SimBa: An Efficient Tool for Approximating Rips-filtration Persistence via **Sim**plicial **Ba**tch-collapse

Tamal K. Dey

Department of Computer Science and Engineering
The Ohio State University

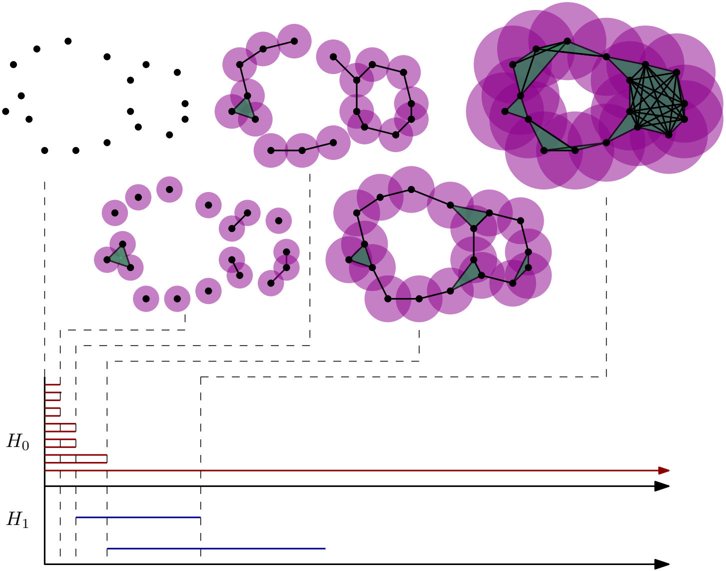
2016

Joint work with Dayu Shi and Yusu Wang

Rips-filtration persistence

- Problem: P from a metric space, $P \in \mathbb{R}^d$, compute Vietoris-Rips (Rips) filtration persistence.
- Rips complex
$$\mathcal{R}^\alpha(P) = \{\langle p_0, \ldots, p_s \rangle \mid \|p_i - p_j\| \leq \alpha, \forall i, j \in [0, s],\ p_i, p_j \in P\}.$$
- Rips filtration
$$\{\mathcal{R}^\alpha(P)\}_\alpha :=\ \mathcal{R}^{\alpha_1}(P) \hookrightarrow \mathcal{R}^{\alpha_2}(P) \cdots \hookrightarrow \mathcal{R}^{\alpha_n}(P) \cdots$$
- Persistent homology
$$H_p(\mathcal{R}^{\alpha_1}(P)) \to H_p(\mathcal{R}^{\alpha_2}(P)) \to \ldots \to H_p(\mathcal{R}^{\alpha_n}(P)).$$

Persistence barcode



Sparsification of Rips filtration

- Size of Rips complex become prohibitively large as α increases.
- Sparse Rips filtration
  - Inclusion: [Sheehy2012]

Sparsification of Rips filtration

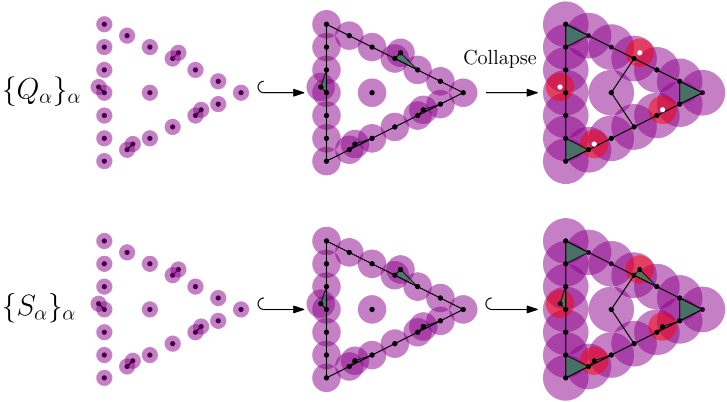
- Size of Rips complex become prohibitively large as α increases.
- Sparse Rips filtration
  - Inclusion: [Sheehy2012]
  - Batch-collapsed Rips [D.-Feng-Wang 2014]

Sparsification of Rips filtration

- Size of Rips complex become prohibitively large as α increases.
- Sparse Rips filtration
  - Inclusion: [Sheehy2012]
  - Batch-collapsed Rips [D.-Feng-Wang 2014]
  - Simple collapse: [Cavanna et al.2015]
- New work: SimBa
  - Use batch-collapse
  - Use set distance

Sparse Rips complexes [Sheehy 12]

- Intuition:



- White points' contribution can be ignored. Stop growing those balls so that they don't contribute to later complexes. Even delete them later.

