KAIST

THE WEB CONFERENCE ACM

Characterization of Simplicial Complexes by Counting Simplets Beyond Four Nodes

Hyunju Kim

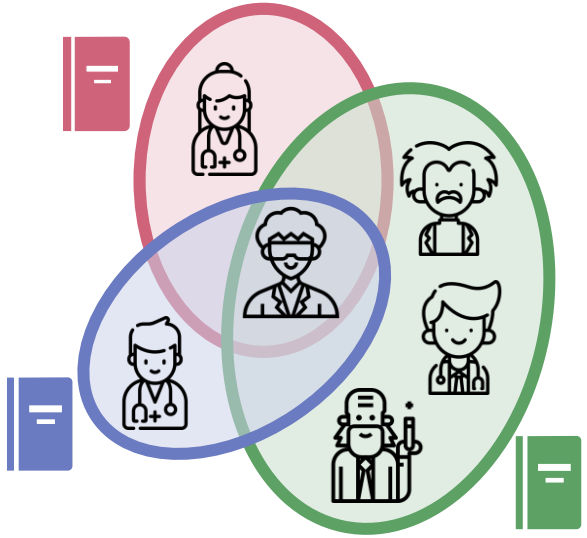
Jihoon Ko

Fanchen Bu

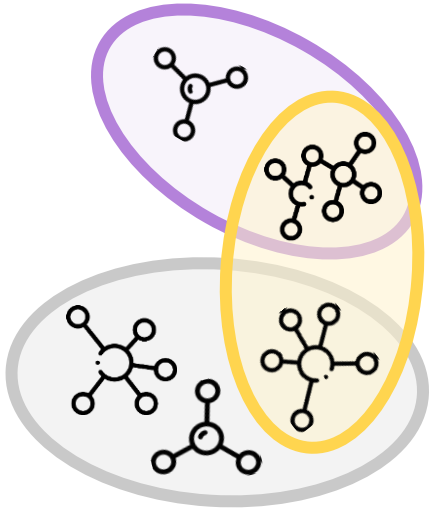
Kijung Shin

Complex Systems Exhibit Group Relations

- A simplicial complex naturally models complex systems with group relations.
- Our goal: explore local connectivity patterns in real simplicial complexes for patterns and applications.



