Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Persistent homology of unweighted complex networks via discrete Morse theory

by

Harish Kannan, Emil Saucan, Indrava Roy, and Areejit Samal

Persistent homology of unweighted complex networks via discrete Morse theory

Harish Kannan,[1] Emil Saucan,[2,3] Indrava Roy,[1,*] and Areejit Samal[1,4,†]

[1]*The Institute of Mathematical Sciences (IMSc),*
Homi Bhabha National Institute (HBNI), Chennai 600113 India
[2]*Department of Applied Mathematics, ORT Braude College, Karmiel 2161002 Israel*
[3]*Department of Electrical Engineering, Technion,*
Israel Institute of Technology, Haifa 3200003 Israel
[4]*Max Planck Institute for Mathematics in the Sciences, Leipzig 04103 Germany*

Topological data analysis can reveal higher-order structure beyond pairwise connections between vertices in complex networks. We present a new method based on discrete Morse theory to study topological properties of unweighted and undirected networks using persistent homology. Leveraging on the features of discrete Morse theory, our method not only captures the topology of the clique complex of such graphs via the concept of critical simplices, but also achieves close to the theoretical minimum number of critical simplices in several analyzed model and real networks. This leads to a reduced filtration scheme based on the subsequence of the corresponding critical weights, thereby leading to a significant increase in computational efficiency. We have employed our filtration scheme to explore the persistent homology of several model and real-world networks. In particular, we show that our method can detect differences in the higher-order structure of networks, and the corresponding persistence diagrams can be used to distinguish between different model networks. In summary, our method based on discrete Morse theory further increases the applicability of persistent homology to investigate the global topology of complex networks.

INTRODUCTION

In recent years, the field of topological data analysis (TDA) has rapidly grown to provide a set of powerful tools to analyze various important features of data [1]. In this context, persistent homology has played a key role in bringing TDA to the fore of modern data analysis. It not only gives a way to visualize data efficiently, but also to extract relevant information from both structured and unstructured datasets. This crucial aspect has been used effectively in various applications from astrophysics (e.g., determination of inter-galactic filament structures [2]) to imaging analysis (e.g., feature detection in 3D gray-scale images [3]) to biology (e.g., detection of breast cancer type with high survival rates [4]). Informally, the essence of the theory is its power to extract the *shape of data*, as well as infer higher-order correlations between various parts of the data at hand which are missed by other classical techniques [1]. The basic mathematical theory used in this subject is that of algebraic topology, and in particular the study of homology, developed by the French mathematician Henri Poincaré at the turn of the 20th century. The origins of persistent homology lie in the ideas of Morse theory [5], which gives a powerful tool to detect the topological features of a given space through the computation of homology using real-valued functions on the space. We refer the reader to the survey article [6] for further details.

On the other hand, the discretized version of Morse theory developed by Robin Forman [7–9], gives a way to characterize the homology group of a simplicial complex in terms a real-valued function with certain properties, known as a discrete Morse function. Examples of such simplicial complexes associated with discrete spaces are the Vietoris-Rips complex corresponding to a discrete metric space, or the clique complex of a graph. Forman [8, 9] showed that given such a function, the so-called critical simplices completely determine the Euler characteristic of the space, which is a fundamental topological invariant.

The study of complex networks in the last few decades has also significantly raised our ability to understand various kinds of interactions arising in both natural and artificial realms [10–13]. Understanding how different parts of networks behave and influence each other is therefore an important problem [10–13]. However, for large networks, detecting higher-order structures remains a difficult task [14]. Moreover, recent studies [15–17] indicate that these higher-order correlations are not captured by usual network measures such as clustering coefficients. While a graph representation captures binary relationships among vertices of a network, simplicial complexes also reflect higher-order relationships in a complex network [15, 16, 18–25]. In this context, persistent homology has been employed to explore the topological properties of complex networks [18–21, 23, 26]. In this work, we present a systematic method to study the persistent homology of unweighted and undirected graphs or networks.

---

[*] Correspondence to: indrava@imsc.res.in
[†] Correspondence to: asamal@imsc.res.in

Previous work has investigated the persistent homology of weighted and undirected networks by creating a filtration of the clique complexes corresponding to threshold graphs obtained via decreasing sequence of edge weights [20, 23]. However, the lack of edge weights in unweighted networks does not permit a filtration based on threshold graphs [20, 23]. Thus for unweighted networks, Horak *et al.* [19] propose a filtration scheme based on the dimension of the simplices in the clique complex corresponding to the unweighted network. Horak *et al.* [19] do not assign weights to vertices, edges or higher-dimensional simplices in the clique complex corresponding to an unweighted graph. An unexplored filtration scheme involves transforming an unweighted network into a weighted network by assigning edge weights based on some network property, such as edge betweenness centrality [27, 28] or discrete edge curvature [29, 30], and then employing the filtration scheme based on threshold graphs [20, 23]. As an alternative, we here use discrete Morse theory [7–9] to create a filtration scheme for unweighted networks by assigning weights to vertices, edges, triangles and higher-dimensional simplices in the clique complex of the graph. In our method, the weight of a simplex is chosen such that it reflects the degree of the vertices which constitute the simplex while simultaneously satisfying the conditions for the weighing function to be a discrete Morse function. Moreover, as explained in the Results section, an equally important intuition behind the choice of these weights is based on the goal of reducing the number of so-called critical simplices.

In the context of TDA, classical Morse theory, which involves smooth functions defined on topological spaces that admit a smooth structure has been used to compute persistent homology, e.g. in statistical topology [31], astrophysics [2]. Since the clique complex of a weighted or unweighted graph does not permit a smooth structure in general, applying classical Morse theory is not possible. However, discrete Morse theory [7–9] provides an efficient way of computing persistent homology. A discrete Morse function not only captures higher-order topological information of the underlying space, a "preprocessing" with respect to a suitable discrete Morse function leads to significant simplification of the topological structure. This makes computation of persistent homology groups or homology groups of filtered complexes much more efficient, see e.g. [32]. This is especially useful for large datasets where computationally efficient methods are key to compute their persistent homology.

We have applied discrete Morse theory to compute persistent homology of unweighted simple graphs. This is done by using the values given by the discrete Morse function to pass from an unweighted graph to a weighted simplicial complex (Figure 1). This transformation automatically produces a filtration that is needed for the computation of persistent homology, through the so-called level subcomplexes associated with weights of critical simplices (See Theory section and Figure 2). Moreover, this filtration is consistent with the topology of the underlying space and reveals finer topological features than the dimensional filtration scheme used in Horak *et al.*[19] The combination of these techniques have been used [3, 33] with applications for image processing. However, to the best of our knowledge, this method has not been used for studying persistent homology in unweighted complex networks to date. Discrete Morse theory gives a theoretical lower bound on the number of critical simplices which can be attained by an optimal choice of the function on a simplicial complex. Interestingly, our method achieves close to the theoretical minimum number of critical simplices in several model and real networks analyzed here (See Results section). Furthermore, our algorithm for computing the discrete Morse function is easy to implement for complex networks.

Our results underline the potence of persistent homology to detect inherent topological features of networks which are not directly captured by homology alone. For instance, the p-Betti numbers of the clique complexes corresponding to small-world [10] and scale-free [11] networks with similar size and average degree, respectively, are of comparable magnitude and thus, homology reveals no deep insight into the differences between the topological features of these two model networks. On the other hand, our observations on the persistent homology of these two networks indicate a clear demarkation with respect to the evolution of topological features in the clique complexes corresponding to these model networks during the filtration process. This dissimilarity in the evolution of topological characteristics that resonates across dimensions and the average degree of the underlying network, indicates an inherent disparity in the persistent homology of small-world and scale-free graphs. In addition to unravelling higher-order relationships in networks, this ability to capture inherent topological differences between two dissimilar networks thus motivates the application of our methods to study the persistent homology of real-world networks.

The remainder of the paper is organized as follows. We begin with a **Theory** section which gives a brief overview of concepts in persistent homology and discrete Morse theory. We then proceed to describe the model networks and real-world networks that have been studied in this work in the **Network datasets** section. In the subsequent section on **Results and Discussion**, we present our algorithm to construct a discrete Morse function on a simplical complex associated with a network. In the same section, we present our results for model networks and real-world networks. The final section on **Conclusions** gives a summary and outlook of our findings. In Supplementary Information (SI) **Appendix**, we give a brief review of the mathematical theory of homology groups of a simplicial complex. In SI Appendix, we also provide a rigorous proof of concept for our algorithm to construct a discrete Morse function. We then follow up with two algorithms both of which illustrate key procedures that are essential to construct the filtration of a simplicial complex associated with the investigated networks. In SI Appendix, time complexity for our algorithms as well as some theoretical results on stability of persistent homology and persistence diagrams are also given.

THEORY

Graphs and Simplicial Complexes

Consider a finite simple graph $G(\mathcal{V}, \mathcal{E})$ having vertex set $\mathcal{V} = \{v_0, v_1, \cdots, v_n\}$ and the edge set $\mathcal{E}$. Note that a simple graph does not contain self-loops or multi-edges [34]. Such a simple graph G can be viewed as a *clique* complex K [35]. A clique simplicial complex K is a collection of simplices where a p-dimensional simplex (or p-simplex) in K is a set of $p + 1$ vertices that form a complete subgraph. In other words, vertices correspond to 0-simplices, edges to 1-simplices, and triangles to 2-simplices in the clique complex of a graph. Note that the dimension p of simplices contained in K is restricted to the range 0 to $(|\mathcal{V}| - 1)$ in the graph G. The dimension d of the clique complex K is given by the maximum dimension of its constituent simplices. A *face* γ of a p-simplex α is a subset of α with cardinality less than $p + 1$. Note that by definition, a face γ of a p-simplex α is a l-simplex where $0 \leq l < p$ and this relationship is denoted as $\gamma^l < \alpha^p$. Formally, the clique complex K corresponding to the simple graph G satisfies the following condition which defines an abstract simplicial complex, namely, K is a collection of non-empty finite sets or simplices such that if α is an element (simplex) of K then so is every non-empty subset of α. For additional details, the interested reader is referred to standard text in algebraic topology [36].

Figure 1 displays an example of the correspondence between a simple graph and its clique complex. The ordering of the vertex set $\{v_0, v_1, v_2, \cdots, v_p\}$ of a p-simplex α determines its *orientation*. Moreover, two orderings of the vertex set of α are considered to be equivalent if and only if they differ by an even permutation. If the dimension of a p-simplex is greater than 1, then all possible orderings of its vertex set fall under two equivalence classes, with each class being assigned an *orientation* [36]. An exception is the 0-simplex with one vertex which has exactly one equivalence class and orientation. An oriented p-simplex α specifies the orientation of its $p + 1$ vertices and is represented by $[v_0, v_1, v_2, \cdots, v_p]$ [36]. In figure 1, the oriented 2-simplices $[v_2, v_3, v_4]$ and $[v_2, v_4, v_3]$ have opposite orientations, i.e., $[v_2, v_3, v_4] = -[v_2, v_4, v_3]$.

Persistent homology of a simplicial complex

In SI Appendix, we give a brief review of the mathematical theory of homology groups of a simplicial complex. In particular, we define p-chain group, p-boundary operator, p-boundary, p-cycle, p-hole, p-homology group and p-Betti number.

A subset K^i of a simplicial complex K is called a subcomplex of K if K^i by itself is an abstract simplicial complex. Then a filtration of a simplicial complex K is defined as a nested sequence of subcomplexes K^i of K where:

$$\emptyset \subseteq K^0 \subseteq K^1 \subseteq ... \subseteq K^q = K \tag{1}$$

Note that each subcomplex K^i has an associated index i in the filtration. Moreover, each subcomplex K^i in the filtration has corresponding p-chain complexes C_p^i, p-boundary operators ∂_p^i, p-boundaries B_p^i and p-cycles Z_p^i.

The j-persistent p-homology group of K^i denoted as $H_p^{i,j}$ is defined as:

$$H_p^{i,j} = Z_p^i / (B_p^{i+j} \cap Z_p^i). \tag{2}$$

In the above equation, B_p^{i+j} is the subgroup of C_p^{i+j} which constitutes the p-boundaries of the subcomplex K^{i+j}. The j-persistent p-Betti number of K^i denoted as $\beta_p^{i,j}$ is defined as:

$$\beta_p^{i,j} = \dim(H_p^{i,j}). \tag{3}$$

An intuitive explanation of the above definitions of the j-persistent p-homology group and the corresponding Betti number is as follows. A p-hole of the subcomplex K^i can potentially become the boundary of a $(p+1)$-chain of a later subcomplex K^{i+j} with $j > 0$, and thus, no longer constitute a p-hole of K^{i+j}. The j-persistent p-Betti number of K^i represents the number of p-holes at the filtration index i that persist at the filtration index $j + i$. Therefore, each p-hole that appears across the filtration has a unique index that corresponds to its *birth* and *death*, and the persistence of such a p-hole can thus be characterized by its corresponding birth and death indices. Studying persistent homology allows us to quantify the longevity of such p-holes during filtration, and thus, measures the importance of these topological features which appear and disappear across the filtration.
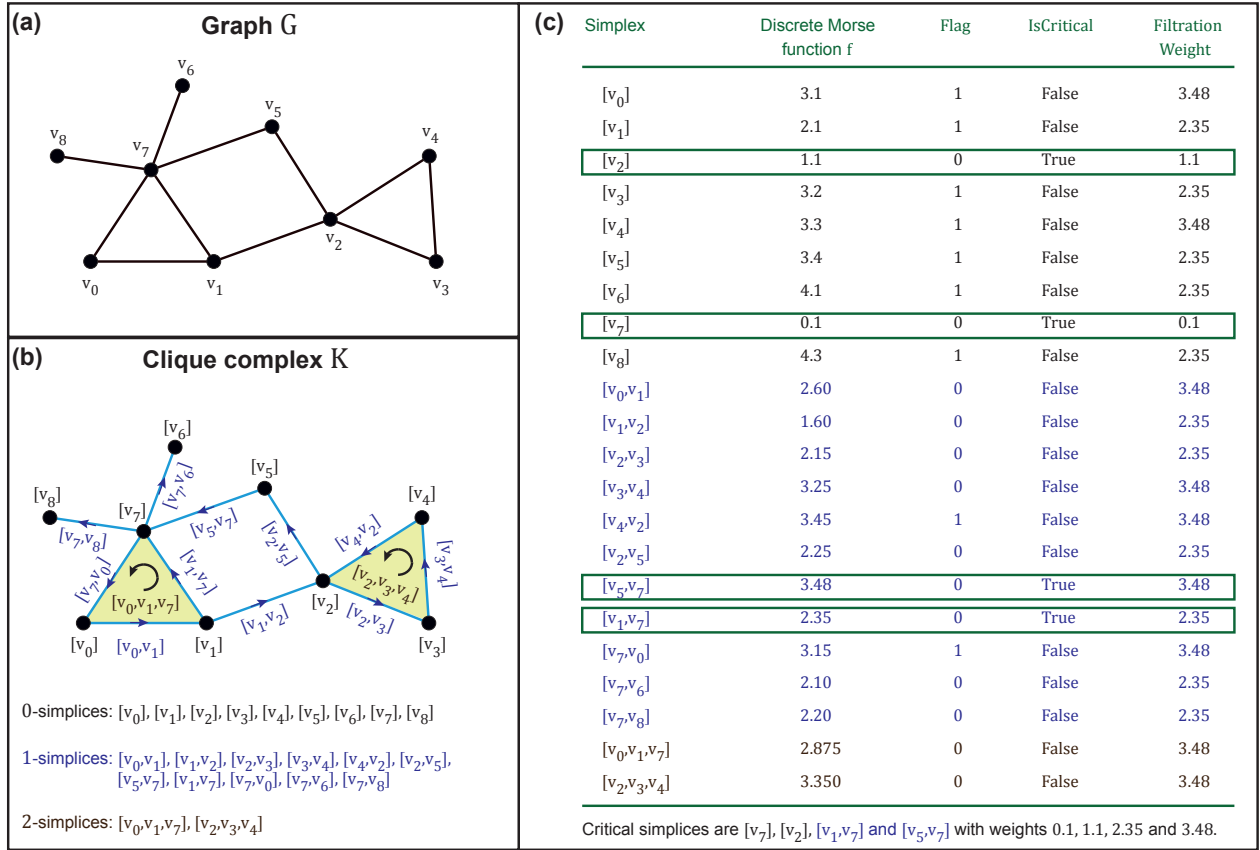
(a) Graph G

(b) Clique complex K

0-simplices: $[v_0]$, $[v_1]$, $[v_2]$, $[v_3]$, $[v_4]$, $[v_5]$, $[v_6]$, $[v_7]$, $[v_8]$

1-simplices: $[v_0,v_1]$, $[v_1,v_2]$, $[v_2,v_3]$, $[v_3,v_4]$, $[v_4,v_2]$, $[v_2,v_5]$, $[v_5,v_7]$, $[v_1,v_7]$, $[v_7,v_0]$, $[v_7,v_6]$, $[v_7,v_8]$

2-simplices: $[v_0,v_1,v_7]$, $[v_2,v_3,v_4]$

(c)

Simplex	Discrete Morse function f	Flag	IsCritical	Filtration Weight
$[v_0]$	3.1	1	False	3.48
$[v_1]$	2.1	1	False	2.35
$[v_2]$	1.1	0	True	1.1
$[v_3]$	3.2	1	False	2.35
$[v_4]$	3.3	1	False	3.48
$[v_5]$	3.4	1	False	2.35
$[v_6]$	4.1	1	False	2.35
$[v_7]$	0.1	0	True	0.1
$[v_8]$	4.3	1	False	2.35
$[v_0,v_1]$	2.60	0	False	3.48
$[v_1,v_2]$	1.60	0	False	2.35
$[v_2,v_3]$	2.15	0	False	2.35
$[v_3,v_4]$	3.25	0	False	3.48
$[v_4,v_2]$	3.45	1	False	3.48
$[v_2,v_5]$	2.25	0	False	2.35
$[v_5,v_7]$	3.48	0	True	3.48
$[v_1,v_7]$	2.35	0	True	2.35
$[v_7,v_0]$	3.15	1	False	3.48
$[v_7,v_6]$	2.10	0	False	2.35
$[v_7,v_8]$	2.20	0	False	2.35
$[v_0,v_1,v_7]$	2.875	0	False	3.48
$[v_2,v_3,v_4]$	3.350	0	False	3.48

Critical simplices are $[v_7]$, $[v_2]$, $[v_1,v_7]$ and $[v_5,v_7]$ with weights 0.1, 1.1, 2.35 and 3.48.