Sparse Rips complexes [Sheehy 12]

- Greedy permutation $\{p_1, .., p_n\}$:
  - Let $p_1 \in P$ be any point and define p_i recursively as $p_i = \text{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$, where $P_{i-1} = \{p_1, ..., p_{i-1}\}$.
  - Insertion radius of p_i: $\lambda_{p_i} = d(p_i, P_{i-1})$.

Sparse Rips complexes [Sheehy 12]

- Greedy permutation $\{p_1, .., p_n\}$:
  - Let $p_1 \in P$ be any point and define p_i recursively as $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$, where $P_{i-1} = \{p_1, ..., p_{i-1}\}$.
  - Insertion radius of p_i: $\lambda_{p_i} = d(p_i, P_{i-1})$.
- The weight of a point [BCOS15]:

$$w_p(\alpha) = \begin{cases} 0 & \text{if } \alpha \leq \frac{\lambda_p}{\varepsilon} \\ \alpha - \frac{\lambda_p}{\varepsilon} & \text{if } \frac{\lambda_p}{\varepsilon} < \alpha \leq \frac{\lambda_p}{\varepsilon(1-\varepsilon)} \\ \varepsilon\alpha & \text{if } \frac{\lambda_p}{\varepsilon(1-\varepsilon)} \leq \alpha \end{cases}$$

Sparse Rips complexes [Sheehy 12]

- Greedy permutation $\{p_1, .., p_n\}$:
  - Let $p_1 \in P$ be any point and define p_i recursively as $p_i = \mathrm{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$, where $P_{i-1} = \{p_1, ..., p_{i-1}\}$.
  - Insertion radius of p_i: $\lambda_{p_i} = d(p_i, P_{i-1})$.
- The weight of a point [BCOS15]:

$$w_p(\alpha) = \begin{cases} 0 & \text{if } \alpha \leq \frac{\lambda_p}{\varepsilon} \\ \alpha - \frac{\lambda_p}{\varepsilon} & \text{if } \frac{\lambda_p}{\varepsilon} < \alpha \leq \frac{\lambda_p}{\varepsilon(1-\varepsilon)} \\ \varepsilon\alpha & \text{if } \frac{\lambda_p}{\varepsilon(1-\varepsilon)} \leq \alpha \end{cases}$$

- Perturbed distance between two points:

$$\hat{d}_\alpha(p, q) = d(p, q) + w_p(\alpha) + w_q(\alpha).$$

Sparse Rips filtration

- Sparse Rips complex:
$$\mathcal{Q}^\alpha = \{\sigma \subset N_{\varepsilon(1-\varepsilon)\alpha} \mid \forall p, q \in \sigma,\ \hat{d}_\alpha(p,q) \leq 2\alpha\}.$$

Sparse Rips filtration

- Sparse Rips complex:

$$\mathcal{Q}^\alpha = \{\sigma \subset N_{\varepsilon(1-\varepsilon)\alpha} \mid \forall p, q \in \sigma,\ \hat{d}_\alpha(p,q) \leq 2\alpha\}.$$

- Sparse Rips filtration:

$$\{\mathcal{S}^\alpha\}_\alpha, \text{ where } \mathcal{S}^\alpha = \bigcup_{\alpha' \leq \alpha} \mathcal{Q}^{\alpha'}.$$

Sparse Rips filtration

- Sparse Rips complex:

$$\mathcal{Q}^{\alpha} = \{\sigma \subset N_{\varepsilon(1-\varepsilon)\alpha} \mid \forall p, q \in \sigma,\ \hat{d}_{\alpha}(p,q) \leq 2\alpha\}.$$

- Sparse Rips filtration:

$$\{\mathcal{S}^{\alpha}\}_{\alpha}, \text{ where } \mathcal{S}^{\alpha} = \bigcup_{\alpha' \leq \alpha} \mathcal{Q}^{\alpha'}.$$

- Persistence barcode of Sparse Rips filtration approximates that of Rips filtration. Use **GUDHI** to compute its persistence.

