On Edge Collapse Of Random Simplicial Complexes

Soumik Dutta, Joint work with Jean-Daniel Boissonnat, Kunal Dutta, and Siddharth Pritam

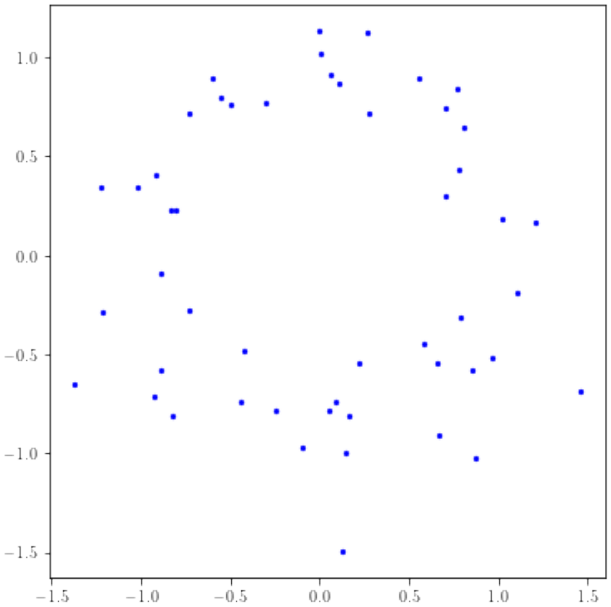
Algorithms seminar, MIMUW

24.5.2024

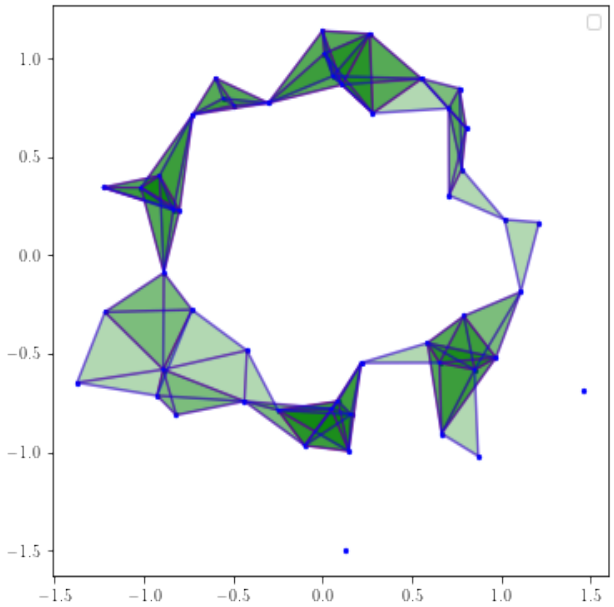
Table of Contents

Introduction

- Topological Data Analysis (TDA) aims to extract topological features of spatially distributed data.
- First the point cloud is turned into a topological object suitable for computation.
- Then homological information is extracted.



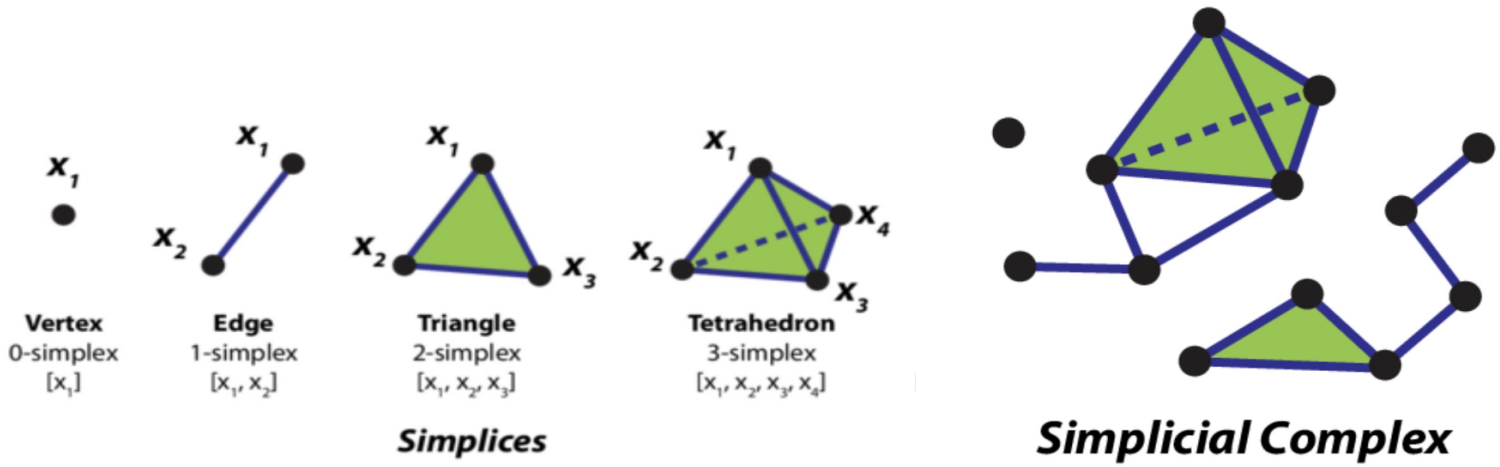
(a) Point cloud



(b) Vietoris-Rips complex

Simplicial Complex

- Simplicial Complex is a discrete version of topological spaces.
- They are made of vertices and 'faces'(an edge being a face of dimension 1) of different dimensions satisfying that the intersection of two faces is either another face or empty.
- An abstract simplicial complex is a collection of faces that is closed under subset operation.
- They may contain holes, voids, and so on.



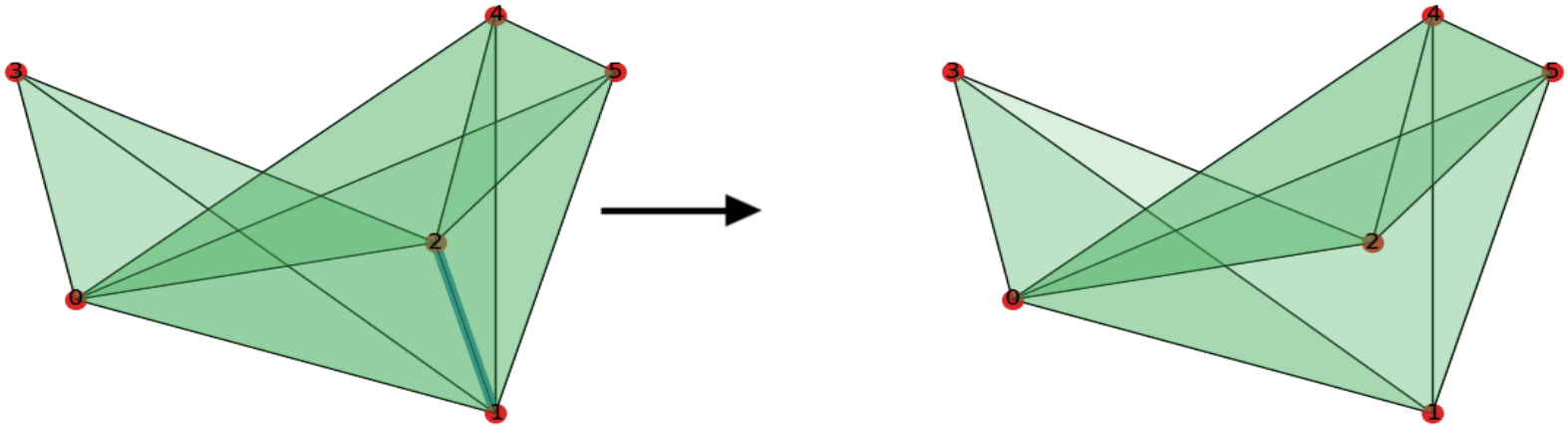
Collapses

- Calculating homology and persistent homology is computationally costly, so preprocessing is needed.