FIG. 1. An illustration of the construction of a discrete Morse function f on a clique complex K corresponding to an unweighted and undirected graph G using our algorithm. (a) A simple example of an unweighted and undirected graph G containing 9 vertices and 11 edges. (b) The clique simplicial complex K corresponding to the simple graph G shown in (a). The clique complex K consists of 9 vertices or 0-simplices, 11 edges or 1-simplices and 2 triangles or 2-simplices. The figure also displays the orientation of the 1- and 2-simplices using arrows. (c) Generation of a discrete Morse function f on the clique complex K shown in (b) using our algorithm. The figure lists the state of the `Flag` variable in algorithm 1 and `IsCritical` variable in algorithm 2 (See SI Appendix) for each simplex in K. In this example, the clique complex has 4 critical simplices and their respective critical weights correspond to the filtration steps. The figure also lists the `FiltrationWeight` for each simplex in K obtained using algorithm 3 (See SI Appendix).

Discrete Morse Theory

Recalling from the preceeding section, to study the persistent homology corresponding to the clique complex K of a simple graph G, the primary requirement is a filtration of K. We here present a systematic method to study the persistent homology of unweighted and undirected networks by utilizing a refined filtration of the clique complex K based on discrete Morse theory [7–9]. It is important to note that the order in which the simplices of the clique complex K are added during the filtration affects the evolution of topological features which are observed by studying the persistent homology. Our proposed scheme which is based on discrete Morse theory developed by Forman [7–9] tackles this by assigning weights to 0-simplices (vertices), 1-simplices (edges), 2-simplices (triangles) and higher-dimensional simplices appearing in the clique complex corresponding to an unweighted and undirected network. Assigning weights to higher-dimensional simplices captures important higher-order correlations in addition to edges or 1-simplices. Moreover, we leverage the following important features of the framework of discrete Morse theory in our new scheme. Firstly, the framework enables assignment of weights to p-simplices which are concordant with weights of $(p-1)$-simplices. Secondly, it captures the topology of a simplicial complex via the concept of *critical* simplices described below. Most importantly, the framework provides a natural way to create a filtration scheme to study persistent homology based upon the weights of the aforementioned *critical* simplices as will be described below.

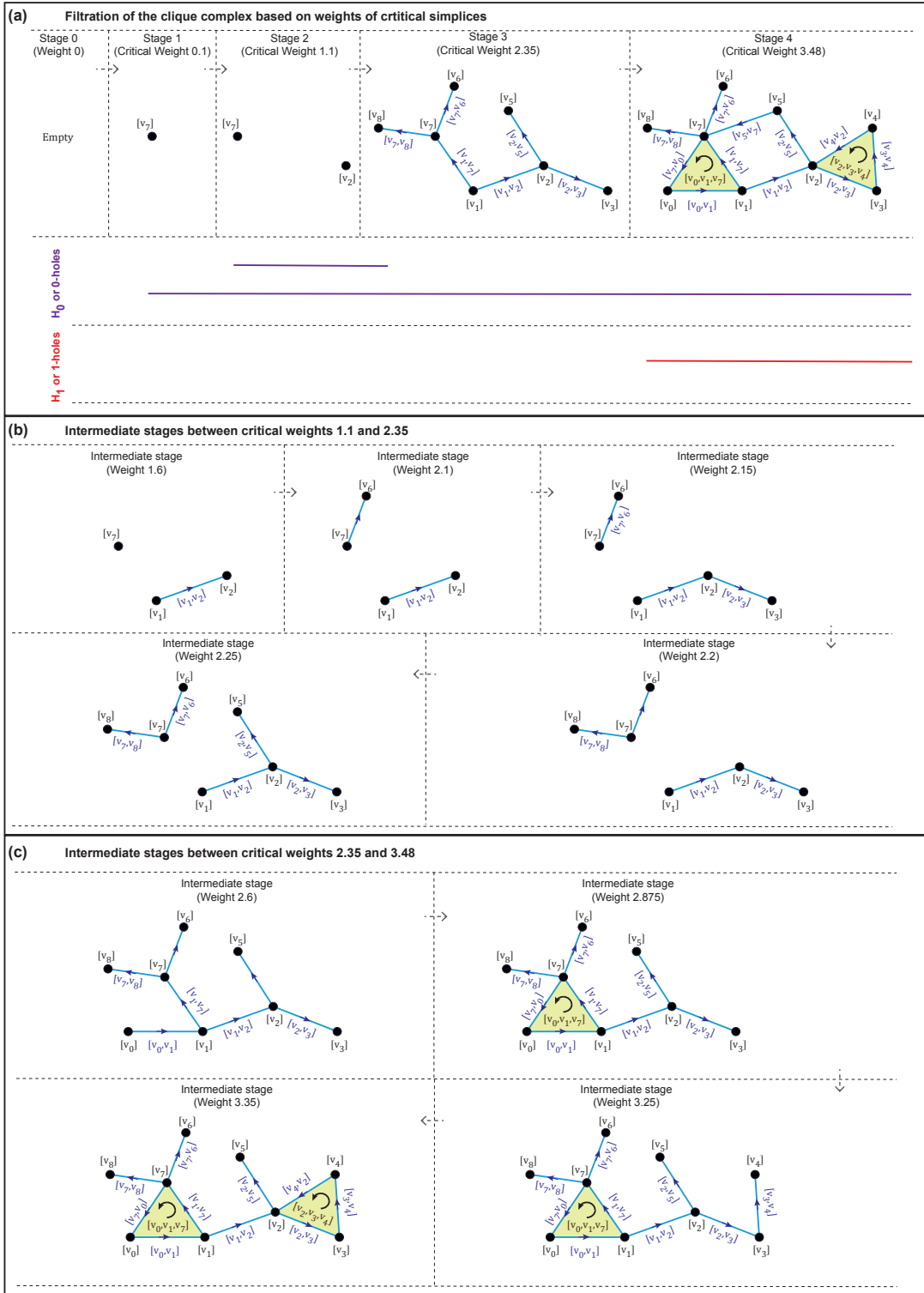We next provide the fundamental definitions in discrete Morse theory [8, 9]. We remark that a p-dimensional

FIG. 2. Filtration based on the entire sequence of weights satisfying discrete Morse function is equivalent to filtration based only on the subsequence of critical weights in terms of persistent homology. (a) Filtration of the network shown in Figure 1 based on weights of the 4 critical simplices. There is a 0-hole (or connected component) that persists across the 4 stages of the filtration while another 0-hole is born at stage 2 on addition of critical vertex v_2 but dies at the stage 3 which corresponds to the weight of the critical edge $[v_1, v_7]$. Moreover, a 1-hole is born at the stage 4 on addition of the critical edge $[v_5, v_7]$. (b) Five intermediate stages during the filtration between critical weights 1.1 (stage 2) and 2.35 (stage 3). (c) Four intermediate stages during the filtration between critical weights 2.35 (stage 3) and 3.48 (stage 4). It is seen that the homology of the clique complex remains unchanged during the intermediate stages of the filtration whereby the birth and death of holes occur only at stages which correspond to critical weights.

simplex α in a simplicial complex K is denoted by $\alpha^p \in K$. Also, if a p-simplex α^p in K is a face of a $(p+1)$-simplex β^{p+1} in K then this is represented as $\alpha^p < \beta^{p+1}$ in the sequel. Given a function $f : K \to \mathbb{R}$, for each simplex $\alpha^p \in K$, two sets U_α^f and V_α^f are defined as follows:

$$U_\alpha^f = \{\beta^{p+1} \mid \alpha^p < \beta^{p+1} \text{ and } f(\beta) \leq f(\alpha)\} \tag{4}$$

$$V_\alpha^f = \{\gamma^{p-1} \mid \gamma^{p-1} < \alpha^p \text{ and } f(\alpha) \leq f(\gamma)\} \tag{5}$$

Simply stated, the set U_α^f contains any $(p+1)$-simplex β^{p+1} of which α^p is a face and the function value on β is less than or equal to the function value on α. The set V_α^f contains any $(p-1)$-simplex γ^{p-1} which is a face of α^p and the function value on α is less than or equal to the function value on γ. A function $f : K \to \mathbb{R}$ is a discrete Morse function [8, 9] if and only if for each simplex $\alpha^p \in K$:

$$|U_\alpha^f| \leq 1 \text{ and } |V_\alpha^f| \leq 1. \tag{6}$$

Given a discrete Morse function f on the simplicial complex K, a simplex $\alpha^p \in K$ is critical [8, 9] if and only if:

$$|U_\alpha^f| = 0 \text{ and } |V_\alpha^f| = 0. \tag{7}$$

Simply stated, a p-simplex $\alpha^p \in K$ is critical if the following conditions are simultaneously satisfied. The first condition being that if β^{p+1} is any $(p+1)$-simplex in K of which α^p is a face, then $f(\alpha) < f(\beta)$. The second condition being that if γ^{p-1} is any $(p-1)$-simplex in K which is a face of α^p then $f(\alpha) > f(\gamma)$. The concept of critical simplices in discrete Morse theory is in spirit a discrete analogue to the concept of critical points in classical Morse theory wherein the critical points corresponding to a smooth real valued function on $X \subseteq \mathbb{R}^d$ are the points where the gradient of the function vanishes.

We remark that once a discrete Morse function f on a simplicial complex K is fixed, the sets U_α^f and V_α^f are denoted by U_α and V_α, respectively, to simplify the notation. A simple example for a discrete Morse function on a simplicial complex K is the dimension function used in Horak *et al* [19]. The value of the dimension function on a given simplex α is the dimension of the simplex α. By a direct consequence of the definitions presented above in this section, for every simplex $\alpha \in K$, the sets U_α and V_α corresponding to the dimension function are empty. Thus, the dimension function is indeed a discrete Morse function and every simplex in K is critical. In the results section, we present our new scheme and algorithm 1 to assign a discrete Morse function f to a clique complex K of an unweighted graph G.

We next describe the filtration of the clique simplicial complex K based on the discrete Morse function f. Given a discrete Morse function f on a simplicial complex K and a real number r, a level subcomplex $K(r)$ is defined [8, 9] as follows:

$$K(r) = \bigcup_{f(\beta) \leq r} \bigcup_{\alpha \leq \beta} \alpha \tag{8}$$

Simply stated, $K(r)$ contains all simplices β in K with the value of the discrete Morse function or assigned weight $f(\beta) \leq r$ along with any face α of β. Note that a face α of β is included in $K(r)$ even if the discrete Morse function or assigned weight to a face α is greater than r.

Let $\{f(\sigma)\}_{\sigma \in K}$ denote the entire set of values assigned to simplices in K using the discrete morse function f. Then, let $\{w_k\}_{k=0,\cdots,n}$ denote the finite increasing sequence of the unique values in the set $\{f(\sigma)\}_{\sigma \in K}$ associated with the finite simplicial complex considered here. We now have a sequence of inclusions of level subcomplexes corresponding to this increasing sequence $\{w_k\}$ as follows:

$$\emptyset \subseteq K(w_0) \subseteq K(w_1) \subseteq \cdots \subseteq K(w_{n-1}) \subseteq K(w_n) = K. \tag{9}$$

This nested sequence gives a filtration of the simplicial complex K which enables the study of persistent homology in the context of unweighted networks.

According to *Lemma 2.6* by Forman [9], if there are no critical simplices α with $f(\alpha) \in (a, b]$, then $K(b)$ is *homotopy equivalent* to $K(a)$.

The implications of this Lemma are as follows. Let $\{f(\sigma_c)\}$ denote the set of values assigned to critical simplices σ_c in K by the discrete Morse function f and $\{w_{c_k}\}_{k=0,\cdots,m}$ denote the increasing sequence of the unique values in $\{f(\sigma_c)\}$. We refer to the function values $\{f(\sigma_c)\}$ assigned to the critical simplices σ_c in K as *critical weights*. Note that the set $\{f(\sigma_c)\}$ defined for critical simplices is a subset of $\{f(\sigma)\}$ defined for all simplices in K and the increasing sequence $\{w_{c_k}\}_{k=0,\cdots,m}$ is a subsequence of $\{w_k\}_{k=0,\cdots,n}$ with $m \leq n$. The above definition implies that there are no critical simplices α with $f(\alpha) \in (w_{c_i}, w_{c_{i+1}})$. As homology is invariant under homotopy equivalence, Forman's Lemma 2.6 gives us that for any x and y belonging to the real number interval $(w_{c_i}, w_{c_{i+1}})$, the homology groups of $K(x)$ and $K(y)$ are isomorphic. Thus, in order to observe the changes in homology as the filtration proceeds, it suffices to study

the persistent homology of a filtration which corresponds to the subsequence $\{w_{c_k}\}_{k=0,\cdots,m}$ of $\{w_k\}_{k=0,\cdots,n}$, where $m \leq n$, and this results in a potential decrease in the required number of filtration steps. The new filtration sequence can be represented as:

$$\emptyset \subseteq K(w_{c_0}) \subseteq K(w_{c_1}) \subseteq \cdots \subseteq K(w_{c_{m-1}}) \subseteq K(w_{c_m}) \subseteq K. \tag{10}$$

Note that each simplex α in the clique complex K is first introduced as part of certain level subcomplex $K(w_{c_i})$ in the above nested filtration sequence. Therefore, each simplex α in K can be associated with a unique weight w_{c_i} referred to as the filtration weight of α. In SI Appendix, we present algorithm 2 and algorithm 3 which depict the procedure to compute the filtration weights of simplices in the clique complex K of a graph G. In SI Appendix, we also give a sufficient condition for two discrete Morse functions to induce the same filtration and thus the persistent homology groups.

Using an example network in figure 2, we also show that the persistent homology observed using the filtration based on the entire sequence of weights satisfying discrete Morse function is equivalent to that observed using the filtration based on the subsequence of critical weights.

Let m_p represent the number of critical p-simplices in a simplicial complex K and let β_p denote the p-Betti number of K. Then *Theorem 2.11* by Forman [9] can be stated as follows.

(i) For each $p = 0, 1, 2, \cdots, d$ (where d is the dimension of K), $m_p \geq \beta_p$.

(ii) $m_0 - m_1 + m_2 - \cdots + (-1)^d m_d = \beta_0 - \beta_1 + \beta_2 - \cdots + (-1)^d \beta_d$.

In other words, the above theorem gives a lower bound of the number of p-critical simplices m_p for each dimension p as the p-Betti number β_p of K. In results section, we present our algorithm 1 to assign weights satisfying discrete Morse function to simplices in the clique complex K of a graph G. Our choice of the function in algorithm 1 to assign weights to simplices in the clique complex K tries to minimize the number of critical simplices (which has a lower bound given by Forman's Theorem 2.11 [9]), and thus, reduces the number of filtration steps required to compute the persistent homology without loss of information. In the results section, we will show that our algorithm achieves near-optimal number of critical weights in clique complexes corresponding to many model and real networks analyzed here.

Comparing Persistence diagrams

Given a discrete Morse function f and its associated filtration $\{K(w_{c_k})\}$ of the clique complex K of a graph G (equation 10), each p-hole has a critical weight $w_{c_{birth}}$ which corresponds to its birth index and $w_{c_{death}}$ which corresponds to its death index, with $w_{c_{birth}} < w_{c_{death}}$.

Persistence diagram D(f) for a d-dimensional simplicial complex K is the collection of points in $\mathbb{R}^2$ whose first and second coordinates, x and y, respectively, correspond to the birth weight and death weight of a p-hole where $0 \leq p \leq d$ [37]. Since two different holes can have the same birth and death weights, each point in the persistence diagram has a corresponding multiplicity, we refer the reader to the SI Appendix for more details. Thus, the persistence diagram is a multiset of points in $\mathbb{R}^2$. The persistence of a p-hole which has birth and death weights, $w_{c_{birth}}$ and $w_{c_{death}}$, respectively, is defined as $w_{c_{birth}} - w_{c_{death}}$. Thus, the persistence diagram for a clique complex K corresponding to a graph G is a compact representation of the persistent homology of a network.

