KAIST AI
Kim Jaechul Graduate School

THE WEB CONFERENCE ACM

MiDaS: Representative Sampling from Real-world Hypergraphs


Minyoung Choe
KAIST


Jaemin Yoo
SNU


Geon Lee
KAIST


Woonsung Baek
KAIST

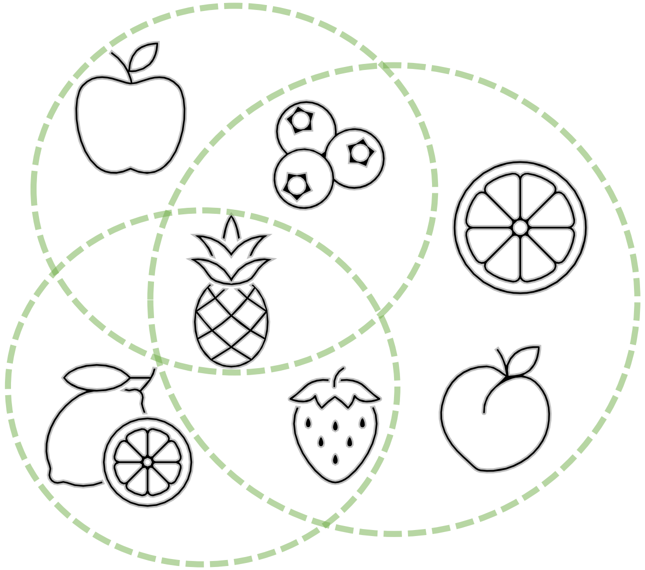

U Kang
SNU


Kijung Shin
KAIST

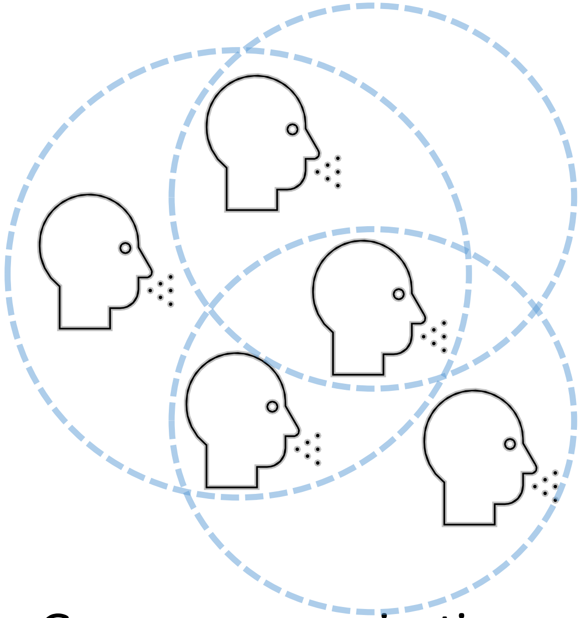
Hypergraphs represent group interactions

- Group interactions exist in many complex systems
- **Hypergraphs** consist of nodes and hyperedges, and each **hyperedge** is a subset of any number of nodes

Collaborations of researchers



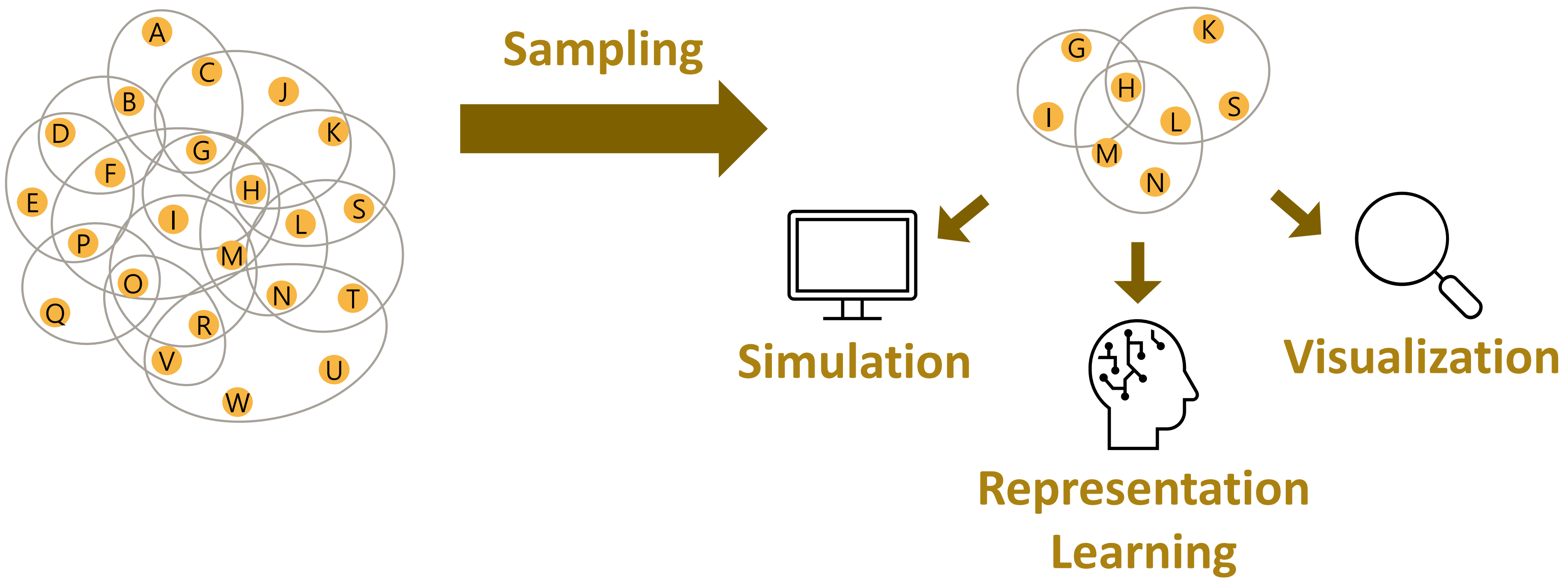
Co-purchases of items



Group communications

Sampling is indispensable

- Analyzing large-scale hypergraphs is time-consuming
- Sampling can mitigate difficulties in various tasks

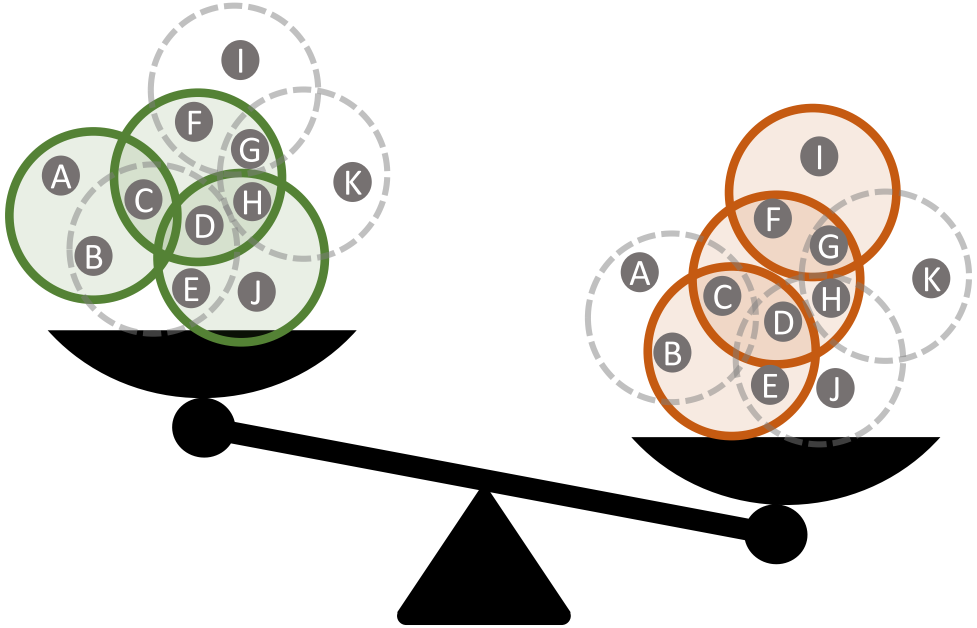


Representative Hypergraph Sampling Problem

Q1 What is a **'representative'** sample?

How can we measure the quality of a sub-hypergraph?

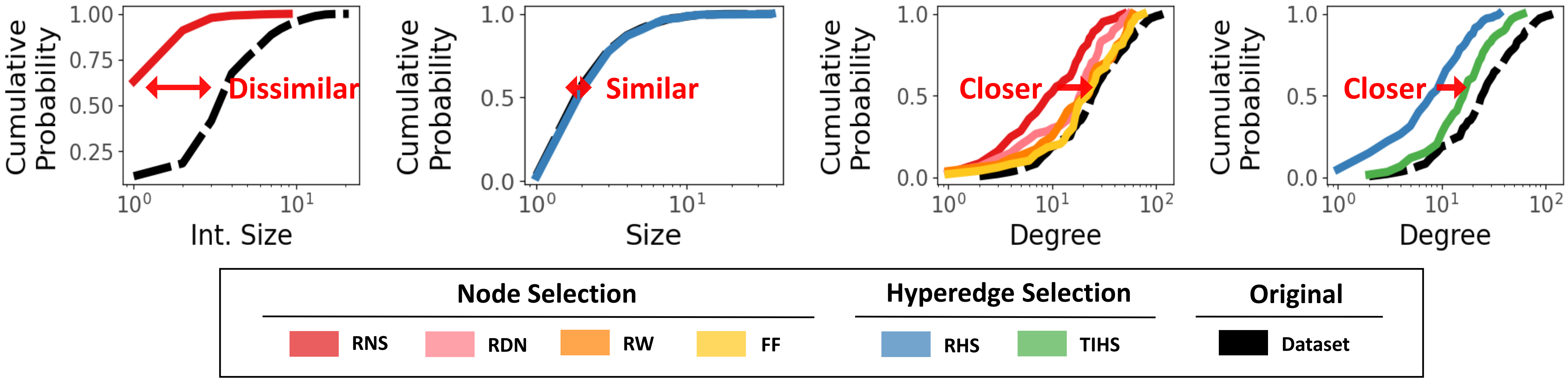
A1 We compare sampled and entire hypergraphs using **10 statistics**



Representative Hypergraph Sampling Problem

Q2 What are the **benefits and limitations** of simple and intuitive approaches for representative hypergraph sampling?

