Approximating Simplet Frequency Distribution for Simplicial Complexes

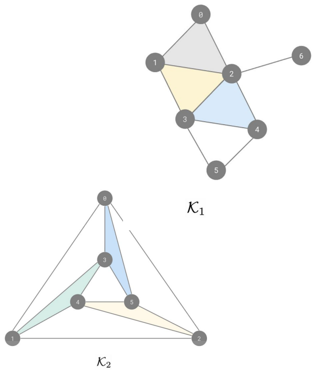
Hamid Beigy[1], Mohammad Mahini[2], Salman Qadami[3], and Morteza Saghafian[4]

1 Sharif University of Technology
`beigy@sharif.edu`
2 Sharif University of Technology
`m_mahini@ce.sharif.edu`
3 Amirkabir University of Technology
`salmanqadami@gmail.com`
4 Institute of Science and Technology Austria
`morteza.saghafian@ist.ac.at`

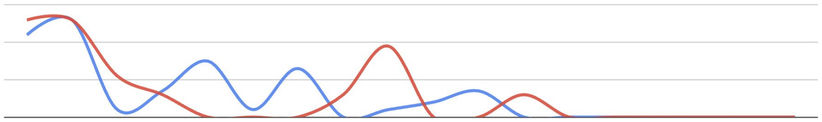
Approximating
Simplet Frequency Distribution
for
Simplicial Complexes

EuroCG 2024
13-15 March
Ioannina, Greece

Thursday
12:15-12:30



Simplet Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\lvert\mathcal{S}_{\mathcal{K}_1}\rvert$	**10**	**12**	**1**	**3**	**7**	**1**	**6**	0	**1**	**2**	**3**	0	0	0	0	0	0	0
$\lvert\phi_{\mathcal{K}_1}\rvert$	**0.22**	**0.26**	**0.02**	**0.07**	**0.15**	**0.02**	**0.13**	0	**0.02**	**0.04**	**0.07**	0	0	0	0	0	0	0
$\lvert\mathcal{S}_{\mathcal{K}_2}\rvert$	**12**	**12**	**5**	**3**	0	0	0	**3**	**9**	0	0	**3**	0	0	0	0	0	0
$\lvert\phi_{\mathcal{K}_2}\rvert$	**0.26**	**0.26**	**0.11**	**0.06**	0	0	0	**0.06**	**0.19**	0	0	**0.06**	0	0	0	0	0	0



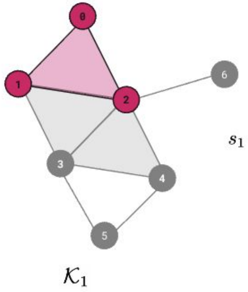
Creating a vector from a Simplicial Complex (SC)
Based on its **Small Building Blocks (Simplets)**

That can be helpful for ML applications such as classification.

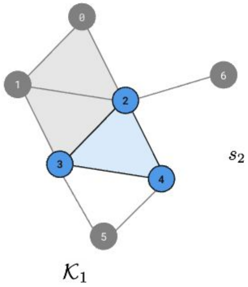
Hamid Beigy, Mohammad Mahini, Salman Qadami, Morteza Saghafian
Presenting by Mohammad Mahini

Simplet

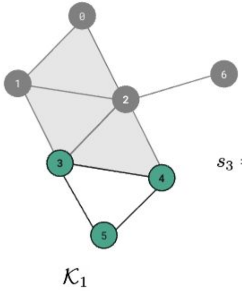
- Simplets are **small induced connected sub-complexes** of a massive complex that appear at any frequency.
- Every simplet can be identified by its vertices.
- ***Simplet Types*** are isomorphic classes of simplets.
- We denoted $\mathcal{S}_{\mathcal{K}}(i)$ as a set of all simplets of type i in $\mathcal{K}$.



$$\begin{aligned} s_1 = \{&(0),(1),(2), && (vertices)\\ &(0,1),(0,2),(1,2), && (edges)\\ &(0,1,2)\} && (triangles) \end{aligned}$$