Given two persistence diagrams X and Y (which may correspond to two different networks), the ∞-Wasserstein distance between X and Y, also known as the *bottleneck distance* [37], is defined as follows:

$$W_\infty(X, Y) = \inf_{\eta:X \to Y} \sup_{x \in X} ||x - \eta(x)||_\infty. \tag{11}$$

Similarly, given two persistence diagrams X and Y, the q-Wasserstein distance [38] between X and Y is defined as follows:

$$W_q(X, Y) = \left[\inf_{\eta:X \to Y} \sum_{x \in X} ||x - \eta(x)||_\infty^q \right]^{\frac{1}{q}}. \tag{12}$$

In the above equations, η ranges over all bijective maps from X to Y, and given $(a, b) \in \mathbb{R}^2$, $||(a, b)||_\infty = \max\{|a|, |b|\}$ is the L_∞ norm. In this work, we use Dionysus 2 package (http://www.mrzv.org/software/dionysus2/) to compute the Wasserstein distance between two persistence diagrams corresponding to two different model networks (See results section). Note that it is not generally true that two persistence diagrams X and Y have the same number of off-diagonal points, i.e., features with non-zero persistence, and we refer the readers to Kerber *et al.* [38] for details on

circumventing this issue and further information regarding how the computation of the Wasserstein distance is reduced to a bipartite graph matching problem in Dionysus 2 package. We remark that the bottleneck distance between two persistence diagrams which are subsets of the unit square is in the range 0 to 1.

Stability of persistence diagrams with respect to small changes in the discrete Morse function is a key property of persistent homology. The first such stability theorem was given by Cohen-Steiner *et al.* [37], and later refined by Chazal *et al.* [39]. In the SI Appendix, using results from Chazal *et al.*[39], we give a stability result for persistence diagrams of discrete Morse functions with respect to the bottleneck distance.

NETWORK DATASETS

Model networks. We have investigated the following models of unweighted and undirected networks, namely, the Erdös-Rényi (ER) [40], the Watts-Strogatz (WS) [10], the Barabási-Albert (BA) [11] and the Hyperbolic Graph Generator (HGG) [41]. The ER model [40] is characterized by the property that the probability p of the existence of each possible edge between any two vertices among the n vertices in the graph G is constant. The existence of edges in the ER model are independent of each other, and thus, the model produces random graphs $G(n, p)$ with average vertex degree $p(n - 1)$. The WS model [10] produces small-world graphs as follows. The WS model starts with an initial regular graph with n vertices where each vertex is connected to its k nearest neighbours. Next, the endpoint of each edge in the initial regular graph of the WS model is randomly chosen for rewiring based on a fixed rewiring probability p and is rewired to another vertex in the graph which is chosen with uniform probability. The BA model [11] produces scale-free graphs which are characterized by a degree distribution that follows a power law decay. The BA model utilizes a preferential attachment scheme to produce scale-free graphs. The BA model generates an initial graph of m_0 vertices, and then, at each successive iteration a new vertex is added with edges to m already existing vertices which are chosen with probability proportional to their degree at that particular iteration. The iterations in the BA model cease when the graph has attained the requisite number n of vertices. The HGG model [41, 42] produces a random graph of n vertices by initially fixing n vertices to n points on a hyperbolic disk. In the HGG model, the probability of existence of an edge between two vertices is proportional to the hyperbolic distance between the two points on the hyperbolic disk that correspond to these two vertices. By tuning the input parameter γ, the HGG model can produce either a hyperbolic or a spherical random graph [41, 42]. Specifically, the HGG model produces hyperbolic random graphs for $\gamma \in [2, \infty)$ whereas spherical random graphs for $\gamma = \infty$.

Real networks. We have also studied seven real-world networks which are represented as unweighted and undirected graphs. We have considered two biological networks, namely, the *Yeast protein interaction* network [43] with 1870 vertices and 2277 edges, and the *Human protein interaction* network [44] with 3133 vertices and 6726 edges. In both biological networks, each vertex represents a protein and an edge represents an interaction between the two proteins. We have considered two infrastructure networks, namely, the *US Power Grid* network [45] and the *Euro road* network [46]. In the US Power Grid network, the 4941 vertices represent the generators, transformers and substations in the Western states of USA and the 6594 edges represent power links between them. The 1174 vertices of the Euro road network correspond to cities in Europe and the 1417 edges correspond to roads linking the cities. We have also studied the *Email* network [47] of the University of Rovira i Virgili with 1133 vertices representing users and 5451 edges, each representing the existence of at least one Email communication between the two users corresponding to the vertices anchoring the edge. We have also studied the *Route views* network [45] which has 6474 autonomous systems as vertices and 13895 edges representing communication between the systems that are represented as vertices. We have considered a social network, the *Hamsterster friendship* network [48], containing 1858 vertices which represent the users and 12534 edges which represent friendships between the users. Note that we omit self-loops while constructing the clique complex K corresponding to the undirected graph G of a real-world network.

RESULTS AND DISCUSSION

Algorithm to construct discrete Morse function on a simplicial complex

From an unweighted and undirected graph $G(\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, it is straightforward to construct a clique simplicial complex K with dimension d (See Theory section). Figure 1 shows the construction of a clique complex starting from an example network. Given a simplicial complex K, its dimension d and a non-negative real-valued function g on the 0-simplices of K, the algorithm 1 assigns weights to any simplex in K, producing a discrete Morse function f defined in equation 6. In the pseudocode of the algorithm 1, lines 2-6 initialize a variable Flag[α] for every simplex α in clique complex K with the value 0. We remark that the variable Flag[α] associated with a simplex α in K serves as a counter for the size of the set U_α defined in equation 4. Lines 7-9 assign weights to every

0-simplex in K based on the input non-negative function g. Lines 10-24 assign weights to 1- or higher-dimensional simplices in K in a manner which is consistent with the definition in equation 6 of a discrete Morse function. In summary, algorithm 1 outputs a discrete Morse function f on K, and in SI Appendix, we present a rigorous proof for the following theorem which states the same.

Theorem. *Algorithm 1 produces a discrete Morse function f on any simplicial complex K of finite dimension d.*

Algorithm 1 Algorithm to construct a discrete Morse function on a d-dimensional simplicial complex K

1: **function** DISCRETEMORSEFUNCTION(K, d, g)

2:   **for** $p = 0, \cdots, d$ **do**   ▷ Initialize `Flag` variable associated with each simplex in K
3:     **for** each p-simplex $\alpha \in K$ **do**
4:       `Flag`$[\alpha] = 0$ ▷ The variable `Flag` keeps track of the size of the set U_α for each simplex α
5:     **end for**
6:   **end for**

7:   **for** each 0-simplex $\alpha \in K$ **do**   ▷ Assign weights to 0-simplices in K
8:     $f(\alpha) = g(\alpha)$
9:   **end for**

10:   **for** $p = 1, \cdots, d$ **do**
11:     **for** each p-simplex $\alpha \in K$ **do**   ▷ Assign weights to p-simplices in K with $p \geq 1$
12:       Let `Faces`[ ] be an array of all $(p-1)$-dimensional faces of α
13:       Sort `Faces`[ ] such that $f(\texttt{Faces}[i]) \geq f(\texttt{Faces}[i+1])$ for each $i \in \{0, 1, \cdots, p-1\}$
14:       Let $\gamma_0 = \texttt{Faces}[0]$
15:       Let $\gamma_1 = \texttt{Faces}[1]$
16:       **if** `Flag`$[\gamma_0] = 0$ and $f(\gamma_0) > f(\gamma_1)$ **then**
17:         $f(\alpha) = (f(\gamma_0) + f(\gamma_1))/2$
18:         `Flag`$[\gamma_0] = 1$
19:       **else**
20:         $\epsilon = \text{random}(0, 0.5)$ ▷ ϵ is generated randomly during runtime using a uniform distribution on $(0, 0.5)$
21:         $f(\alpha) = f(\gamma_0) + \epsilon$
22:       **end if**
23:     **end for**
24:   **end for**

25:   **return** f

26: **end function**

Given a simplicial complex K, its dimension d and a discrete Morse function f on K, the algorithm 2 in SI Appendix determines the weights of critical simplices in K. Given an unweighted and undirected graph G, we restrict the construction of clique complex K by including simplices up to a maximum dimension d. Then, the algorithm 3 in SI Appendix creates the filtration of clique complex K based on weights of critical simplices as described in the Theory section. In SI Table S1, we describe the role of key variables which appear in algorithms 1, 2 and 3. In SI Appendix, we also give a time complexity analysis for these algorithms.

<div align="center">Rationale for the choice of function on vertices</div>

In order to construct a discrete Morse function f on clique complex K corresponding to a graph G using our algorithm 1, a real-valued function g has to be fixed on the 0-simplices of K (See lines 7-9 in algorithm 1). Let deg_{max} denote the maximum degree of a vertex in the graph $G(\mathcal{V}, \mathcal{E})$. Our choice for the function value on the vertices or 0-simplices, $g : \mathcal{V} \to \mathbb{R}$, is as follows:

$$g(v) = deg_{max} - \text{degree}(v) + \epsilon \tag{13}$$

where degree(v) is the degree of the vertex $v \in G$ and ϵ corresponding to each vertex is a random number (noise) generated using the uniform distribution on the interval $(0, 0.5)$.

In the Theory section, we had highlighted the Theorem 2.11 by Forman [9] which gives a lower bound on the number of critical p-simplices, m_p, in a simplicial complex K as the p-Betti number β_p. The choice of the real-valued

function g in algorithm 1 plays a key role in determining if m_p is close to the theoretical minimum β_p stated above. In the Theory section, we have shown that the number of critical simplices determines the effective number of filtration weights to study the persistent homology of a clique complex (See equation 10). This motivated our choice for the real-valued function (equation 13) which determines the weights of 0-simplices, and the rationale for this choice is as follows.

Ignoring the noise term ϵ in equation 13, the reader can discern our intuition for choosing the function $g(v) = deg_{max} - \text{degree}(v)$ for any vertex v in G with the following example. Consider the simple example of the clique complex K corresponding to a graph G in figure 1. Here, we would like to obtain a discrete Morse function f on K such that the number of critical simplices is close to the theoretical minimum. This requirement applies to simplices of any dimension in K, and in the context of this example, we would like the number of critical 1-simplices (edges) to be as close as possible to the 1-Betti number β_1 of K. Note that $\beta_1 = 1$ for the example clique complex K in figure 1.

Let us now examine lines 11-23 in algorithm 1. Consider any edge $e_{vw} = [v, w]$ such that $g(v) > g(w)$. While assigning the function value to the edge e_{vw} in algorithm 1, the edge e_{vw} and the vertex v are guaranteed to be not critical provided that the *if condition* in the line 16 is satisfied. This is a consequence of the definition of a critical simplex (See equation 7). Thus, we would like to force this *if condition* to be True for as many edges as possible. Moreover, once the function value of the 1-simplex e_{vw} is set, we set the variable `Flag`$[v]$ to 1 in line 18, and this subsequently forces the *if condition* in the line 16 to fail for all other edges $e_{vz} = [v, z]$ in the graph that contain v and have function value $g(v) > g(z)$.

Let us now examine the edge $e_{78} = [v_7, v_8]$ in figure 1 which is anchored by vertices v_7 and v_8 with degree 5 and 1, respectively. As the degree of a vertex gives the number of edges that contain the vertex, v_7 is part of 4 other edges apart from e_{78} while v_8 is part of only the edge e_{78}. Suppose e_{78} is the first edge chosen for the function assignment in line 11 of algorithm 1 and both `Flag`$[v_7]$ and `Flag`$[v_8]$ for the anchoring vertices are 0. We would then prefer that the *if condition* in the line 16 is satisfied for e_{78}, and as v_7 is part of 4 other edges apart from e_{78} while v_8 is part of only e_{78}, ideally `Flag`$[v_8]$ is set to 1 instead of `Flag`$[v_7]$, in other words, we need the function value $g(v_8) > g(v_7)$. We emphasize that this choice of v_8 over v_7 prevents the forced failure (described in the previous paragraph) of the *if condition* for the 4 other edges apart from e_{78} that contain v_7.

The above example suggests a need for a function g on the vertices that has an inverse relationship with the degree of the vertices. Hence, our choice $g(v) = deg_{max} - \text{degree}(v)$ provides a simple and effective solution for the above requirement. As a consequence of this choice for the function g, the weights assigned to the simplices by algorithm 1 reflect the degree of its constituent vertices. The degree of a vertex can be thought of as a measure of its importance in the network. Hence, the intuition behind the assignment of weights to simplices by our method is to have an inverse or opposite relationship between 'weight' and 'importance' of a simplex while simultaneously satisfying the definition of a discrete Morse function. This inverse relationship instead of a proportional relationship between 'importance' and 'weight' plays a key role in our filtration scheme which is based on the sequence of level subcomplexes corresponding to the increasing sequence of critical weights (See Theory section and algorithm 3 in SI Appendix). Thus, within the constraint of ensuring that the definition for a simplicial complex is satisfied at each stage of the filtration, our scheme prioritizes the addition of simplices with higher 'importance' at earlier stages of the filtration due to their lower weights.

We now provide a rationale for the addition of a random noise ϵ in equation 13. As reasoned above, we would like to force the *if condition* in line 16 of algorithm 1 to be True for as many edges as possible. Consider an edge $e_{vw} = [v, w]$ such that $degree(v) = degree(w)$. The absence of a random noise ϵ in equation 13 forces $g(v) = g(w)$. Thus, irrespective of the state of `Flag`$[v]$ and `Flag`$[w]$, the *if condition* fails. This implies that the set $V_{e_{vw}}$ (See equation 5) corresponding to the edge e_{vw} is forced to be empty since the only other possibility is $|V_{e_{vw}}| = 2$ which cannot be the case because algorithm 1 produces a discrete Morse function (See Theorem and equation 6). Thus, provided e_{vw} is not a face of any higher dimensional simplex, e_{vw} would be a critical simplex irrespective of the state of `Flag`$[v]$ and `Flag`$[w]$. Hence, we would like $g(v) \neq g(w)$ while also retaining the inverse relationship of the function with the degree. Thus generating a small random noise ϵ in the range $(0, 0.5)$ for each vertex as in equation 13 provides a simple resolution. We remark that the above argument can be generalized to higher-dimensional simplices, and thus, provides the intuition for the addition of noise ϵ in line 21 of algorithm 1.

We remind the readers that our initial motivation was not to develop a scheme to construct the optimal discrete Morse function on a clique complex corresponding to a graph. Rather, our main goal is to develop a systematic filtration scheme to study persistent homology in unweighted and undirected networks. In fact, constructing an optimal discrete Morse function in the general case has been shown to be MAX-SNP Hard [49]. The primary utility of our scheme is to create a filtration by assigning weights to simplices in the clique complex K of a graph G. However, we next report our empirical results from an exploration of model and real-world networks which underscore the following. Although our scheme is not optimal in the sense of minimizing the number of critical simplices, in practice, it achieves near-optimal results in several model and real-world networks (Table I). Hence, our scheme based on discrete Morse theory reduces the number of filtration steps and increases the applicability of persistent homology to

study complex networks.

TABLE I. The table lists the number of p-simplices (n_p), the number of critical p-simplices (m_p) and the p-Betti number β_p for clique complexes corresponding to model and real networks. Note that the dimension p of simplices ranges from 0 to 3. In case of model networks, the statistics is reported for a particular instance of ER graph with $n = 1000$ and $p = 0.004$, WS graph with $n = 1000$, $k = 4$ and $p = 0.5$, BA graph with $n = 1000$ and $m = 2$, Spherical random graph produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$, and Hyperbolic random graph produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$. For the statistics corresponding to a more comprehensive list of model networks with different parameters, we refer the readers to SI Table S2. Note that we omit self-loops in the real networks considered here.

Network	n_0	m_0	β_0	n_1	m_1	β_1	n_2	m_2	β_2	n_3	m_3	β_3
ER model with $n = 1000$ and $p = 0.004$	1000	90	21	2007	1090	1021	7	0	0	0	0	0
WS model with $n = 1000$, $k = 4$ and $p = 0.5$	1000	123	1	2000	991	864	137	5	0	0	0	0
BA model with $n = 1000$ and $m = 2$	1000	8	1	1996	949	942	55	0	0	0	0	0
Spherical model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$	1000	172	126	2028	118	0	2029	180	0	1321	554	446
Hyperbolic model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$	1000	144	144	2593	20	0	5440	426	0	11456	8159	7753
US Power Grid	4941	573	1	6594	1671	1080	651	21	0	90	15	13
Email communication	1133	6	1	5451	1694	1186	5343	871	53	3419	1577	1262
Route views	6474	17	1	12572	2459	2157	6584	627	19	5636	3335	3013
Yeast protein interaction	1870	272	173	2203	424	318	222	12	0	41	12	7
Hamsterster frienship	1858	33	23	12534	4484	2970	16750	5324	1880	10015	4814	2874
Euro road	1174	213	26	1417	425	237	32	1	0	0	0	0
Human protein interaction	3133	269	210	6149	2454	2298	1047	109	1	142	35	24

Application to model and real networks

Given a model or real network G, we limit our study of persistent homology to the 3-dimensional clique simplicial complex K corresponding to G. In other words, during the construction of the clique complex, we only include p-simplices which have dimension $0 \leq p \leq 3$ (See Theory section). On this 3-dimensional clique complex K, we create the corresponding filtration based on the assigned weights to simplices using discrete Morse theory. In SI Appendix, we present algorithm 3 which outlines the procedure to compute the filtration weights of simplices. Thereafter, we make use of GUDHI [50], a C++ based library for Topological Data Analysis (http://gudhi.gforge.inria.fr/) to study the persistent homology of this filtration of K.