Sparse Rips with collapse [CJS2015]

- The size of $\{\mathcal{S}^\alpha\}_\alpha$ is still large due to union operation.

Sparse Rips with collapse [CJS2015]

- The size of $\{\mathcal{S}^\alpha\}_\alpha$ is still large due to union operation.
- Consider $\{\mathcal{Q}^\alpha\}_\alpha$ connected by simplicial maps $\mathcal{Q}^\alpha \to \mathcal{Q}^{\alpha'}$ for $\alpha < \alpha'$ originated from vertex collapses.

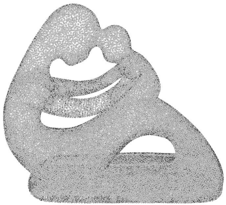
Sparse Rips with collapse [CJS2015]

- The size of $\{\mathcal{S}^\alpha\}_\alpha$ is still large due to union operation.
- Consider $\{\mathcal{Q}^\alpha\}_\alpha$ connected by simplicial maps $\mathcal{Q}^\alpha \to \mathcal{Q}^{\alpha'}$ for $\alpha < \alpha'$ originated from vertex collapses.
- collapse p to its nearest neighbor at its deletion time (scale) $\alpha_p = \frac{\lambda_p}{\varepsilon(1-\varepsilon)}$.

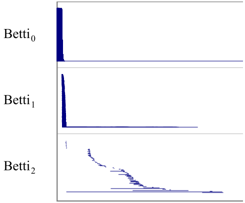
Sparse Rips with collapse [CJS2015]

- The size of $\{\mathcal{S}^\alpha\}_\alpha$ is still large due to union operation.
- Consider $\{\mathcal{Q}^\alpha\}_\alpha$ connected by simplicial maps $\mathcal{Q}^\alpha \to \mathcal{Q}^{\alpha'}$ for $\alpha < \alpha'$ originated from vertex collapses.
- collapse p to its nearest neighbor at its deletion time (scale) $\alpha_p = \frac{\lambda_p}{\varepsilon(1-\varepsilon)}$.
- The persistence of $\{\mathcal{Q}^\alpha\}_\alpha$ is exactly the same as that of $\{\mathcal{S}^\alpha\}_\alpha$. Use **Simpers** [DFW2014] to compute its persistence.

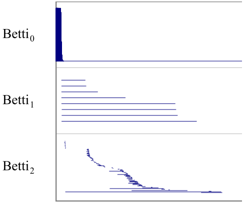
A snapshot experiment



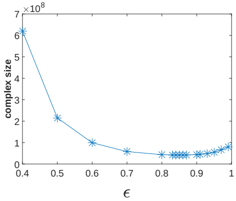
MotherChild model



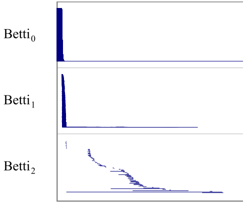
S.R. + GUDHI (original)



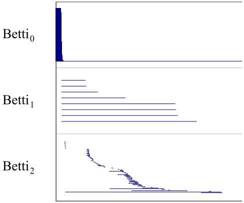
S.R. + GUDHI (denoised)



Cumulative complex size



S.R. + Simpers (original)



S.R. + Simpers (denoised)

Limitation of Sparse Rips

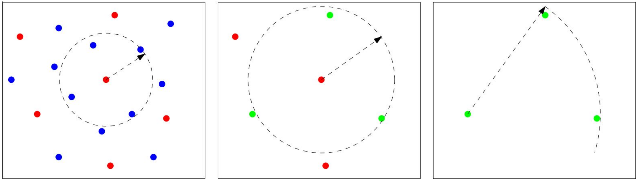
- Linear-size guarantee contains a hidden constant factor which depends exponentially on the doubling dimension of the space.