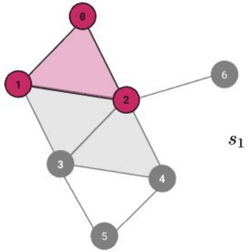
$$\begin{aligned} s_2 = \{&(2),(3),(4), && (vertices)\\ &(2,3),(2,4),(3,4), && (edges)\\ &(2,3,4)\} && (triangles) \end{aligned}$$



$$\begin{aligned} s_3 = \{&(3),(4),(5), && (vertices)\\ &(3,4),(3,5),(4,5)\} && (edges) \end{aligned}$$

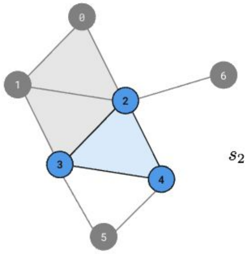
Simplet

- Simplets are **small induced connected sub-complexes** of a massive complex that appear at any frequency.
- Every simplet can be identified by its vertices.
- ***Simplet Types*** are isomorphic classes of simplets.
- We denoted $\mathcal{S}_{\mathcal{K}}(i)$ as a set of all simplets of type i in $\mathcal{K}$.



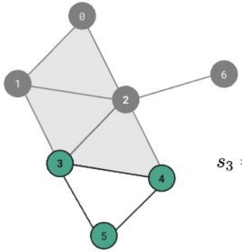
$\mathcal{K}_1$

$$\begin{aligned} s_1 = \{&(0),(1),(2), && (vertices)\\ &(0,1),(0,2),(1,2), && (edges)\\ &(0,1,2)\} && (triangles)\end{aligned}$$



$\mathcal{K}_1$

$$\begin{aligned} s_2 = \{&(2),(3),(4), && (vertices)\\ &(2,3),(2,4),(3,4), && (edges)\\ &(2,3,4)\} && (triangles)\end{aligned}$$

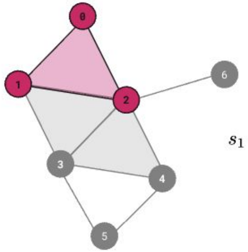


$\mathcal{K}_1$

$$\begin{aligned} s_3 = \{&(3),(4),(5), && (vertices)\\ &(3,4),(3,5),(4,5)\} && (edges)\end{aligned}$$

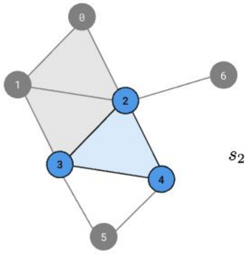
Simplet

- Simplets are **small induced connected sub-complexes** of a massive complex that appear at any frequency.
- Every simplet can be identified by its vertices.
- ***Simplet Types*** are <u>isomorphic classes</u> of simplets.
- We denoted $\mathcal{S}_{\mathcal{K}}(i)$ as a set of all simplets of type i in $\mathcal{K}$.



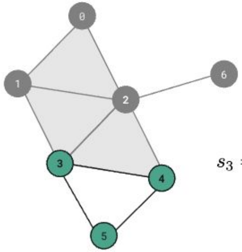
$\mathcal{K}_1$

$$\begin{aligned} s_1 = \{&(0), (1), (2), && (vertices) \\ &(0,1), (0,2), (1,2), && (edges) \\ &(0,1,2)\} && (triangles) \end{aligned}$$



$\mathcal{K}_1$

$$\begin{aligned} s_2 = \{&(2), (3), (4), && (vertices) \\ &(2,3), (2,4), (3,4), && (edges) \\ &(2,3,4)\} && (triangles) \end{aligned}$$