Each hole of any dimension in K has a corresponding *birth* filtration weight and a *death* filtration weight (See Theory section). We normalize the *birth* and *death* filtration weights of all holes in K by dividing with $w_N = 1 + \max\{f(\alpha) \mid \alpha \in K\}$. In other words, w_N is 1 plus the maximum value among weights assigned to the simplices in K. We also make the convention that a normalized *death* filtration weight of 1 for a hole in K represents that the particular hole never *dies*.

A H_p-barcode diagram corresponding to a filtration of the clique complex K is a graphical representation containing horizontal line segments, each of which represents a p-hole in K, plotted against the x-axis ranging from 0 to 1 which corresponds to the normalized filtration weights of simplices in K [51]. A horizontal line in the H_p-barcode diagram of K is referred to as a *barcode*. Thus, a barcode in the H_p-barcode diagram of K which begins at a x-axis value of w_1 and ends at a x-axis value of w_2 represents a p-hole in K whose *birth* and *death* weights are w_1 and w_2, respectively. In figures 3, 4 and 5, we display the barcode diagrams for model and real-world networks analyzed here.

Model Networks

In this work, we have investigated the persistent homology of unweighted and undirected graphs corresponding to five model networks, namely, ER, WS, BA, hyperbolic random graphs and spherical random graphs. We have considered model networks with 1000 vertices and expected average degree 4, 6 and 8. In main text, we report results for model networks with expected average degree 4, and in SI, those with expected average degree 6 and 8.

The H_0 barcode diagram of BA networks indicate a low number of 0-holes in BA networks across the entire filtration (Figure 3 and SI Figures S1-S2). A standard result in algebraic topology [36] gives that the 0-Betti number of a simplicial complex K is equal to the number of connected components in K. In other words, the above observation

TABLE II. The table lists the value of optimality indicator μ for various model and real networks analyzed here. Equation 14 gives the definition of μ which is an indicator for the proximity of our algorithm to the optimal case. The value of μ ranges from 0 to 1 with $\mu = 1$ indicating that our algorithm achieves exactly the theoretical minimum number of critical simplices while $\mu = 0$ indicating that all simplices are critical. For model networks, the value reported in this table is the average of μ across 10 samples of each model network for a chosen set of parameter values along with the corresponding standard deviations.

Network	Optimalitiy indicator μ
ER model with $n = 1000, p = 0.004$	0.924 ± 0.004
ER model with $n = 1000, p = 0.006$	0.947 ± 0.004
ER model with $n = 1000, p = 0.008$	0.959 ± 0.006
WS model with $n = 1000, k = 4, p = 0.5$	0.890 ± 0.003
WS model with $n = 1000, k = 6, p = 0.5$	0.917 ± 0.007
WS model with $n = 1000, k = 8, p = 0.5$	0.906 ± 0.003
BA model with $n = 1000, m = 2$	0.989 ± 0.003
BA model with $n = 1000, m = 3$	0.985 ± 0.004
BA model with $n = 1000, m = 4$	0.964 ± 0.006
Spherical model with $n = 1000, \gamma = \infty, T = 0, k = 4$	0.925 ± 0.007
Spherical model with $n = 1000, \gamma = \infty, T = 0, k = 6$	0.901 ± 0.003
Spherical model with $n = 1000, \gamma = \infty, T = 0, k = 8$	0.887 ± 0.003
Hyperbolic model with $n = 1000, \gamma = 2, T = 0, k = 4$	0.939 ± 0.013
Hyperbolic model with $n = 1000, \gamma = 2, T = 0, k = 6$	0.927 ± 0.007
Hyperbolic model with $n = 1000, \gamma = 2, T = 0, k = 8$	0.921 ± 0.005
US Power Grid	0.893937
Email communication	0.871847
Route views	0.952140
Yeast protein interactions	0.942157
Hamsterster friendship	0.793236
Euro road	0.840678
Human protein interaction	0.957924

indicates that the scale-free BA network has a strong tendency to maintain a low number of connected components during filtration (Figure 3 and SI Figures S1-S2). In contrast, both the random ER network and small-world WS network have a relatively high number of connected components at initial phase of the filtration and then progress towards a more connected network at later stages of the filtration (Figure 3 and SI Figures S1-S2). This indicates that the simplices in the clique complex which are key to the connectivity of the model networks are introduced very early into the filtration for the scale-free BA network while this is not the case for the random ER network or small-world WS network. The H_1 barcode diagram of BA networks also indicate late introduction of 1-holes during filtration in contrast to both ER and WS networks where 1-holes appear across a wider range of the filtration (Figure 3 and SI Figures S1-S2). Moreover, in ER and WS networks, it is interesting to observe that the 1-holes start to appear at roughly the same stage of filtration which corresponds to a sharp reduction in the number of connected components (Figure 3 and SI Figures S1-S2).

In contrast to ER, WS and BA networks, the spherical and hyperbolic networks are characterized by a relatively high 0-Betti number β_0 and a low 1-Betti number β_1 (Table I and SI Table S2). Simply stated, this observation on the magnitude of β_0 indicates that both hyperbolic and spherical networks have a higher number of connected components in comparison to the ER, WS and BA networks of similar size, i.e., number of vertices, and average vertex degree (Table I and SI Table S2). Although, both spherical and hyperbolic networks exhibit a higher number of connected components, they differ from each other with respect to the evolution of these connected components during filtration (Figure 3 and SI Figures S1-S2). The hyperbolic model maintains a relatively low number of connected components until very late in the filtration wherein there is a sharp increase in the number of connected components (Figure 3 and SI Figures S1-S2). This is in contrast with the behavior of the H_0 barcode diagram of the spherical model which exhibits a more distributed evolution of connected components during filtration (Figure 3 and SI Figures S1-S2). In addition, the low β_1 for spherical and hyperbolic networks conveys the lack of 1-holes in these networks. A possible reason for this observation is the incidence of higher number of 2-simplices in the clique complex K of the spherical and hyperbolic networks in comparison to the ER, WS and BA networks (Table I and SI Table S2). Note that the formation of a 2-simplex can potentially fill in a 1-hole, and thus, result in a low value for β_1 (See Theory section).
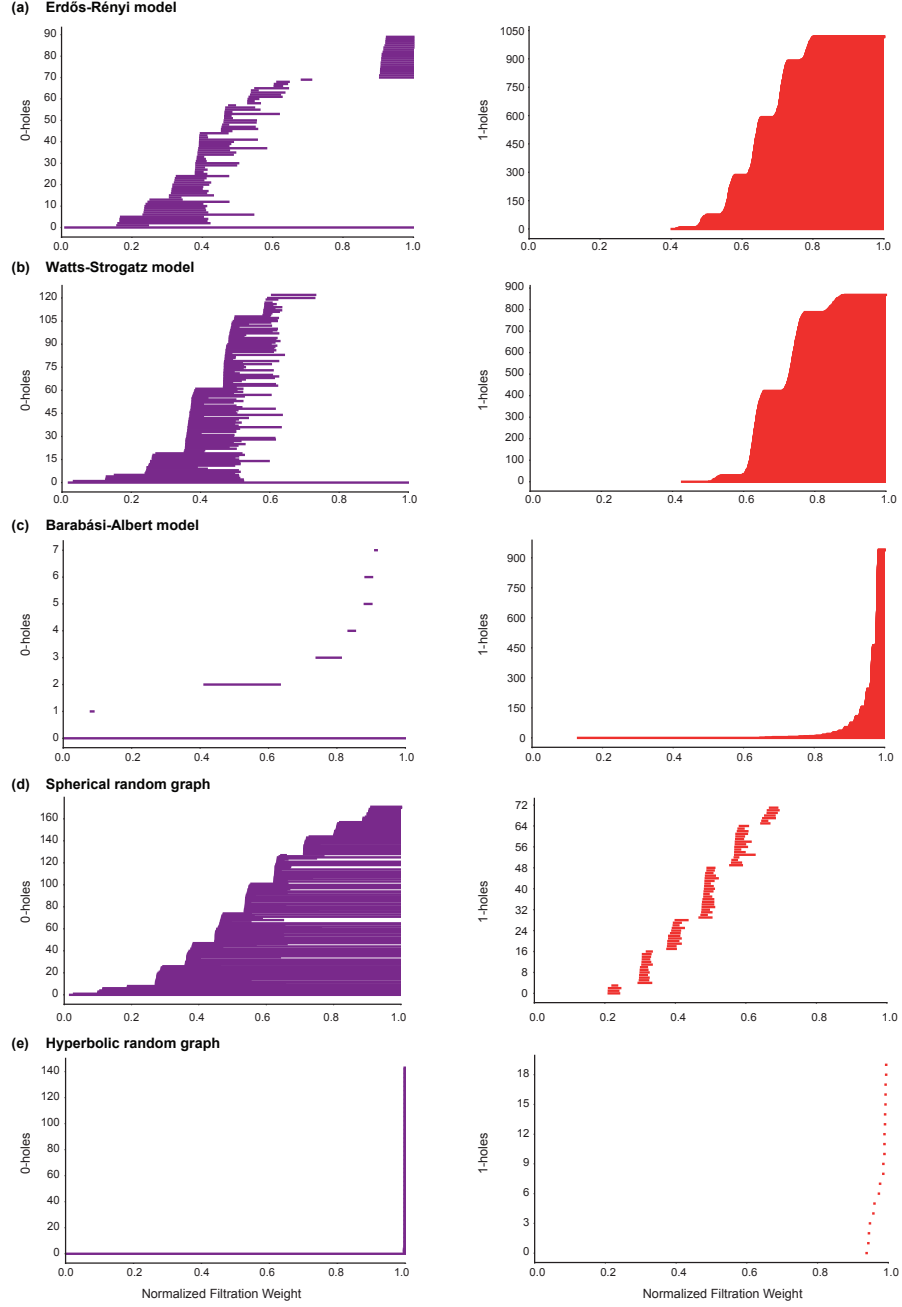
FIG. 3. Barcode diagrams for H_0 and H_1 in model networks. (a) ER model with $n = 1000$ and $p = 0.004$. (b) WS model with $n = 1000$, $k = 4$ and $p = 0.5$. (c) BA model with $n = 1000$ and $m = 2$. (d) Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$. (e) Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$.

Such a behaviour is also seen in the H_2 barcode diagrams of spherical and hyperbolic networks wherein the 2-holes have very short persistence since the addition of 3-simplices successively fill in the 2-holes (Figure 5 and SI Figure S3). The H_3 barcode diagrams of spherical and hyperbolic networks (See SI Figures S3-S4) also indicate a clear difference in the evolution of their corresponding topological features during the filtration. The H_2 and H_3 barcode diagrams of ER, WS and BA networks do not provide any insight into network structure primarily due to a lack of higher-order correlations in these model networks that is essential for the formation of 2-holes and 3-holes.

A visual inspection of the barcode diagrams for the five model networks (Figures 3, 5 and SI Figures S1-S4) suggests that the different models can be distinguished based on their persistent homology. In Theory section, we had
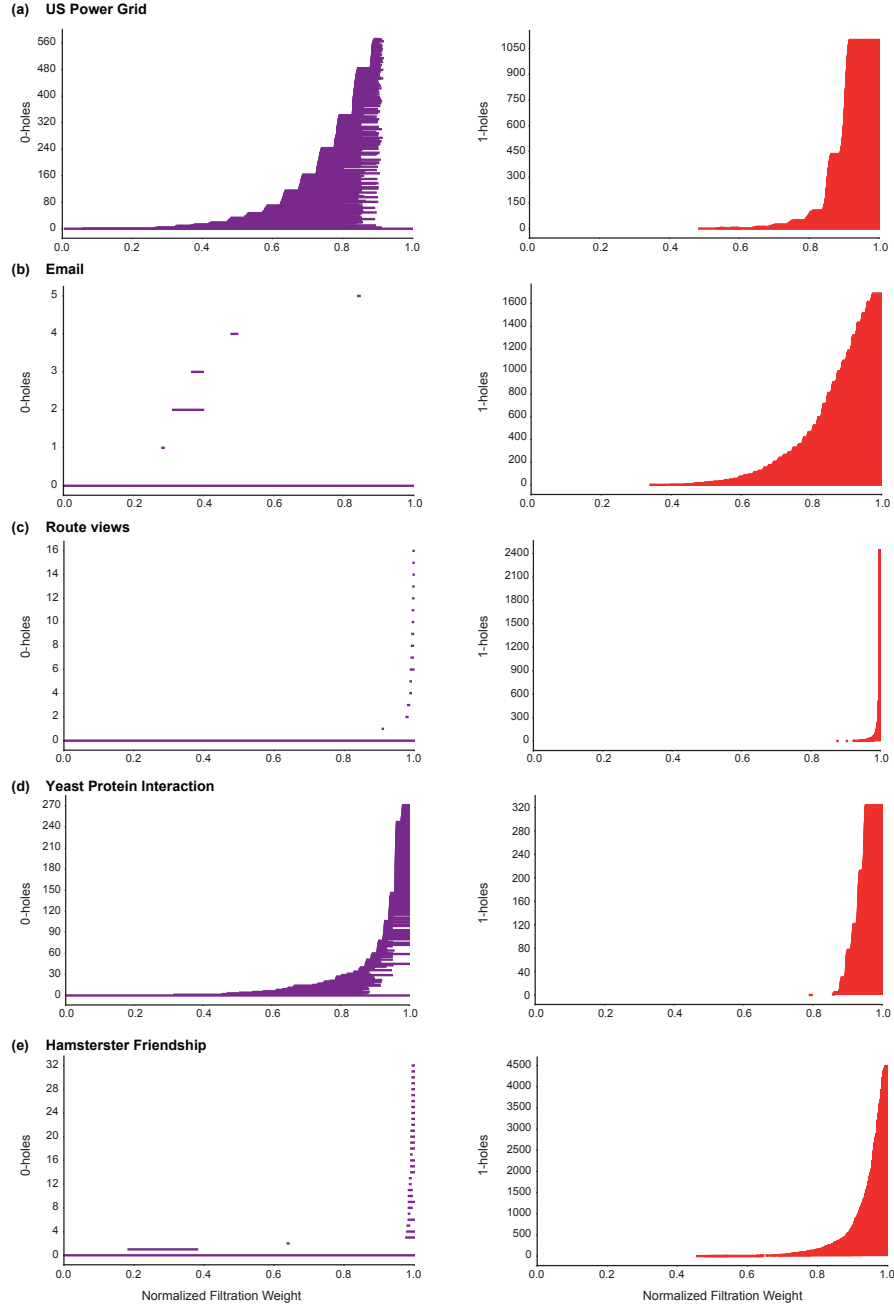
FIG. 4. Barcode diagrams for H_0 and H_1 in real networks. (a) US Power Grid. (b) Email communication. (c) Route views. (d) Yeast protein interaction. (e) Hamsterster friendship.

introduced the bottleneck distance which can be employed to quantify the differences between persistence diagrams from filtration of clique complexes corresponding to different model networks. Recall that the persistence diagram of a d-dimensional simplicial complex K is a compact representation of the persistent homology of K which encompasses topological information across all d dimensions. Figure 6 and SI Table S3 give the bottleneck distance between different model networks with the same number of vertices and similar average vertex degree. For each of the five model networks, 10 random samples are generated by fixing the number of vertices n and other parameters of the model. In figure 6 and SI Table S3, we report the distance between two different models as the average of the distance between each of the possible pairs of the 10 sample networks corresponding to the two models along with the standard error. We find a relatively higher distance between a random instance of a BA network and a random instance of a ER
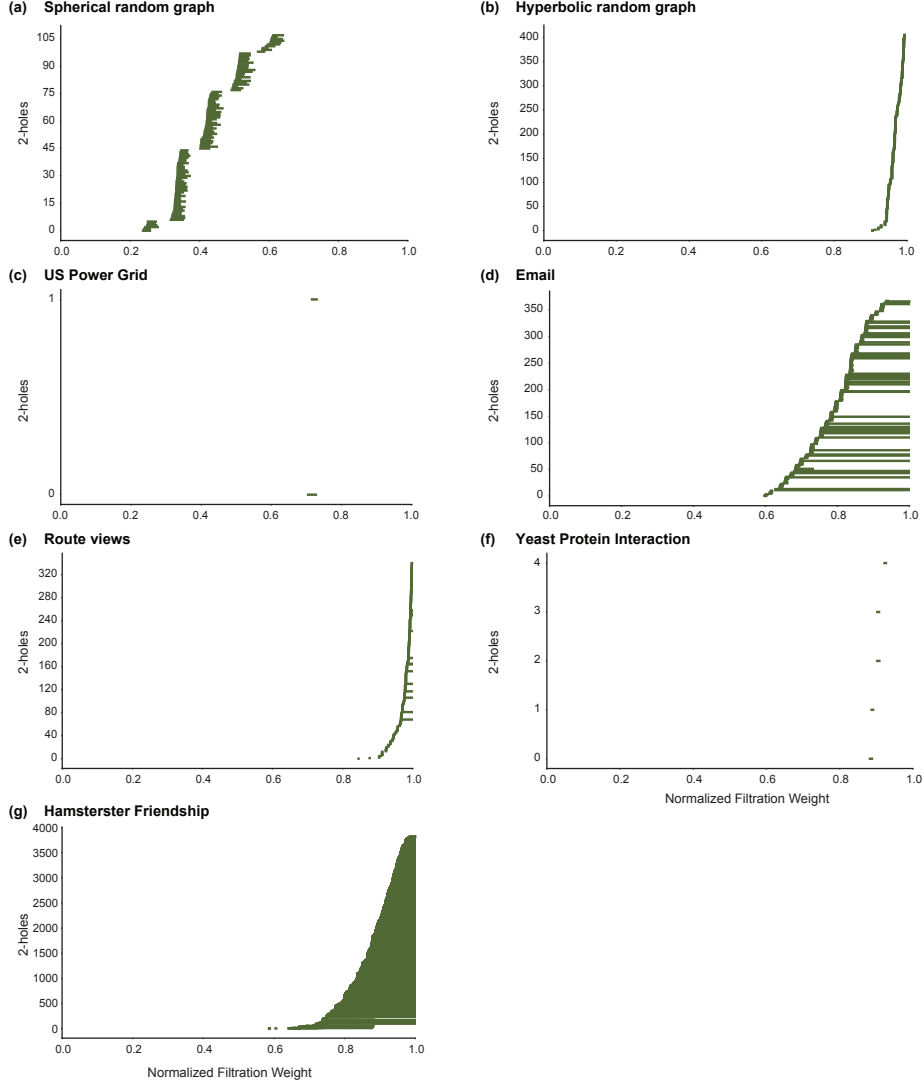
FIG. 5. Barcode diagrams for H_2 in model and real networks. (a) Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$. (b) Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$. (c) US Power Grid. (d) Email communication. (e) Route views. (f) Yeast protein interaction. (g) Hamsterster friendship.