- Simple collapse is a combinatorial process that simplifies a simplicial complex without changing its topology. It can be expressed as a series of elementary moves of removals of pair of simplices σ and τ, such that σ is uniquely contained in τ. The notion of simple collapse was introduced by J.H.C Whitehead to study homotopy types of cell complexes.
- Strong Collapse, as a variant of simple collapse, was introduced by Barmak and Minian. In this process, vertices are removed.
- Edge collapse was introduced by Boissonnat and Pritam. In this process, edges are removed.

Edge Collapse

- An edge e in the simplicial complex K is dominated by another vertex $v \in K$ if all the maximal simplices of K that contain e also contain v.
- Removing the dominated edge along with all its cofaces does *not* change the homotopy type of the complex.
- The process is called edge collapse.

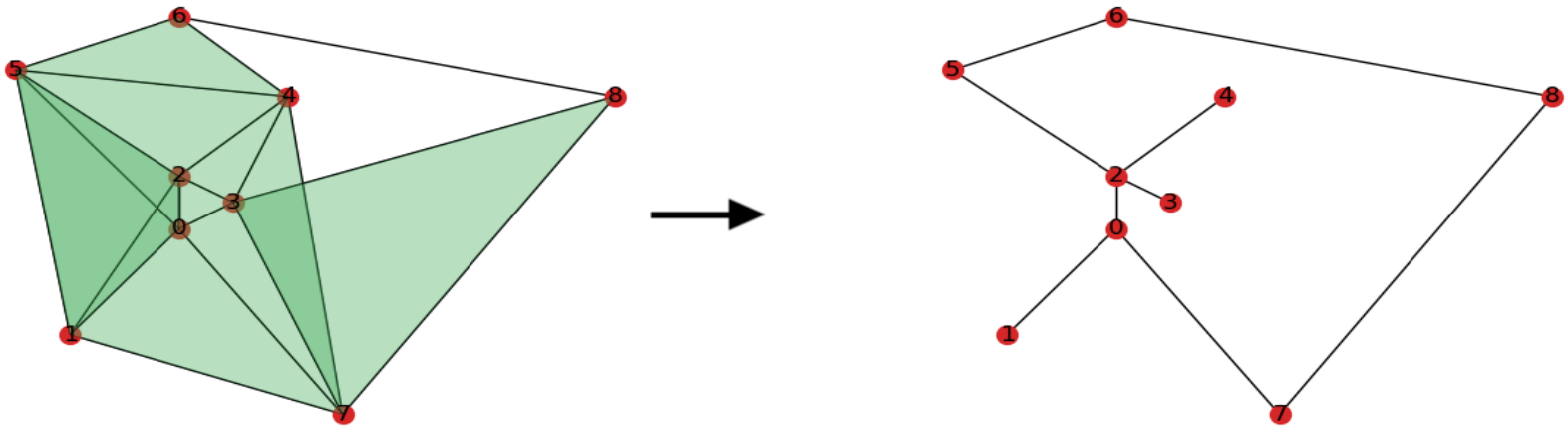


(a) Initial simplicial complex

(b) simplicial complex after collapsing the edge $(1, 2)$

Example of size reduction in simplicial complex

- Even though the following complexes have different sizes they have the same homotopy type.
- Same homotopy type implies the same homology and Betti numbers.
- The complex on the right can be achieved by a sequence of edge collapse.



(a) Initial simplicial complex

(b) simplicial complex after applying edge collapses

Motivations

- The number of d-simplices is typically $O(m^{d-2})$ (m= # edges) and time complexity of persistent homology algorithm is $O(N^{\omega})$ (N = # simplices, ω = matrix multiplication exponent).
- Edge collapse accelerates PH computation.
- Our study is motivated by the need for theoretical explanation of the experimental success of [Wilkerson et al. ICASSP'14; Boissonnat, Pritam SoCG'19, SoCG'20].
- Our results also provide an average case analysis for edge collapse when no extra information about the data set is provided.
- Observed experimental and practical successes have led to its deployment in software packages like GUDHI.

Models of random simplicial complex

- The study of random simplicial complexes was initiated in the seminal paper of Linial and Meshulam. They have introduced the d-dimensional LM model of random simplicial complex.
- Simple collapses on random complexes were studied by Aronshtam, Linial, Łuczak, and Meshulam.
- Linial and Peled obtained precise asymptotic bounds on the size of the core of such complexes.
- The random clique complex model was introduced by Kahle.

Our goal is to study the effect of edge collapse in reducing the size of the complex for random clique complexes.