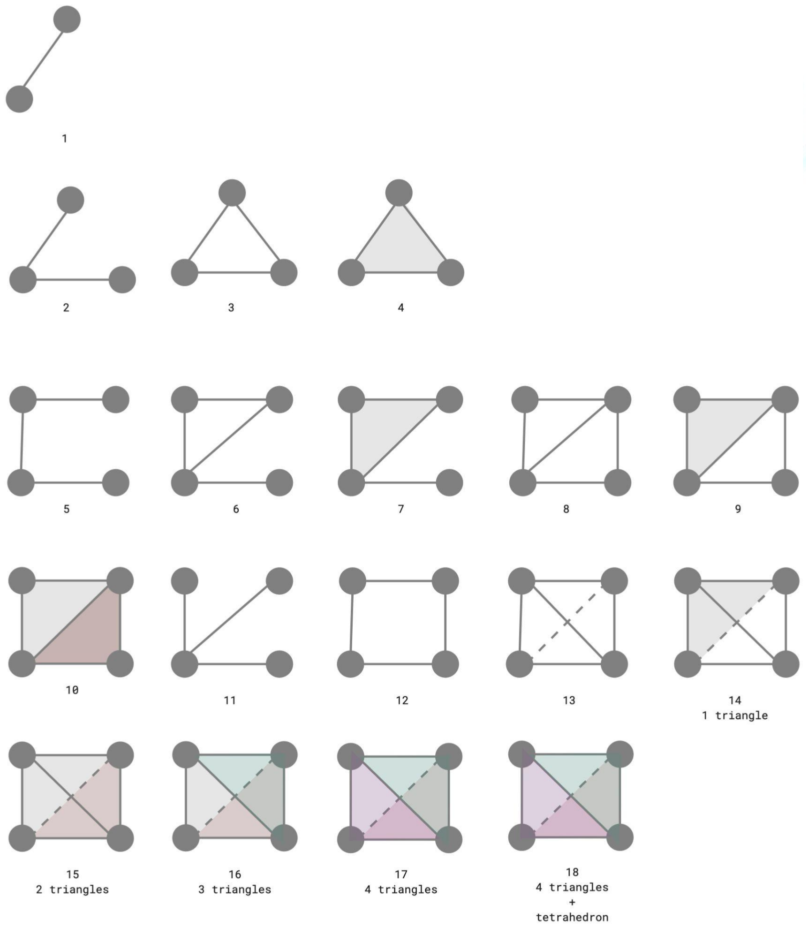
$\mathcal{K}_1$

$$\begin{aligned} s_3 = \{&(3), (4), (5), && (vertices) \\ &(3,4), (3,5), (4,5)\} && (edges) \end{aligned}$$

Simplet Types

- Simplet types with two to four vertices

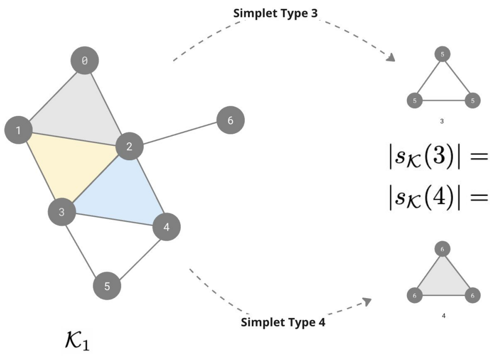
- Example:



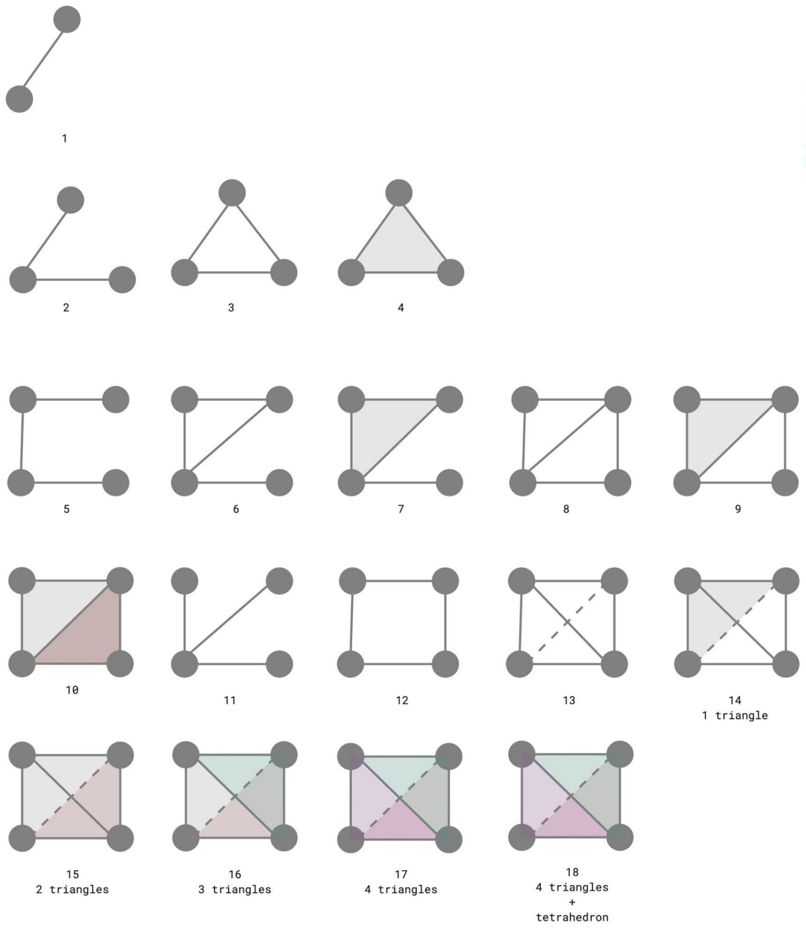
Simplet Types

- Simplet types with two to four vertices

- Example:



$$|s_{\mathcal{K}}(3)| = 1,$$
$$|s_{\mathcal{K}}(4)| = 3$$



Simplet Frequency Distribution (SFD) Vector

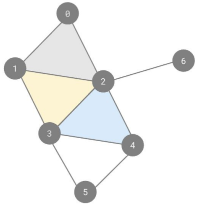
- The relative frequencies of various simplet types in $\mathcal{K}$
- The frequency denoted by $\phi_{\mathcal{K}}(i)$ is obtained by dividing $|\mathcal{S}_{\mathcal{K}}(i)|$ by $\sum_{j=1}^{N_m} |\mathcal{S}_{\mathcal{K}}(j)|$
- The vector $(\phi_{\mathcal{K}}(1), \ldots, \phi_{\mathcal{K}}(N_m))$ is called the SFD vector of the $\mathcal{K}$

- Example:

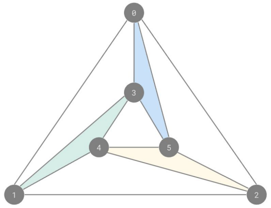
Simplet Frequency Distribution (SFD) Vector

- The relative frequencies of various simplet types in $\mathcal{K}$
- The frequency denoted by $\phi_{\mathcal{K}}(i)$ is obtained by dividing $|\mathcal{S}_{\mathcal{K}}(i)|$ by $\sum_{j=1}^{N_m} |\mathcal{S}_{\mathcal{K}}(j)|$
- The vector $(\phi_{\mathcal{K}}(1), \ldots, \phi_{\mathcal{K}}(N_m))$ is called the SFD vector of the $\mathcal{K}$

- Example:



$\mathcal{K}_1$



$\mathcal{K}_2$

Simplet Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\|\mathcal{S}_{\mathcal{K}_1}\|$	**10**	**12**	**1**	**3**	**7**	**1**	**6**	0	**1**	**2**	**3**	0	0	0	0	0	0	0
$\|\phi_{\mathcal{K}_1}\|$	**0.22**	**0.26**	**0.02**	**0.07**	**0.15**	**0.02**	**0.13**	0	**0.02**	**0.04**	**0.07**	0	0	0	0	0	0	0
$\|\mathcal{S}_{\mathcal{K}_2}\|$	**12**	**12**	**5**	**3**	0	0	0	**3**	**9**	0	0	**3**	0	0	0	0	0	0
$\|\phi_{\mathcal{K}_2}\|$	**0.26**	**0.26**	**0.11**	**0.06**	0	0	0	**0.06**	**0.19**	0	0	**0.06**	0	0	0	0	0	0