network with the same number of vertices and similar average degree (Figure 6 and SI Table S3). Similarly, we observe a relatively higher distance between a random instance of a BA network and a random instance of a WS network with similar size and average degree. In contrast, a relatively lower average distance is observed between a random instance of a ER network and a random instance of a WS network with similar size and average degree (Figure 6 and SI Table S3). These observations indicate a similarity between networks generated by ER and WS models in terms of their corresponding persistence diagrams and also show that the BA model exhibits topological properties that are different from the other two model networks with the same number of vertices and similar average degree. Moreover, the average distance between a random instance of a spherical network and a random instance of a hyperbolic network with similar size and average degree is very high (Figure 6 and SI Table S3). The last observation is a reflection of the differences in the persistent homology of the clique complexes corresponding to spherical and hyperbolic networks. Finally, the nature of differences observed between persistence diagrams of different model networks using bottleneck distance as shown in figure 6 remain consistent if the 1-Wasserstein or 2-Wasserstein distance is employed in place of the bottleneck distance (data not shown).

Table I and SI Table S2 list the empirical data on the number of p-simplices n_p, the number of critical p-simplices m_p, that our algorithm achieves and the p-Betti number β_p of the clique complexes corresponding to the five model
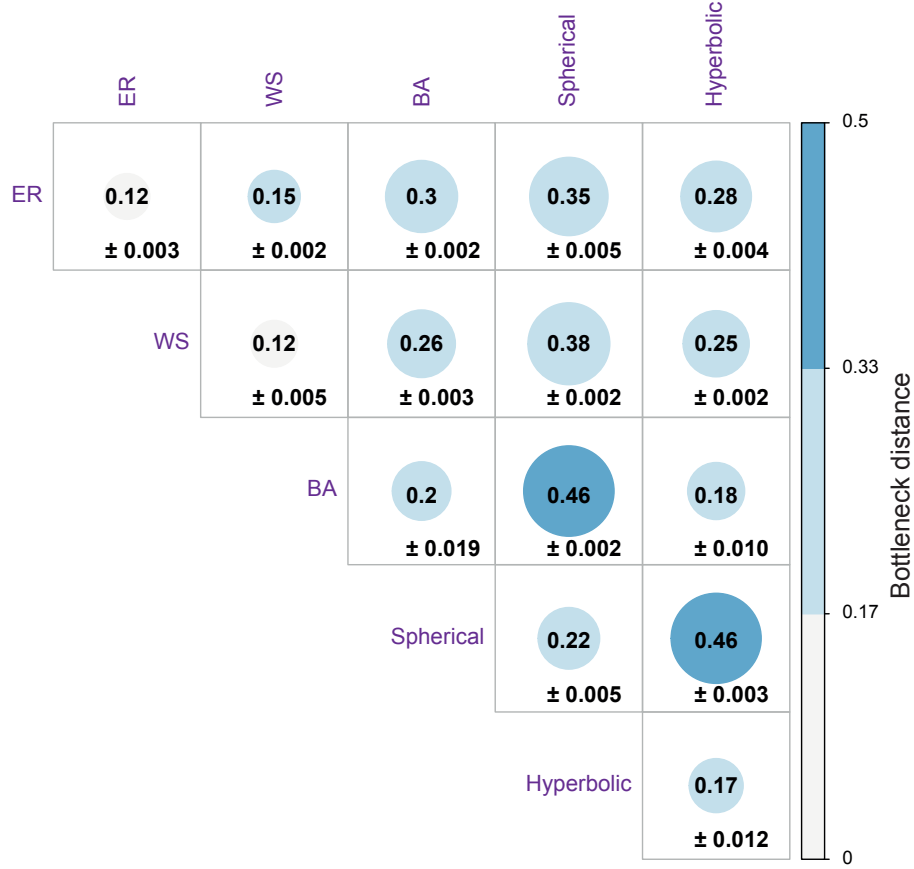
FIG. 6. Bottleneck distance between persistence diagrams of model networks, namely, ER model with $n = 1000$ and $p = 0.004$, WS model with $n = 1000$, $k = 4$ and $p = 0.5$, BA model with $n = 1000$ and $m = 2$, Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$, and Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$. For each of the five model networks, 10 random samples are generated by fixing the number of vertices n and other parameters of the model. We report the distance (rounded to two decimal places) between two different models as the average of the distance between each of the possible pairs of the 10 sample networks corresponding to the two models along with the standard error.

networks. In Table II we report a value μ which indicates the optimality of our algorithm with respect to reducing the number of critical simplicies for each model network that has been analyzed. The definition for μ is as follows.

$$\mu = \frac{\sum_{p=0}^{d} n_p - \sum_{p=0}^{d} m_p}{\sum_{p=0}^{d} n_p - \sum_{p=0}^{d} \beta_p} \tag{14}$$

Here d is the dimension of the corresponding clique complex. The value μ corresponding to a particular discrete Morse function on the clique complex of a given network is an indicator of its optimality with respect to minimizing the number of critical simplices. The value of μ ranges from 0 to 1. $\mu = 1$ indicates that the discrete Morse function achieves exactly the minimum number of critical simplices thereby corresponding to the most optimal situation, while $\mu = 0$ indicates that all simplices are critical thereby corresponding to the least optimal case. Note that the value of μ for a particular network increases linearly with a decrease in the number of critical simplices. In Table II, for model networks, the value reported is the average of μ across 10 samples of each model network for a chosen set of parameter values along with the corresponding standard deviations. Based on the data presented in Table II, we report that our algorithm achieves near-optimal results in terms of reducing the number of critical simplices for each of the five model networks analyzed. Moreover for these model networks, data presented in SI Table S2 underlines the close proximity of m_p to the theoretical lower bound β_p across all dimensions $p = 0, 1, 2, 3$ (See Theory section). We remark that the

worst case scenario corresponds to all simplices in the clique complex being critical which is never attained by our algorithm in the analyzed networks.

Comparison of our method with dimension based filtration in model networks

In SI Figures S5 and S6, we show the H_0, H_1, H_2 and H_3 barcode diagrams for model networks with expected average degree 4 obtained using the dimensional filtration scheme used in Horak *et al.* [19] We restrict our investigation to the three-dimensional clique complex while computing the barcode diagrams for model networks using the dimensional filtration scheme of Horak *et al.* [19] In SI Figures S5 and S6, we normalize the filtration index to be in the range 0 to 1, and p-holes with normalized filtration index 1 indicate that they never die. Moreover, we also report in SI Table S4 and SI Figure S7, the bottleneck distances between persistence diagrams of model networks obtained by the dimensional filtration scheme used in Horak *et al.* [19], and we find that the resultant barcode diagrams and the bottleneck distances between the persistence diagrams are inconclusive to distinguish between the five model networks which have similar size and average vertex degree.

The discrete Morse function produced by our algorithm is empirically near-optimal for the model networks analyzed here, in the sense of minimizing the number of critical simplices (See Table II). On the other hand, it should be noted that the dimension function used in Horak *et al.* [19] is the most non-optimal in this regard. Moreover, in terms of persistent homology, our methods based on the discrete Morse function constructed using algorithms 1, 2 and 3 have the distinct feature of being able to distinguish between various model networks with an efficient filtration.

Although the dimension function used in Horak *et al.* [19] is algorithmically efficient in terms of having a low number of filtration steps, unlike our method it does not conclusively distinguish between different model networks based on the barcode diagrams and also the bottleneck distances between the corresponding persistence diagrams. Thus our discrete Morse function is a good candidate for applications in both, computational aspects of discrete Morse theory as well as in persistent homology of unweighted networks, and achieves a tradeoff between efficiency and applicability.

Real networks

In this work, we have investigated the persistent homology of unweighted and undirected graphs corresponding to seven real-world networks and the barcode diagrams for five of these real networks is shown in figures 4, 5 and SI figure S4. Based on the H_0 barcode diagrams, the behavior of most real networks considered here can be broadly classified into two categories. Real networks such as the Email communication, the Hamsterster friendship and the Route views exhibit a relatively low number of connected components across the entire range of filtration (Figure 4). On the other hand, the two biological networks, namely, the Yeast protein interaction and the Human protein interaction, exhibit a sharp increase in the number of connected components at the later stages of filtration. The H_0 barcode diagrams for the US Power Grid and Euro road do not conform with either of the above characterizations. The H_0 barcode diagram of the US Power Grid network reveals that though there exists only a single connected component at the end of the filtration, there are a considerable number of non-persisting connected components that appear and subsequently disappear during filtration (Figure 4). The H_0 barcode diagram of the Euro road network shows a more distributed increase in the number of connected components (data not shown).

In the context of the H_1 barcode diagrams, the real networks considered here exhibit similar properties with 1-holes appearing late in the filtration (Figure 4). The H_2 barcode diagrams reveal a lack of 2-holes with long persistence in both biological networks, as well as the Route views network, the Euro road network and the US Power Grid network (Figure 5). In contrast, from the H_2 and H_3 barcode diagrams (Figure 5, SI Figure S4) we find that the social network, Hamsterster friendship, and the Email communication network exhibit a relatively high number of 2-holes and 3-holes with longer persistence.

Table II lists the value of the optimality indicator μ (See Equation 14) for each of the seven real networks analyzed here. This data indicates near-optimal performance of our algorithm with respect to minimizing the number of critical simplices for each of these seven real networks. Table I also lists the empirical data on the number of critical p-simplices, m_p, that our algorithm achieves and the p-Betti number β_p of the clique complexes across each dimension p, corresponding to the seven real networks analyzed here.

CONCLUSIONS

To conclude, we have proposed a systematic scheme based on discrete Morse theory to study the persistent homology of unweighted and undirected networks. Our methods leverage the concept of *critical* simplices to permit a reduced filtration scheme while simultaneously admitting a finer inspection of the changes in topology across the filtration of a clique complex corresponding to an unweighted network. Moreover, our proposed algorithm to construct a discrete Morse function on the clique complex of a simple graph achieves close to optimal number of critical simplices for several model and real networks that have been studied here. Furthermore, based on visual representations of persistent homology such as the barcode diagrams as well as quantitative information in the form of distance between persistence diagrams, our methods successfully distinguish various model networks that exhibit inherently different properties. This motivates the application of our methods to real-world networks. We report the results obtained for seven real-world networks that are well studied in the network science community and observe certain patterns in the evolution of their topological features across the filtration. For instance both biological networks, namely the Yeast protein interaction network and the Human protein interaction network exhibit similar characteristics with respect to the H_0, H_1, H_2 and H_3 barcode diagrams. Similarly, both the Email network and the Hamsterster friendship network, exhibit shared features with respect to H_0, H_1, H_2 and H_3 barcode diagrams that vary from the characteristics of the two biological networks considered here. Our observations hint at the ability and possible applications of our methods to detect and classify real-world networks that are inherently different.

Future directions and ongoing work include examining the significance of critical simplices in the context of real-world networks. In other words, we aim to determine whether a critical edge in the context of discrete Morse theory holds any key significance when it is viewed as a link between two real entities in a real-world network. We also intend to explore the presence or absence of a correlation between the notion of critical simplices and network curvature. Since discrete Morse theory captures information about the Euler characteristic of the clique complex corresponding to a graph, the presence of such a correlation could potentially signify a close relationship between the discrete curvature of a graph and its topology, much like in the case of smooth, compact surfaces wherein the Gauss-Bonnet theorem relates the Gaussian curvature of a surface to its Euler characteristic.

Author contributions

H.K., I.R. and A.S. designed the study. H.K. performed the simulations. H.K., E.S., I.R. and A.S. analyzed results. H.K., I.R. and A.S. wrote the manuscript. All authors reviewed and approved the manuscript.

[1] Carlsson, G. Topology and data. Bulletin of the American Mathematical Society **46**, 255–308 (2009).

[2] Pranav, P. et al. The topology of the cosmic web in terms of persistent betti numbers. Monthly Notices of the Royal Astronomical Society **465**, 4281–4310 (2016).

[3] Günther, D., Reininghaus, J., Hotz, I. & Wagner, H. Memory-efficient computation of persistent homology for 3d images using discrete morse theory. In 2011 24th SIBGRAPI Conference on Graphics, Patterns and Images, 25–32 (IEEE, 2011).

[4] Nicolau, M., Levine, A. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences USA **108**, 7265–7270 (2011).

[5] Morse, M. The calculus of variations in the large, vol. 18 (American Mathematical Society, 1934).

[6] Edelsbrunner, H. & Harer, J. Persistent homology-a survey. Contemporary Mathematics **453**, 257–282 (2008).

[7] Forman, R. A discrete morse theory for cell complexes. In Yau, S.-T. (ed.) Geometry, Topology and Physics for Raoul Bott (International Press of Boston, 1995).

[8] Forman, R. Morse theory for cell complexes. Advances in Mathematics **134**, 90–145 (1998).

[9] Forman, R. A user's guide to discrete morse theory. Sém. Lothar. Combin. **48**, 1–35 (2002).

[10] Watts, D. J. & Strogatz, S. H. Collective dynamics of small-world networks. Nature **393**, 440–442 (1998).

[11] Barabási, A. L. & Albert, R. Emergence of scaling in random networks. Science **286**, 509–512 (1999).

[12] Albert, R. & Barabási, A. L. Statistical mechanics of complex networks. Reviews of Modern Physics **74**, 47–97 (2002).

[13] Newman, M. E. J. Networks: An Introduction (Oxford University Press, 2010).

[14] Bianconi, G. Interdisciplinary and physics challenges of network theory. Europhysics Letters **111**, 56001 (2015).

[15] Kartun-Giles, A.P. & Bianconi, G. Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks. Chaos, Solitons and Fractals: X **1(1)**, 100004 (2019).

[16] Iacopini, I., Petri, G., Barrat, A. & Latora, V. Simplicial models of social contagion. Nature Communications **10(1)**, 2485 (2019).

[17] Ritchie, M., Berthouze, L., & Kiss, I. Generation and analysis of networks with a prescribed degree sequence and subgraph family: higher-order structure matters. Journal of Complex Networks **5(1)**, 1–31 (2017).

[18] De Silva, V. & Ghrist, R. Homological sensor networks. Notices of the American mathematical society **54** (2007).

[19] Horak, D., Maletić, S. & Rajković, M. Persistent homology of complex networks. Journal of Statistical Mechanics: Theory and Experiment P03034 (2009).

[20] Petri, G., Scolamiero, M., Donato, I. & Vaccarino, F. Topological strata of weighted complex networks. PloS One **8**, e66506 (2013).

[21] Petri, G. et al. Homological scaffolds of brain functional networks. Journal of The Royal Society Interface **11**, 20140873 (2014).

[22] Wu, Z., Menichetti, G., Rahmede, C. & Bianconi, G. Emergent complex network geometry. Scientific Reports **5**, 10073 (2015).

[23] Sizemore, A., Giusti, C. & Bassett, D. Classification of weighted networks through mesoscale homological features. Journal of Complex Networks **5**, 245–273 (2016).

[24] Courtney, O. & Bianconi, G. Weighted growing simplicial complexes. Physical Review E **95**, 062301 (2017).

[25] Courtney, O. & Bianconi, G. Dense power-law networks and simplicial complexes. Physical Review E **97**, 052303 (2018).

[26] Lee, H., Kang, H., Chung, M., Kim, B.-N. & Lee, D. Persistent brain network homology from the perspective of dendrogram. IEEE transactions on medical imaging **31**, 2267–2277 (2012).

[27] Freeman, L. C. A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977).

[28] Girvan, M. & Newman, M. Community structure in social and biological networks. Proceedings of the National Academy of Sciences USA **99**, 7821–7826 (2002).

[29] Sreejith, R.P., Mohanraj, K., Jost, J., Saucan, E. & Samal, A. Forman curvature for complex networks. Journal of Statistical Mechanics: Theory and Experiment P063206 (2016).