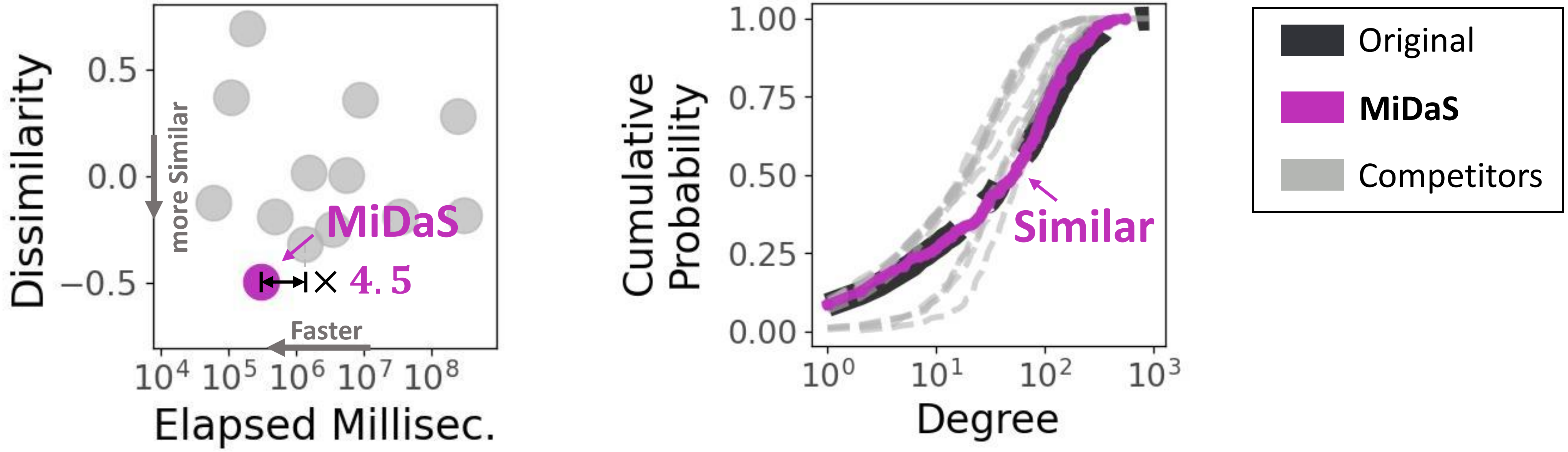
A2 We **analyze six** intuitive approaches in **11** real-world hypergraphs



Representative Hypergraph Sampling Problem

Q3 How can we find a **representative sub-hypergraph rapidly** without extensively exploring the search space?

A3 We propose **MiDaS** (Minimum Degree Biased Sampling of Hyperedges)



Roadmap

1. Introduction
2. **Problem Formulation <<**
3. Simple and Intuitive Approaches
4. MiDaS : Proposed Approach
5. Evaluation
6. Conclusions

Problem Definition

- **Given:** – a large hypergraph $\boldsymbol{G} = (\mathcal{V}, \mathcal{E})$
  – a sampling portion $\boldsymbol{p} \in (0, 1)$
- **Find:** a subhypergraph $\widehat{\boldsymbol{G}} = (\hat{\mathcal{V}}, \hat{\mathcal{E}})$ $where\ \hat{\mathcal{V}} \subseteq \mathcal{V}\ and\ \hat{\mathcal{E}} \subseteq \mathcal{E}$
- **To preserve:** "structural properties" of $\boldsymbol{G}$
- **Subject to:** $|\hat{\mathcal{E}}| = \lfloor |\mathcal{E}| \cdot p \rfloor$

Structural properties of Hypergraphs

- We consider following ten statistics,

Node-Level

P1. Degree
P2. Pair Degree

Hyperedge-Level

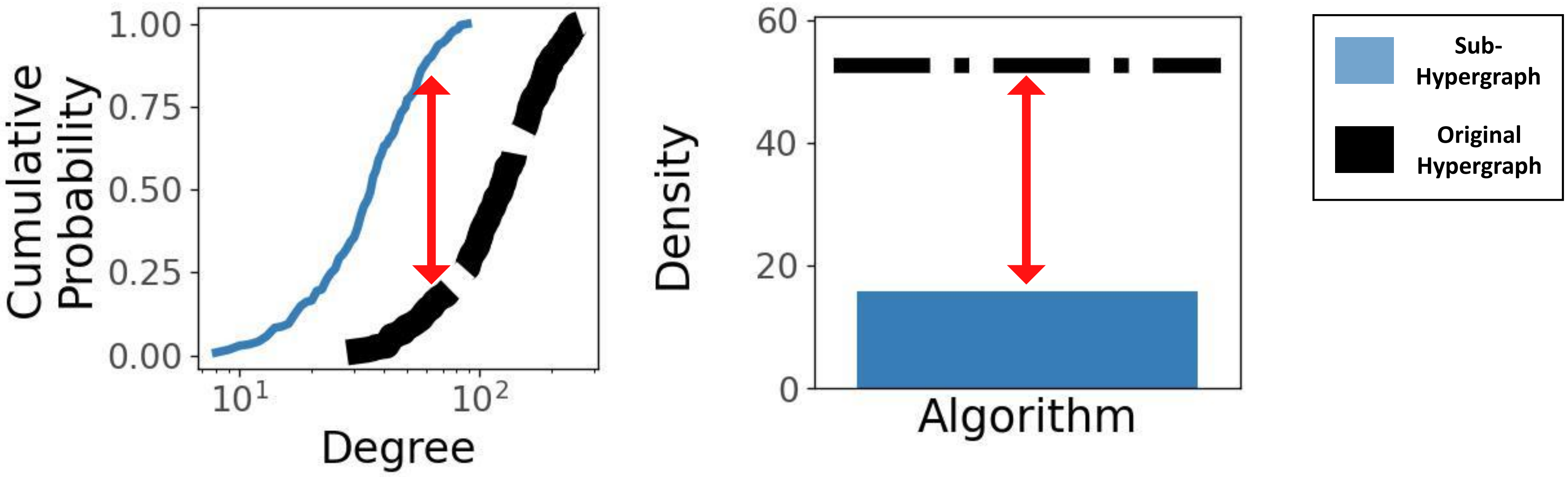
P3. Size
P4. Intersection Size

Graph-Level

P5. Singular Values
P6. Connected Component Size
P7. Global Clustering Coefficient
P8. Density
P9. Overlapness
P10. Effective Diameter

Measuring the quality of a sub-hypergraph

- For probability density functions (**P1 - P6**), we measure the **distance called Kolmogorov-Smirnov D-statistics**
- For scalar values (**P7 - P10**), we measure the **relative difference**



Evaluation Procedures

- To directly compare the qualities of sub-hypergraphs in overall P1 – P10, we average ten distances by **rankings** and **Z-Scores**

Scales are Different

	Distance in Size Dist.	Difference in Density
subhypergraph $\widehat{G_1}$	0.2	7
subhypergraph $\widehat{G_2}$	0.01	13
subhypergraph $\widehat{G_3}$	0.02	1

1. Ranking

	Size Rank	Density Rank	Average
$\widehat{G_1}$	3	2	2.5
$\widehat{G_2}$	**1**	3	2
$\widehat{G_3}$	2	**1**	**1.5**

2. Z-Score

	Size Z-Score	Density Z-Score	Average
$\widehat{G_1}$	1.6	0	0.8
$\widehat{G_2}$	**-0.7**	1.2	0.25
$\widehat{G_3}$	-0.6	**-1.2**	**-0.9**

Datasets

Domain	Dataname	Hyperedge	Node
Email	email-Enron, email-Eu	email	sender and receivers
Contact	contact-primary, contact-high	group interaction	individuals
Drugs	NDC-classes, NDC-substances	NDC code for a drug	classes or substances
Tags	tags-ubuntu, tags-math	post	tags
Threads	threads-ubuntu	question	answerers
Co-authorship	coauth-geology, coauth-history	publication	co-authors

Roadmap

1. Introduction
2. Problem Formulation
3. **Simple and Intuitive Approaches <<**
4. MiDaS : Proposed Approach
5. Evaluation
6. Conclusions

