Persistent Path Homology of Directed Networks

Samir Chowdhury[1] and Facundo Mémoli

January 8, 2018

The Ohio State University
research.math.osu.edu/networks
https://github.com/samirchowdhury/pph-matlab

[1] *presenting author*

Introduction

Introduction

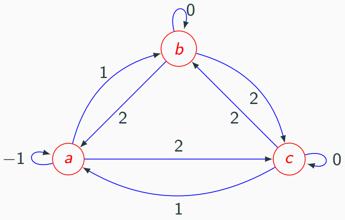
The advent of sophisticated data mining tools has led to rapid growth of **network datasets** in the sciences.

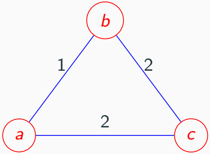
The availability of such network data coincides with a time of steady growth of the mathematical theory of **persistent homology (PH)**, which aims to study the "shape" of data.

Persistent homology has achieved significant success in the past two decades, and is increasingly finding applications in biology, medicine, materials science, and so on.

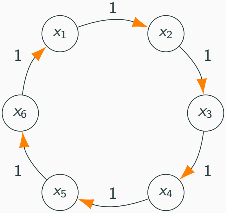
Our goal is to develop PH as a tool for analyzing **directed** network structure.

Examples of networks





(not our setup; common in literature)

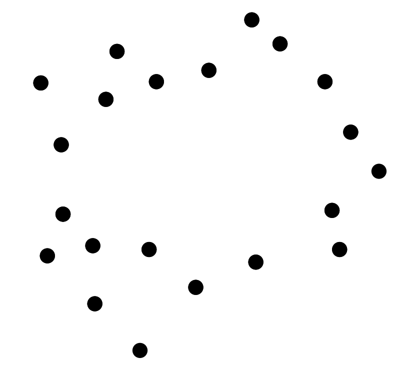


0	1	2	3	4	5
5	0	1	2	3	4
4	5	0	1	2	3
3	4	5	0	1	2
2	3	4	5	0	1
1	2	3	4	5	0

The extant "network-PH" approaches can be categorized by the input data types they accept.

Networks as finite metric/symmetric spaces

- Typically: build a filtration of weighted, undirected graphs, compute clique complexes, apply standard PH.
- Petri, Scolamiero, Donato, & Vaccarino. (2013). *Topological strata of weighted complex networks*. PloS one.
- Sizemore, Giusti, & Bassett. (2016). *Classification of weighted networks through mesoscale homological features*. Journal of Complex Networks.



Gif courtesy of Henry Adams

Networks that are (possibly) purely directed

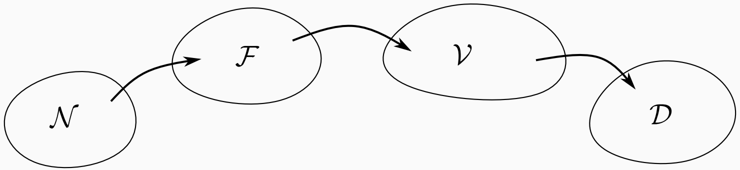
- Chowdhury & Mémoli. (2016). *Persistent Homology of Asymmetric Networks: An Approach based on Dowker Filtrations*. arXiv preprint. https://arxiv.org/abs/1608.05432
  - Setup: network is (X, ω_X), where X is a finite set and $\omega_X : X \times X \to \mathbb{R}$ is any (weight) function.
  - Developed analogues of Rips and Čech (**Dowker**) complexes.

Networks that are (possibly) purely directed

- Chowdhury & Mémoli. (2016). *Persistent Homology of Asymmetric Networks: An Approach based on Dowker Filtrations*. arXiv preprint. https://arxiv.org/abs/1608.05432
  - Setup: network is (X, ω_X), where X is a finite set and $\omega_X : X \times X \to \mathbb{R}$ is any (weight) function.
  - Developed analogues of Rips and Čech (**Dowker**) complexes.
- Dlotko, Hess, Levi, Turner, et al. (2016). *Topological analysis of the connectome of digital reconstructions of neural microcircuits*. arXiv preprint.
- Masulli & Villa. (2016). *The topology of the directed clique complex as a network invariant*. SpringerPlus.
- Turner. (2016). *Generalizations of the Rips filtration for quasi-metric spaces with persistent homology stability results*. arXiv preprint.