[30] Samal, A. et al. Comparative analysis of two discretizations of ricci curvature for complex networks. Scientific Reports **8**, 8650 (2018).

[31] Bubenik, P., Carlsson, G., Kim, P. & Luo, Z. Statistical topology via morse theory persistence and nonparametric estimation. Algebraic methods in statistics and probability II **516**, 75–92 (2010).

[32] Mischaikow, K. & Nanda, V. Morse theory for filtrations and efficient computation of persistent homology. Discrete & Computational Geometry **50(2)**, 330–353 (2013).

[33] Delgado-Friedrichs, O., Robins, V. & Sheppard, A. Morse theory and persistent homology for topological analysis of 3d images of complex materials. In 2014 IEEE International Conference on Image Processing (ICIP), 4872–4876 (IEEE, 2014).

[34] Bollobas, B. Modern Graph Theory (Springer, 1998).

[35] Zomorodian, A. & Carlsson, G. Computing persistent homology. Discrete & Computational Geometry **33**, 249–274 (2005).

[36] Munkres, J. Elements of algebraic topology (CRC Press, 2018).

[37] Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. Stability of persistence diagrams. Discrete & Computational Geometry **37**, 103–120 (2007).

[38] Kerber, M., Morozov, D. & Nigmetov, A. Geometry helps to compare persistence diagrams. J. Exp. Algorithmics **22**, 1.4:1–1.4:20 (2017).

[39] Chazal, F., Cohen-Steiner, D., Guibas, L. J. & Oudot, S., Stability of persistence diagrams revisited, INRIA Research report RR-6568 available at: `https://hal.inria.fr/inria-00292566v1/` (2008).

[40] Erdös, P. & Rényi, A. On the evolution of random graphs. Bull. Inst. Internat. Statist **38**, 343–347 (1961).

[41] Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A. & Boguná, M. Hyperbolic geometry of complex networks. Physical Review E **82**, 036106 (2010).

[42] Aldecoa, R., Orsini, C. & Krioukov, D. Hyperbolic graph generator. Computer Physics Communications **196**, 492–496 (2015).

[43] Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. Nature **411**, 41–42 (2001).

[44] Rual, J. F. et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature **437**, 1173–1178 (2005).

[45] Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (TKDD) **1**, 2 (2007).

[46] Šubelj, L. & Bajec, M. Robust network community detection using balanced propagation. European Physical Journal B **81**, 353–362 (2011).

[47] Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. Physical Review E **68**, 065103 (2003).

[48] Kunegis, J. Konect: the Koblenz network collection. In Proceedings of the 22nd International Conference on World Wide Web compani 1343–1350 (ACM, New York, NY, USA, 2013).

[49] Lewiner, T., Lopes, H. & Tavares, G. Toward optimality in discrete morse theory. Experimental Mathematics **12**, 271–285 (2003).

[50] Maria, C., Boissonnat, J.-D., Glisse, M. & Yvinec, M. The GUDHI library: Simplicial complexes and persistent homology. In International Congress on Mathematical Software, 167–174 (Springer, 2014).

[51] Ghrist, R. Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society **45**, 61–75 (2008).

[52] Dummit, D. & Foote, R. Abstract algebra (Wiley, 2003), 3 edn.

[53] Forman, R. Some applications of combinatorial differential topology. In Lyubich, M. & Takhtajan, L. (eds.) Graphs and patterns in mathematics and theoretical physics, 73, 281–313 (American Mathematical Society (AMS), 2005).

[54] Ayala, R., Fernández, L. M. & Vilches, J.A., Characterizing equivalent discrete Morse functions, Bulletin of the Brazilian Mathematical Society, New Series **40 (2)**, 225–235 (2009).

[55] Bauer, U., Lange, C. & Wardetzky, M., Optimal Topological Simplification of Discrete Functions on Surfaces, Discrete & Computational Geometry **47(2)**, 347–377 (2012).

[56] Di Fabio, B. & Ferri, M. Comparing Persistence Diagrams Through Complex Vectors. In Image Analysis and Processing — ICIAP 2015, 294–305 (Springer International Publishing, Cham, 2015).

Supplementary Information (SI)
Appendix

HOMOLOGY OF A SIMPLICIAL COMPLEX

In this section, we provide an overview of the mathematical theory of homology groups of a simplicial complex.

We first describe a mathematical group which provides the machinery to represent paths in a simplicial complex. The p-chain group C_p of a simplicial complex K is the Abelian group generated by the oriented p-simplices in K with coefficients in a field $\mathbb{F}$. Elements c of the p-chain group C_p are referred to as p-chains and have the form $c = \sum_i n_i \alpha_i$ where we use the same notation to represent both the oriented p-simplex and its corresponding generator in C_p and n_i are scalars from the field $\mathbb{F}$. Note that the identity element 0 in C_p is the unique p-chain for which all the coefficients n_i are zero in $\mathbb{F}$. Note also that if two p-simplices, α and β, have the same vertex set but opposite orientations, then the generators of C_p corresponding to the two simplices are inverse of each other (i.e., $\alpha = -\beta$). In figure 1 of the main text, the cycle of length 4 formed by the edges or 1-simplices, $[v_1, v_2]$, $[v_2, v_5]$, $[v_5, v_7]$ and $[v_1, v_7]$, can be represented as an element of the 1-chain group which is $c = [v_1, v_2] + [v_2, v_5] + [v_5, v_7] - [v_1, v_7]$.

The boundary operator ∂_p on the generator corresponding to an oriented p-simplex $\alpha = [v_0, v_1, \cdots, v_p]$ is defined as follows [36]:

$$\partial_p(\alpha) = \sum_{i=0}^{p} (-1)^i [v_0, v_1, \cdots, \hat{v}_i, \cdots, v_p], \tag{A.1}$$

where $\hat{v}_i$ refers to the absence of v_i in the $(p-1)$-simplex. By linear extension, the boundary operator ∂_p on any element $c = \sum_i n_i \alpha_i$ of p-chain group C_p gives:

$$\partial_p(c) = \sum_i n_i \partial_p(\alpha_i) \tag{A.2}$$

Note that the boundary operator in the above equation maps a p-chain to a $(p-1)$-chain. In figure 1 of the main text, the boundary operator applied to the 1-chain $c = [v_0, v_1]$ representing an edge gives $v_1 - v_0$ while the boundary operator applied to the 1-chain $c = [v_1, v_2] + [v_2, v_5] + [v_5, v_7] - [v_1, v_7]$ representing a cycle of length 4 gives 0.

This motivates the definition of p-cycles Z_p and p-boundaries B_p. The p-cycles Z_p are the elements of the p-chain group C_p which are mapped to 0 by the boundary operator ∂_p, and thus:

$$Z_p = \ker(\partial_p) = \{ c \in C_p \mid \partial_p(c) = 0 \}. \tag{A.3}$$

The p-boundaries B_p is defined as follows:

$$B_p = \mathrm{img}(\partial_{p+1}) = \{ c \in C_p \mid \exists\, b \in C_{p+1}, \partial_{p+1}(b) = c \}. \tag{A.4}$$

Thus, p-boundaries B_p are the elements of the p-chain group C_p which also happen to be the boundary of an element in the $(p+1)$-chain group C_{p+1}. Note that both Z_p and B_p are subgroups of the p-chain group C_p. It can be shown

that the composition $\partial_p \circ \partial_{p+1} = 0$. This implies that B_p is a subgroup of Z_p [36]. Simply stated, the p-boundary operator ∂_p when applied on a p-boundary gives 0. Hence, it follows that every p-boundary is a p-cycle but not necessarily vice versa.

The p-homology group is defined as [36]:

$$H_p = Z_p/B_p, \tag{A.5}$$

where Z_p/B_p is the quotient group [52] of Z_p over B_p. We informally refer to the elements of p-homology group H_p as p-holes. To provide an intuition for the definition of homology groups, a natural way to describe a p-hole would be to characterize it as a p-cycle which is not a p-boundary. In figure 1 of the main text, the 1-cycle $c = [v_1, v_2] + [v_2, v_5] + [v_5, v_7] - [v_1, v_7]$ is a 1-hole as it is not a 1-boundary of any 2-chain while the 1-cycle $c = [v_2, v_3] + [v_3, v_4] + [v_4, v_2]$ is not a 1-hole as it is the 1-boundary of the 2-simplex $[v_2, v_3, v_4]$. Thus, the concept of quotient groups provide the mathematical machinery to characterize such p-holes in simplicial complexes.

We remark that though it is customary to call H_p a group, since we use coefficients from a field $\mathbb{F}$, it satisfies additional properties which make it a vector space over $\mathbb{F}$. The p-Betti number is defined as the dimension of the homology group H_p viewed as a vector space over field $\mathbb{F}$ [36]. Informally, the p-Betti number β_p represents the number of p-holes of the simplicial complex. We remark that the Euler characteristic of the clique complex K with dimension d corresponding to a graph G is given by the alternating sum of Betti numbers [36], namely,

$$\chi(K) = \beta_0 - \beta_1 + \beta_2 - \cdots + (-1)^d \beta_d. \tag{A.6}$$

In this work, we use the finite field $\mathbb{F} = \mathbb{Z}/2$, i.e., the field with two elements.

PROOF OF CONCEPT FOR ALGORITHM 1

In this section, we present a rigorous proof for the following theorem.

Theorem 1. *Algorithm 1 produces a discrete Morse function f on any simplicial complex K of finite dimension d.*

Proof. Let K denote a simplicial complex of dimension d. Recall that for the function f on K which is constructed by algorithm 1, for each simplex $\alpha^p \in K$, the two sets U_α and V_α are defined as follows:

$$U_\alpha = \{\beta^{p+1} \mid \alpha^p < \beta^{p+1} \text{ and } f(\beta) \leq f(\alpha)\}$$
$$V_\alpha = \{\gamma^{p-1} \mid \gamma^{p-1} < \alpha^p \text{ and } f(\alpha) \leq f(\gamma)\}$$

To prove that f is a discrete Morse function we need to show that for each simplex $\alpha \in K$, both $|V_\alpha| \leq 1$ and $|U_\alpha| \leq 1$.

Firstly, we show that for each simplex $\alpha \in K$, $|V_\alpha| \leq 1$. Consider a 0-simplex $\alpha \in K$. Since, the dimension of a simplex cannot be less than 0, for each 0-simplex $\alpha \in K$ the corresponding set V_α is empty. In other words, for each 0-simplex $\alpha \in K$, we have that $|V_\alpha| = 0$. Also, for each p-simplex $\alpha^p \in K$ such that $1 \leq p \leq d$, Lemma 6 below shows that $|V_\alpha| \leq 1$. Thus, for every simplex $\alpha \in K$, we have shown that $|V_\alpha| \leq 1$.

Secondly, we show that for each simplex $\alpha \in K$, $|U_\alpha| \leq 1$. For each p-simplex $\alpha^p \in K$ such that $0 \leq p \leq (d-1)$, we prove in Lemma 5 below that $|U_\alpha| \leq 1$. Now consider a d-simplex $\alpha^d \in K$. Since, by assumption K is a d-dimensional simplicial complex, there are no $(d+1)$-simplices in K, and thus, the set U_α for each d-simplex α^d in K is empty. In other words, for each d-simplex $\alpha^d \in K$, we have that $|U_\alpha| = 0$. Thus, for every simplex $\alpha \in K$ we have shown that $|U_\alpha| \leq 1$.

In summary, we have shown that for each simplex $\alpha \in K$, $|V_\alpha| \leq 1$ and $|U_\alpha| \leq 1$. Thus, f satisfies the definition of a discrete Morse function on the simplicial complex K.

$\square$

We next prove the Lemmas used in the proof of the theorem above. This is done in the following sequence of Lemmas 2 to 6. We assume that K is a d-dimensional simplicial complex and f is the output function on K obtained from algorithm 1. We *remark* that a p-dimensional simplex α of a simplicial complex K is denoted by $\alpha^p \in K$. Also, if p-simplex α is a face of a $(p+1)$-simplex β then this is represented as $\alpha^p < \beta^{p+1}$ in the sequel.

Lemma 2. *For each p where $0 \leq p \leq (d-1)$, if $\alpha^p \in K$ and $\beta^{p+1} \in K$ such that $\alpha^p < \beta^{p+1}$, then, $f(\alpha) \neq f(\beta)$.*

Proof. Let $\gamma_0, \gamma_1, \cdots, \gamma_{p+1}$ denote the p-dimensional faces of β^{p+1} such that $f(\gamma_0) \geq f(\gamma_1) \geq \cdots \geq f(\gamma_{p+1})$. Note that α^p is one such γ_i since by assumption α^p is a p-dimensional face of β^{p+1}. Based on lines 11-23 in algorithm 1, we have that:

$$f(\beta) = \begin{cases} (f(\gamma_0) + f(\gamma_1))/2 & \text{if } \texttt{Flag}[\gamma_0] = 0 \text{ and } f(\gamma_0) > f(\gamma_1) \quad (\text{Case } A) \\ f(\gamma_0) + \epsilon & \text{otherwise} \hspace{3.5cm} (\text{Case } B) \end{cases} \tag{A.7}$$

where $\epsilon > 0$.

Case A implies $f(\gamma_0) > f(\beta) > f(\gamma_1) \geq f(\gamma_2) \cdots \geq f(\gamma_p) \geq f(\gamma_{p+1})$.

Case B implies $f(\beta) > f(\gamma_0) \geq f(\gamma_1) \geq f(\gamma_2) \cdots \geq f(\gamma_p) \geq f(\gamma_{p+1})$.

Thus, for both cases we have that $f(\beta) \neq f(\gamma_i)$ for each $i \in \{0, 1, 2, \cdots, (p+1)\}$. Since α^p is one such γ_i for some $i \in \{0, 1, 2, \cdots, (p+1)\}$, we have $f(\alpha) \neq f(\beta)$.

$\square$

Lemma 3. *For each p where $0 \leq p \leq (d-1)$, if $\alpha^p \in K$ and $\beta^{p+1} \in K$ such that $\alpha^p < \beta^{p+1}$, then $f(\alpha) > f(\beta)$ if and only if $\texttt{Flag}[\alpha]$ changes value from 0 to 1 while assigning function value for β.*

Proof. Given $\alpha^p \in K$ and $\beta^{p+1} \in K$ such that $\alpha^p < \beta^{p+1}$, we first assume $f(\alpha) > f(\beta)$. Let $\gamma_0, \gamma_1, \cdots, \gamma_{p+1}$ denote the p-dimensional faces of β^{p+1} such that $f(\gamma_0) \geq f(\gamma_1) \geq \cdots \geq f(\gamma_{p+1})$. Then, the value $f(\beta)$ is given by equation A.7.

Case A in equation A.7 implies $f(\gamma_0) > f(\beta) > f(\gamma_1) \geq f(\gamma_2) \cdots \geq f(\gamma_p) \geq f(\gamma_{p+1})$.

Case B in equation A.7 implies $f(\beta) > f(\gamma_0) \geq f(\gamma_1) \geq f(\gamma_2) \cdots \geq f(\gamma_p) \geq f(\gamma_{p+1})$.

Since by assumption, α^p is a face of β^{p+1} and $f(\alpha) > f(\beta)$, Case A is applicable, and we have γ_0 equals α^p. Thus, based on line 18 in algorithm 1, while assigning the function value on β^{p+1}, $\texttt{Flag}[\alpha]$ changes value from 0 to 1.

Now, given $\alpha^p \in K$ and $\beta^{p+1} \in K$ such that $\alpha^p < \beta^{p+1}$, we assume that $\texttt{Flag}[\alpha]$ changes value from 0 to 1 while assigning function value for β^{p+1}. Let $\gamma_0, \gamma_1, \cdots, \gamma_{p+1}$ denote the p-dimensional faces of β^{p+1} such that $f(\gamma_0) \geq f(\gamma_1) \cdots \geq f(\gamma_{p+1})$. Based on lines 11-23 in algorithm 1, $\texttt{Flag}[\alpha]$ changes value from 0 to 1 while assigning function value on β^{p+1} implies that γ_0 equals α^p. Thus, we have that $f(\alpha) > f(\beta) = (f(\alpha) + f(\gamma_1))/2$.

$\square$

Lemma 4. *Let α be a simplex of K. Then, the number of times $\texttt{Flag}[\alpha]$ changes value from 0 to 1 is ≤ 1.*

Proof. $\texttt{Flag}[\alpha]$ is initially set to 0 in algorithm 1. In algorithm 1, if $\texttt{Flag}[\alpha]$ transitions to 1, its value never changes. In other words, there is no procedure in our algorithm 1 which changes the $\texttt{Flag}$ of a simplex from 1 to 0. Thus, either $\texttt{Flag}[\alpha]$ remains 0 throughout or changes value from 0 to 1 exactly once in algorithm 1.