Limitation of Sparse Rips

- Linear-size guarantee contains a hidden constant factor which depends exponentially on the doubling dimension of the space.
- Size is still large and becomes worse as the dimension of data increases.
  - E.g., for a gesture phase data of 1747 points in $\mathbb{R}^{18}$ from UCI machine learning repository, the cumulative complex size is 45.6 million (up to tetrahedra).

Batch-collapsed Rips [DFW2014]

- Idea:
  - keep doing sub-sampling and collapsing points to their nearest sub-sample points.
  - build Rips complex only on the new sub-samples.



Batch-collapsed Rips [DFW2014]

- V_{k+1} is αc^{k+1}-net of V_k. Parameter $c > 1$: input constant (scale increase ratio).

Batch-collapsed Rips [DFW2014]

- V_{k+1} is αc^{k+1}-net of V_k. Parameter $c > 1$: input constant (scale increase ratio).
- Vertex map $\pi_k : V_k \to V_{k+1}$, for $k \in [0, m-1]$, such that for any $v \in V_k$, $\pi_k(v)$ is v's nearest neighbor in V_{k+1}.

Batch-collapsed Rips [DFW2014]

- V_{k+1} is αc^{k+1}-net of V_k. Parameter $c > 1$: input constant (scale increase ratio).
- Vertex map $\pi_k : V_k \to V_{k+1}$, for $k \in [0, m-1]$, such that for any $v \in V_k$, $\pi_k(v)$ is v's nearest neighbor in V_{k+1}.
- The sequence ($V_0 = P$):

$$\mathcal{R}^0(V_0) \to \mathcal{R}^{\alpha c \frac{3c-1}{c-1}}(V_1) \cdots \to \mathcal{R}^{\alpha c^m \frac{3c-1}{c-1}}(V_m).$$

Batch-collapsed Rips [DFW2014]

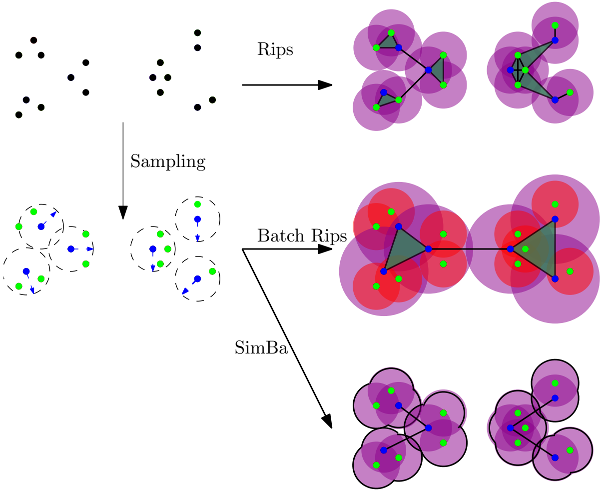
- V_{k+1} is αc^{k+1}-net of V_k. Parameter $c > 1$: input constant (scale increase ratio).
- Vertex map $\pi_k : V_k \to V_{k+1}$, for $k \in [0, m-1]$, such that for any $v \in V_k$, $\pi_k(v)$ is v's nearest neighbor in V_{k+1}.
- The sequence ($V_0 = P$):
$$\mathcal{R}^0(V_0) \to \mathcal{R}^{\alpha c \frac{3c-1}{c-1}}(V_1) \cdots \to \mathcal{R}^{\alpha c^m \frac{3c-1}{c-1}}(V_m).$$
- The persistence barcode of batch-collapsed Rips filtration approximates that of Rips filtration.