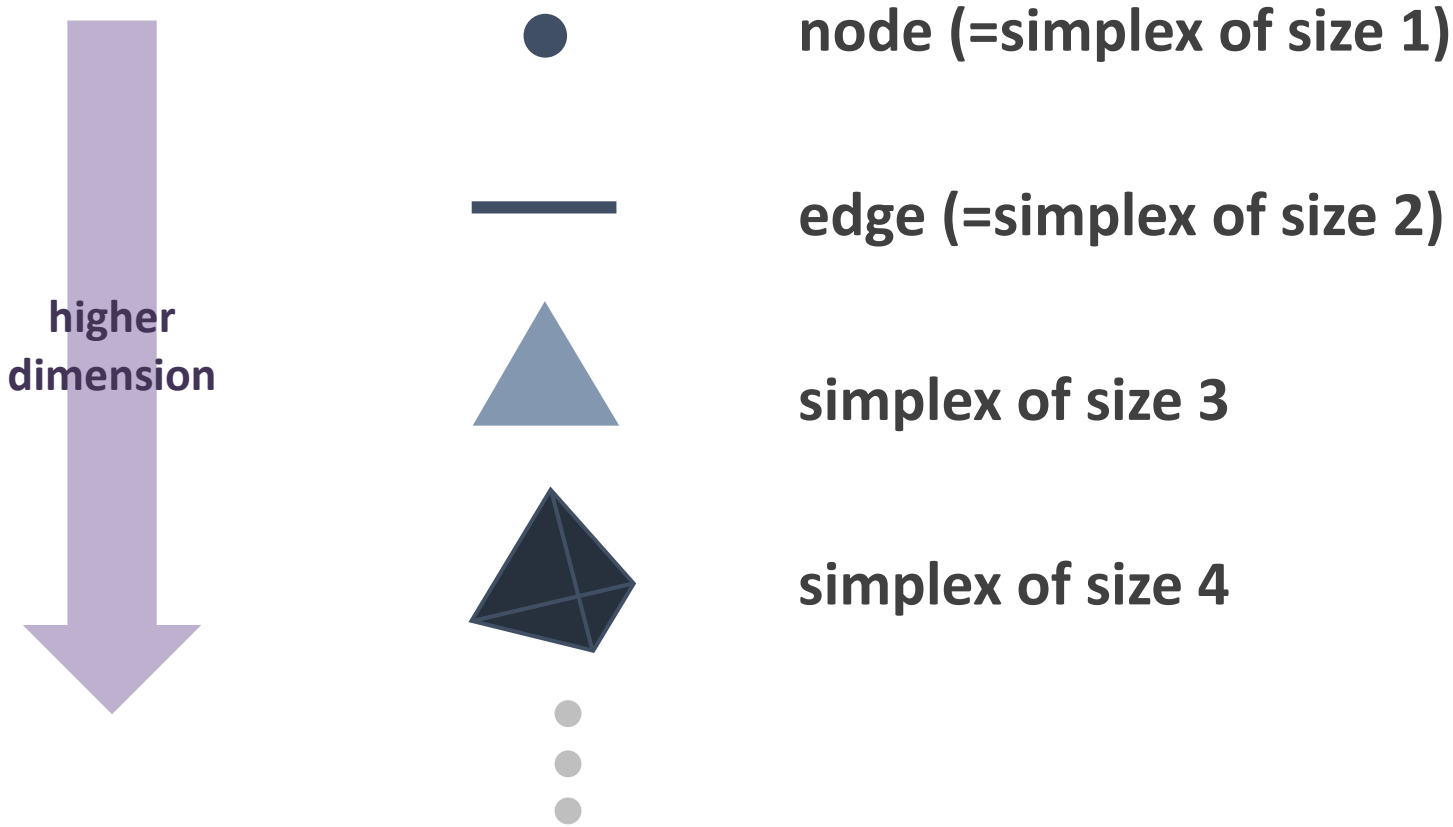
Co-authors of papers



Components of drugs

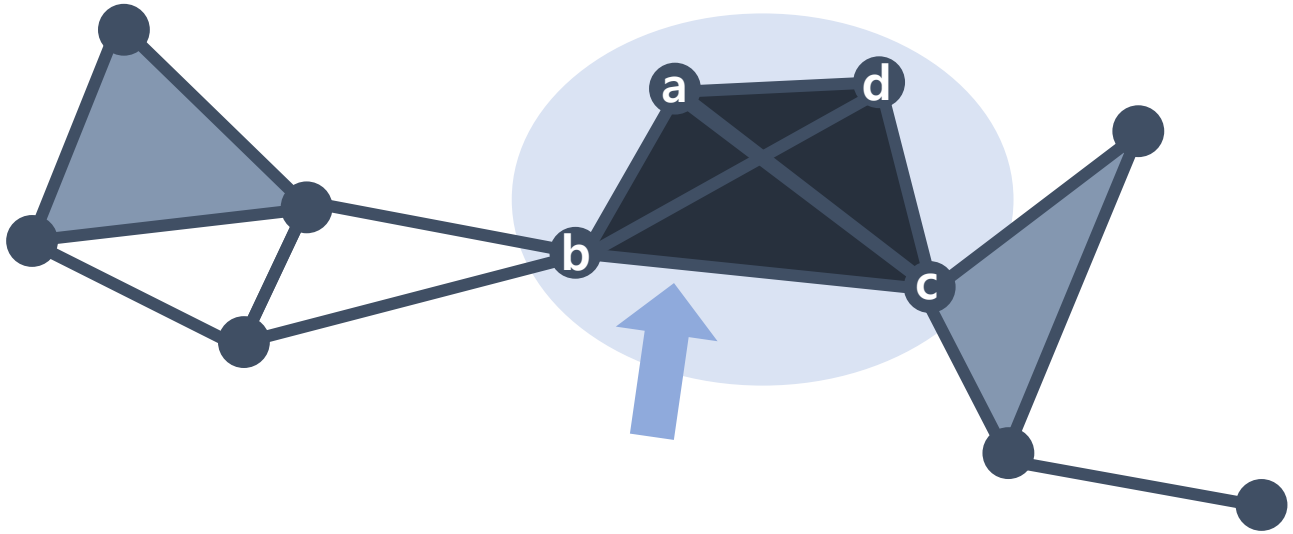
Dataset: Simplicial Complexes

- A simplicial complex is a higher-order structure consisting of simplices.



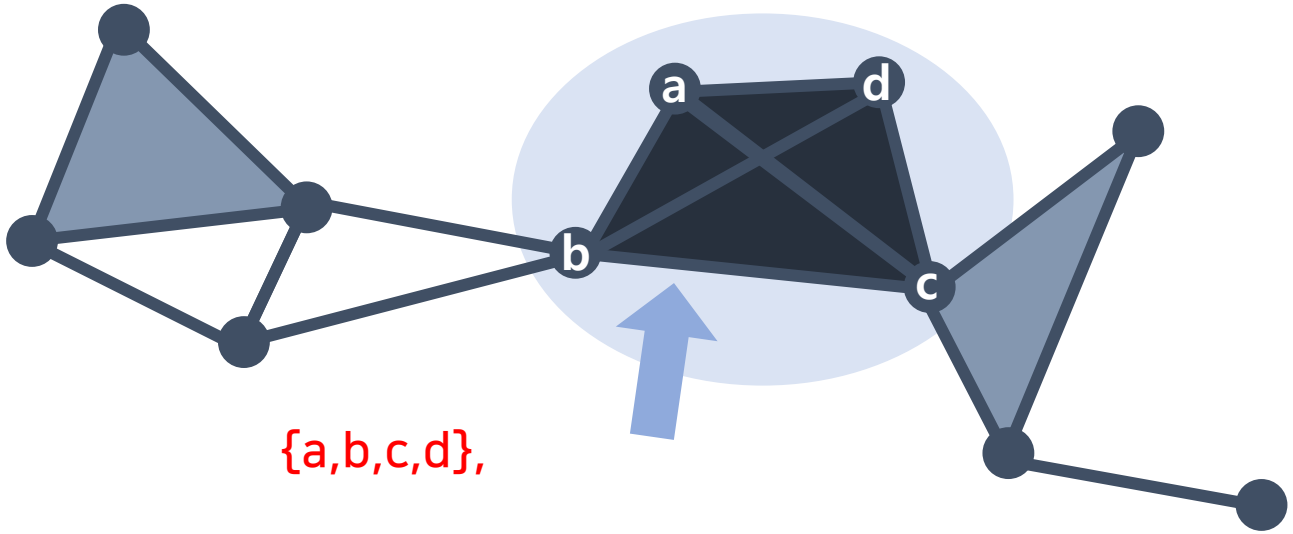
Dataset: Simplicial Complexes

- A simplicial complex G has the downward closure property.
  - If a simplex $\sigma \in G$, then for every $\sigma' \subset \sigma$, $\sigma' \in G$.



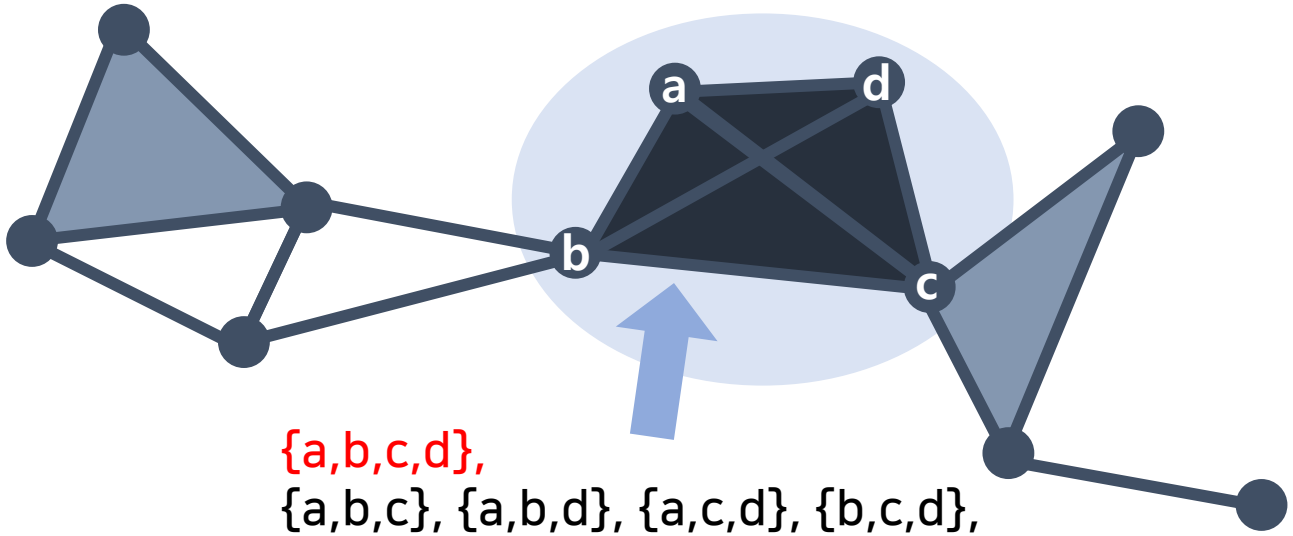
Dataset: Simplicial Complexes

- A simplicial complex G has the downward closure property.
  - If a simplex $\sigma \in G$, then for every $\sigma' \subset \sigma$, $\sigma' \in G$.