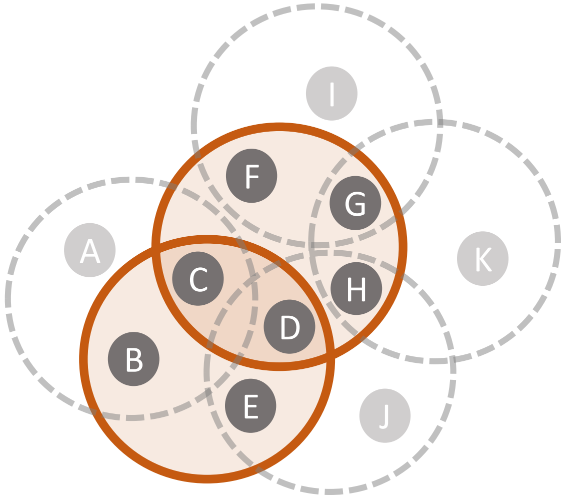
Simple and Intuitive Approaches

❑ **Node Selection**

Choose a subset of nodes and then return the induced sub-hypergraph

❑ **Hyperedge Selection**

Choose a subset of hyperedges and return this



Induced sub-hypergraph of $\{B, C, D, E, F, G, H\}$

Simple and Intuitive Approaches

❑ Node Selection

choose a subset of nodes and then return the induced sub-hypergraph

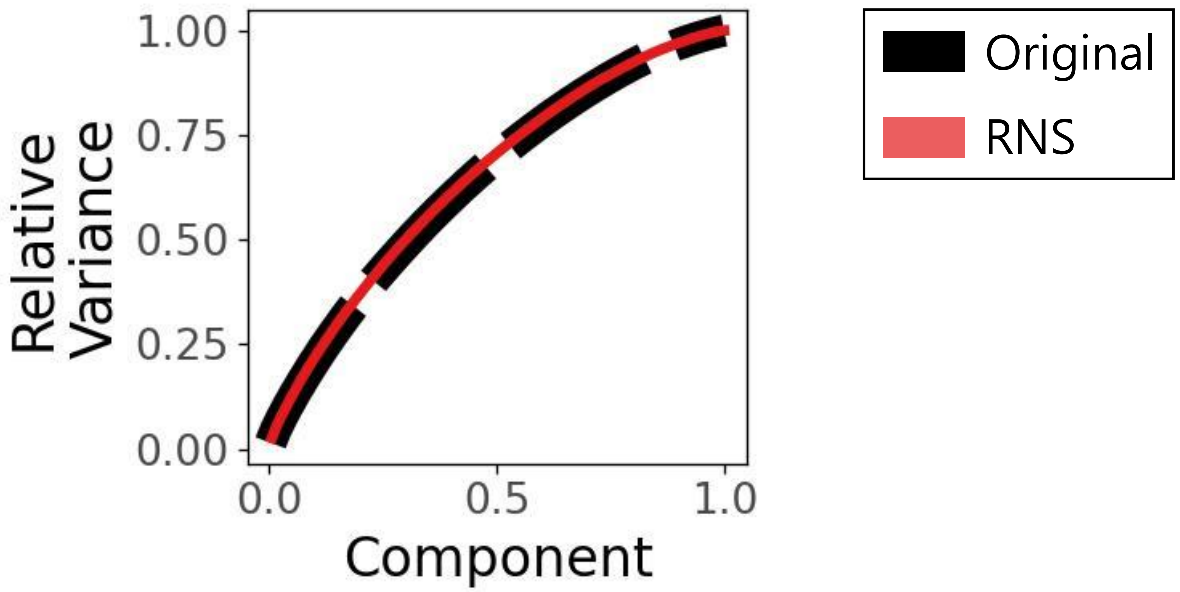
RNS	Random Node Sampling	drawing a node **uniformly** at random
RDN	Random Degree Node	drawing a node with probabilities proportional to node degrees
RW	Random Walk	random walk with restart on clique-expansion
FF	Forest Fire	forest fire in hypergraphs as in HyperFF [1]

[1] Yunbum Kook, Jihoon Ko, and Kijung Shin. 2020. Evolution of Real-world Hypergraphs: Patterns and Models without Oracles. In ICDM.

Characteristics: RNS

- **Pros**

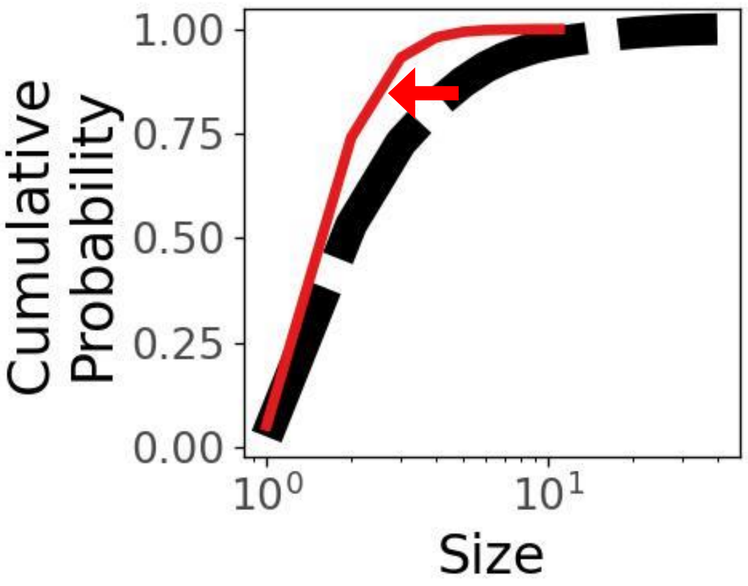
Precise preservation of spectral properties
(i.e., relative singular values)



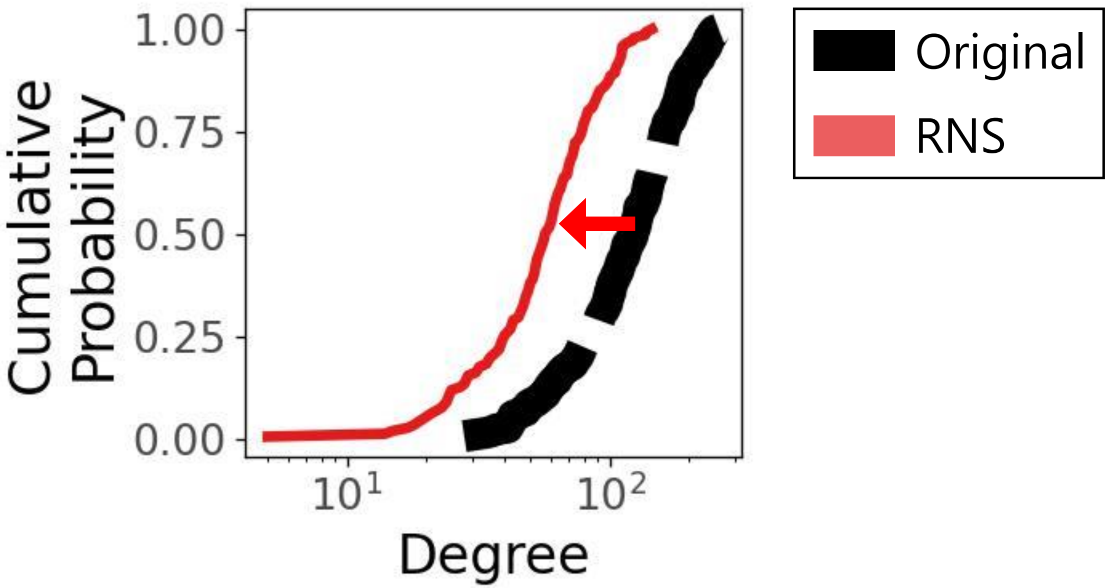
Characteristics: RNS

- **Cons**

Too many Small Hyperedges



Weak Connectivity



Simple and Intuitive Approaches

❑ Node Selection

choose a subset of nodes and then return the induced sub-hypergraph

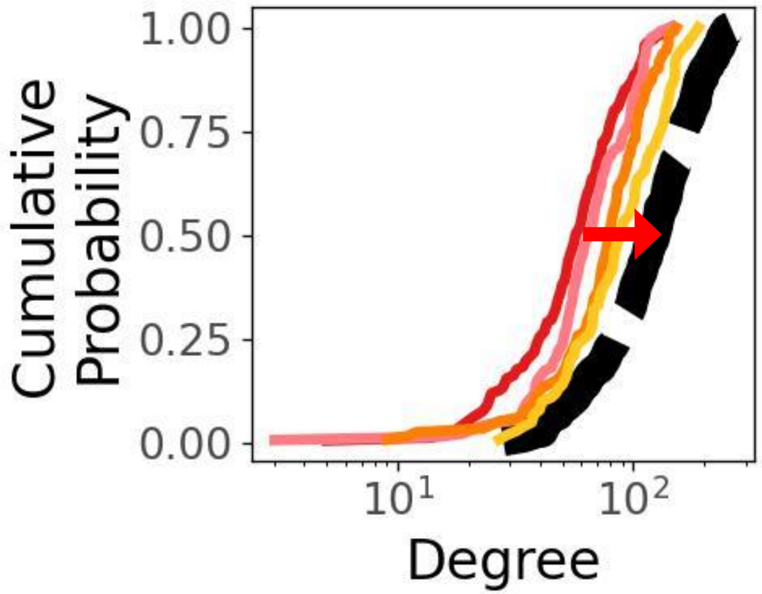
RNS	Random Node Sampling	drawing a node uniformly at random
RDN	Random Degree Node	drawing a node with probabilities **proportional to node degrees**
RW	Random Walk	**random walk** with restart on clique-expansion
FF	Forest Fire	**forest fire** in hypergraphs as in HyperFF [1]

[1] Yunbum Kook, Jihoon Ko, and Kijung Shin. 2020. Evolution of Real-world Hypergraphs: Patterns and Models without Oracles. In ICDM.