$\square$

Lemma 5. *For each p where $0 \leq p \leq (d-1)$, if $\alpha^p \in K$, then $|U_\alpha| \leq 1$.*

Proof. From Lemma 3, we have that for each $\alpha^p \in K$, the number of $(p+1)$-simplices β^{p+1} such that $\alpha^p < \beta^{p+1}$ and $f(\beta) < f(\alpha)$ is equal to the number of times $\texttt{Flag}[\alpha]$ changes value from 0 to 1. Applying Lemma 4, we get that for each $\alpha^p \in K$,

$$\#\{\beta^{p+1} \mid \alpha^p < \beta^{p+1} \text{ and } f(\beta) < f(\alpha)\} \leq 1.$$

Furthermore, Lemma 2 tells us that,

$$\#\{\beta^{p+1} \mid \alpha^p < \beta^{p+1} \text{ and } f(\beta) < f(\alpha)\} = \#\{\beta^{p+1} \mid \alpha^p < \beta^{p+1} \text{ and } f(\beta) \leq f(\alpha)\}$$

Thus, we have that for each $\alpha^p \in K$, $|U_\alpha| = \#\{\beta^{p+1} \mid \alpha^p < \beta^{p+1} \text{ and } f(\beta) \leq f(\alpha)\} \leq 1$.

$\square$

Lemma 6. *For each p where $1 \leq p \leq d$, if $\alpha^p \in K$, then $|V_\alpha| \leq 1$.*

Proof. Let $\gamma_0, \gamma_1, \cdots, \gamma_p$ denote the $(p-1)$-dimensional faces of $\alpha^p \in K$ such that $f(\gamma_0) \geq f(\gamma_1) \geq \cdots \geq f(\gamma_p)$. Based on lines 11-23 in algorithm 1, we have that:

$$f(\alpha) = \begin{cases} (f(\gamma_0) + f(\gamma_1))/2 & \text{if } \texttt{Flag}[\gamma_0] = 0 \text{ and } f(\gamma_0) > f(\gamma_1) \quad (\text{Case } A) \\ f(\gamma_0) + \epsilon & \text{otherwise} \hspace{3.5cm} (\text{Case } B) \end{cases}$$

where $\epsilon > 0$.

Case A implies $f(\gamma_0) > f(\alpha) > f(\gamma_1) \geq f(\gamma_2).... \geq f(\gamma_{p-1}) \geq f(\gamma_p)$, and thus, $|V_\alpha| = 1$.

Case B implies $f(\alpha) > f(\gamma_0) \geq f(\gamma_1) \geq f(\gamma_2).... \geq f(\gamma_{p-1}) \geq f(\gamma_p)$, and thus, $|V_\alpha| = 0$.
Hence, for each $\alpha^p \in K$ with $1 \leq p \leq d$, we have that $|V_\alpha| \leq 1$.

$\square$

FILTRATION OF THE CLIQUE COMPLEX BASED ON WEIGHTS OF CRITICAL SIMPLICES

Given a simplicial complex K, its dimension d and a discrete Morse function f on K, the algorithm 2 determines the weights of critical simplices in K. In the pseudocode of the algorithm 2, lines 2-6 initialize a variable `IsCritical`$[\alpha]$ associated to every simplex α in clique complex K to be True. Lines 7-17 determine the critical simplices in K by checking for the condition in equation 7 of the main text which defines a critical simplex. Lines 18-28 determine the weights of critical simplices or critical weights in K. Finally, the algorithm 2 outputs an array $w_c[]$ which contains an increasing sequence of critical weights in K. Subsequently, this increasing sequence of critical weights will be used for the filtration of the clique complex K.

Given an unweighted and undirected graph G, we restrict the construction of clique complex K by including simplices up to a maximum dimension d. Then, the algorithm 3 creates the filtration of clique complex K based on weights of critical simplices as described in the Theory section of the main text. In the pseudocode of the algorithm 3, lines 2-6 assigns a non-negative function g to 0-simplices in clique complex K. Line 7 calls the algorithm 1 for the assignment of weights satisfying discrete Morse function to every simplex in K. Line 8 calls the algorithm 2 to obtain an increasing sequence of unique weights corresponding to critical simplices in K. Lines 9-11 initialize a variable `IsAdded`$[\alpha]$ associated to every simplex α in K which tracks whether the simplex α has been added to the filtration or not. Lines 12-31 compute the filtration weight of each simplex α in K as described in the Theory section of the main text.

In SI Table S1, we describe the role of key variables which appear in algorithms 2 and 3.

COMPUTATIONAL ASPECTS OF OUR ALGORITHMS

Time complexity of Algorithm 1

We briefly discuss here some computational aspects of the algorithm 1 in the main text. Given a simplicial complex K, its dimension d and a non-negative real-valued function g on the 0-simplices of K, algorithm 1 assigns weights to any simplex in K, producing a discrete Morse function f on K.

We remark that the algorithms 1, 2 and 3 were implemented using the C++ programming language. The simplices of a d-dimensional simplicial complex were stored using the set container which is a part of the C++ Standard Template Library (STL). The association of the discrete Morse function values to the corresponding simplices was implemented using the map container of the C++ STL. Implementing the storage of the `Flag` variable associated with each simplex was also done using the map container of C++ STL. We remark that accessing a particular value of a map container has a complexity which is logarithmic with respect to the size of the container, namely, the number of simplices n.

Assuming that the dimension d of K is constant, we here present a brief discussion about the complexity of algorithm 1 with respect to the number of simplices n in K. We remark that since the dimension d is constant and the complexity of accessing values using map containers is logarithmic with respect to n, lines 2-9 in algorithm 1 have a complexity of $O(n \log n)$. The complexity of the operation corresponding to line 12 in algorithm 1, namely, finding the $(p-1)$-dimensional faces of a p-simplex is independent of n. Such an operation depends on p which has an upper bound of d that has been assumed to be a constant.

The operation in line 13 of algorithm 1 corresponds to arranging the simplices computed in the preceding line based on their function values. As noted earlier, accessing values using map containers is logarithmic with respect to n. Thus, since $p \leq d$ and d is fixed, the operation represented by line 13 of algorithm 1 has a complexity of $O(\log n)$. Lines 14-22 of algorithm 1 except the random number generation in line 20 of algorithm 1 require accessing values of map containers. Random number generation based on a uniform distribution using C++ can be implemented with amortized constant complexity. Thus, the block of pseudocode represented by lines 14-22 in algorithm 1 have complexity of $O(\log n)$. Since d is constant and for any p the number of p-simplices is bounded by the total number of simplices n, the pseudocode represented in lines 10-24 of algorithm 1 has a complexity of $O(n \log n)$. Hence the complexity of algorithm 1 with respect to the number of simplices n can be described as $O(n \log n)$.

Algorithm 2 Algorithm to compute the weights of critical simplices in K corresponding to the discrete Morse function f

---

1: **function** GETCRITICALWEIGHTS(K, d, f)

2:     **for** $p = 0, \cdots, d$ **do**   ▷ Initialize `IsCritical` variable associated with each simplex in K
3:         **for** each p-simplex $\alpha \in K$ **do**
4:             `IsCritical`$[\alpha]=$ True   ▷ `IsCritical` indicates whether a given simplex is critical or not
5:         **end for**
6:     **end for**

7:     **for** $p = 1, \cdots, d$ **do**   ▷ Determine the critical simplices in K
8:         **for** each p-simplex $\alpha \in K$ **do**
9:             Let `Faces`[ ] be an array of all $(p-1)$-dimensional faces of α
10:             Sort `Faces`[ ] such that $f(\texttt{Faces}[i]) \geq f(\texttt{Faces}[i+1])$ for each $i \in \{0, 1, \cdots, p-1\}$
11:             Let $\gamma_0 = $ `Faces`$[0]$
12:             **if** $f(\gamma_0) \geq f(\alpha)$ **then**
13:                 `IsCritical`$[\alpha] = $ False
14:                 `IsCritical`$[\gamma_0] = $ False
15:             **end if**
16:         **end for**
17:     **end for**

18:     Initialize $i = 0$
19:     Declare empty array w_c[ ]
20:     **for** $p = 0, \cdots, d$ **do**   ▷ Determine the weights of critical simplices in K
21:         **for** each p-simplex $\alpha \in K$ **do**
22:             **if** `IsCritical`$[\alpha] = $ True **then**
23:                 $w_c[i] = f(\alpha)$
24:                 $i = i + 1$
25:             **end if**
26:         **end for**
27:     **end for**
28:     Sort array w_c[ ] in increasing order and remove duplicates

29:     **return** w_c[ ]

30: **end function**

Time complexity of algorithms 1, 2 and 3 with respect to the number of critical simplices

Firstly, it is apparent that the complexity of algorithm 1 is independent of the number of critical simplices since the notion of critical simplices is relevant only after the discrete Morse function has been constructed.

Algorithm 2 accepts as inputs a simplicial complex K, its dimension d and a discrete Morse function f on K and outputs the sorted sequence of weights which correspond to the critical simplices in K. Let m denote the total number of critical simplices in K corresponding to f. We assume henceforth that the simplicial complex K and its dimension d are fixed. This implies that the number of simplices n and the dimension d are treated as constants. In such a situation, the complexity of all lines in the pseudocode of algorithm 2 except for line 28 is constant and independent of m. Line 28 of algorithm 2 represents sorting the array which contains the function values of all critical simplices, i.e., the critical weights, and removing any duplicate entries. The number of critical weights is less than or equal to the number of critical simplices m, and thus, the operation represented by line 28 of algorithm 2 can be implemented with a complexity of $O(m^2)$. Hence, assuming that the simplicial complex K and its dimension d are fixed, algorithm 2 can be implemented with a complexity of $O(m^2)$ in respect to the number of critical simplices m.

Algorithm 3 accepts a graph as input and determines the filtration weights of the simplices in the d-dimensional clique complex K of G. Similar to the analysis of algorithm 2, we assume that the graph G and the dimension d of K is fixed which implies that the number of simplices in the d-dimensional clique complex K of G is also constant. In such a situation, lines 1-7 of algorithm 3 have constant complexity which is independent of the number of critical simplices m. As seen in the preceding paragraph, provided K and d are fixed, the complexity of algorithm 2 and hence the complexity of line 8 of algorithm 3 is $O(m^2)$. Lines 9-11, lines 13-24 and lines 26-31 in the pseudocode of

---

Algorithm 3 Algorithm to determine the filtration weights of the simplices in the clique complex of a simple graph G

1:   $K = d$-dimensional clique complex of graph G

2:   deg_{max} = Maximum degree of a vertex in graph G
3: **for** each 0-simplex $\alpha \in K$ **do**   ▷ Assign non-negative function g to 0-simplices in K
4:     $\epsilon = \text{random}(0, 0.5)$ ▷ ϵ is generated randomly during runtime using a uniform distribution on $(0, 0.5)$
5:     $g[\alpha] = deg_{max}$ - degree$(\alpha) + \epsilon$
6: **end for**

7:   $f = \text{DISCETEMORSEFUNCTION}(K, d, g)$ ▷ Call Algorithm 1
8:   $w_c[\,] = \text{GETCRITICALWEIGHTS}(K, d, f)$ ▷ Call Algorithm 2

9:   **for** each simplex $\alpha \in K$ **do**   ▷ Initialize `IsAdded` variable associated with each simplex in K
10:    `IsAdded`$[\alpha] = False$ ▷ `IsAdded` indicates whether a given simplex has been added to the filtration.
11: **end for**

12: **for** $i = 0, \cdots, len(w_c[\,]) - 1$ **do**   ▷ Calculate Filtration weight for each simplex in K
13:    **for** each simplex $\alpha \in K$ **do**
14:      **if** $f(\alpha) \le w_c[i]$ AND `IsAdded`$[\alpha] = False$ **then**
15:        `FiltrationWeight`$[\alpha] = w_c[i]$
16:        `IsAdded`$[\alpha] = True$
17:        **for** each face $\gamma < \alpha$ **do**
18:          **if** `IsAdded`$[\gamma] = False$ **then**
19:            `FiltrationWeight`$[\gamma] = w_c[i]$
20:            `IsAdded`$[\gamma] = True$
21:          **end if**
22:        **end for**
23:      **end if**
24:    **end for**
25: **end for**

26: **for** each simplex $\alpha \in K$ **do**
27:    **if** `IsAdded`$[\alpha] = False$ **then**
28:      `FiltrationWeight`$[\alpha] = w_c[len(w_c[\,]) - 1]$
29:      `IsAdded`$[\alpha] = True$
30:    **end if**
31: **end for**

---

algorithm 3 have constant complexity which is independent of the number of critical simplices m. The number of times the operations of lines 13-24 in algorithm 3 are repeated is equal to the number of critical weights which is less than or equal to the number of critical simplices m. Thus, assuming that the graph G and the dimension d of the corresponding clique complex K are fixed, algorithm 3 can be implemented with a complexity of $O(m^2)$ in respect to the number of critical simplices m.

THEORETICAL RESULTS ON STABILITY OF PERSISTENT HOMOLOGY AND PERSISTENCE DIAGRAMS OF DISCRETE MORSE FUNCTIONS

Stability of persistent homology

In Algorithm 1, some choices are made in assigning values to every simplex in order to get a discrete Morse function. A natural question arises which is the following. Under what conditions do two discrete Morse functions give the same persistent homology groups. We remark that in Forman's theory [9], the actual values of the discrete Morse functions are less important than the *gradient vector field* that the discrete Morse function induces. For the definition of a gradient vector field, we refer the reader to Definition 3.3 in Forman's [9] article. In his expository paper [9], Forman asserts on page 15 that: *"In fact, this gradient vector field contains all of the information that we will need to know about the function for most applications."*

Note that two Morse functions f and g induce the same gradient vector field if and only if they satisfy the following condition (see Forman [53] or Theorem 3.1 in Ayala *et al.*[54]):

C0 : $f(\alpha) < f(\beta) \Leftrightarrow g(\alpha) < g(\beta)$ for a p-simplex α and a $(p+1)$-simplex β such that $\alpha^p < \beta^{p+1}$

We show that for two discrete Morse functions to have the same filtration given by their level subcomplexes on critical weights, the actual values of the discrete Morse function are not as important as the *order relation* induced by the function on the simplices. The discrete Morse functions with the same filtration clearly give the same persistent homology groups. In particular, we show below that a sufficient condition for this to happen is that both functions have the same order relation with respect to the values assigned to the simplices.

Lemma 7. *Let f and g be two discrete Morse functions on a finite simplicial complex K. Assume that the following condition is satisfied:*

C1 : $f(\alpha) \leq f(\beta) \Leftrightarrow g(\alpha) \leq g(\beta)$ for any pair of simplices $\alpha, \beta \in K$.

Then f and g have the same set of critical simplices $C := \{c_1, c_2, \cdots, c_k\}$, and the level subcomplex of f corresponding to the critical cell c_i is the same as the level subcomplex of g corresponding to c_i, for any $i = 1, 2, \cdots, k$.

Proof. The condition **C1** guarantees that the functions f and g have the same gradient vector field, since **C1** implies **C0**. Therefore they have the same set of critical simplices C. Denote the level subcomplex of f (respectively, of g) corresponding to the critical cell c_i by $K^f(c_i)$ (respectively, by $K^g(c_i)$). Now, if $\alpha \in K^f(c_i)$, then there exists $\beta \in K$ such that $\alpha \leq \beta$ and $f(\beta) \leq f(c_i)$. Note that $\alpha \leq \beta$ represents that either two simplices α and β are same or α is a face of β. By assumption **C1** this implies that $g(\beta) \leq g(c_i)$, thus $\alpha \in K^g(c_i)$. So we get $K^f(c_i) \subseteq K^g(c_i)$. Reversing the roles of f and g in the above argument, we also get $K^g(c_i) \subseteq K^f(c_i)$. Thus $K^f(c_i) = K^g(c_i)$ for any critical cell $c_i \in C$. $\square$

A simple example of two discrete Morse functions f and g which satisfy **C1** is the following. Starting with a discrete Morse function f obtain g by *sliding* the values of f by a fixed real number r. A slightly more sophisticated example of two functions which satisfy **C1** is the following. Given a discrete Morse function f on a finite simplicial complex K, one can arrange the values of f on the simplices of K in non-decreasing order. Suppose that α and β are simplices of K which are adjacent in this ordering, i.e., we have $f(\alpha) \leq f(\beta)$ and no other simplex σ exists such that $f(\sigma)$ lies in the interval $[f(\alpha), f(\beta)]$, and let us call such pairs of simplices (α, β) as f-*successive*. If for each f-successive pair of simplices (α, β), one adds a non-negative number $\delta(\alpha, \beta)$ to $f(\alpha)$ such that $\delta(\alpha, \beta) \leq f(\beta) - f(\alpha)$, we obtain a new discrete Morse function g given by $g(\alpha) := f(\alpha) + \delta(\alpha, \beta)$, for which any pair (α, β) is f-successive if and only if it is g-successive. It is easy to deduce that f and g then satisfy the condition **C1**. Thus, f and g will have the same persistent homology groups.

<h3 align="center">Stability of persistence diagrams</h3>

We remark that two discrete Morse functions may have the same filtration (and thus the same persistent homology groups) but their persistence diagrams may differ if their critical weights (i.e., values on the critical simplices) are different. A sufficient condition that assures that two discrete Morse functions f and g have the same persistence diagrams is the following, keeping the notations of Lemma 7:

C2: in addition to **C1**, we have $f(c_i) = g(c_i)$ for any $c_i \in C$

Even if the condition **C2** is not satisfied, there are stability theorems which give conditions under which the bottleneck distance between the respective persistence diagrams does not change much, i.e., the persistence diagrams are *close* in the bottleneck metric. The first such stability theorem was given by Cohen-Steiner *et al.* [37]. This result [37] is not directly applicable to discrete Morse functions, as they are not continuous. However, we show below that stability theorems for discrete Morse functions for finite regular CW-complexes can be inferred from stability results in Chazal *et al.* [39] (see also Bauer *et al.*[55]). For simplicity we give the arguments for discrete Morse functions on a finite simplicial complex, although all the results below are also valid for finite regular CW-complexes.