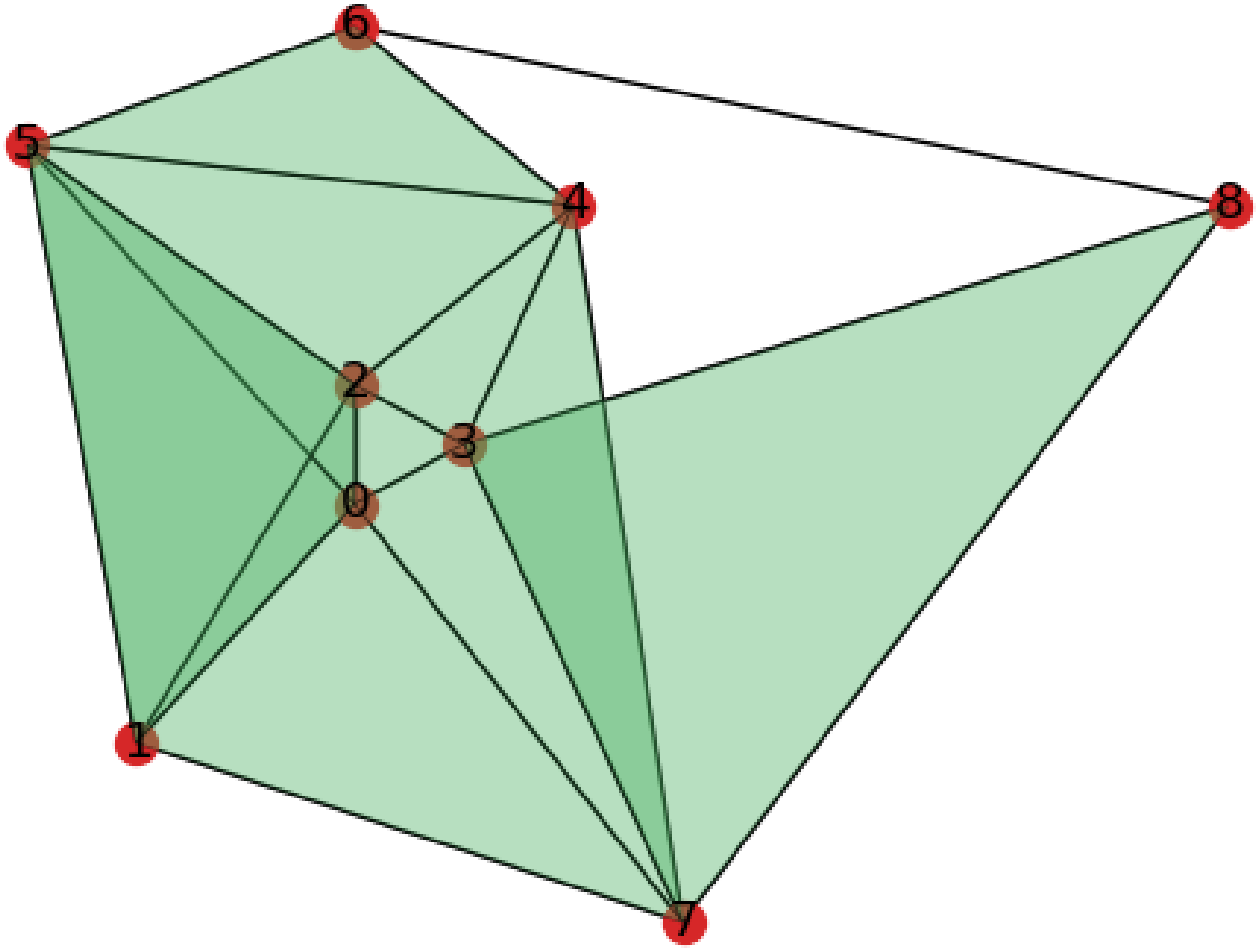
t-core Algorithm

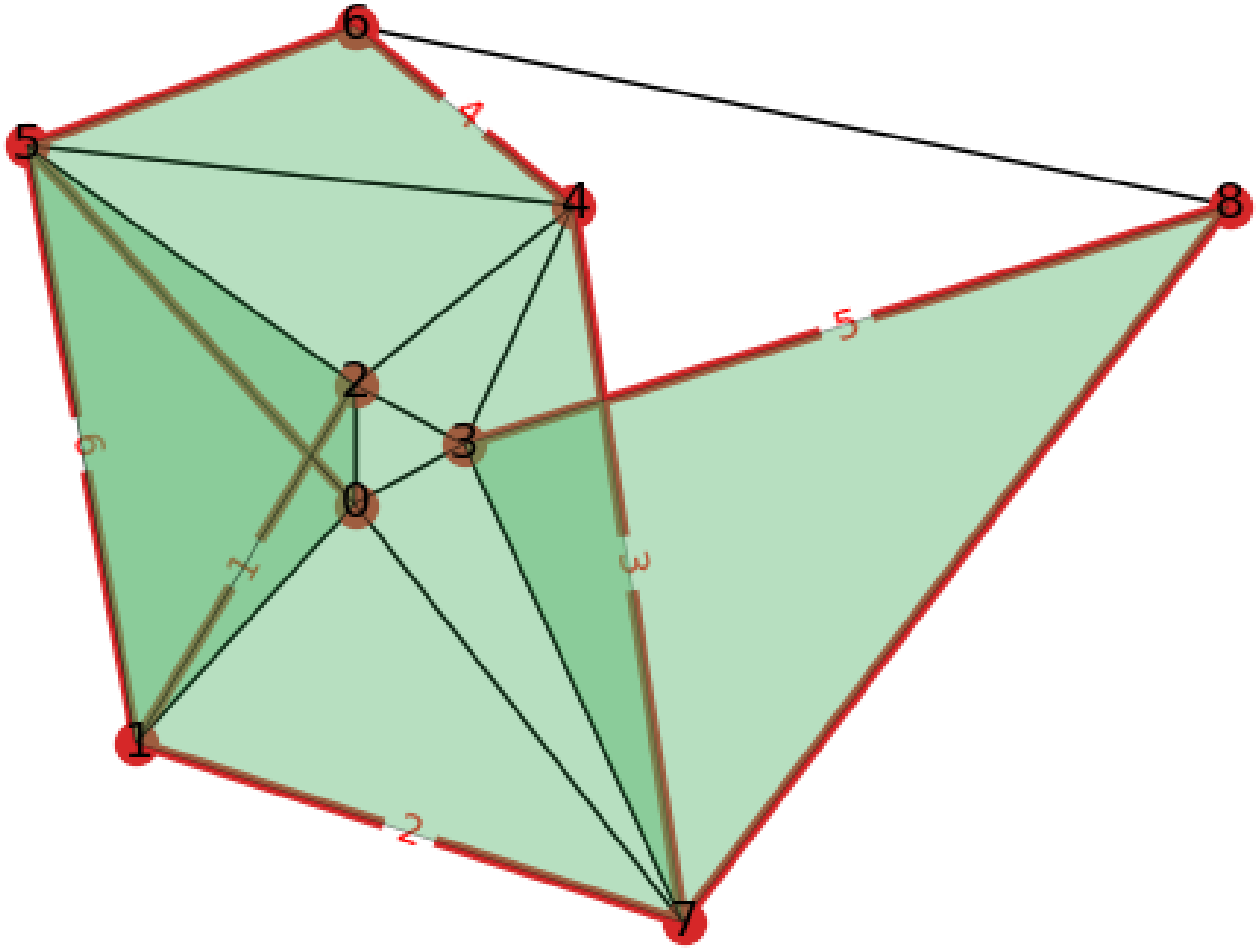
We shall iteratively run the edge collapse procedure according to the following algorithm.

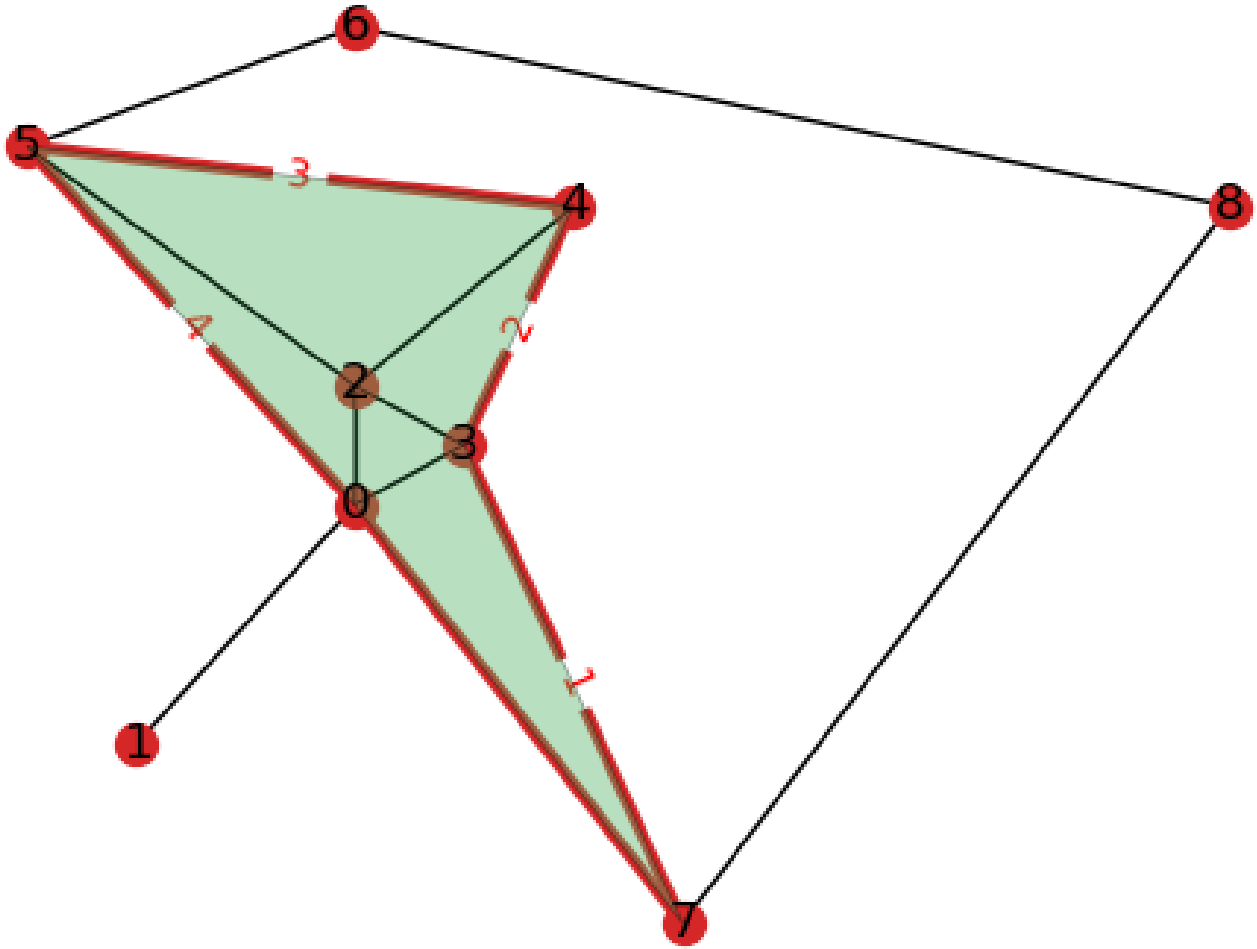
Algorithm 0: Compute t-core under edge collapse