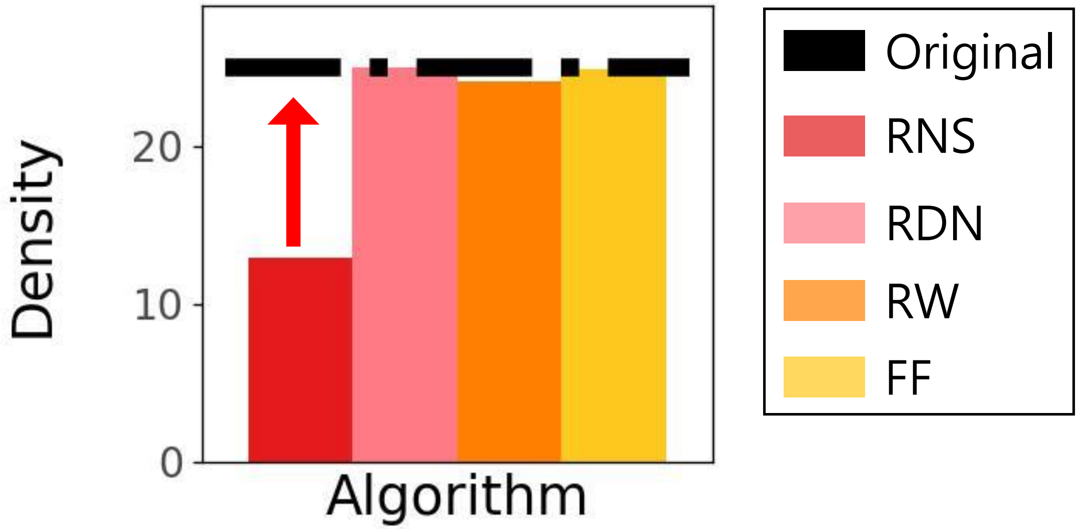
Characteristics: RDN, RW and FF

- **Pros**

More high-degree nodes than RNS



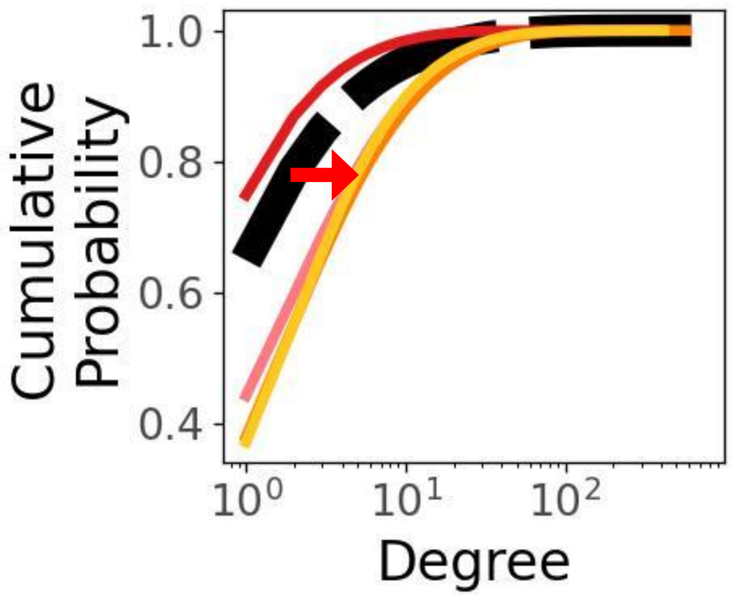
Stronger connectivity than RNS



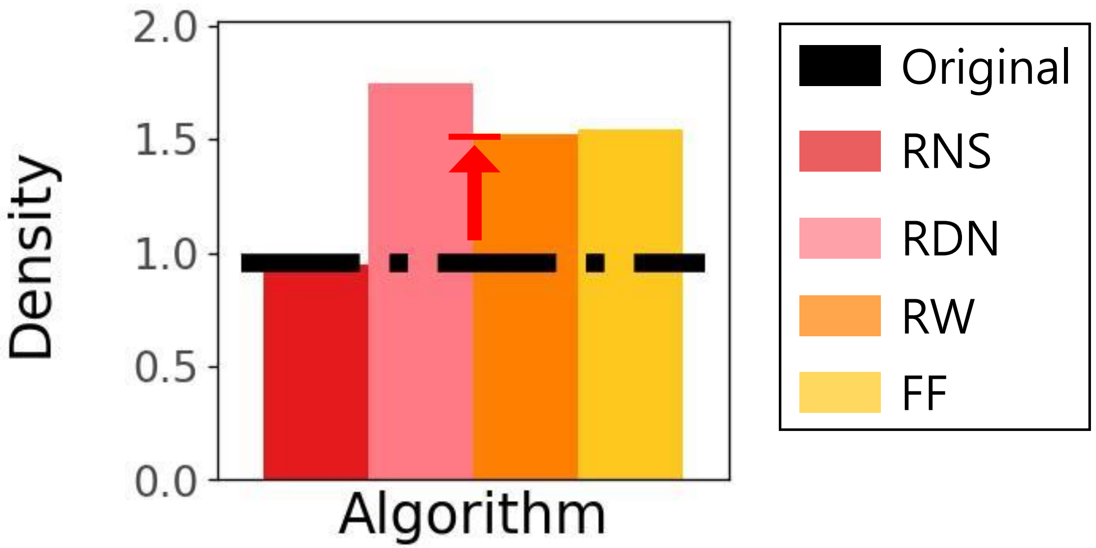
Characteristics: RDN, RW and FF

- **Cons**

Too many high-degree nodes in some datasets



Too strong connectivity in some datasets



Simple and Intuitive Approaches

❑ Hyperedge Selection

choose a subset of hyperedges and return this

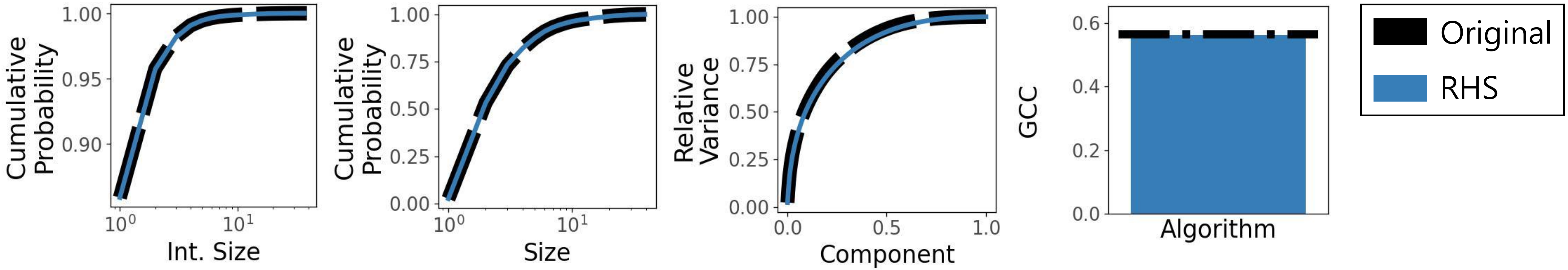
RHS	Random Hyperedge Sampling	draw a target number of hyperedges uniformly at random
TIHS	Totally-Induced Hyperedge Sampling	extend totally-induced edge sampling [2] to hypergraphs

[2] Nesreen Ahmed, Jennifer Neville, and Ramana Rao Kompella. 2011. Network sampling via edge-based node selection with graph induction.

Characteristics: RHS

- **Cons**

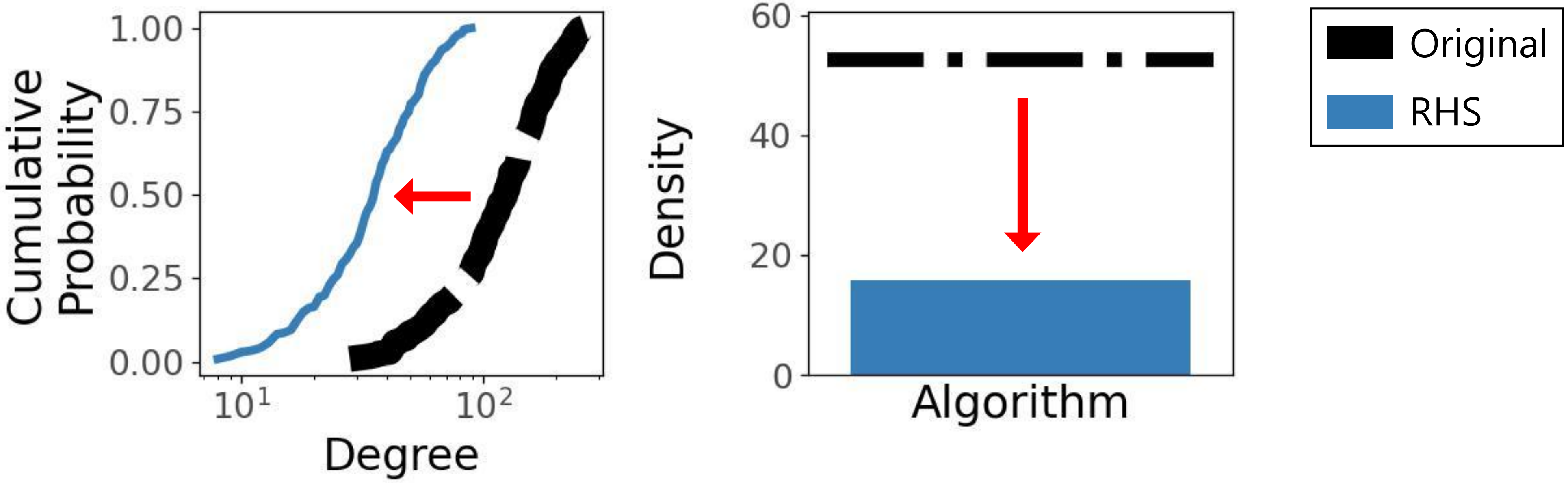
Best preservation of many properties



Characteristics: RHS

- **Cons**

Weak connectivity



Simple and Intuitive Approaches

❑ Hyperedge Selection

choose a subset of hyperedges and return this

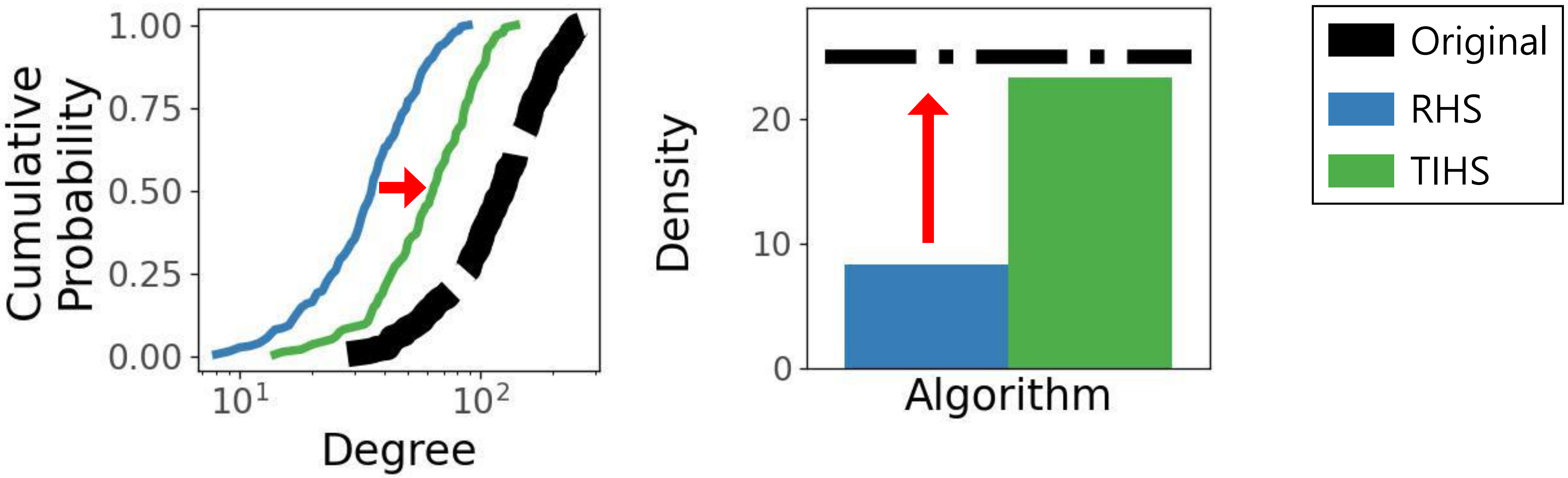
RHS	Random Hyperedge Sampling	draw a target number of hyperedges uniformly at random
TIHS	Totally-Induced Hyperedge Sampling	extend totally-induced edge sampling [2] to hypergraphs

[2] Nesreen Ahmed, Jennifer Neville, and Ramana Rao Kompella. 2011. Network sampling via edge-based node selection with graph induction.