Calculating The SFD Vector for Large SCs

- Simple Simplet counting algorithm is in $\Theta(n^k)$

Our approach:

- Instead of calculating the exact counts we use an approximation on Simplet frequencies.
- Our algorithm is sublinear in the size of $\mathcal{K}$

Calculating The SFD Vector for Large SCs

- Simple Simplet counting algorithm is in $\Theta(n^k)$

Our approach:

- Instead of calculating the exact counts we use an approximation on Simplet frequencies.
- Our algorithm is sublinear for large and sparse SCs in the size of $\mathcal{K}$

Approximating The SFD Vector with Sampling

1. [Number of Samples] With a set of $\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$ simplets sampled uniformly from SC $\mathcal{K}$, we can have an (ϵ,δ)-approximation on the SFD vector of $\mathcal{K}$.

2. [Sampling Algorithm] We propose a uniform sampling algorithm for simplets in a connected simplicial complex that find a sample with complexity in $O(log(n)\cdot\Delta\cdot diam(\mathcal{K})^2)$.

The time complexity of (ϵ,δ)-approximation of SFD vector of $\mathcal{K}$ is

$$O(\frac{1}{\epsilon^2}\cdot(1+\ln\frac{1}{\delta})\cdot\log(n)\cdot\Delta\cdot diam(\mathcal{K})^2)$$

Approximating The SFD Vector with Sampling

1. [Number of Samples] With a set of $\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$ simplets sampled uniformly from SC $\mathcal{K}$, we can have an (ϵ,δ)-approximation on the SFD vector of $\mathcal{K}$.

2. [Sampling Algorithm] We propose a uniform sampling algorithm for simplets in a connected simplicial complex that find a sample with complexity in $O(log(n)\cdot\Delta\cdot diam(\mathcal{K})^2)$.

The time complexity of (ϵ,δ)-approximation of SFD vector of $\mathcal{K}$ is

$$O(\frac{1}{\epsilon^2}\cdot(1+\ln\frac{1}{\delta})\cdot\log(n)\cdot\Delta\cdot diam(\mathcal{K})^2)$$

Number of Samples

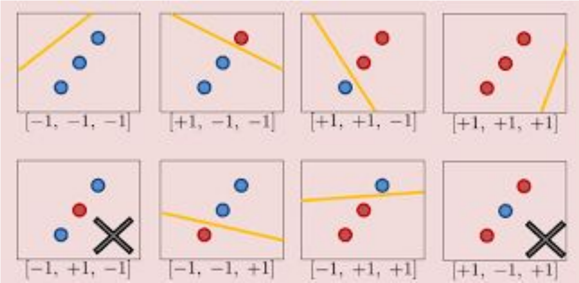
[Number of Samples] With a set of $\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$ simplets sampled uniformly from SC $\mathcal{K}$, we can have an (ϵ,δ)-approximation on the SFD vector of $\mathcal{K}$.

$$\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$$

Number of Samples

[Number of Samples] With a set of $\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$ simplets sampled uniformly from SC $\mathcal{K}$, we can have an (ϵ,δ)-approximation on the SFD vector of $\mathcal{K}$.

- We use VC dimension to prove this bound.
- For a domain D and collection $\mathcal{R}$ of subsets of D, the VC dimension represents the maximum size of a set $X \subseteq D$ that can be shattered by $\mathcal{R}$ which means $\{r \cap X | \forall r \in R\} = 2^{|X|}$.



Number of Samples

[Number of Samples] With a set of $\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$ simplets sampled uniformly from SC $\mathcal{K}$, we can have an (ϵ,δ)-approximation on the SFD vector of $\mathcal{K}$.

- [VC Dimension of Simplets] Let $\mathcal{R}=\{\mathcal{S}_i \mid 1\leq i\leq N_m\}$ be a family of all simplet sets where N_m is the number of simplet types with at most m vertices, and D is all simplets of SC $\mathcal{K}$, Then we have $VC(D,\mathcal{R})=1$.