Limitation of Batch-collapsed Rips

- Over-connection: $\frac{3c-1}{c-1}$ results from the approximation guarantee, which ensures there is no missing link but causes over-connection.
- Trade-off: large c reduces over-connection but results in worse approximation.
- Over-connection becomes worse as data dimension increases.

SimBa

- Idea: use set distance rather than point distance to resolve the over-connection issue while still ensuring no missing link.



SimBa

- Vertex map is the same as that of batch-collapsed Rips.

SimBa

- Vertex map is the same as that of batch-collapsed Rips.
- The set (cluster):

$$B_v^k = \{p \in V_0 \mid \pi_{k-1} \circ \cdots \circ \pi_0(p) = v\}$$

SimBa

- Vertex map is the same as that of batch-collapsed Rips.
- The set (cluster):

$$B_v^k = \{p \in V_0 \mid \pi_{k-1} \circ \cdots \circ \pi_0(p) = v\}$$

- The sequence:

$$\mathcal{B}^0(V_0) \to \mathcal{B}^{\alpha c}(V_1) \to \cdots \mathcal{B}^{\alpha c^m}(V_m)$$

  where $\mathcal{B}^{\alpha c^k}(V_k)$ is the clique complex induces by edges $\{(u,v) \in V_k \mid d(B_u^k, B_v^k) \leq \alpha c^k\}$ and α is chosen to be the minimum pairwise distance of input P.

SimBa

- Vertex map is the same as that of batch-collapsed Rips.
- The set (cluster):

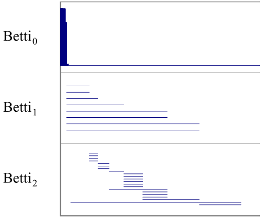
$$B_v^k = \{p \in V_0 \mid \pi_{k-1} \circ \cdots \circ \pi_0(p) = v\}$$

- The sequence:

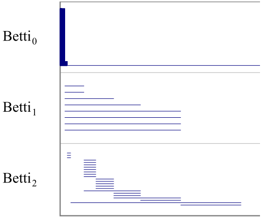
$$\mathcal{B}^0(V_0) \to \mathcal{B}^{\alpha c}(V_1) \to \cdots \mathcal{B}^{\alpha c^m}(V_m)$$

  where $\mathcal{B}^{\alpha c^k}(V_k)$ is the clique complex induces by edges $\{(u, v) \in V_k \mid d(B_u^k, B_v^k) \le \alpha c^k\}$ and α is chosen to be the minimum pairwise distance of input P.
- Approximation of PD: $3\log(\frac{2}{c-1} + 3)$-approximates that of Rips filtration.

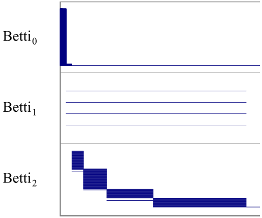
SimBa v.s. batch-collapsed Rips



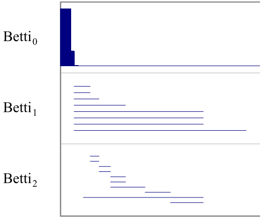
B.R. (c = 1.3)



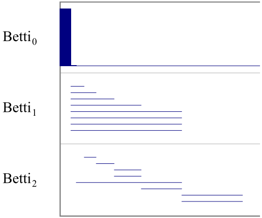
B.R. (c = 1.5)



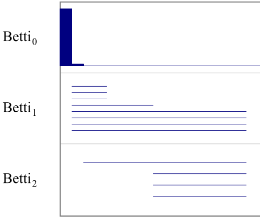
B.R. (c = 2.0)



SimBa (c = 1.3)

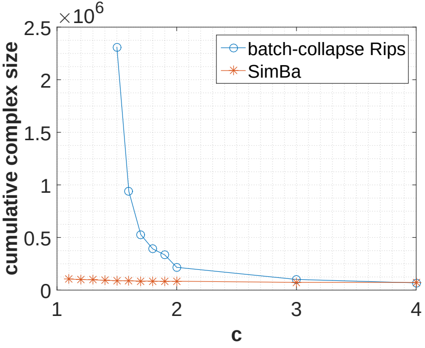


SimBa (c = 1.5)