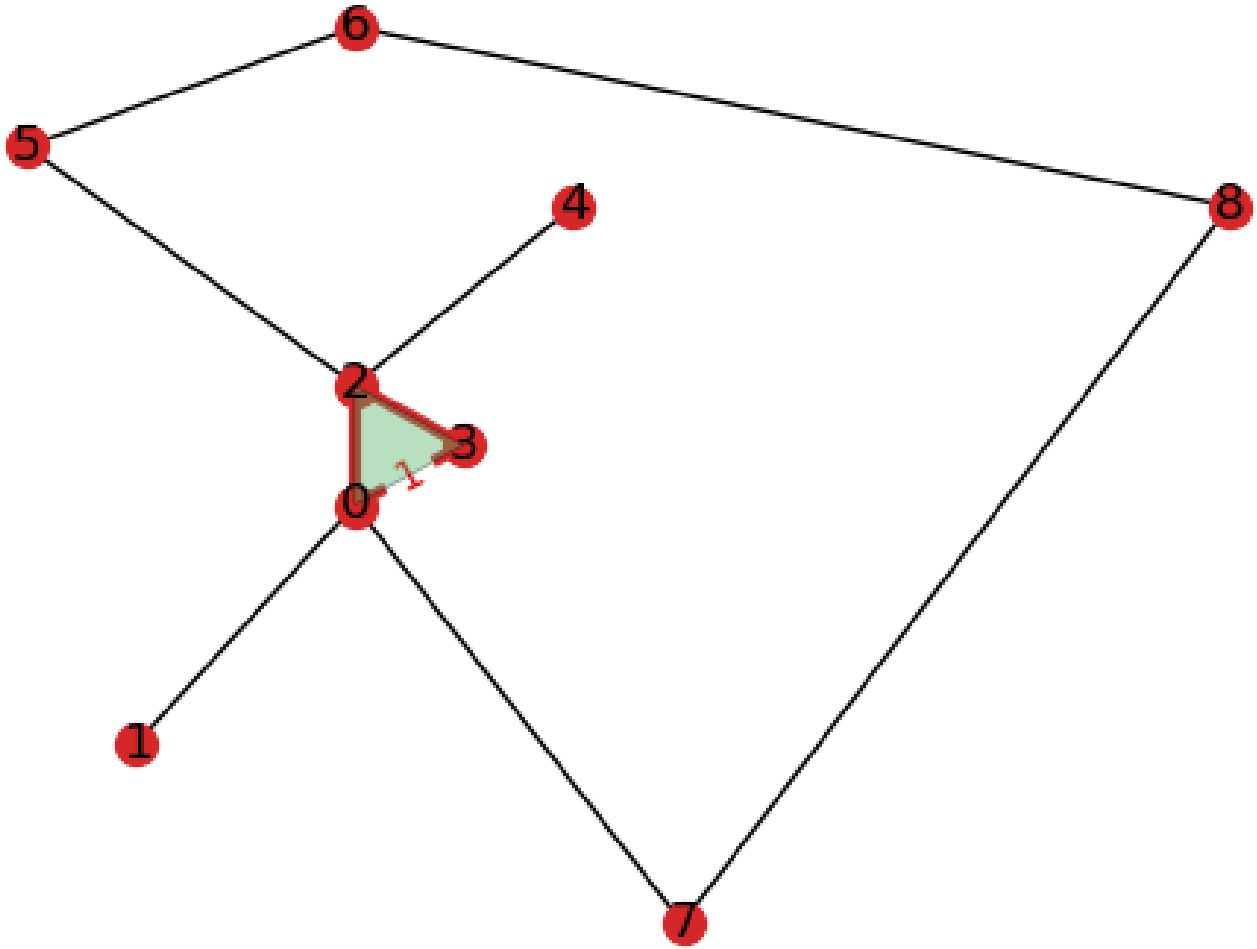
Data: Simplicial complex X
Result: $R_t(X)$, i.e., t-core of X under edge collapse

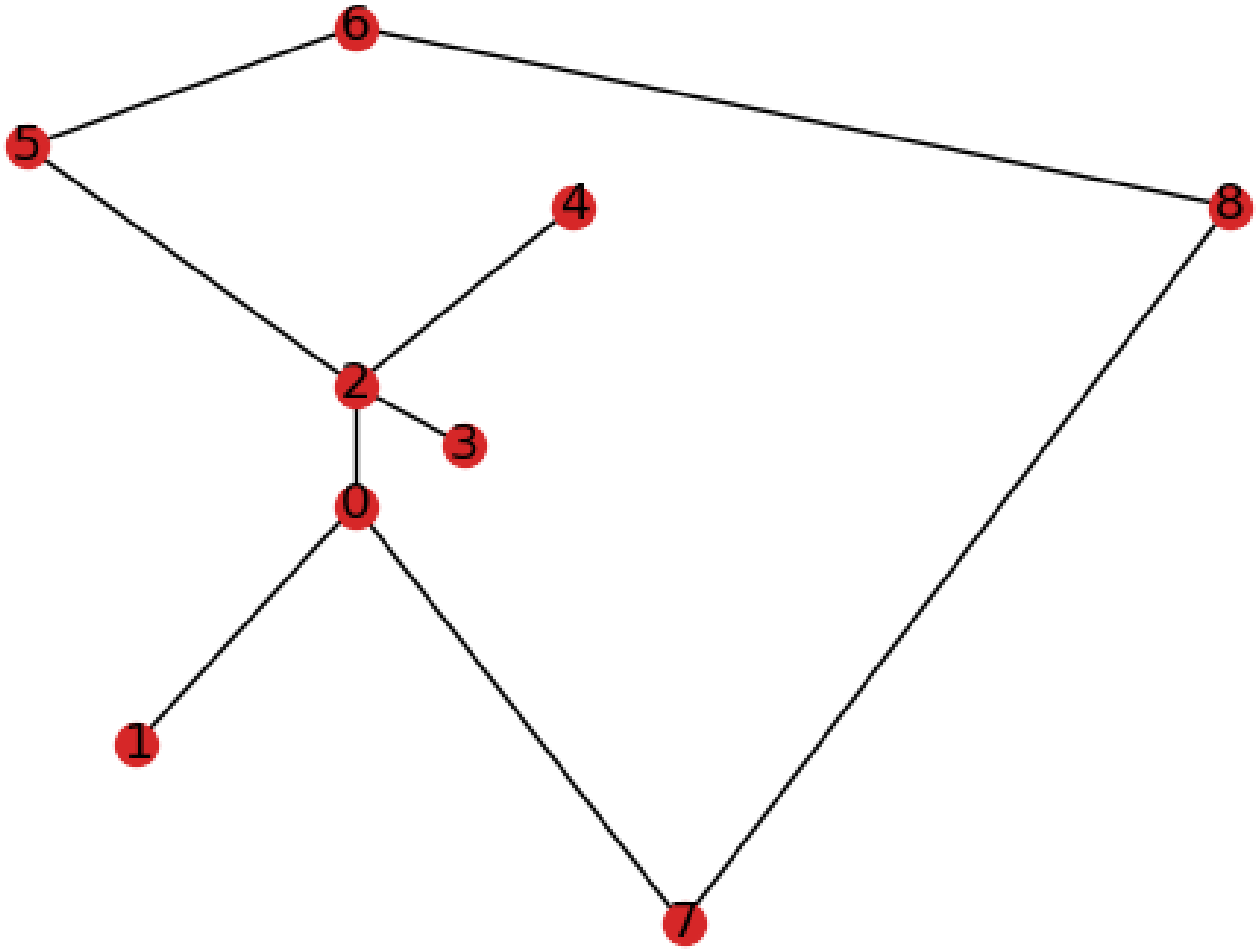
1 Calculate the set D of all the dominated edges of the current complex;
2 Shuffle elements of D uniformly randomly;
3 **for** e *in* D **do**
4 | If e is dominated by some vertex in the current complex remove e and all its cofaces.
5 **end**
6 Repeat the above steps t times.