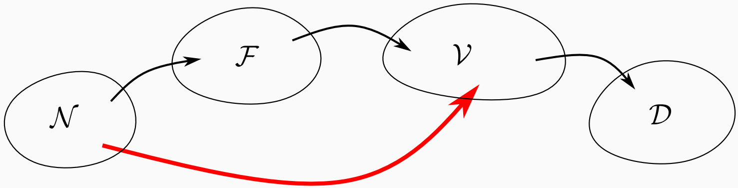
  The latter three papers use the directed clique/directed flag/ordered tuple complex construction.

Problem discussion



- $\mathcal{N}$: collection of finite networks
- $\mathcal{F}$: filtered simplicial complexes
- $\mathcal{V}$: vector spaces with linear maps
- $\mathcal{D}$: *persistence diagrams*—topological summaries of input data

Problem discussion



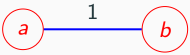
- $\mathcal{N}$: collection of finite networks
- $\mathcal{F}$: filtered simplicial complexes
- $\mathcal{V}$: vector spaces with linear maps
- $\mathcal{B}$: *persistence diagrams*—topological summaries of input data

The red arrow marks our proposed pathway.

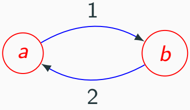
Idea: bypass simplicial setting to avoid information loss!

Problem discussion

Given this two-point metric space, the standard linearization process used in PH maps $\{a, b\}$ to the vector $[a, b] = -[b, a]$ "at time $t = 1$".



However, in the asymmetric network setting, this (simplicial) linearization may be unsatisfactory. The equality $[a, b] = -[b, a]$ at the vector space level ignores the unequal weights on the arrows $a \to b$ and $b \to a$.



Path homology solves this problem by assigning $[a, b]$ and $[b, a]$ to **linearly independent** components at the vector space level.

- We develop the framework of **persistent path homology (PPH)**—a version of persistent homology that uses a recently-developed (2014) theory of the homology of digraphs.
- We provide an $O(n^3)$ algorithm and a Matlab implementation for computing PPH.
- We prove theoretical guarantees for PPH via stability and characterization results.

Path homology of digraphs

Path homology

We invoke a notion of digraph homology—called **path homology**—to bypass the simplicial setting.

Path homology

We invoke a notion of digraph homology—called **path homology**—to bypass the simplicial setting.

- Developed by A. Grigor'yan, S.-T. Yau, Y. Lin, and Y. Muranov between 2014-2016
- Can be nontrivial in all dimensions
- Compatible with homotopy theories of graphs and digraphs
- Good functorial properties w.r.t. graph-theoretical operations; path homology of a Cartesian product of digraphs satisfies the Künneth formula

Construction of path homology

Let $\mathbb{K}$ be a field that we fix throughout this talk.

Given a finite set X and any integer $p \in Z_+$, an ***elementary* p-path over X** is a sequence $[x_0, \ldots, x_p]$ of $p+1$ elements of X.

One then defines a boundary map on elementary p-paths as follows:

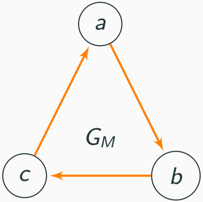
$$\partial_p([x_0, \ldots, x_p]) := \sum_{i=0}^{p} (-1)^i [x_0, \ldots, \widehat{x_i}, \ldots, x_p].$$

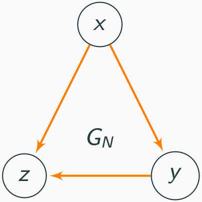
E.g. $\partial_2(abc) = [bc] - [ac] + [ab]$.

$\partial_1(ab) = [b] - [a]$.

Let $G = (X, E)$ be a digraph. For each $p \in \mathbb{Z}_+$, one defines an elementary p-path $[x_0, \dots, x_p]$ on X to be **allowed** if $(x_i, x_{i+1}) \in E$ for each $0 \leq i \leq p-1$.

For each $p \in \mathbb{Z}_+$, the free vector space on the collection of allowed p-paths on (X, E) is denoted $\mathcal{A}_p = \mathcal{A}_p(G) = \mathcal{A}_p(X, E, \mathbb{K})$, and is called the **space of allowed p-paths**.