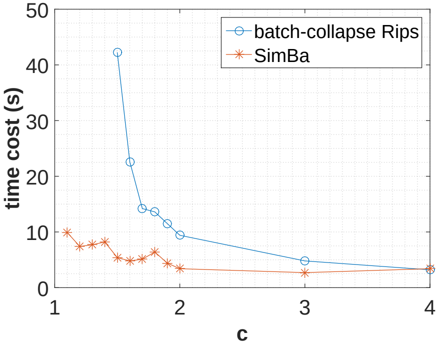


SimBa (c = 2.0)

SimBa v.s. batch-collapsed Rips

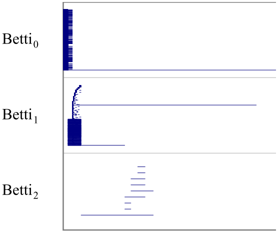


cumulative complex size

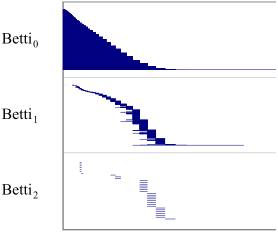


time cost

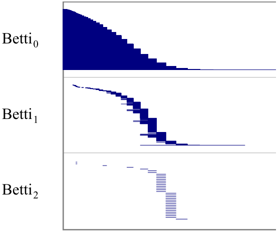
High dimensional data with ground truth



Klein Bottle in $\mathbb{R}^4$

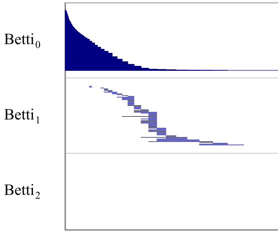


Primary Circle in $\mathbb{R}^{25}$

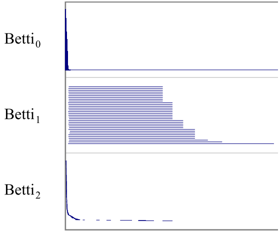


Primary Circle in $\mathbb{R}^{49}$

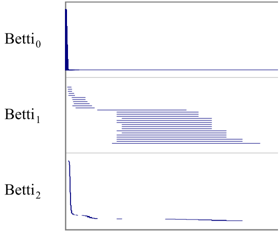
High dimensional data without ground truth



Gesture Phase data in $\mathbb{R}^{18}$



Survivin data in $\mathbb{R}^3$



Survivin data in $\mathbb{R}^{150}$

Performance results

				S.R.+GUDHI		S.R.+Simpers		B.R.+Simpers		SimBa	
Data	n	D	d	size	time(s)	size	time(s)	size	time(s)	size	time(s)
Mother	23075	3	2	$43.5 \cdot 10^6$	350	$43.5 \cdot 10^6$	463.7	$2.3 \cdot 10^6$	42.3	104701	8.8
KlBt	22500	4	2	$20.9 \cdot 10^6$	205.3	$20.9 \cdot 10^6$	303.5	440049	8	78064	6.6
PrCi25	15000	25	?	∞	–	∞	–	–	∞	$4.8 \cdot 10^6$	216
PrCi49	15000	49	?	∞	–	∞	–	–	∞	$10.2 \cdot 10^6$	585
GePh	1747	18	?	$45.6 \cdot 10^6$	282.5	$45.6 \cdot 10^6$	432.8	$1.4 \cdot 10^6$	29	7145	0.83
Sur3	252996	3	?	∞	–	∞	–	$15.7 \cdot 10^6$	1056.4	915110	1079.6
Sur150	252996	150	?	∞	–	∞	–	–	∞	$3.1 \cdot 10^6$	5089.7

- SimBa paper by T.K. Dey, D. Shi, Y. Wang. To appear in ESA 2016.
- SimPers and SimBa software: tamaldey/SimPers/SimPers.html

Thank you !

Questions ?