  Proof. Let $\{s_1,s_2\}\subseteq D$
  - If s_1 and s_2 are belong to the same simplet type, $\{s_1\}$ can't be shattered.
  - Otherwise, $\{s_1,s_2\}$ can't be shattered.

Number of Samples

$$\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$$

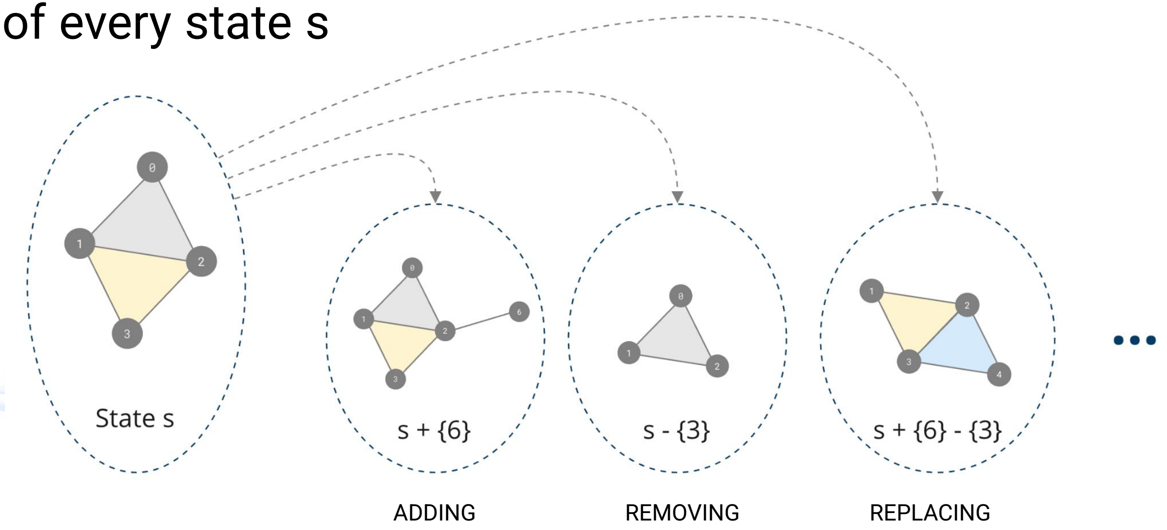
[Number of Samples] With a set of $\frac{c}{\epsilon^2}(1+\ln\frac{1}{\delta})$ simplets sampled uniformly from SC $\mathcal{K}$, we can have an (ϵ,δ)-approximation on the SFD vector of $\mathcal{K}$.

- [VC Dimension of Simplets] Let $\mathcal{R}=\{\mathcal{S}_i \mid 1\leq i\leq N_m\}$ be a family of all simplet sets where N_m is the number of simplet types with at most m vertices, and D is all simplets of SC $\mathcal{K}$, Then we have $VC(D,\mathcal{R})=1$.

> [Lemma] For a domain D and collection $\mathcal{R}$ of subsets of D, with $VC(D,\mathcal{R})\leq d$ and using $\frac{c}{\epsilon^2}\left(d+\ln\frac{1}{\delta}\right)$ uniform samples, we can have an (ϵ,δ)-approximation on distribution of all subsets in $\mathcal{R}$.

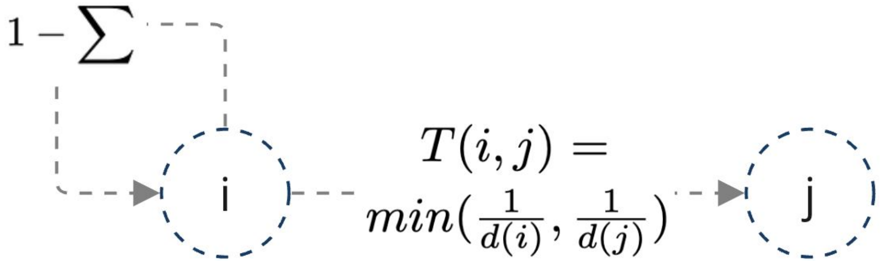
Sampling Algorithm

- A Monte-Carlo Markov-Chain algorithm.
- Random walk on a directed graph $\mathcal{P}_{\mathcal{K}}^{m}$ whose vertex set (states) is a set of all simplets in complex $\mathcal{K}$
- Out-neighbors of every state s