$\mathcal{A}_0(G_M) = \mathbb{K}[\{a, b, c\}]$
$\mathcal{A}_1(G_M) = \mathbb{K}[\{ab, bc, ca\}]$
$\mathcal{A}_2(G_M) = \mathbb{K}[\{abc, bca, cab\}]$
$\mathcal{A}_3(G_M) = \mathbb{K}[\{abca, bcab, cabc\}]$

$\mathcal{A}_0(G_N) = \mathbb{K}[\{x, y, z\}]$
$\mathcal{A}_1(G_N) = \mathbb{K}[\{xy, yz, xz\}]$
$\mathcal{A}_2(G_N) = \mathbb{K}[\{xyz\}]$
$\mathcal{A}_3(G_N) = \{0\}$

The allowed paths do not form a chain complex, because the image of an allowed path under ∂ need not be allowed. This is rectified as follows. Given a digraph $G = (X, E)$ and any $p \in \mathbb{Z}_+$, the **space of ∂-invariant p-paths on G** is defined to be the following subspace of $\mathcal{A}_p(G)$:

$$\Omega_p = \Omega_p(G) = \Omega_p(X, E, \mathbb{K}) := \{c \in \mathcal{A}_p : \partial_p(c) \in \mathcal{A}_{p-1}\}.$$

Now it follows by the definitions that $\partial_p(\Omega_p) \subseteq \Omega_{p-1}$ for any integer $p \geq -1$. Thus we have a chain complex:

$$\ldots \xrightarrow{\partial_3} \Omega_2 \xrightarrow{\partial_2} \Omega_1 \xrightarrow{\partial_1} \Omega_0 \xrightarrow{\partial_0} \mathbb{K} \xrightarrow{\partial_{-1}} 0$$

For each $p \in \mathbb{Z}_+$, the **p-dimensional path homology groups of $G = (X, E)$** are defined as:

$$H_p^{\Xi}(G) = H_p^{\Xi}(X, E, \mathbb{K}) := \ker(\partial_p) / \operatorname{im}(\partial_{p+1}).$$

An example (with biological motivations) i

$$\begin{tikzcd} a \arrow[r] \arrow[d] \arrow[dr, phantom, "G_M"] & b \\ d & c \arrow[l] \arrow[u] \end{tikzcd}$$

$$\begin{tikzcd} w \arrow[r] \arrow[d] \arrow[dr, phantom, "G_N"] & x \arrow[d] \\ z \arrow[r] & y \end{tikzcd}$$

$$\Omega_0(G_M) = \mathbb{K}[\{a, b, c, d\}]$$
$$\Omega_1(G_M) = \mathbb{K}[\{ab, cb, cd, ad\}]$$
$$\Omega_2(G_M) = \{0\}$$

$$\Omega_0(G_N) = \mathbb{K}[\{w, x, y, z\}]$$
$$\Omega_1(G_N) = \mathbb{K}[\{wx, xy, zy, wz\}]$$
$$\Omega_2(G_N) = \mathbb{K}[\{wxy - wzy\}]$$

Note $\partial_2(wxy) = xy - wy + wx$,
$\partial_2(wzy) = zy - wy + wz$

An example (with biological motivations) ii

The crux of the $\Omega_\bullet$ construction lies in understanding $\Omega_2(G_N)$. Note that even though $\partial_2^{G_N}(wxy)$, $\partial_2^{G_N}(wzy) \notin \mathcal{A}_2(G_N)$ (because $wy \notin \mathcal{A}_1(G_N)$), we still have:

$$\partial_2^{G_N}(wxy - wzy) = xy - wy + wx - zy + wy - wz \in \mathcal{A}_1(G_N).$$

One can then verify that

$$\ker(\partial_1^{G_M}) = \mathbb{K}[\{ab - cb + cd - ad\}] \neq \{0\} = \operatorname{im}(\partial_2^{G_M}),$$

$$\ker(\partial_1^{G_N}) = \mathbb{K}[\{wx + xy - zy - wz\}] = \operatorname{im}(\partial_2^{G_N}).$$

Thus $\dim(H_1^{\Xi}(G_M)) = \mathbf{1}$, and $\dim(H_1^{\Xi}(G_N)) = \mathbf{0}$.

An example (with biological motivations) iii

- In systems biology, G_M, G_N are referred to as the *bi-fan* and *bi-parallel motifs*, respectively. Distinguishing between these two motifs is an important task in that domain.
- The directed clique complex homology referenced earlier *cannot* distinguish between these two motifs.
- **The challenge of finding a natural basis for $\Omega_\bullet$.** G_N is a minimal example showing that it is nontrivial to compute bases for the vector spaces $\Omega_\bullet$. Specifically, while it is trivial to read off bases for the allowed paths $\mathcal{A}_\bullet$ from a digraph, one needs to consider *linear combinations* of allowed paths in a systematic manner to obtain bases for the ∂-invariant paths.**This raises a red flag for computations!**

Persistent Path Homology (PPH)

In the rest of the talk, we highlight our main contributions:

- a **persistent** framework for path homology which develops path homology signatures at different edge weight resolutions
- a quantitative stability result showing that this method is robust to small changes in input data
- a polynomial time algorithm for computing PPH
- characterization results for developing intuition about PPH

Persistent Path Homology

Let (X, ω_X) be a network. Fix $p \in \mathbb{Z}_+$. For any $\delta \in \mathbb{R}$, the digraph $G_X^\delta = (X, E_X^\delta)$ is defined as follows:

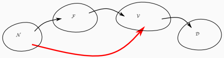
$$E_X^\delta := \left\{(x, x') \in X \times X : x \neq x', \omega_X(x, x') \leq \delta\right\}.$$

Note that for any $\delta' \geq \delta \in \mathbb{R}$, we have a natural inclusion map $G_X^\delta \hookrightarrow G_X^{\delta'}$. Thus we may associate to (X, ω_X) the **digraph filtration** $\{G_X^\delta \hookrightarrow G_X^{\delta'}\}_{\delta \leq \delta' \in \mathbb{R}_+}$.

Applying path homology then yields a **persistent vector space**

$$\mathcal{H}_p^{\Xi}(X, \omega_X) := \{H_p^{\Xi}(G_X^\delta) \xrightarrow{(\iota_{\delta,\delta'})_\#} H_p^{\Xi}(G_X^{\delta'})\}_{\delta \leq \delta' \in \mathbb{R}_+}.$$

We denote the associated **persistence diagram** by $\mathrm{Dgm}_p^{\Xi}(X)$.



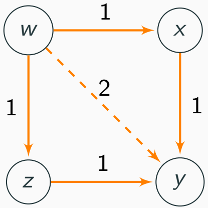
Algorithm and implementation

Red flag raised earlier: there is no natural choice of basis for $\Omega_\bullet$.

This is an obstruction to using any of the standard [2,3] PH algorithms.

[2] Edelsbrunner, Letscher, & Zomorodian. (2000). *Topological persistence and simplification*. FoCS.
[3] Zomorodian, & Carlsson. (2005). *Computing persistent homology*. DCG.

Algorithm and implementation



G_N

self weights $= 0$

other weights $= 2$

wx, xy, wz, zy, wxy, and wzy all have *allow time* ($\mathfrak{at}$) $= 1$.

wxy and wzy have *entry time* ($\mathfrak{et}$) $= 2$. This is the first time $wxy, wzy \in \Omega_2(G_N)$.

However, recall from before that $\mathfrak{et}(wxy - wzy) = 1$.

How do we arrive at this algorithmically?

Recall: $\Omega_p = \{c \in \mathcal{A}_p : \partial_p(c) \in \mathcal{A}_{p-1}\} \subseteq \mathcal{A}_p$.

Algorithm and implementation

		$\mathfrak{at}=1$		$\mathfrak{at}\geq 2$		
		wxy	wzy	.	.	.
$\mathfrak{at}=1$	wx	1		.	.	.
	xy	1		.	.	.
	wz		1	.	.	.
	zy		1	.	.	.
$\mathfrak{at}\geq 2$	wy	-1	-1	.	.	.
	.	.	.	.	.	.

rearrange rows $\longrightarrow$

		$\mathfrak{at}=1$		$\mathfrak{at}\geq 2$		
		wxy	wzy	.	.	.
	.	.	.	.	.	.
$\mathfrak{at}\geq 2$	wy	-1	-1	.	.	.
$\mathfrak{at}=1$	wx	1		.	.	.
	xy	1		.	.	.
	wz		1	.	.	.
	zy		1	.	.	.

		$\mathfrak{at}=1$		$\mathfrak{at}\geq 2$		
		wxy	wzy	.	.	.
	.	.	.	.	.	.
$\mathfrak{at}\geq 2$	wy	-1	-1	.	.	.
$\mathfrak{at}=1$	wx	1		.	.	.
	xy	1		.	.	.
	wz		1	.	.	.
	zy		1	.	.	.

left to right column reduction $\longrightarrow$

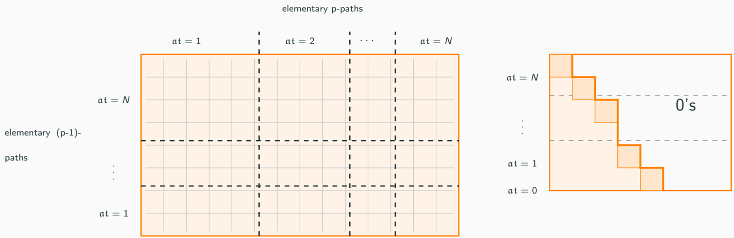
		$\mathfrak{at}=1$		$\mathfrak{at}\geq 2$		
		wxy	wzy - wxy	.	.	.
	.	.	.	.	.	.
$\mathfrak{at}\geq 2$	wy	-1	.	.	.	.
$\mathfrak{at}=1$	wx	1	-1	.	.	.
	xy	1	-1	.	.	.
	wz		1	.	.	.
	zy		1	.	.	.