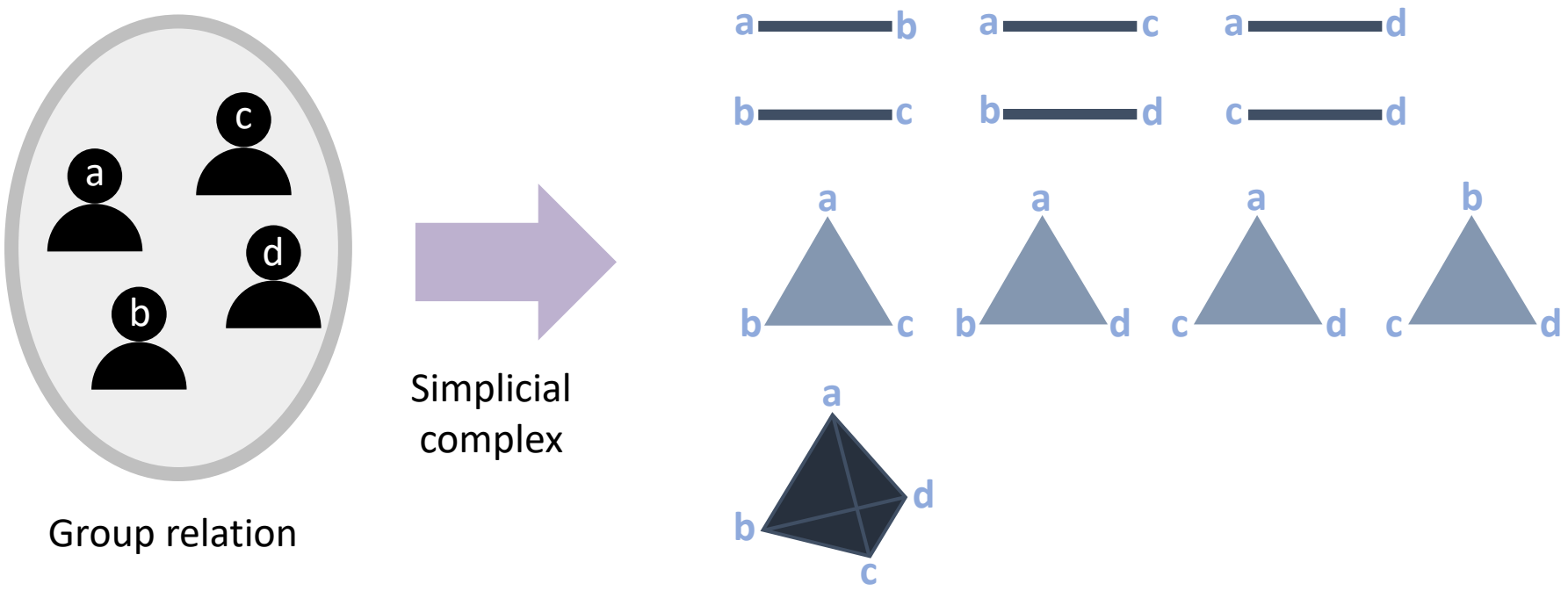
Dataset: Simplicial Complexes

- A simplicial complex G has the downward closure property.
  - If a simplex $\sigma \in G$, then for every $\sigma' \subset \sigma$, $\sigma' \in G$.



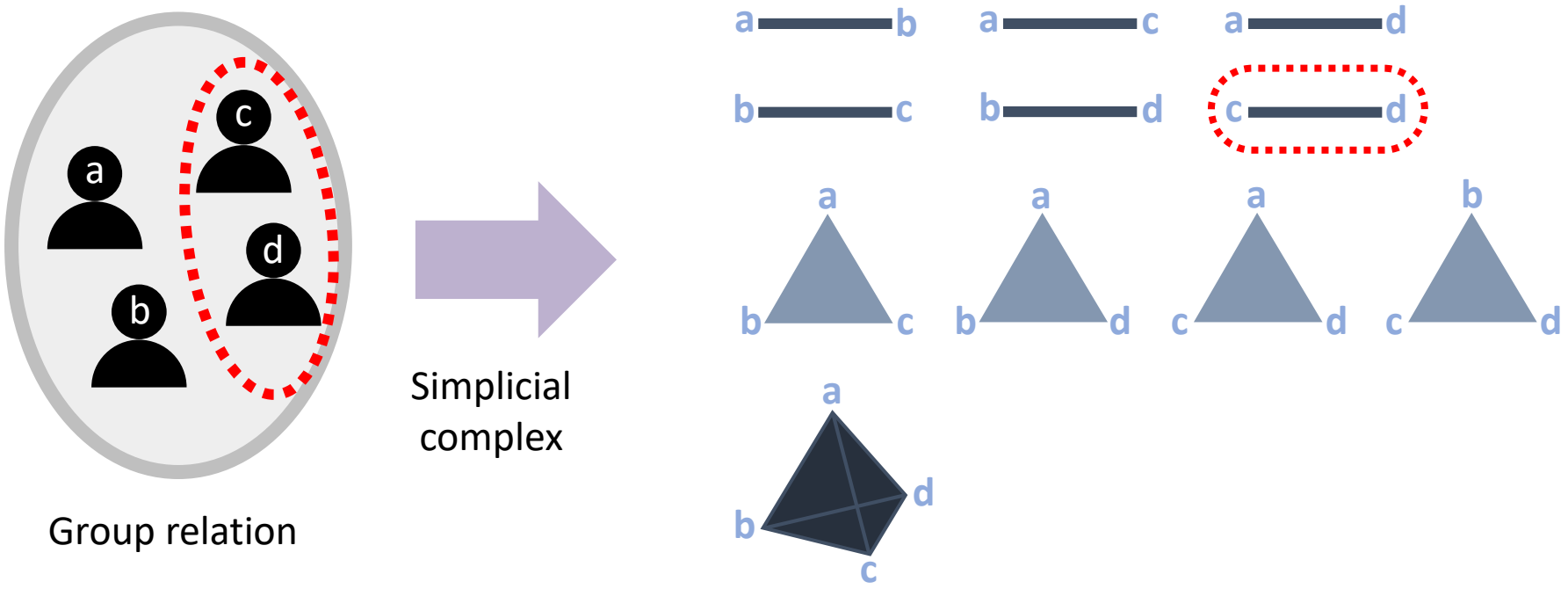
Dataset: Simplicial Complexes

- A group relation is modeled as a set of simplices in a simplicial complex.
- Every sub-relation is preserved as a simplex.
- This is useful when analyzing local connectivity among a subset of entities.