Characteristics: TIHS

- **Pros**

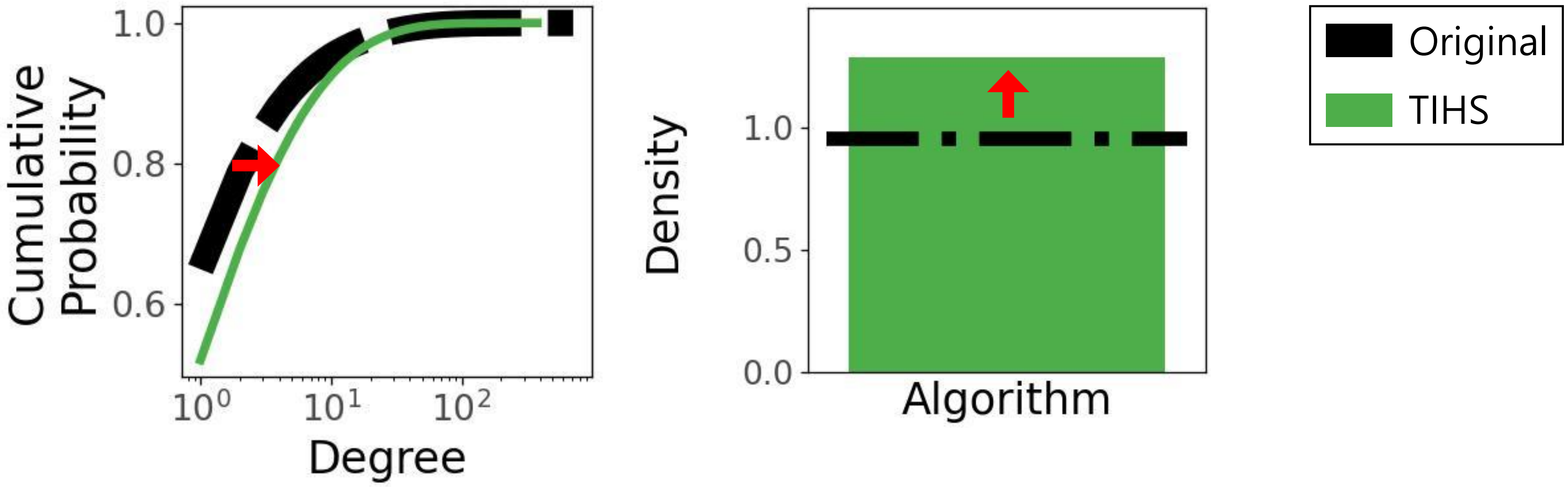
Complementarity to RHS



Characteristics: TIHS

- **Cons**

Too strong connectivity in some datasets



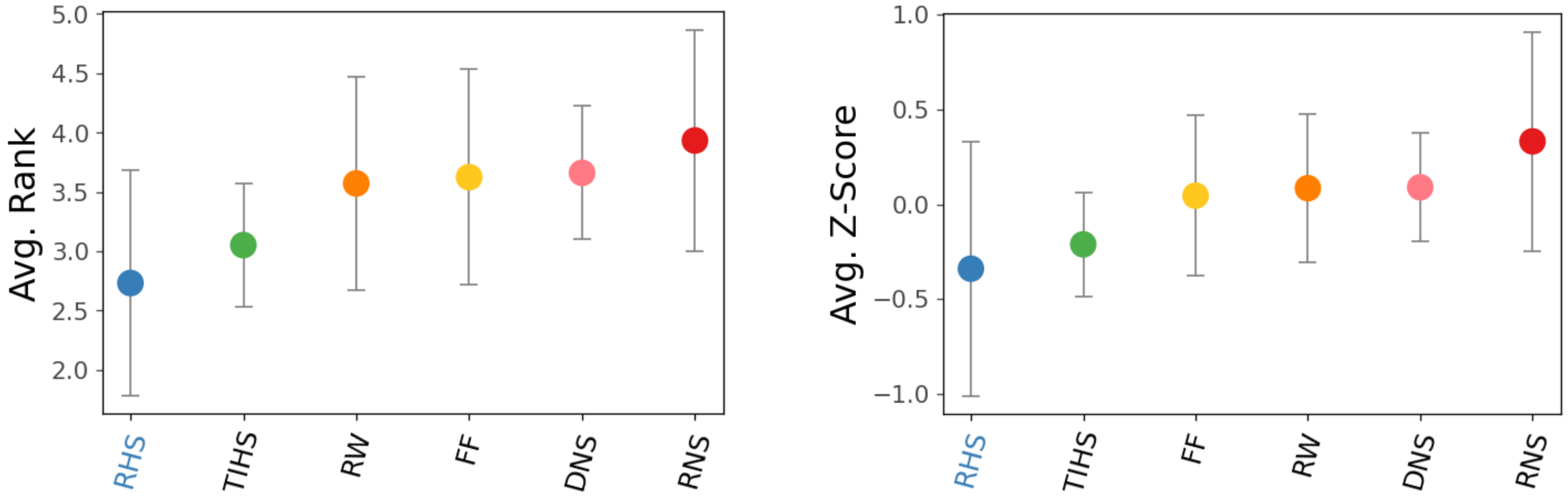
Roadmap

1. Introduction
2. Problem Formulation
3. Simple and Intuitive Approaches
4. **MiDaS: Proposed Approach <<**
5. Evaluation
6. Conclusions

Intuitions behind MiDaS

- Analyzing the simple approaches motivates us to come up with MiDaS

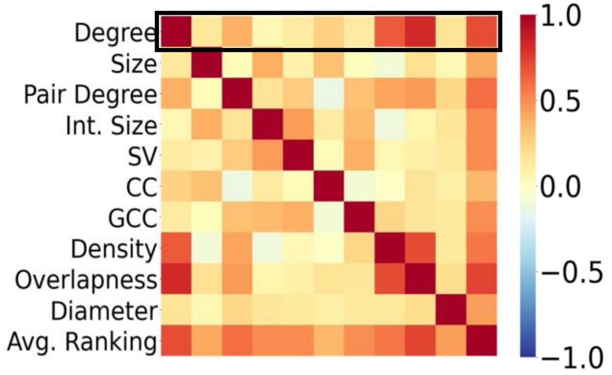
  1) **RHS** performs *best* overall, but its samples suffer from *weak connectivity*, including lack of high-degree nodes



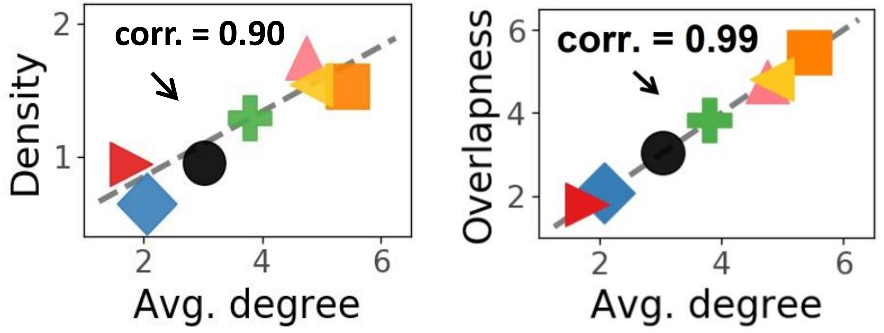
Intuitions behind MiDaS

- Analyzing the simple approaches motivates us to come up with MiDaS

2) **Degree preservation** is strongly correlated with the abilities to preserve other properties and thus the overall performance



Pearson correlation coefficients between rankings w.r.t. P1 - P10



Correlation between (a) the average degree and (b) overlapness and density

Intuitions behind MiDaS

- Analyzing the simple approaches motivates us to come up with MiDaS.

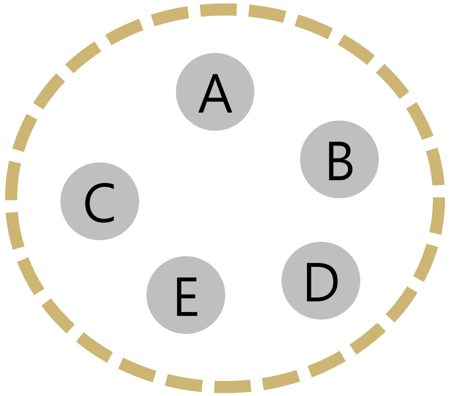
→ **Aim to overcome the lack of high-degree nodes in *RHS***

→ **Focus on better preserving node degrees**

MiDaS-Basic: Preliminary version

- To increase the fraction of high-degree nodes, **prioritize** hyperedges composed only of high-degree nodes.