6
5
4
8
2
3
0
1
7


6
5
4
8
2
3
0
1
7
4
5
6
1
3
2


6
5
3
4
8
4
2
2
3
0
1
1
7


6
5
4
8
2
3
1
0
1
7


6
5
4
8
2
3
0
1
7

Flag Complex Of Random Graph

In the Erdos-Renyi model of random graph on a fixed number of vertices(n), any of the possible edges can occur independently with some fixed probability(p). This model is denoted as $G(n, p)$. Then, every k-clique is replaced with a $k-1$ face. This model of random simplicial complex is denoted by $X(n, p)$.

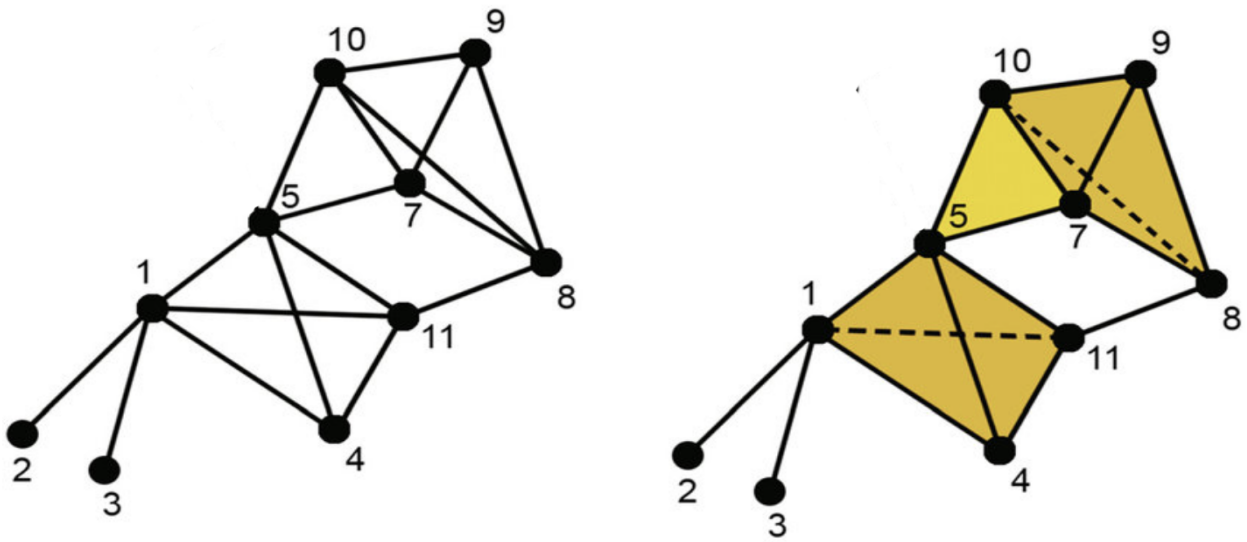


Figure: Flag Complex

Results

- For $X \sim X(n,p)$, let $P(X)$ denote the number of dominated edge-dominating vertex pairs.It can be shown that $\mathbb{E}[P(X)] = O(n^3p^3\left(1-p^2(1-p)\right)^{\binom{n-3}{}}) = O(n^3p^3\exp(-np^2(1-p)))$.
- Let $p = c/\sqrt{n}$. If $c > 2\log(n)$ then, with high probability, there is no dominated edge to start the collapsing procedure.
- If $c \lessapprox 1.57$ then the expected number of 2-simplices in the core becomes $o(n\sqrt{n})$. So we focus on the regime where $c \gtrapprox 1.57$ is constant with respect to n.
- Newman'24 studies the same in sparse settings with $c \leq 0.00625$.
- Define $\eta_2 = \inf\{\eta | x = e^{-\eta(1-x)^2} \text{has a solution in } (0,1)\}$. In particular, $\eta_2 \approx 2.45$ and $c = \sqrt{\eta_2} \approx 1.57$

Results

- We give bounds on the size of the *t-core* (i.e. the remaining complex after t phases of edge collapse) of a random simplicial complex.
- We show that for ER clique complexes, the size of the core after t phases of maximal edge collapses is a.a.s. a constant fraction of initial number of edges.
- The constant depends only on the edge probability, and is bounded away from 1.

Theorem

Let $X \sim X(n,p)$ such that $p = c/\sqrt{n}$ with $c > \sqrt{\eta_2}$. Let $\{\gamma_t\}_{t\geq 0}$ be defined recursively by $\gamma_{t+1} = e^{-c^2(1-\gamma_t)^2}$ and $\gamma_0 = 0$. Fix $t \leq \log n/(8\log\log n) - 1$. Then with probability greater than $(1 - O(1/n^{11/2}))$

$$(1+o(1))\binom{n}{2}p(1-\gamma_{t+1}-c^2\gamma_t(1-\gamma_t)^2) \leq |f_1(R_t(X))| \tag{1}$$

$$|f_1(R_t(X))| \leq (1+o(1))\binom{n}{2}p(1-(c^2/3)(1-(1-\gamma_t)^3)). \tag{2}$$

We also provide upper and lower bounds on the size of the final core

Theorem

Define $\eta_2 = \inf\{\eta | x = e^{-\eta(1-x)^2}$ *has a solution in* $(0,1)\}$. *Let* $X \sim X(n,p)$ *with* $p = c/\sqrt{n}$ *and* $c > \sqrt{\eta_2}$. *Let* γ *be the smallest positive solution of the equation* $x = e^{-c^2(1-x)^2}$. *Then a.a.s.*