Reminder: a p-path has $\mathfrak{at}=t$ if it enters $\mathcal{A}_p$ (i.e. becomes allowed) at time t.

A p-path has $\mathfrak{et}=t$ if it enters Ω_p (i.e. becomes ∂-invariant) at time t.

Theorem

Left-to-right column reduction is sufficient to obtain compatible bases for $\Omega_\bullet$. This is precisely the operation used in the classical PH algorithm, so the two steps (obtaining bases and computing PH) can be combined into one.



The running time for the algorithm is the same as that of Gaussian elimination, i.e. is cubic in the number of paths. The complete algorithm is available in the proceedings.

Matlab implementation:
https://github.com/samirchowdhury/pph-matlab

Stability

The classical result guaranteeing viability of PH in the setting of metric spaces states that the *bottleneck distance* between two Rips persistence diagrams arising from metric spaces is bounded above by the Gromov-Hausdorff distance between the metric spaces.

We obtain a similar result with respect to a particular **network distance**[4] $\mathrm{d}_{\mathcal{N}}$.

Theorem

Let $(X, \omega_X), (Y, \omega_Y)$ *be two networks. Then for any* $p \in \mathbb{Z}_+$,

$$\mathrm{d}_{\mathrm{B}}(\mathrm{Dgm}_p^{\Xi}(X), \mathrm{Dgm}_p^{\Xi}(Y)) \leq 2\, \mathrm{d}_{\mathcal{N}}(X, Y).$$

[4] Carlsson, Mémoli, Ribeiro, & Segarra. (2013). *Axiomatic construction of hierarchical clustering in asymmetric networks*. ICASSP.

The proof of stability is via a now-standard tool called the Algebraic Stability Theorem[5], but there is an interesting deviation from the arguments in simplicial settings.

[5] Chazal, Cohen-Steiner, Glisse, Guibas, & Oudot. (2009). *Proximity of persistence modules and their diagrams*. SoCG.

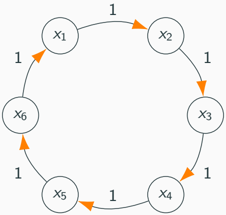
The proof of stability is via a now-standard tool called the Algebraic Stability Theorem[5], but there is an interesting deviation from the arguments in simplicial settings.

It becomes necessary to invoke results from the **homotopy theory of digraphs**, which was also developed recently by Grigor'yan et. al.

[5] Chazal, Cohen-Steiner, Glisse, Guibas, & Oudot. (2009). *Proximity of persistence modules and their diagrams*. SoCG.

Characterization results

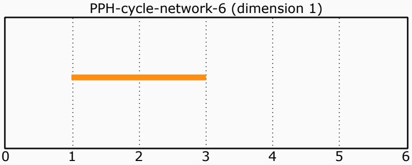
Cycle networks



0	1	2	3	4	5
5	0	1	2	3	4
4	5	0	1	2	3
3	4	5	0	1	2
2	3	4	5	0	1
1	2	3	4	5	0

Theorem

Let G_n be a cycle network for some integer $n \geq 3$. Fix a field $\mathbb{K} = \mathbb{Z}/p\mathbb{Z}$ for some prime p. Then $\mathrm{Dgm}_1^{\Xi}(G_n) = \{(1, \lceil n/2 \rceil)\}$.

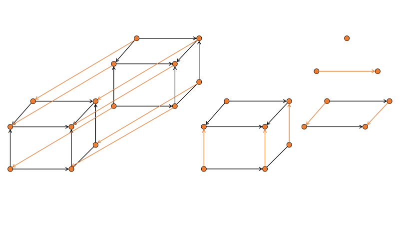


Future work

- Code for a `Matlab` implementation is currently available: https://github.com/samirchowdhury/pph-matlab
- We are presently working on a C++ implementation
- Many directions in which to proceed from here, with regards to improved algorithms, applications, and interpretations.

Acknowledgments

This work was supported by NSF grants IIS-1422400 and CCF-1526513. Facundo Mémoli was supported by the Mathematical Biosciences Institute at The Ohio State University. Samir Chowdhury was supported by a SIAM travel award funded by an NSF grant to Columbia University.



Thank you!