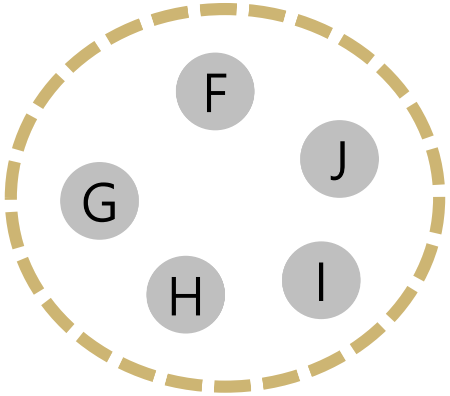
e_1



Node	Degree
A	1
B	2
C	3
D	3
E	2

<

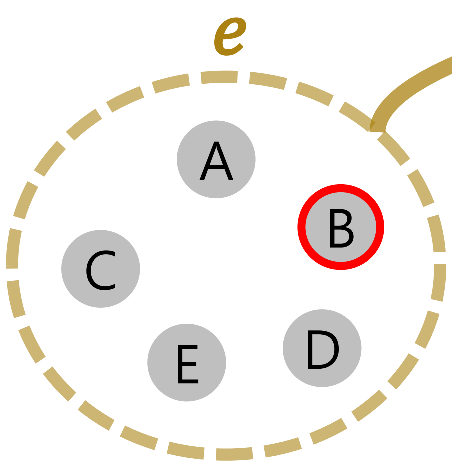
e_2



Node	Degree
F	10
G	11
H	8
I	15
J	10

MiDaS-Basic: Preliminary version

Sampling a target number of hyperedges with probability proportional to ***the minimum degree of nodes* in each hyperedge** to the power of $\boldsymbol{\alpha}$



Probability of Sampling $e \propto \mathbf{2}^{\alpha}$

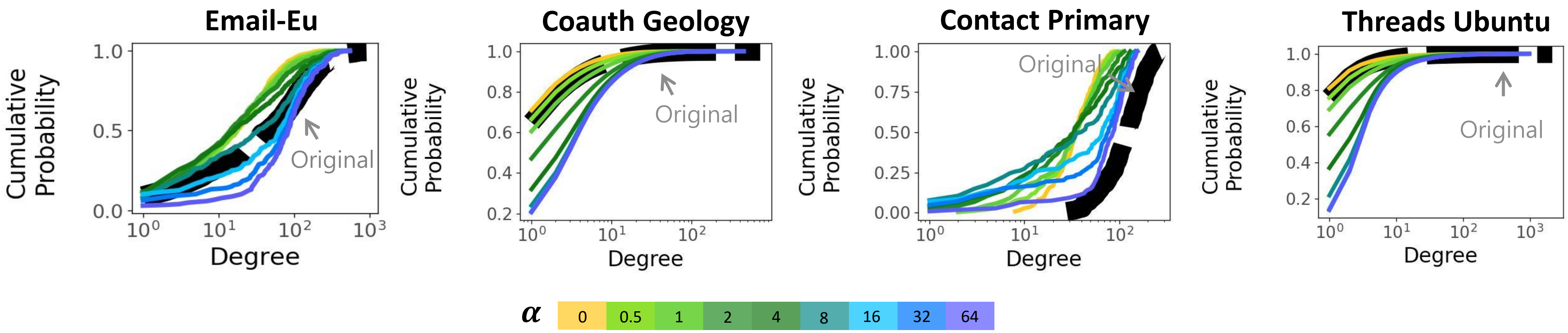
Node	Degree
A	4
B	**2**
C	3
D	8
E	6

: Minimum degree of nodes in e

Empirical Properties of MiDaS-Basic

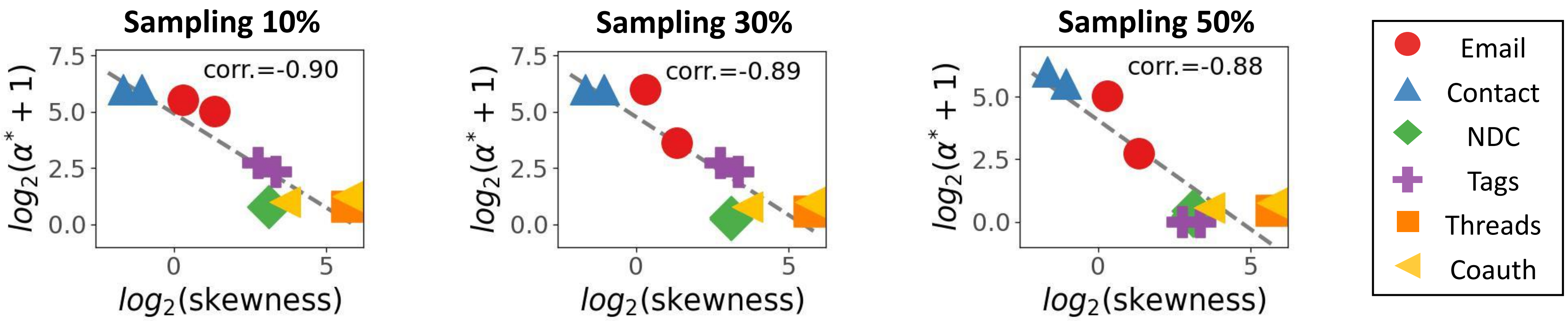
Obs. 1 As α **increases**, the degree distributions in samples tend to be **more biased towards high-degree nodes**

→ *the bias in degree distributions can be directly controlled by α*



Empirical Properties of MiDaS-Basic

Obs. 2 As degree distributions in original hypergraphs are **more skewed, larger α** values are required to preserve the distributions

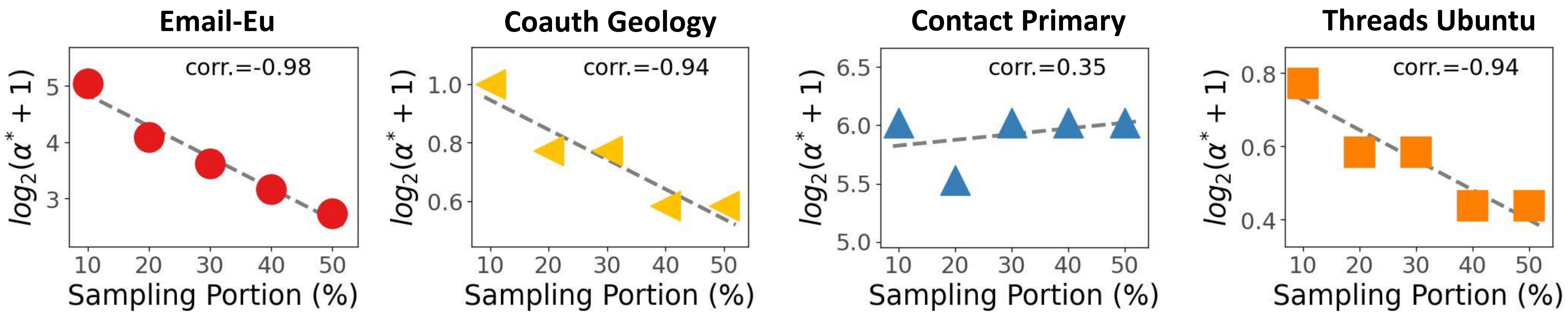


Empirical Properties of MiDaS-Basic

Obs. 2 As degree distributions in original hypergraphs are **more skewed**, **larger α** values are required to preserve the distributions

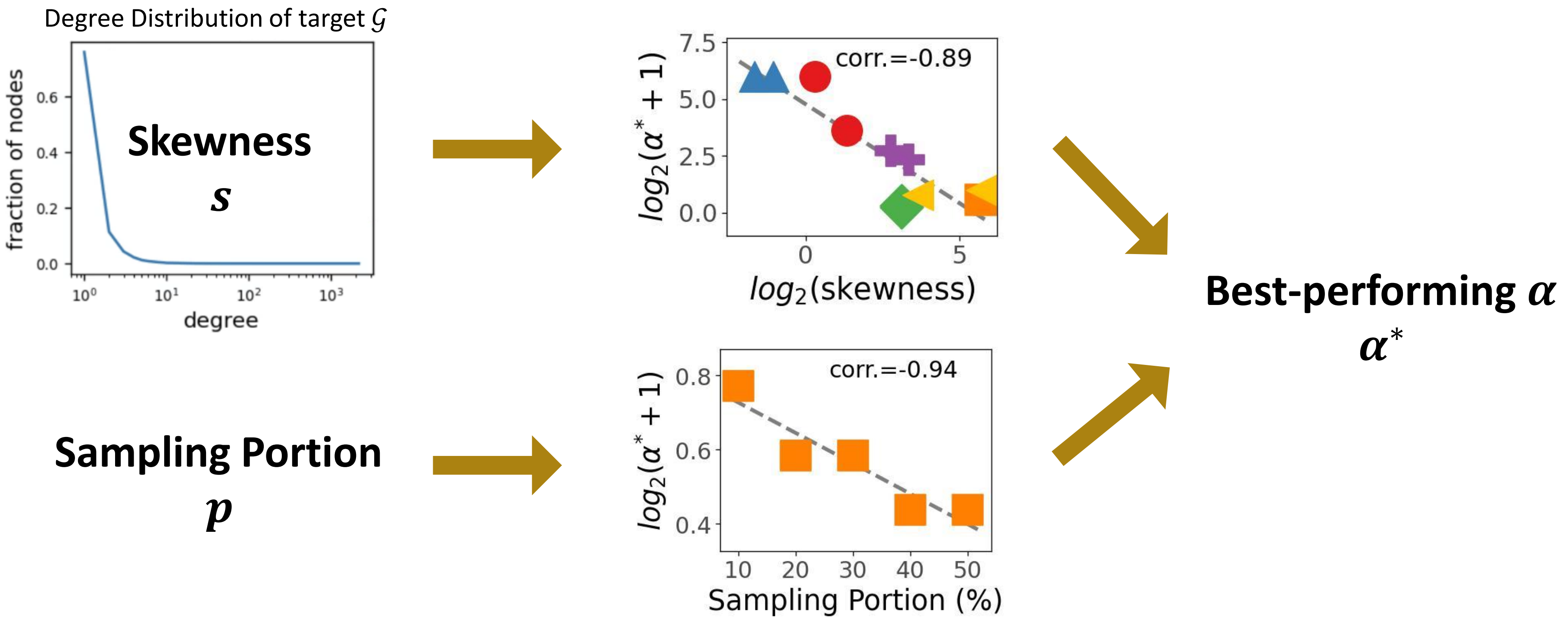
Obs. 3 As we sample **fewer hyperedges**, **larger α** values are required to preserve degree distributions