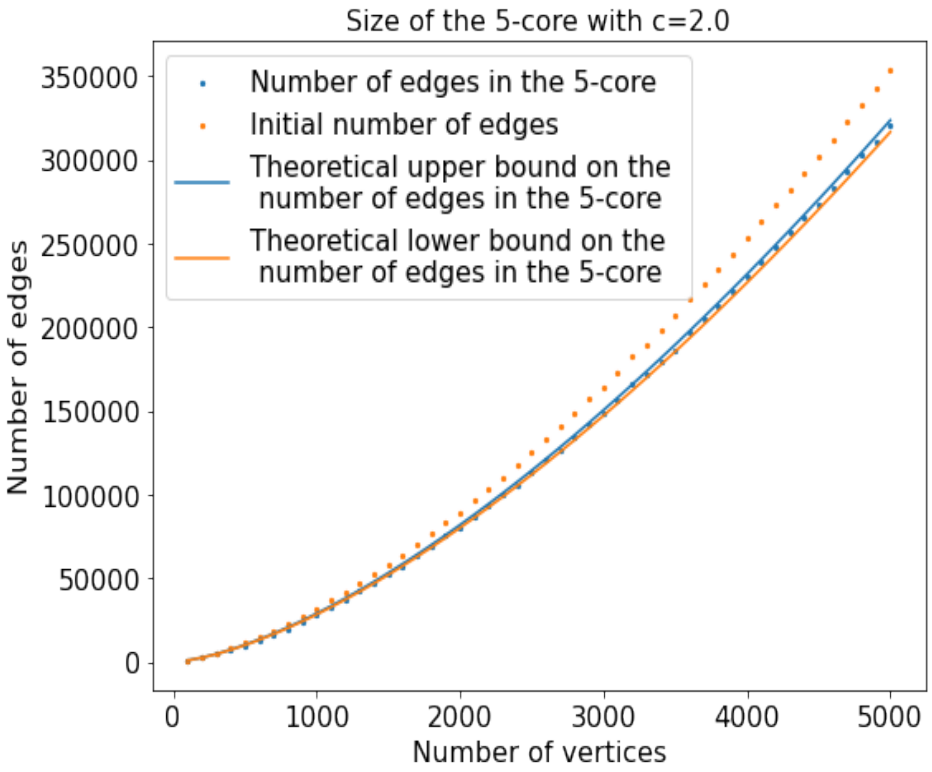
$$|f_1(R_\infty(X))| \leq (1+o(1))\binom{n}{2}(c/\sqrt{n})(1-(c^2/3)(1-(1-\gamma)^3)) \tag{3}$$

and

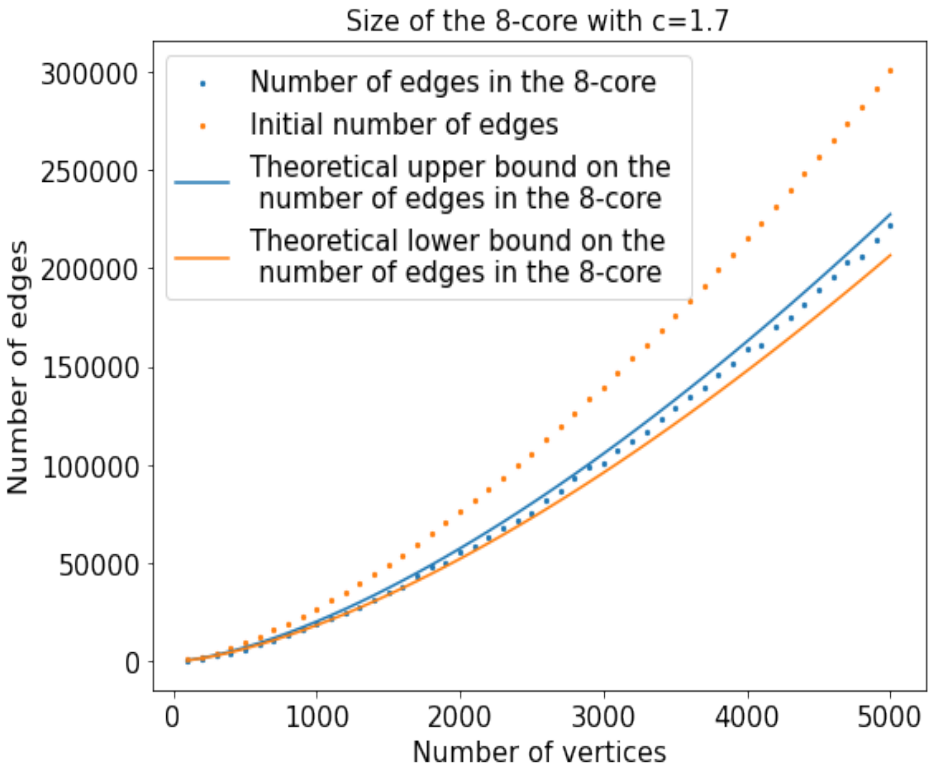
$$(1+o(1))\binom{n}{2}p(1-\gamma-c^2\gamma(1-\gamma)^2) \leq |f_1(R_\infty(X))|$$

Simulation

https://github.com/soumikdt/Edge-Collapse-On-Random-Clique-Complex

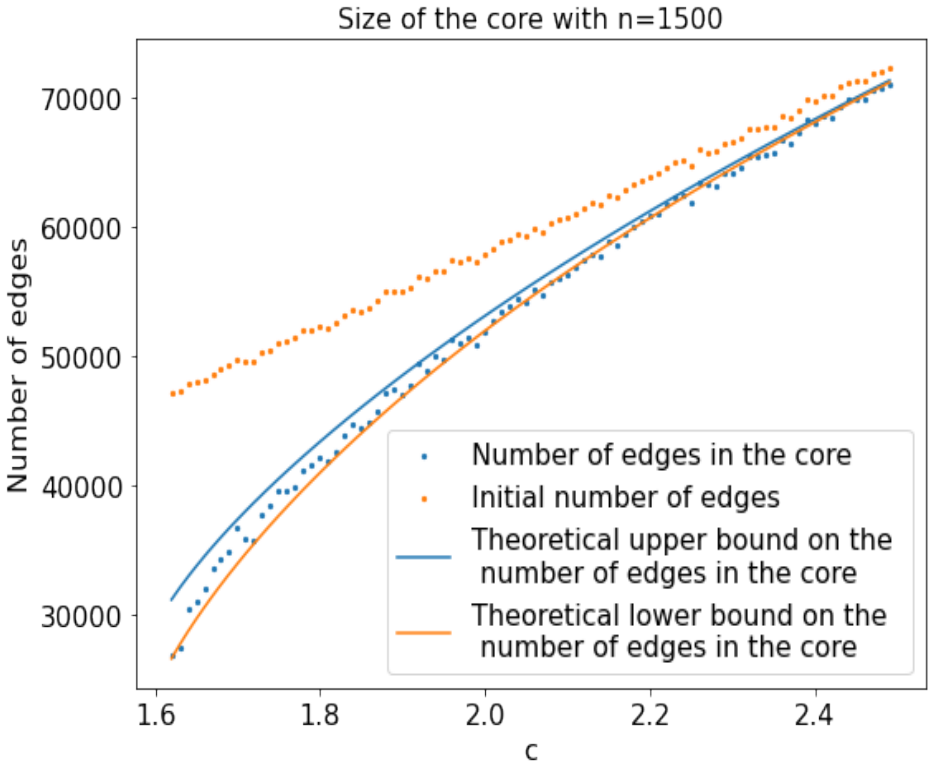


(a) Size of the 5-core with $c = 2.0$.