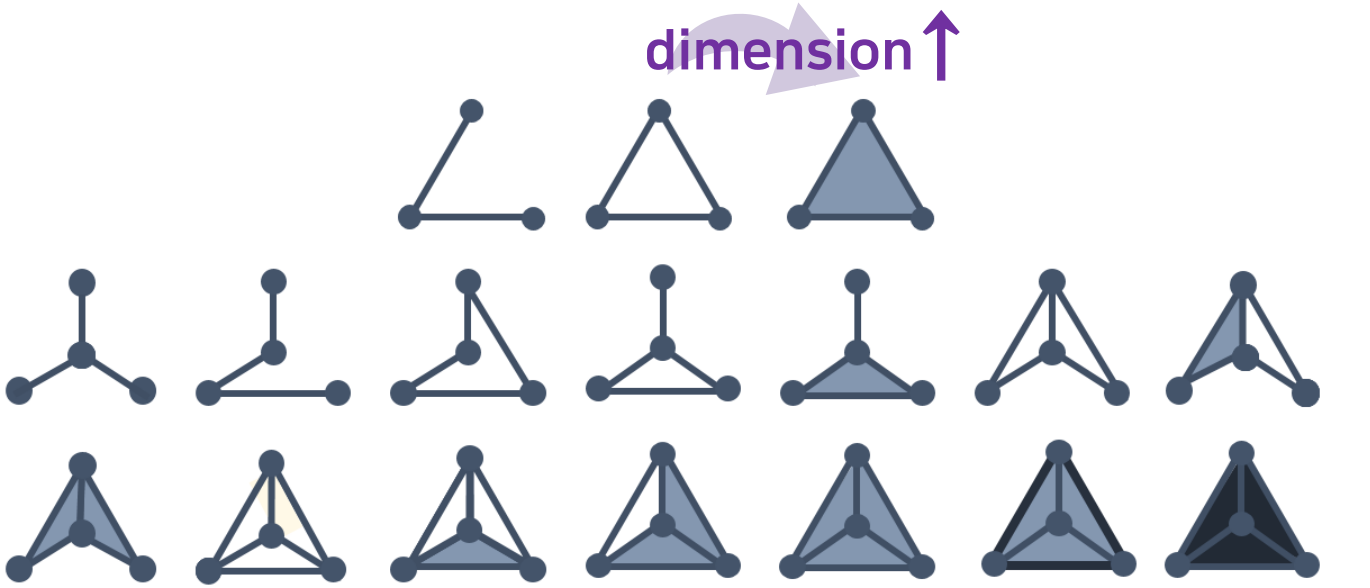
Dataset: Simplicial Complexes

- A group relation is modeled as a set of simplices in a simplicial complex.
- Every sub-relation is preserved as a simplex.
- This is useful when analyzing local connectivity among a subset of entities.



Patterns: Simplets

- Simplets of size k are isomorphic classes of connected simplicial complexes composed of k nodes.
- Three simplets of size 3, 14 simplets of size 4, 157 simplets of size 5



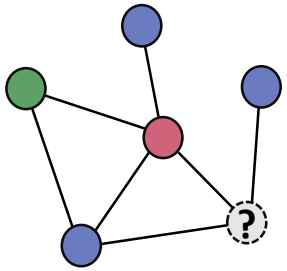
All simplets of size 3 and 4

Our goal

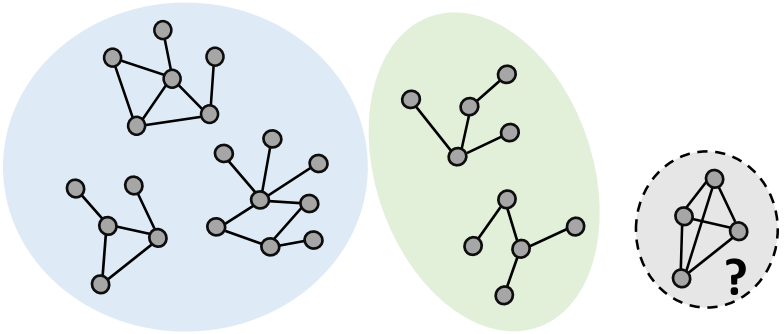
Given: a simplicial complex G and simplet size k,

To count: induced sub-complexes in G isomorphic to each simplet $s \in S^k$.

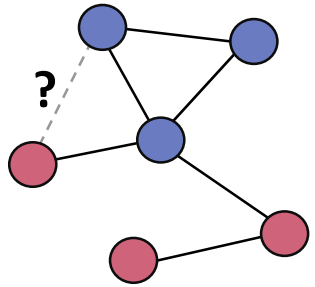
- In this work, we set simplet size $k \in \{4,5\}$.
- Q. Applications?
- A. The counts of simplets can serve as features for application/ML tasks.



node classification



graph classification/clustering



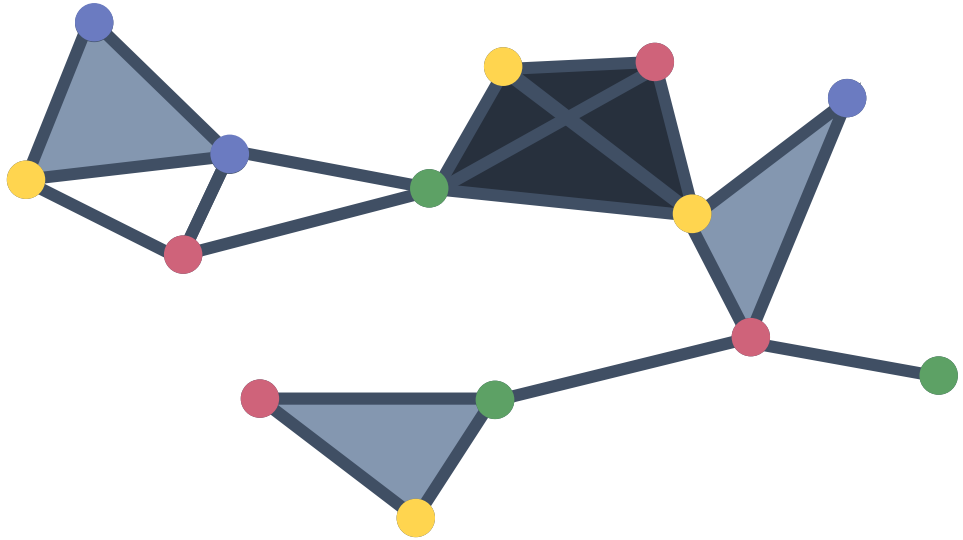
link prediction

Roadmap

1. **Our method: SC3**
2. Experiments
   - Q1. Accuracy of SC3
   - Q2. Speed of SC3
   - Q3. Characterization power of simplets
3. Conclusions

