Importance

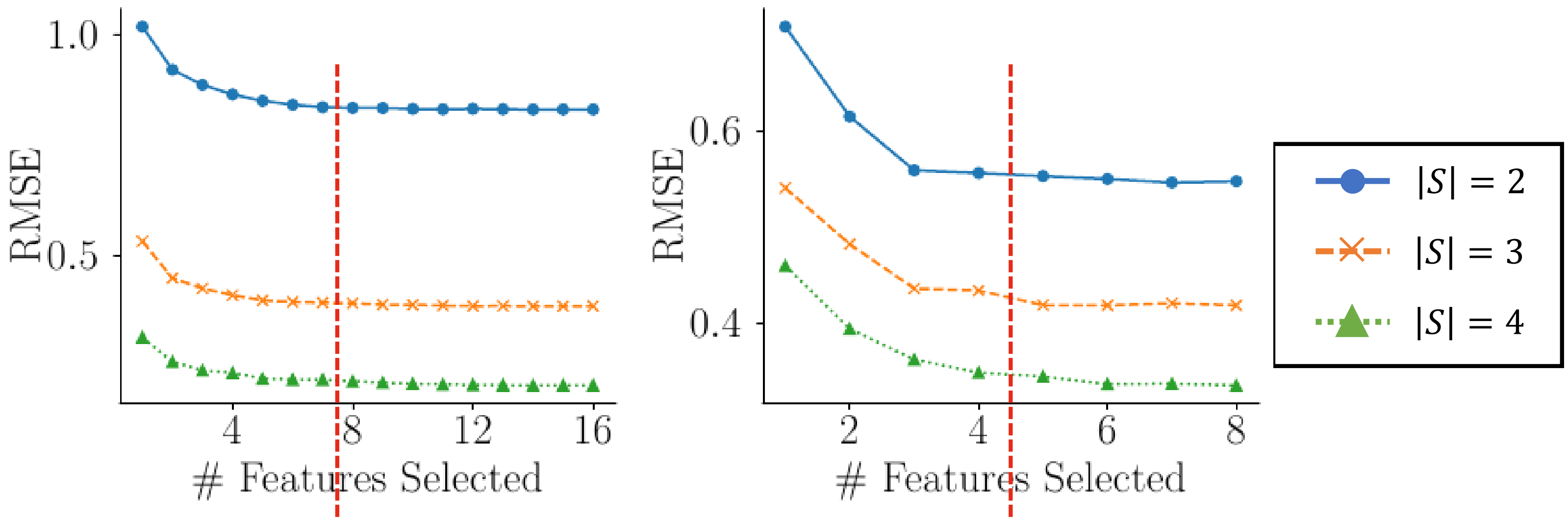
Obs. 8: In predicting the k-node persistence, its **PageRank (i.e., r)** and the **average (weighted) degree of its neighbors (i.e., $\bar{w}$ and $\bar{d}$)** are most useful.

Size of HOIs	d	w	c	r	$\bar{d}$	$\bar{w}$	l	o
2	6.7	4.3	7.2	<u>3.2</u>	3.4	**2.9**	5.3	<u>3.2</u>
3	6.6	4.1	7.3	**2.7**	*3.5*	**2.7**	5.0	4.3
4	6.1	4.0	6.6	**2.6**	*3.5*	<u>3.1</u>	5.3	4.9
Avg.	6.4	4.1	7.0	**2.8**	*3.5*	<u>2.9</u>	5.2	4.1

Feature Importance Ranking

Exp. 2: Effect of Number of Features

Obs. 9: About a **half of the considered structural features** based on their importance yields similar performance.



(1) Persistence (2) k-Node Persistence

Exp. 3: Effect of Observation Periods

Obs. 10: Observing HOIs for **longer periods of time** enables us to better predict their persistence.

Target	Persistence of HOIs						k-Node Persistence of Nodes					
Measure	RMSE* of RF			Improvement (in %)			RMSE* of RF			Improvement (in %)		
T_s	2**	3	4	2	3	4	2	3	4	2	3	4
1	0.96	0.48	0.32	31.6	42.3	50.7	0.62	0.46	0.43	18.5	25.5	31.8
3	0.88	0.42	0.28	34.1	45.4	55.0	0.55	**0.41**	**0.38**	24.8	**29.1**	**34.4**
5	**0.83**	**0.38**	**0.24**	**36.0**	**47.7**	**59.4**	**0.54**	0.41	0.39	**27.4**	26.4	27.5

*The lower, the better. **The size of HOIs (i.e., $|S|$).

Roadmap

- Introduction
- Observations
  - Hypergraph-Level Analysis
  - Group-Level Analysis
  - Node-Level Analysis
- Predictions
- **Conclusions <<**

Conclusions

- We empirically examined the **persistence of HOIs** at hypergraph-, group-, and node- levels in 13 real-world hypergraphs to answer the following questions:

✓ How is the persistence of HOIs **distributed**?
✓ Which **structural features** govern the persistence of HOIs?
✓ How accurately can we **forecast the persistence of HOIs?**

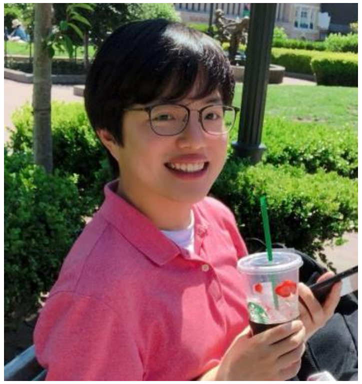
- Github link: https://github.com/jin-choo/persistence

KAIST

SIAM 2022 | International Conference on Data Mining

On the Persistence of Higher-Order Interactions in Real-World Hypergraphs

Hyunjin Choo



Kijung Shin