Sampling Algorithm

- $T(i,j)$ is Transition probability matrix T on $\mathcal{P}_{\mathcal{K}}^{m}$



- <u>The random walk is</u>
  - Irreducible
  - Aperiodic
  - Converges to the uniform stationary distribution

Sampling Algorithm (Time Complexity)

- The mixing time of the markov chain on $\mathcal{P}_{\mathcal{K}}^{m}$ is in

$$O(\log(n) \cdot \Delta \cdot diam(\mathcal{K})^2)$$

Proof.

- $\Delta(\mathcal{P}_{\mathcal{K}}^{m}) \in O(m^2 \cdot \Delta)$
- $diam(\mathcal{P}_{\mathcal{K}}^{m}) = diam(\mathcal{K}) + m$

(**m** is the maximum number of simplet vertices; and is constant in size of SC)

[Lemma] The mixing time t_{mix}^{G} of a random walk on graph G with n vertices is in

$$O(\log(n) \cdot \Delta(G) \cdot diam(G)^2)$$

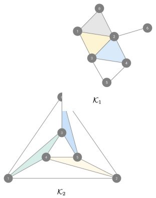
Approximating

Simplet Frequency Distribution

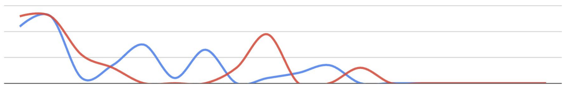
for

Simplicial Complexes

Conclusion



Simplet Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\lvert\mathcal{S}_{\mathcal{K}_1}\rvert$	**10**	**12**	**1**	**3**	**7**	**1**	**6**	0	**1**	**2**	**3**	0	0	0	0	0	0	0
$\lvert\phi_{\mathcal{K}_1}\rvert$	**0.22**	**0.26**	**0.02**	**0.07**	**0.15**	**0.02**	**0.13**	0	**0.02**	**0.04**	**0.07**	0	0	0	0	0	0	0
$\lvert\mathcal{S}_{\mathcal{K}_2}\rvert$	**12**	**12**	**5**	**3**	0	0	0	**3**	**9**	0	0	**3**	0	0	0	0	0	0
$\lvert\phi_{\mathcal{K}_2}\rvert$	**0.26**	**0.26**	**0.11**	**0.06**	0	0	0	**0.06**	**0.19**	0	0	**0.06**	0	0	0	0	0	0



Creating a vector from a Simplicial Complex (SC) using local structures (Simplets)

(ϵ, δ) - approximation algorithm

GitHub

$=$

Upper-bound on the # of samples

$$\frac{c}{\epsilon^2}\left(1 + \ln\frac{1}{\delta}\right)$$

$\times$

Uniform simplet sampling using random walk

$$O(\log(n) \cdot \Delta \cdot diam(\mathcal{K})^2)$$

Sub-linear for Real World SCs

Hamid Beigy, Mohammad Mahini, Salman Qadami, Morteza Saghafian

Presenting by Mohammad Mahini

Future Directions

1. [The SFD vector for Simplexes] $S_{\mathcal{K}}(s,i)$

$$SFD(s) = (\phi_{\mathcal{K}}(s,1), \ldots, \phi_{\mathcal{K}}(s,N_m))$$

2. [Centrality measure for simplexes]

$$\sum w_i \times \phi_{\mathcal{K}}(s,i)$$

3. [Alpha Complexes] - Filtering
4. [Simplicial Complex Similarity Metric]
5. [Classification Applications]

Thanks for your attention