Our method: SC3

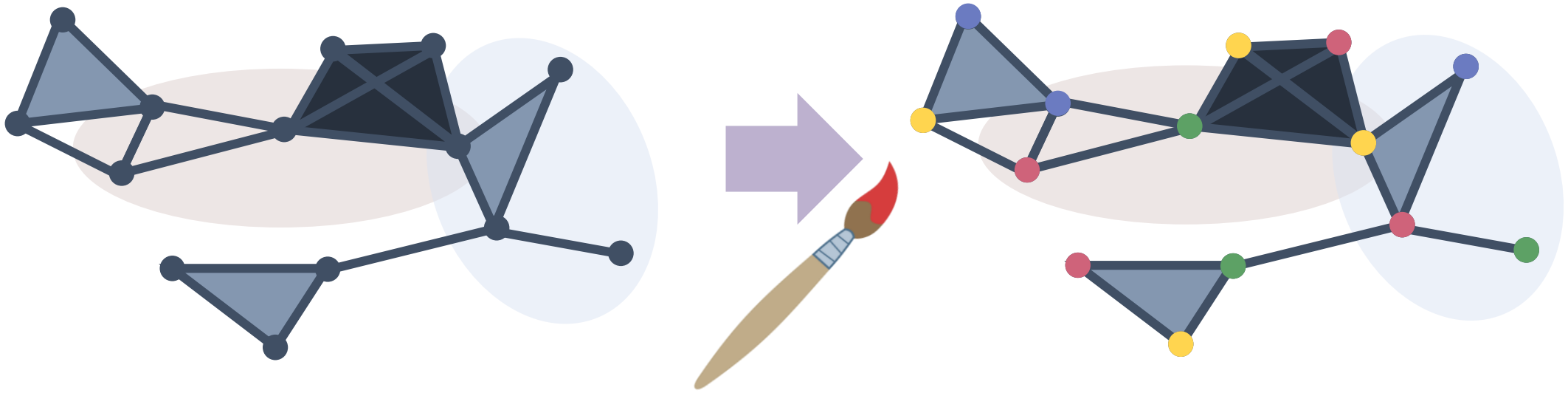
- SC3 (**S**implet **C**ounting using **C**olor **C**oding) is an algorithm using sampling based on Color Coding [1,2].
- It consists of four steps: (1) building, (2) sampling, (3) scanning, and (4) matching.



[1] Marco Bressan et al. Counting graphlets: Space vs time. WSDM'17
[2] Marco Bressan et al. Motif counting beyond five nodes. TKDD'18

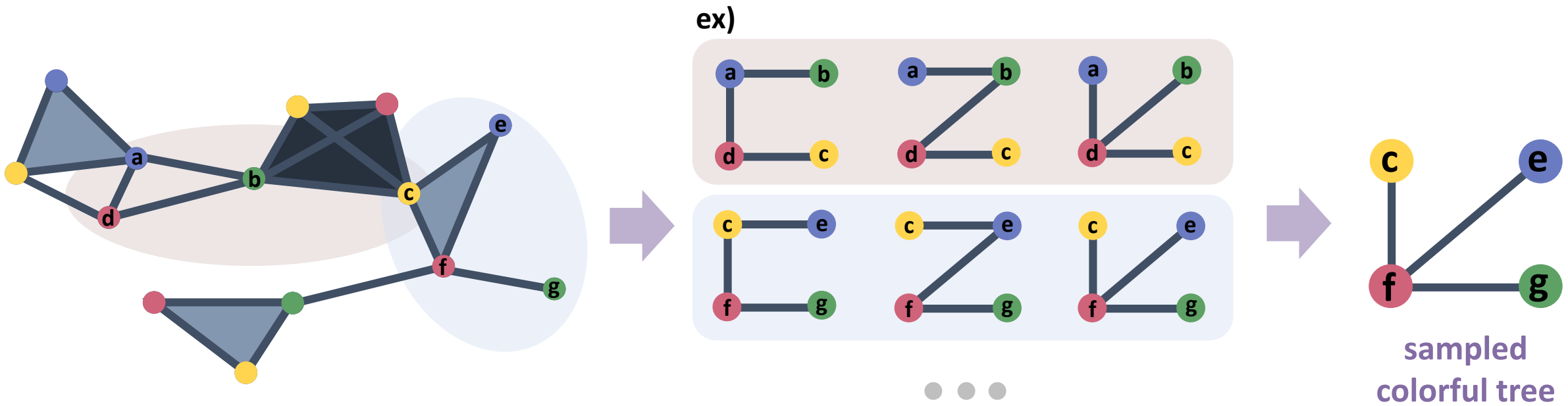
Our method: SC3

- **Given:** a simplicial complex G, a simplet size k, the number of samples N
- **(1) Building (Pre-processing):**
  - Color each node among k colors uniformly at random.
  - Pre-compute intermediate results for speeding up the next steps.



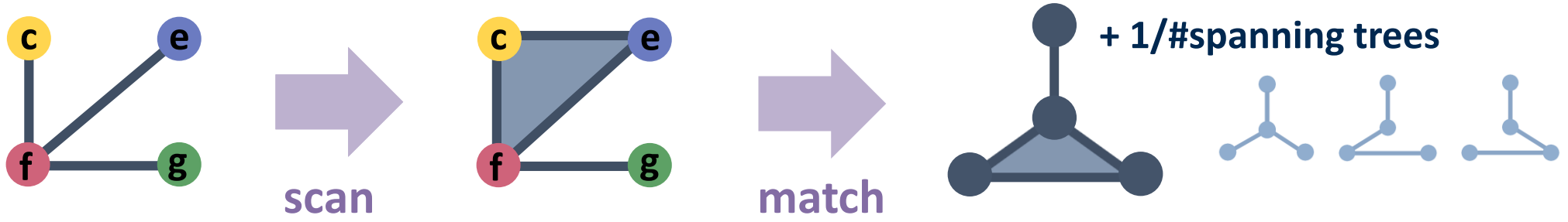
Our method: SC3

- **Given:** a simplicial complex G, a simplet size k, the number of samples N
- **(2) Sampling:** Sample N colorful trees uniformly at random



Our method: SC3

- **(3) Scanning:** Recover the sub-complex induced by each sampled tree.
- **(4) Matching:** Find the matching simplet and update its count.
- **Output:** unbiased estimate of count of each simplet of size k.