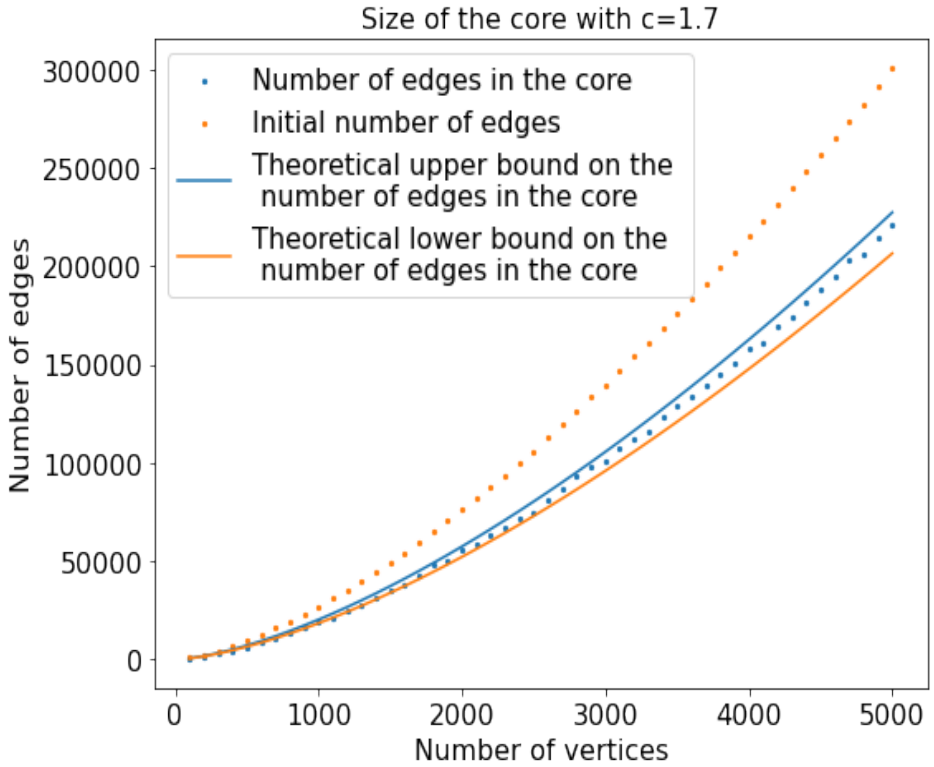


(b) Size of the 8-core with $c = 1.7$.

Figure: Fig. 5a shows the number of edges in the 5-core of $X \sim X(n, 2/\sqrt{n})$ along with the initial number of edges. The solid line indicates the theoretical upper and lower bounds on the expected number of edges. Fig. 5b shows the same thing for the 8-core of $X \sim X(n, 1.7/\sqrt{n})$.



(a) Size of the core plotted against c.



(b) Upper and lower bound of the size of the core.

Figure: In Fig. 6a, we plotted the number of edges in the core of $X \sim X(1500, c/\sqrt{1500})$ for $c \in [1.6, 2.5]$ along with the initial number of edges. The solid lines indicate the theoretical upper and lower bounds on the expected number of edges. Note that these bounds tend to coincide as c increases as shown in the figure. In Fig. 6b, we ran the pruning procedure until there were no more dominated edges. The figure shows the number of edges in the final core of $X \sim X(n, 1.7/\sqrt{n})$ along with the initial number of edges. The solid line indicates the theoretical upper bound on the expected number of edges.

Outline of the proof:

- Upper and lower bound on t-core.
  - Expectation
    - We first identify the local structure.
    - We then exploit the tree like property of the local structure.
  - Concentration
    - We first identify a class of critical structures responsible for large deviation from the expectation.
    - Then we apply a concentration inequality that we developed for typically bounded functions.
- Upper and lower bound on core.

The Local Structure: Random 2−tree

- Due to the sparsity of the graph, 2-dimensional cycles are rare.
- Thus the complex locally looks like a 2-tree.
- The complex locally looks like a random 2-tree with high probability.

Lemma

Let $X \sim X(n, c/\sqrt{n})$ *and* $e \in f_1(X)$. *Then, for* $t \leq \log n/(8 \log\log n) - 1$,
$Pr\{\{\mathcal{N}_t \in \mathcal{T}\} \cap E\} = 1 - O(n^{-1/4})$.

- Each edge has $Binomial(c^2/n, n-2) \approx Poisson(c^2)$ children.

The Local Structure: Random 2−tree

A 2-dimensional tree is built recursively as follows:
1. Start with a single edge(root).
2. In the nth iteration, add children to all the leaves (edges) at distance $n-1$ from the root from the Poisson distribution with parameter $c' = c^2$.

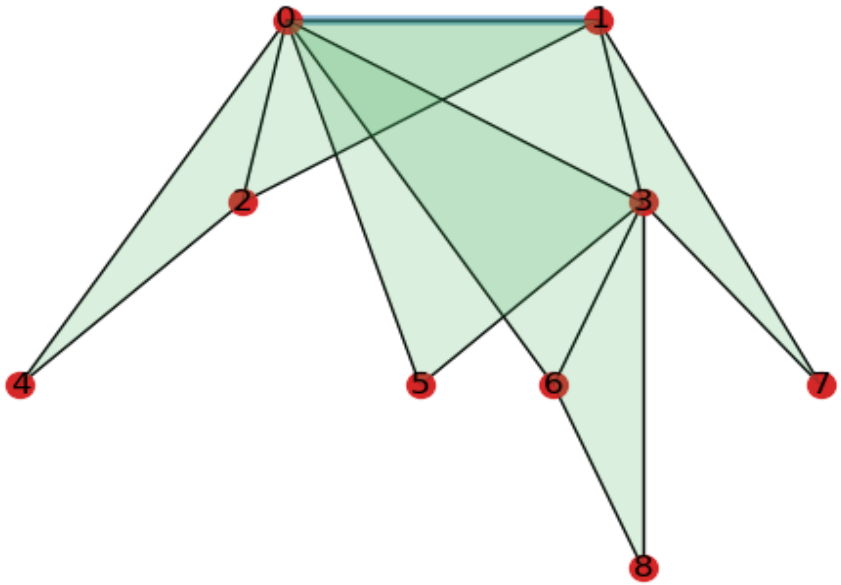


Figure: 2-dimensional Tree

- Collapsing starts form the leaves.
- Let γ_t be the probability that a tree $T \in \mathcal{T}_t$ is pruned to the root in *no more than* $t-1$ steps.
- $\gamma_{t+1} = e^{-c'(1-\gamma_t)^2}$ and $\gamma_0 = 0$.
- γ is the limit of γ_t.
- Surviving probability $\beta_t = 1 - \gamma_{t+1}$.