→ ***exploited to automatically tune α***



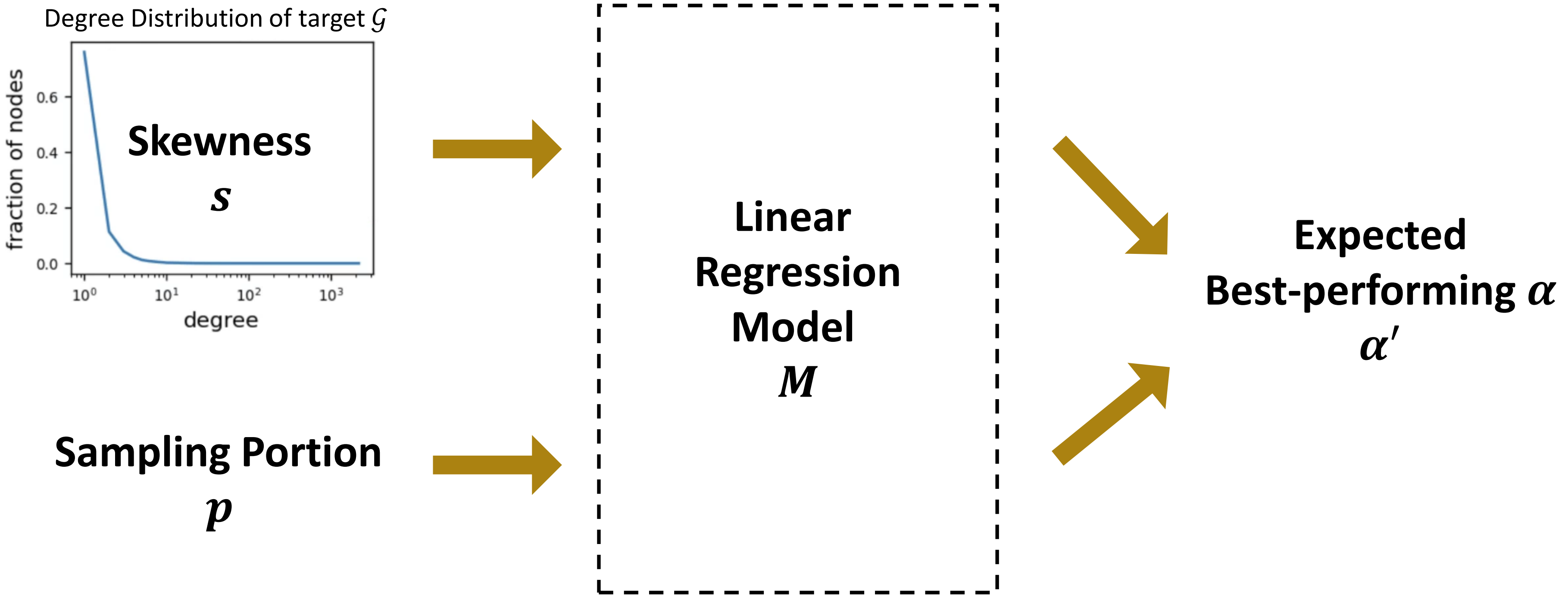
MiDaS: Full-fledged version

- **MiDaS** is our final sampling algorithm that **automatically tunes α**
- Based on the strong correlations in **Observations 2** and **3**, *the best-performing α* can be expected from skewness and sampling portions



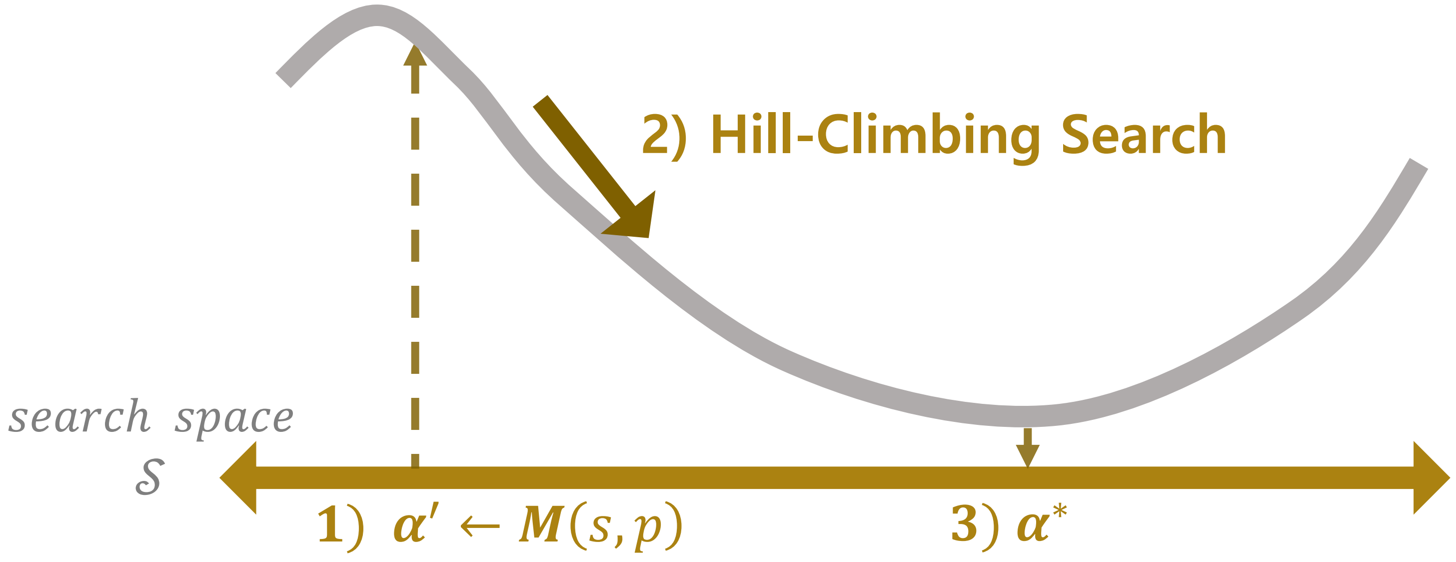
MiDaS: Full-fledged version

- We fit **a linear regressor $\boldsymbol{M}$** that maps **(a)** the skewness of the degree distribution in the input hypergraph and **(b)** the sampling portion to **(c)** a best-performing α value



MiDaS: Full-fledged version

- The α value obtained by the linear regression model $\boldsymbol{M}$ is further tuned using **hill climbing**
- A search aims to **minimize the distance** between the input hypergraph and sub-hypergraph sampled from MiDaS-Basic with α **in the degree distribution**



Roadmap

1. Introduction
2. Problem Formulation
3. Simple and Intuitive Approaches
4. MiDaS : Proposed Approach
5. **Evaluation <<**
6. Conclusions

Evaluation

- We aim to demonstrate that **MiDaS is …**

❑ **Quality**: preserves the ten structural properties of real-world hypergraphs

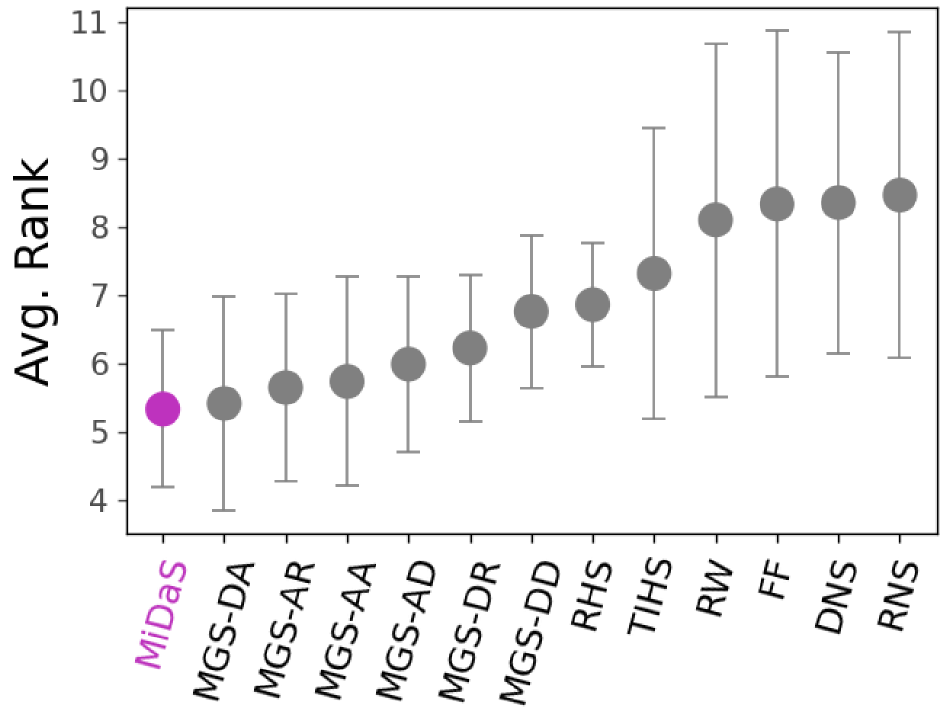
❑ **Robustness**: performs well regardless of the sampling portions

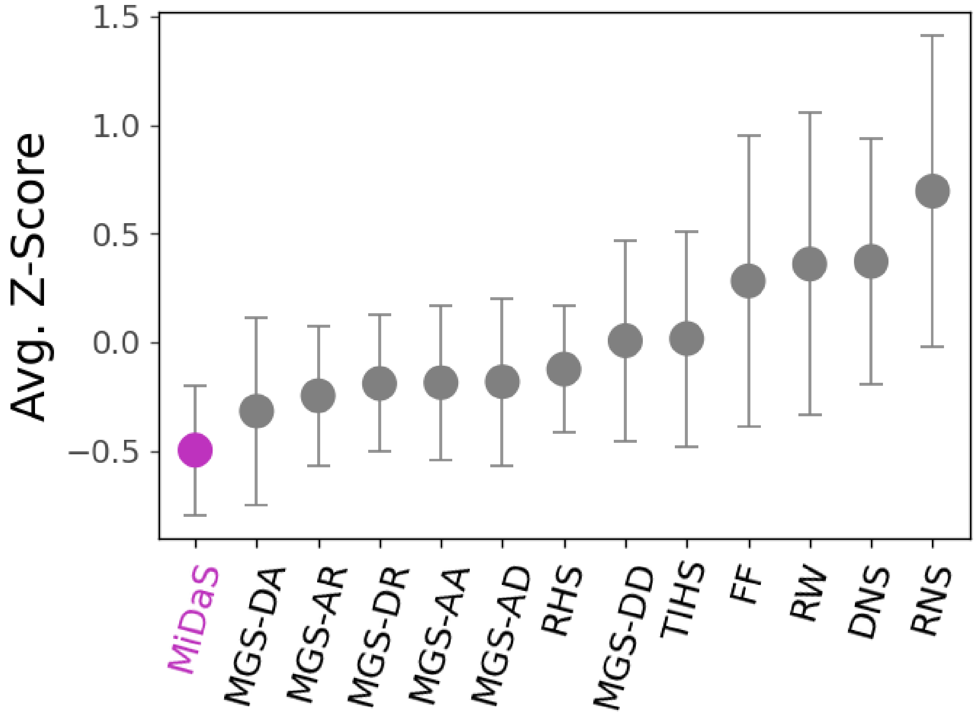
❑ **Speed:** runs fast compared to the competitors