Roadmap

1. Our method: SC3
2. **Experiments**
   - Q1. Accuracy of SC3
   - Q2. Speed of SC3
   - Q3. Characterization power of simplets
3. Conclusions

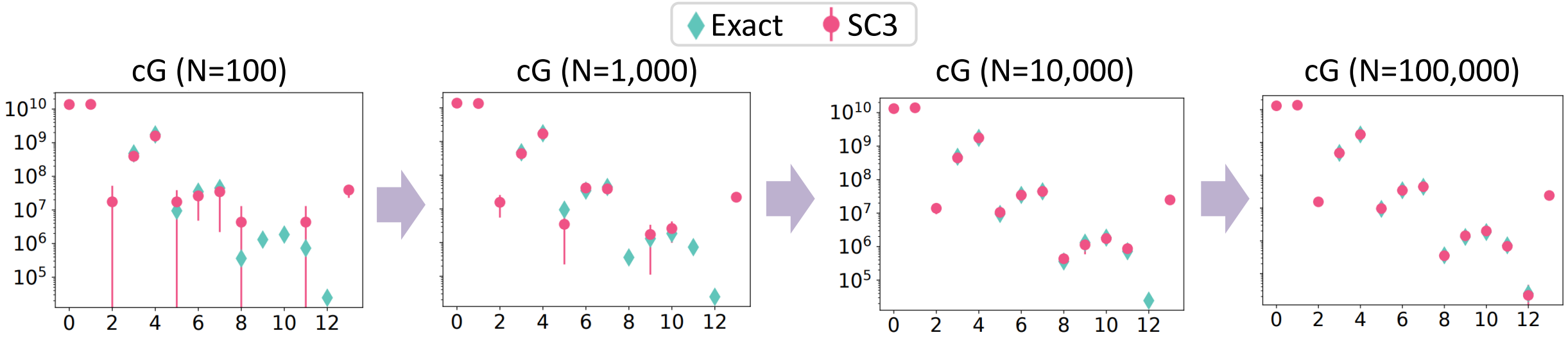
Datasets

- We used 16 real-world simplicial complexes in experiments.

Domain	Datasets
Email	email-Enron, email-Eu
Contact	contact-primary, contact-high
Tags	tags-ubnutu, tags-stack-overflow
Threads	threads-ubuntu, threads-math, threads-stack-overflow
Co-authorship	coauth-DBLP, coauth-geology, coauth-history
Others	congress-bills, DAWN, NDC-substances, NDC-classes

Experiments: Q1. Accuracy

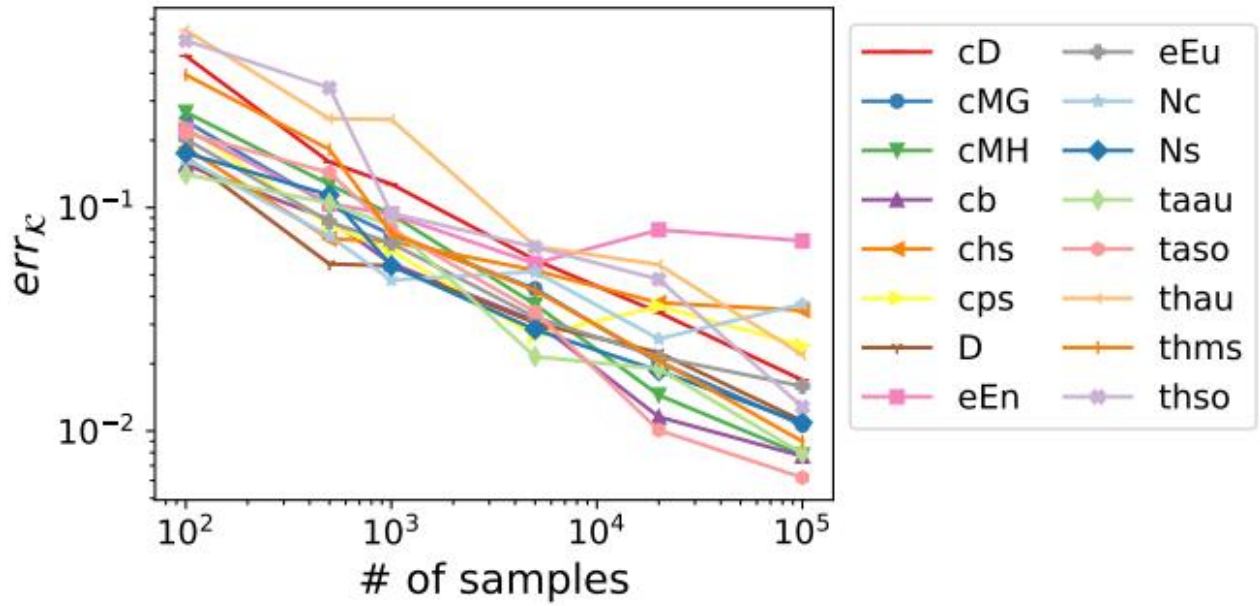
- **Q1.** How **accurate** is SC3?
- dataset: **coauth-geology**



⇒The number of samples ↑, estimation error ↓

Experiments: Q1. Accuracy

- **Q1.** How **accurate** is SC3?