Expected size of t-core

Theorem

Define $\eta_2 = \inf\{\eta | x = e^{-\eta(1-x)^2}$ *has a solution in* $(0,1)\}$. *Let* $X \sim X(n,p)$ *such that* $p = c/\sqrt{n}$ *with* $c > \sqrt{\eta_2}$. *Let* γ_t *be defined recuresively by* $\gamma_{m+1} = e^{-c^2(1-\gamma_m)^2}$ *and* $\gamma_0 = 0$. *Then*

$$(1+o(1))\binom{n}{2}p(1-\gamma_{t+1}-c^2\gamma_t(1-\gamma_t)^2) \leq \mathbb{E}[|f_1(R_t(X))|] \tag{4}$$

$$\mathbb{E}[|f_1(R_t(X))|] \leq (1+o(1))\binom{n}{2}p(1-(c^2/3)(1-(1-\gamma_t)^3)). \tag{5}$$

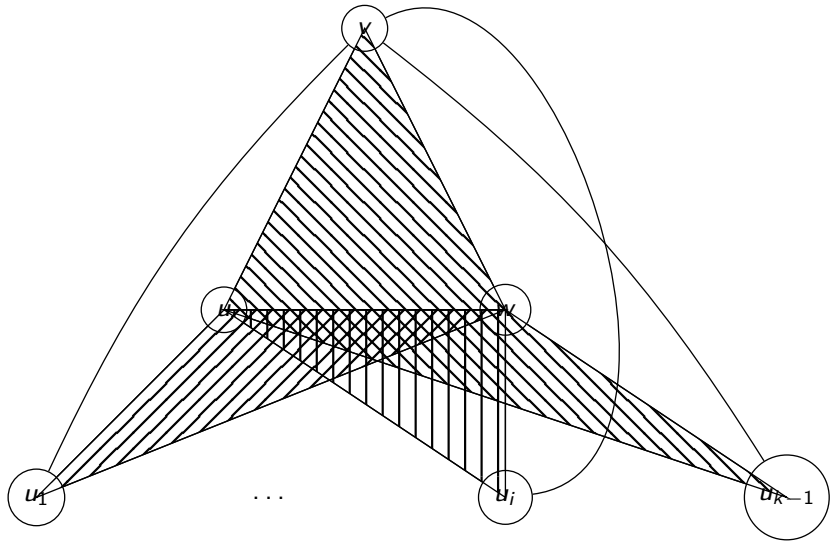
- Note that the upper and lower bound do not coincide.
- This is because not all dominated simplices are collapse during a pruning phase.
- We look at the higher (3-dimensional) simplices and prove that

$$d_2 \leq d_3 + d_1$$

where d_i is the number of collapsed i-dimensional simplices.

Concentration of t-core size: Identifying the critical structures

- Deleting one dominated edge can potentially generate $O(n^2)$ dominated edges. We call the structure responsible for this *Critical Complex*.



- Thus the actual size of the t-core can drastically vary from its expected size depending on the random simplicial complex and route that our randomized algorithm takes.
- We prove that a critical complex must be present as a subcomplex for a critical event to occur.
- This problem is not present in LM model.

- Critical complex of order k requires $k+2$ vertices and $3k$ edges. Expected number is $O(\frac{1}{(\sqrt{n})^{k-4}})$. Thus, the expected number is vanishingly low for large enough k.
- Let f denote the function that gives the number of edges in the t-core of a complex.
- If our initial complex does not contain any critical complex then the f is not much affected by flipping one edge of the initial complex.
- If our initial complex contains a critical complex then the f may be hugely affected by flipping one edge of the initial complex.
- So, we can not use concentration inequality for bounded functions, e.g., McDiarmid's inequality.
- As f is typically well behaved, we developed an concentration inequality that improves upon an inequality of Warnke'16.

Concentration of t-core size: Formulating the concentration inequality

Fix n to be the number of vertices of a graph G, $p < 0.5$, and set $m = \frac{n(n-1)}{2}$. We order every pair of vertices G in an arbitrary manner. Let e_i denote the i-th such pair. We set $e_i = 1$ if the corresponding edge exists and $e_i = 0$ otherwise. For $1 \leq i \leq m$, let $e_i \sim Bernoulli(p)$ be i.i.d. random variables. Define $e_i(G)$ be the graph obtained after *flipping* the value e_i in G, i.e., if the potential edge e_i does not exist in G it exists in $e_i(G)$ and vice versa. Let $C_G(H)$ denote the number of copies of H in G as a subgraph.

Theorem

Assume notations as above. Let $\mathcal{G}$ be the set of labeled graphs on n vertices and H be an arbitrary but fixed graph on less than n vertices. Let $f : \mathcal{G} \to \mathbb{R}$ be a function which satisfies the following modified Lipschitz conditions:

1. *If $C_G(H) = 0$ and $C_{e_i(G)}(H) = 0 \ \forall i \in \{1, \cdots, m\}$ then $|\Delta_{e_i}(f(G))| := |f(G) - f(e_i(G))| \leq C \ \ \forall i$.*
2. *Otherwise $|\Delta_{e_i}(f(G))| \leq D \ \ \forall i \in \{1, \cdots, m\}$.*

Now let $G \sim G(n, p)$ with $p \leq 1/2$ and $B := \mathbb{E}[C_G(H)]$. Then for all $T > B$

$$Pr\{|f - \mathbb{E}[f]| > s\} \leq 2 \exp\left(-\frac{s^2}{2(C + DT/p)(mpC + mDT + s/3)}\right) + B/T. \tag{6}$$

Applying the last theorem along with the previous lemma on the critical structures we get our desired concentration result.

Theorem

Let $X \sim X(n,p)$. Let $|f_1(R_t(X))|$ be the number of edges after t edge collapsing phases and $Y_0 = \mathbb{E}(|f_1(R_t(X))|)$ be its expected value. Then for any $s \geq 0$, we have,

$$Pr\{||f_1(R_t(X))| - Y_0| \geq s \cdot n\} \leq 2\exp(-s^2\Theta(\sqrt{n})) + O(1/n^{5.5})$$

Upper bound of core

- First we observe that, for a fixed flag complex X, $|f_1(R_t(X))| \geq |f_1(R_{t+1}(X))|$ for any $t \geq 1$.
- Thus $|f_1(R_t(X))| \geq |f_1(R_\infty(X))|$ for any $t \geq 1$. Thus for any $t \geq 1$, $|f_1(R_t(X))|$ is an upper bound of the size of the final core.
- In particular, we shall take $t = \rho := \log n/16 \log\log n$. It can be shown that a.a.s. $(1 + o(1))\binom{n}{2}(c/\sqrt{n})(1 - (c^2/3)(1 - (1 - \gamma)^3)) \geq (|f_1(R_\rho(X))|$.
- Then the result follows immediately as $|f_1(R_\rho(X))| \geq |f_1(R_\infty(X))|$.

Lower bound of core

- The lower bound of the size of the core can not be obtained directly form the lower bound of the size of t-core.
- This is because $t = O(\log n / \log \log n)$ and edges can be collapsed even after t iterations.
- So the problem is about inferring a global property from the knowledge of a local property.
- The proof for the lower bound on the size of the core follows a different approach.
- The proof is via the recursive technique of Riordan, 2008, and Linial and Peled, 2017 to simulate cycles in infinite trees.

Proof idea

- The idea is, roughly speaking, to craft two properties, $\mathcal{P}$ and $\mathcal{A}$, defined on the space of rooted random 2-trees. The property $\mathcal{A}$ says that a 2-tree can not be pruned within a fixed finite number of steps.
- The property $\mathcal{P}$ implies that sufficiently lower descendant leaves of the root of the tree have property $\mathcal{A}$. Then one shows that property $\mathcal{A}$ implies property $\mathcal{P}$ with high probability.

$$\mathcal{P} \implies^{\text{(lower descendants have)}} \mathcal{A} \implies^{w.h.p} \mathcal{P}$$

- By construction

$$\mathcal{A} \implies \text{can not be collapsed within } \theta(\log n / \log\log n) \text{ phases}$$

- Due to this recursive nature of the property $\mathcal{A}$, an edge with property $\mathcal{A}$ never gets collapsed a.a.s. Thus estimating the number of edges with property $\mathcal{A}$ gives a lower bound on the size of the core.
- Thus by studying $\mathcal{A}$, which is a property defined on a local neighborhood, we can infer about a global property.

THANK YOU

A full version of the paper is available at
https://www.researchgate.net/publication/379119064_On_Edge_Collapse_of_Random_Simplicial_Complexes