- *Settings*: 11 Datasets under five different sampling portions (0.1 - 0.5)
- *Baseline*: six simple methods and six versions of metropolis graph sampling

Evaluation: Quality

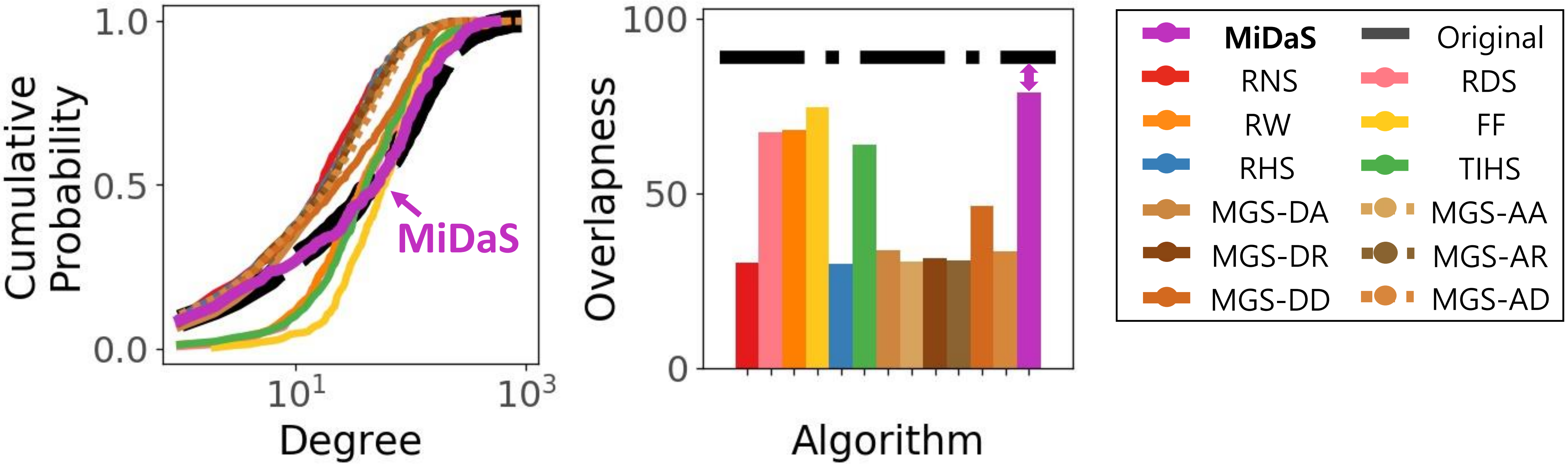
- **MiDaS** provides overall **the most representative samples** in terms of both average rankings and average Z-Scores





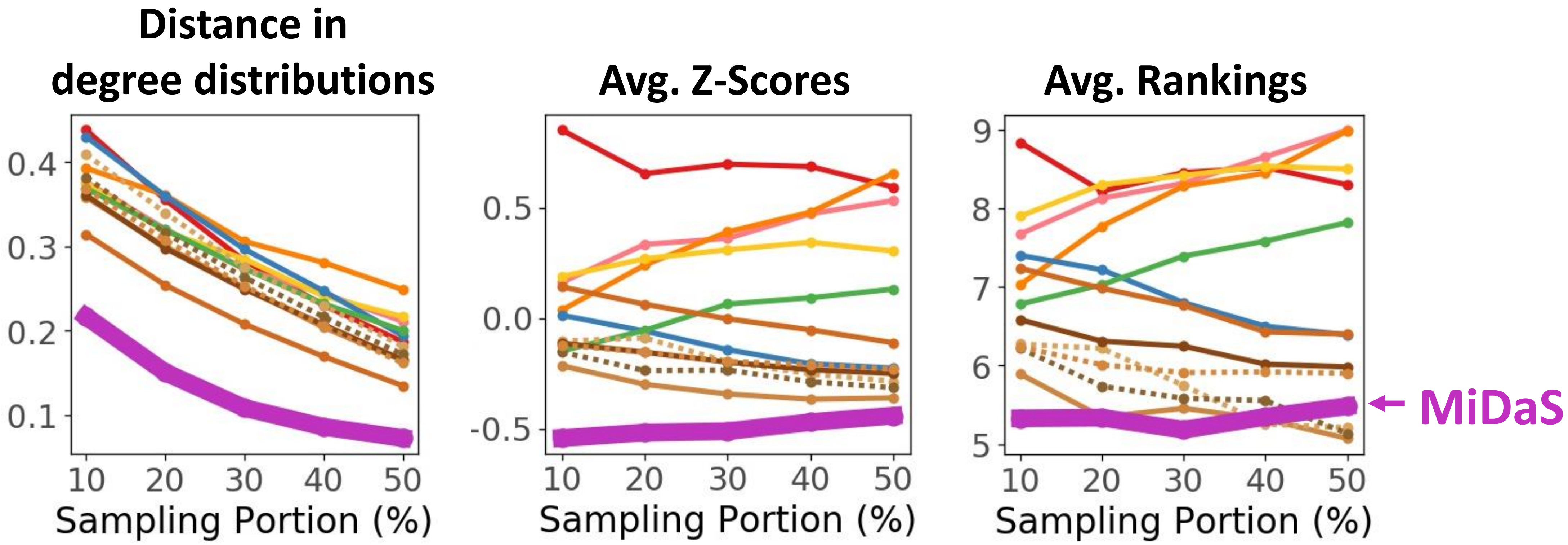
Evaluation: Quality

- Especially, MiDaS **best preserves** node degrees, density, overlapness, and diameter



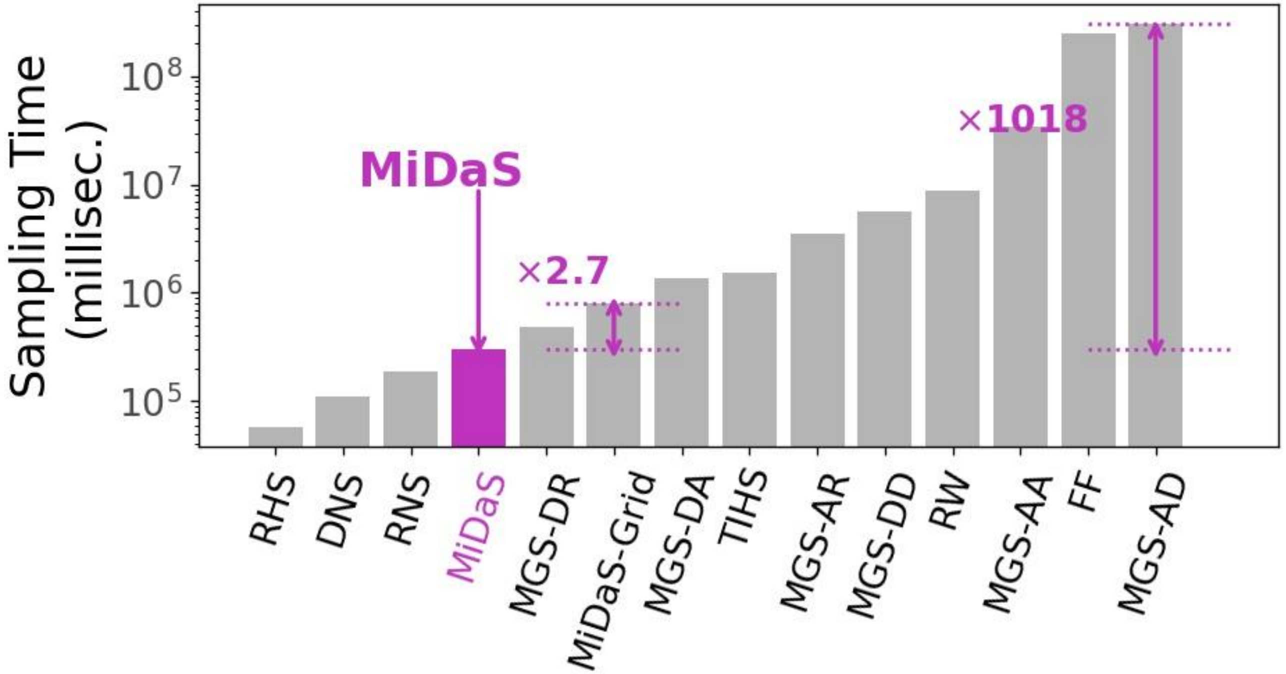
Evaluation: Robustness

- **MiDaS is consistently best regardless of sampling portions** in degree distributions, average Z-Scores, and average rankings



Evaluation: Speed

- **MiDaS is the fastest** except for the simplest methods
- It is $\mathbf{2.7\times}$ **faster** than MiDaS-Basic with grid search



Roadmap

1. Introduction
2. Problem Formulation
3. Simple and Intuitive Approaches
4. MiDaS : Proposed Approach
5. Evaluation
6. **Conclusions <<**

Conclusion

- We propose **MiDaS, a representative sub-hypergraphs sampling method**

Our contributions are:

- ❑ We **formulated** the problem of representative sampling from hypergraphs
- ❑ We **observed** the characteristics of six intuitive sampling approaches
- ❑ We **designed** *MiDaS* which is fast while sampling overall the best sub-hypergraphs

- Code & Datasets: https://github.com/young917/MiDaS

KAIST AI
Kim Jaechul Graduate School

THE WEB CONFERENCE ACM

MiDaS: Representative Sampling from Real-world Hypergraphs


Minyoung Choe
KAIST


Jaemin Yoo
SNU


Geon Lee
KAIST


Woonsung Baek
KAIST


U Kang
SNU


Kijung Shin
KAIST