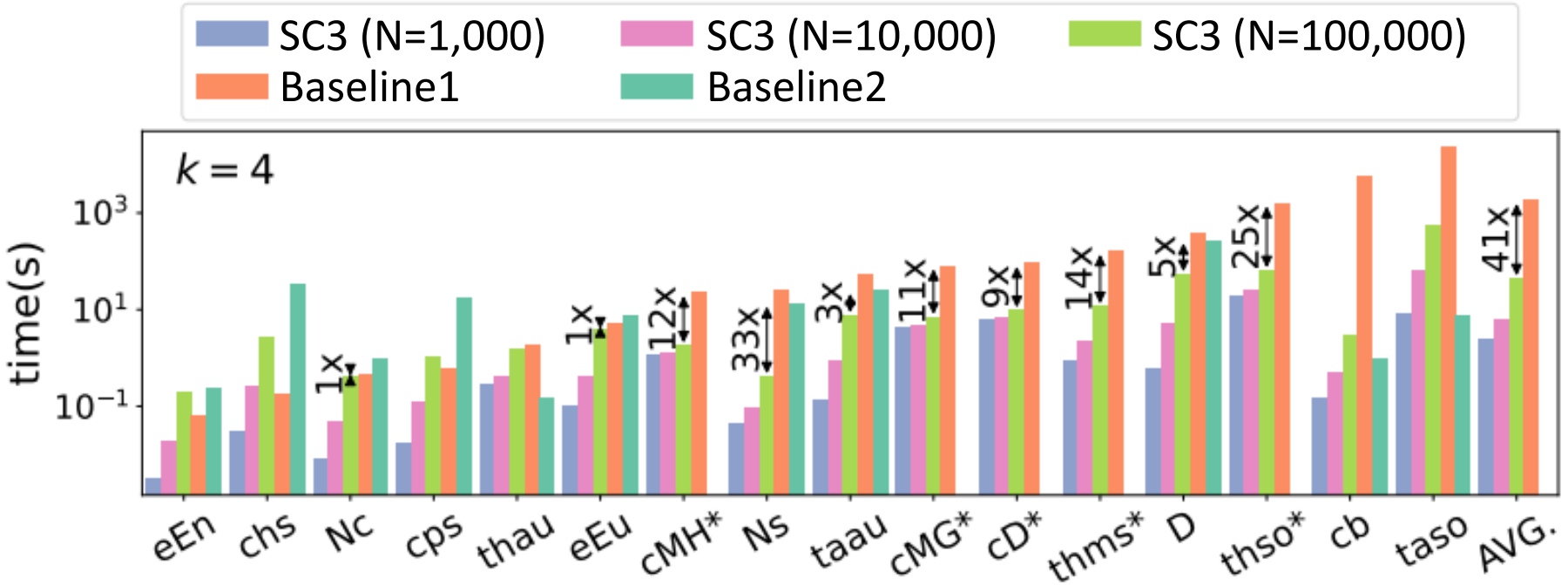


⇒The number of samples ↑, estimation error ↓

⇒The same trends have been found in all other datasets.

Experiments: Q2. Speed

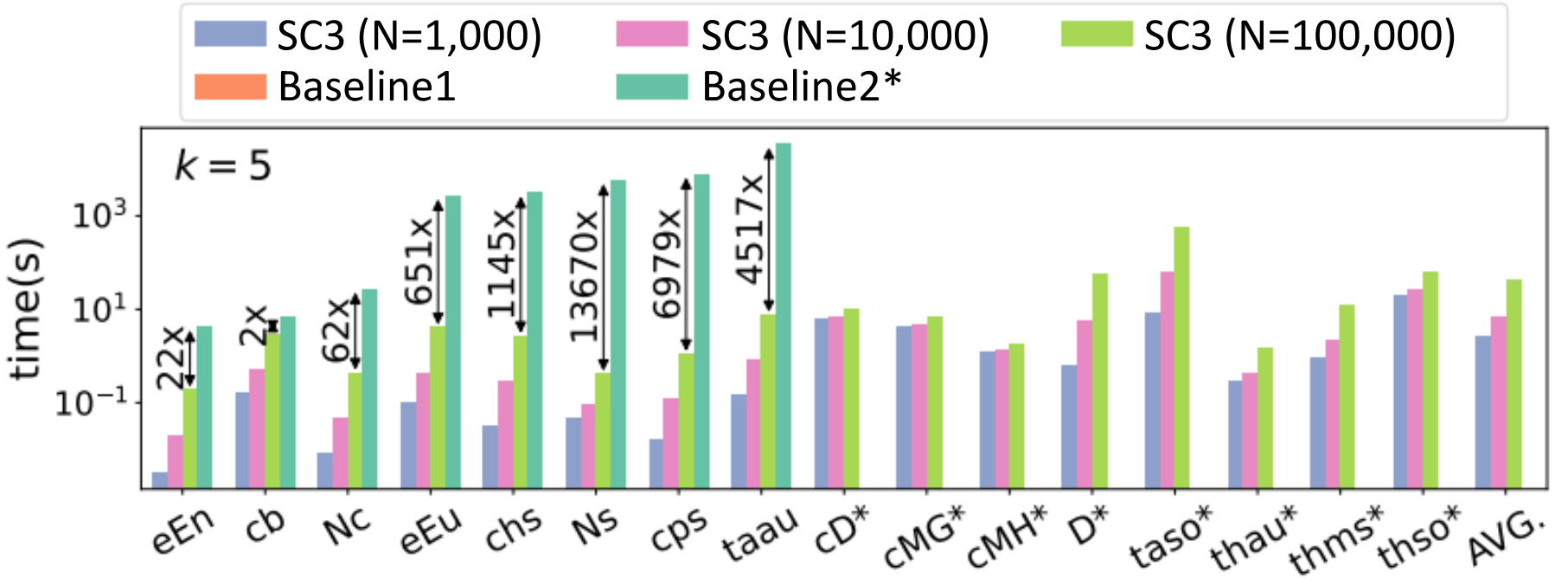
- **Q2.** How **fast** is **SC3** compared to baseline algorithms?



⇒For k=4, **SC3** was **41× faster** than **Baseline1** on average.

Experiments: Q2. Speed

- **Q2.** How **fast** is SC3 compared to baseline algorithms?

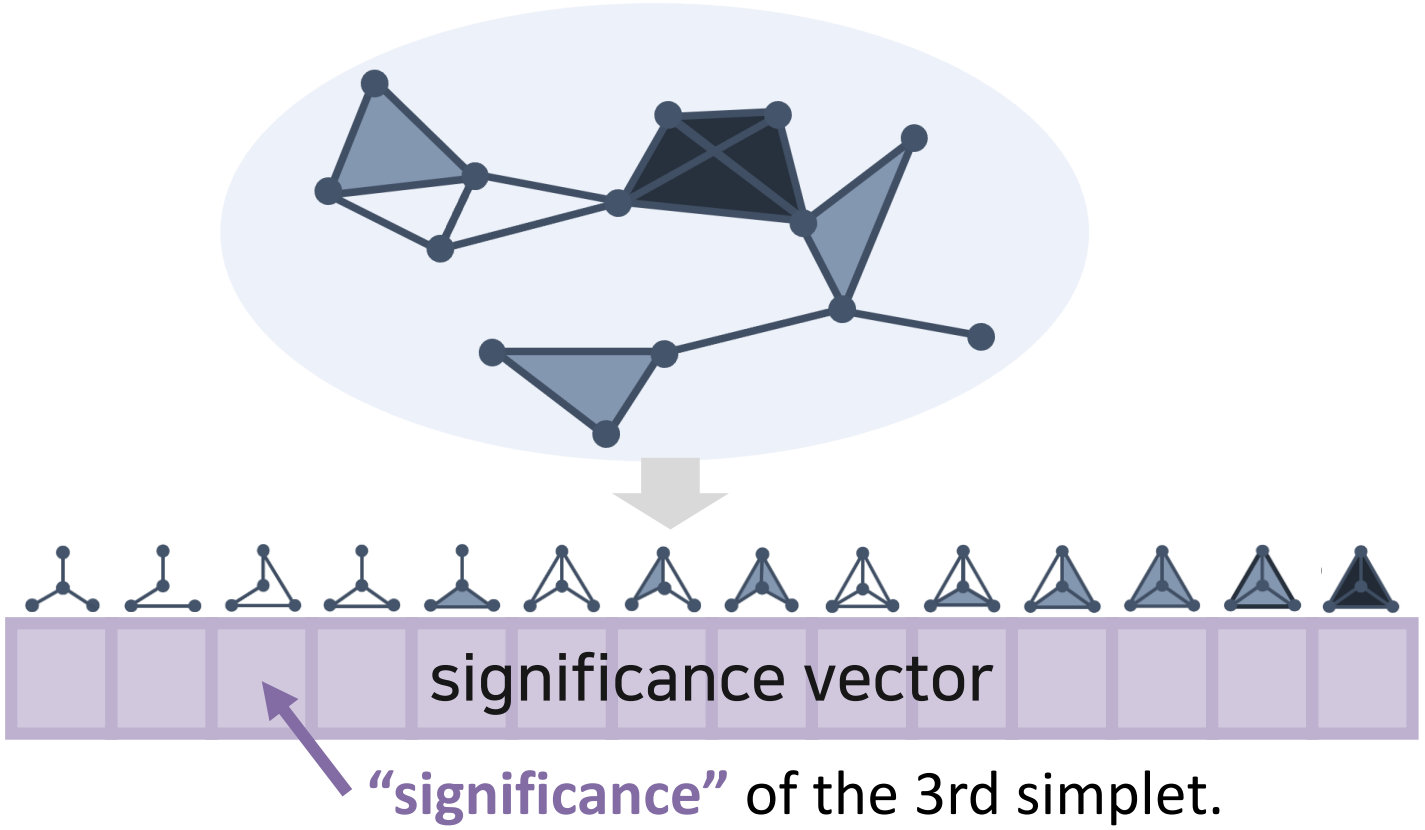


* Estimates supports instead of counts

⇒For k=5, SC3 successfully counted simplets on all datasets.

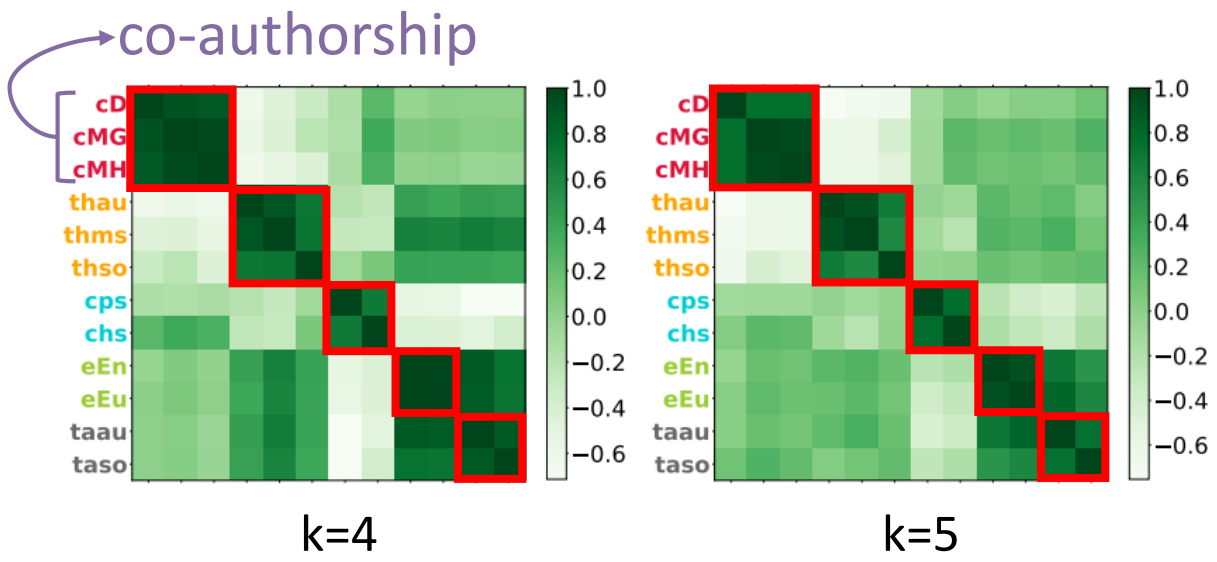
Experiments: Q3. Characterization Power

- **Q3.** How similar are simplet counts in real-world simplicial complexes?
- **Our tool:** count-based significance vector

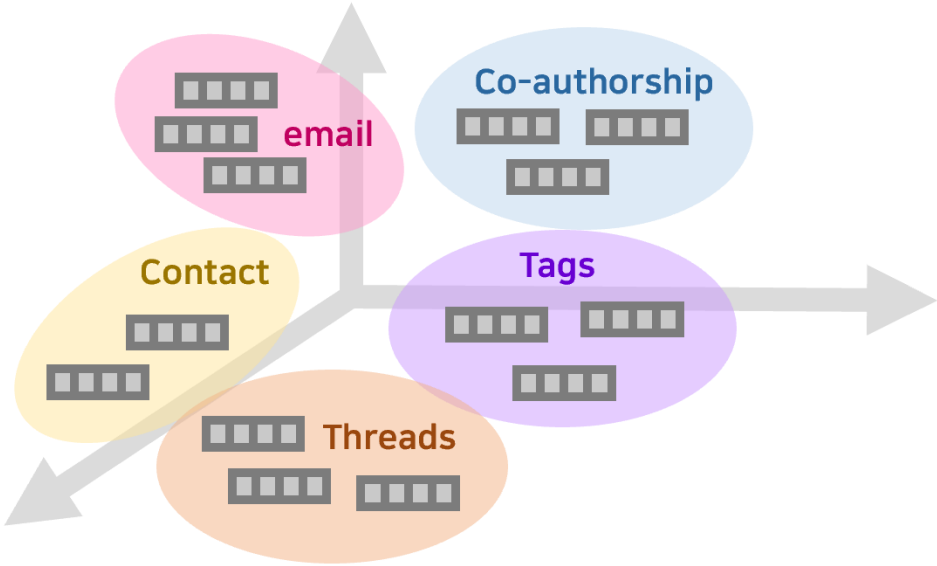


Experiments: Q3. Characterization Power

- **Q3.** How similar are simplet counts in real-world simplicial complexes?
- **A.** We found domain-based similarity.



Similarity matrix of the significance vector



Simplicial complex clustering

Roadmap

1. Our method: SC3
2. Experiments
   - Q1. Accuracy of SC3
   - Q2. Speed of SC3
   - Q3. Characterization power of simplets
3. **Conclusions**

Conclusions

- **Our contributions in this work:**

- **First Trials to Count Simplets beyond Four Nodes**
- **Accurate and Fast Algorithm**
- **Characterization of Real-World Simplicial Complexes**

The code and datasets used in the paper are available at
https://github.com/hhyy0401/SC3

KAIST

THE WEB CONFERENCE ACM

Characterization of Simplicial Complexes by Counting Simplets Beyond Four Nodes

Hyunju Kim

Jihoon Ko

Fanchen Bu

Kijung Shin