Recall the ∞-Wasserstein distance or bottleneck distance between two multisets X and Y in $\mathbb{R}^2$:

$$W_\infty(X, Y) = \inf_{\eta : X \to Y} \sup_{x \in X} ||x - \eta(x)||_\infty. \tag{A.8}$$

In the above equation, the supremum is taken over all bijections $\eta : X \to Y$ (with the convention that a point with multiplicity $k \in \mathbb{N}$ is considered as k individual points) and for $(a, b) \in \mathbb{R}^2$, $||(a, b)||_\infty := \max\{|a|, |b|\}$.

For a discrete Morse function f with critical values $w_i := f(c_i), i = 1, 2, \cdots, n$, the kth persistence diagram $D^k f$ (or simply Df when the value of k is fixed), is defined as follows. Consider the multiset of points $W_f^k := \{(w_i, w_j) : w_i < w_j, i, j = 1, 2, \cdots, n\}$ with each point (w_i, w_j) endowed with the multiplicity $\mu_k(w_i, w_j)$ given by (see e.g. Di Fabio *et al.* [56]):

$$\mu_k(w_i, w_j) := \lim_{\epsilon \to 0^+} (\beta_{w_i+\epsilon}^{w_j-\epsilon} - \beta_{w_i+\epsilon}^{w_j+\epsilon} + \beta_{w_i-\epsilon}^{w_j+\epsilon} - \beta_{w_i-\epsilon}^{w_j-\epsilon})$$

where $\beta_x^y := \mathrm{rank}(H_k(K^f(x)) \to H_k(K^f(y)))$ for $x, y \in \mathbb{R}$ with $x < y$. Denote by Δ the diagonal in $\mathbb{R}^2$ considered as a multiset with infinite multiplicity given to each of its points.

Definition 8. *The persistence diagram $D^k f$ is the subset of $W_f^k \cup \Delta$ consisting of points (u, v) with $\mu_k(u, v) > 0$.*

We state below the stability results from Chazal *et al.* [39] that we shall need, and to do so, we recall the concepts of *persistence modules* and *ϵ-interleaving*:

Definition 9. (Persistence modules (Definition 2.2. in Chazal *et al.* [39])) *Let R be a commutative ring with unity. A collection $\mathcal{F} = \{F^\alpha\}_{\alpha \in \mathbb{R}}$ of R-modules F^α together with homomorphisms $\phi_{\alpha\beta} : F^\alpha \to F^\beta$ for $\alpha \le \beta$, is called a persistence module if*

$$\forall \alpha \le \beta \le \gamma, \qquad \phi_{\alpha\alpha} = \mathrm{id}_{F^\alpha} \text{ and } \qquad \phi_{\alpha\beta} \circ \phi_{\beta\gamma} = \phi_{\alpha\gamma}$$

A persistence module is called δ-tame if for any $\alpha < \alpha + \delta < \beta$ one has $\mathrm{rank}\, \phi_{\alpha\beta} < \infty$.

Definition 10. (ϵ-interleaving of functions (Definition 4.1 in Chazal *et al.* [39])) *Let $\epsilon > 0$, and let f and g be two real-valued functions on a finite simplicial complex K. For $x \in \mathbb{R}$, denote the sub-level sets $f^{-1}(-\infty, x]$ of f (respectively, g) by F^x (respectively, G^x). The functions f and g are said to be strongly ϵ-interleaved if for any $a \in \mathbb{R}$ and $n \in \mathbb{Z}$, one has*

$$F^{a+2n\epsilon} \subseteq G^{a+(2n+1)\epsilon} \subseteq F^{a+(2n+2)\epsilon}$$

The main result for ϵ-interleaved functions that we shall need is as follows:

Lemma 11. (Lemma 4.2(iii) in Chazal *et al.* [39]) *Given $\epsilon > 0$, two functions f and g are strongly ϵ-interleaved if and only if $||f - g||_\infty < \epsilon$.*

Definition 12. (ϵ-interleaving of persistence modules (Definition 4.3 in Chazal *et al.* [39])) *Let $\mathcal{F} = \{F^\alpha\}_{\alpha \in \mathbb{R}}$ and $\mathcal{G} = \{G^\alpha\}_{\alpha \in \mathbb{R}}$ be persistence modules over a commutative unital ring R. Let $\epsilon > 0$, then $\mathcal{F}$ and $\mathcal{G}$ are said to be strongly ϵ-interleaved if for all $a \in \mathbb{R}$ and $n \in \mathbb{Z}$, there exist homomorphisms $\Phi_{a+2n\epsilon} : F^{a+2n\epsilon} \to G^{a+(2n+1)\epsilon}$ and $\Psi_{a+(2n+1)\epsilon} : G^{a+(2n+1)\epsilon} \to F^{a+(2n+2)\epsilon}$ such that the following diagram commutes:*

$$\cdots \longrightarrow F^{a+2n\epsilon} \longrightarrow F^{a+(2n+1)\epsilon} \longrightarrow F^{a+(2n+2)\epsilon} \longrightarrow \cdots$$
$$\cdots \longrightarrow G^{a+2n\epsilon} \longrightarrow G^{a+(2n+1)\epsilon} \longrightarrow G^{a+(2n+2)\epsilon} \longrightarrow \cdots$$

If f and g are strongly ϵ-interleaved, then the associated persistence modules given by their k^{th} singular homology groups are also strongly ϵ-interleaved for any $k \in \mathbb{N} \cup \{0\}$, the associated homomorphisms in Definition 10 are induced by the corresponding inclusion maps. Note that these persistence modules are 0-tame since the functions are defined on a finite simplicial complex K.

One can define persistence diagrams for δ-tame persistence modules (see Definitions 3.6 and 3.10 in Chazal *et al.* [39]). Denote the persistence diagram of a δ-tame persistence module $\mathcal{F}$ by $D_\delta \mathcal{F}$. The persistence diagram of persistence modules generalizes the classical persistence diagrams associated with real-valued functions, i.e., if f is a tame real-valued function on a finite simplicial complex K, the associated 0-tame persistence module given by the kth singular homology groups of its sublevel sets F^α, denoted $\mathcal{F}_k := \{H_k(F^\alpha)\}_{\alpha \in \mathbb{R}}$, has the same persistence diagram as the kth persistence diagram of f, i.e. $D^k f = D_0 \mathcal{F}_k$ (see Remark 1, page 12 in Chazal *et al.* [39] for a short proof).

The next lemma shows that if f is a discrete Morse function and $K^f(\alpha), \alpha \in \mathbb{R}$ are the level subcomplexes of f, then the persistence diagram of the 0-tame persistence module $\mathcal{K}_k = \{H_k(K^f(\alpha))\}_{\alpha \in \mathbb{R}}$ is the same as the kth

persistence diagram $D^k f$ of f corresponding to the filtration of K given by the level subcomplexes at critical weights, i.e., $D^k f = D_0 \mathcal{K}_k$. The proof is essentially the same as the arguments given in Remark 1, page 12 of Chazal *et al.* [39], except that we use Forman's Lemma 2.6 [9] for comparing homology groups across critical weights. Note that due to Forman's Lemma 2.6 [9], the set of k-homological critical values of f, i.e., points $c \in \mathbb{R}$ such that for any $\epsilon > 0$ the homomorphism on the k-homology groups induced by inclusion map $K^f(c - \epsilon) \hookrightarrow K^f(c + \epsilon)$ is *not* an isomorphism, is a subset of its critical weights in general. However, it is easy to see that only those points (w_i, w_j) appear in the persistence diagram of f for which both w_i and w_j are k-homological critical values of f (otherwise the multiplicity of such points is zero).

Lemma 13. *We have $D^k f = D_0 \mathcal{K}_k$.*

Proof. Let k be fixed, to simplify the notation we denote the k-persistence diagrams of f and $\mathcal{K}_k$ by Df and $D\mathcal{K}$, respectively. From the definition of the persistence diagrams of persistence modules (Definition 3.6 in Chazal *et al.* [39]), it suffices to show that there exists $\epsilon > 0$ $x \in (0, \epsilon)$, such that $W_\infty(Df, D\mathcal{K}_{\frac{\epsilon}{2^n}, x}) < \epsilon/2^n$ for any $n \geq 1$, where $D\mathcal{K}_{\frac{\epsilon}{2^n}, x}$ is the multi-subset of $\mathbb{R}^2$ given by points of non-zero multiplicity in the union of the grid

$$G_{\epsilon/2^n, x} := \{(x + \frac{m\epsilon}{2^n}, x + \frac{m'\epsilon}{2^n}) \in \mathbb{R}^2, m, m' \in \mathbb{Z} : m' > m\}$$

and the diagonal Δ, where the multiplicity $\mu(m, m')$ of a point $(x + \frac{m\epsilon}{2^n}, x + \frac{m'\epsilon}{2^n}) \in G_{\epsilon/2^n, x}$ is given by the following formula from Definition 3.1(ii) in Chazal *et al.* [39]:

$$\mu(m, m') := \mathrm{rank}(H(m) \to H(m' - 1)) - \mathrm{rank}(H(m) \to H(m'))$$
$$+ \mathrm{rank}(H(m - 1) \to H(m')) - \mathrm{rank}(H(m - 1) \to H(m' - 1))$$

where we have used the notation $H(m) := H_k(K^f(x + \frac{m\epsilon}{2^n})), m \in \mathbb{Z}$ for simplicity.

Since the number of critical weights is finite, one can choose $\epsilon > 0$, $x \in (0, \epsilon]$ such that none of the grids $G_{\epsilon/2^n, x}$ have a point outside the diagonal that coincides with a point in Df. Then, for each $n \geq 1$ a point $(w_i, w_j) \in Df$ is contained in exactly one of the cells C in the grid $G_{\epsilon/2^n, x}$. It is easy to check using Forman's Lemma 2.6 [9] that the multiplicity $\mu(w_i, w_j)$ is then equal to the multiplicity of the upper-right corner of C, while all other off diagonal points in $D\mathcal{K}$ which are not upper-right corners of cells containing a point of Df, have multiplicity zero. Therefore the bijection $\Phi_{\epsilon, n} : Df \to D\mathcal{K}_{\epsilon/2^n, x}$ which maps (w_i, w_j) in Df to the upper right corner of the unique cell in $G_{\epsilon/2^n, x}$ containing (w_i, w_j) and any point in the region $\{(x, y) \in \mathbb{R}^2 : y \in [x, x + \epsilon/2^n]\}$ to the nearest point on the diagonal, moves points by a distance at most $\frac{\epsilon}{2^n}$. Thus, we get $W_\infty(Df, D\mathcal{K}_{\frac{\epsilon}{2^n}, x}) < \epsilon/2^n$ for any $n \geq 1$. $\square$

We can now state the stability theorem for ϵ-interleaved persistence modules:

Theorem 14. (Theorem 4.4 in Chazal *et al.* [39]) *Let $\mathcal{F}$ and $\mathcal{G}$ be two strongly ϵ-interleaved persistence modules which are δ-tame for some $\delta \geq 0$. Then,*

$$W_\infty(D_\delta \mathcal{F}, D_\delta \mathcal{G}) < 3\epsilon.$$

We are now ready to state a stability theorem for persistence diagrams for discrete Morse functions:

Theorem 15. *Let f and g be discrete Morse functions on a finite simplicial complex K which induce k^{th} persistence diagrams $D^k f$ and $D^k g$, respectively. Suppose that there exists $\epsilon > 0$, such that $||f - g||_\infty < \epsilon$. Then,*

$$W_\infty(D^k f, D^k g) < 3\epsilon.$$

Proof. Let us show that the level subcomplexes of f and g are strongly ϵ-interleaved. Indeed, since $||f - g||_\infty < \epsilon$, f and g are strongly ϵ-interleaved by Lemma 11. Let $c \in \mathbb{R}$ and let $\sigma \in K^f(c)$ and denote by F^c the sub-level set $f^{-1}((-\infty, c])$ of f; G^c is defined similarly for the discrete Morse function g. Then, by Lemma 3.2 of Forman [8], there exists a co-face τ of σ such that $f(\tau) \leq c$, i.e. $\tau \in F^c$. By the strong ϵ-interleaving property, one gets $F^c \subseteq G^{c+\epsilon}$, so we have $g(\tau) \leq c + \epsilon$, which implies that $\sigma \in K^g(c + \epsilon)$, so that we get $K^f(c) \subseteq K^g(c + \epsilon)$. Similarly one easily gets $K^g(c) \subseteq K^f(c + \epsilon)$ for any $c \in \mathbb{R}$. This implies that the persistence modules of f and g given by the kth homology groups of the level subcomplexes are strongly ϵ-interleaved for any $k \in \mathbb{N} \cup \{0\}$. Hence by Theorem 14, one gets:

$$W_\infty(D^k f, D^k g) < 3\epsilon$$

This concludes the proof. $\square$

As a concluding remark to this section, we observe that the bound of 3ϵ can be improved to ϵ, following the same arguments as given in the proof of Theorem 4.8 in Chazal *et al.* [39], we omit the details here.
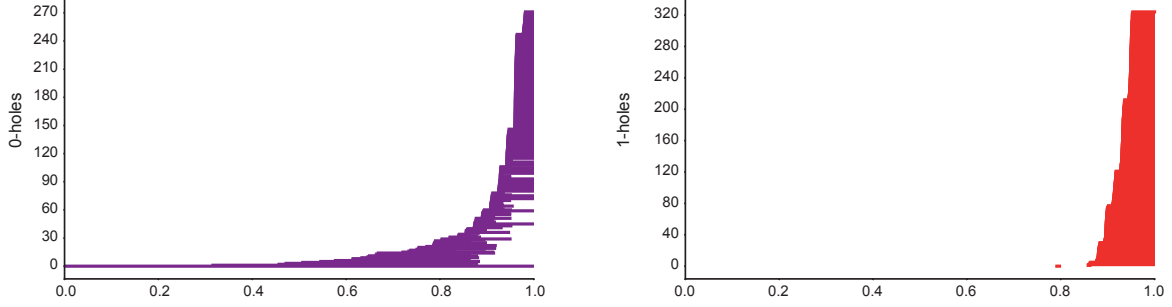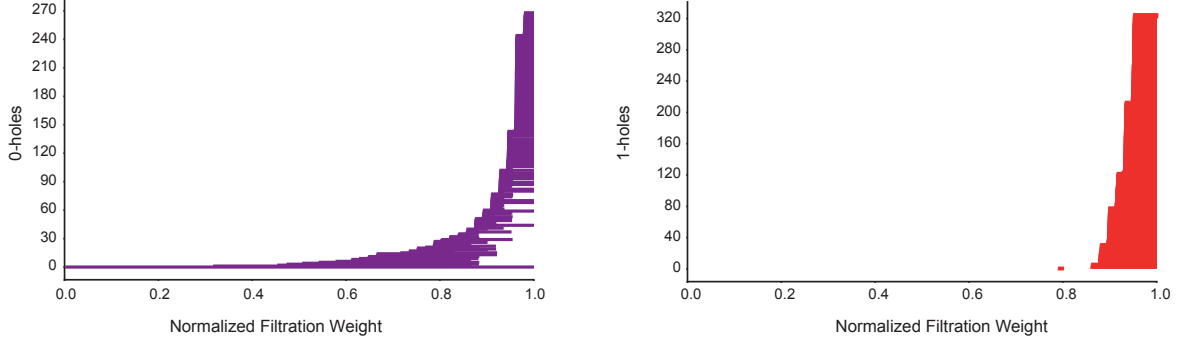
(a) Yeast Protein Interaction with discrete Morse function f$_1$ with uniform distribution

(b) Yeast Protein Interaction with discrete Morse function f$_2$ with normal distribution

FIG. A1. Barcode diagrams for H_0 and H_1 in the Yeast protein interaction network computed using two discrete Morse functions f_1 and f_2 generated by making small variations in Algorithm 1. f_1 uses a uniform distribution on $[0.1, 0.4)$ for generating the values of ϵ in Algorithm 1, f_2 uses a normal distribution with mean $\mu = 0.25$ and standard deviation $\sigma = 0.05$ for generating the values of ϵ in Algorithm 1.

Illustration of the stability of persistence diagrams using a real network example

As an illustration of the stability property, we compute the bottleneck distance between persistence diagrams of two discrete Morse functions f_1 and f_2 on the Yeast protein interaction network [43] generated by making small variations in Algorithm 1. The construction of f_1 uses a uniform distribution on $[0.1, 0.4)$ for generating the values of ϵ in Algorithm 1, while the construction of f_2 uses a normal distribution with mean $\mu = 0.25$ and standard deviation $\sigma = 0.05$ for generating the values of ϵ in Algorithm 1. The L^∞-norm of the difference $||f_1 - f_2||_\infty$ is 18.3782, while the bottleneck distance between their *total* persistence diagrams (i.e., the persistence diagram of the persistence module $\mathcal{K}_* := \{\oplus_{k=0}^{\dim K} H_k(K^f(\alpha))\}_{\alpha \in \mathbb{R}}$), is computed to be $W_\infty(Df_1, Df_2) = 0.0061$. The barcode diagrams for the 0th and 1st homology groups in the Yeast protein interaction network [43] with the two discrete Morse functions, f_1 and f_2, are shown in Appendix figure A1.

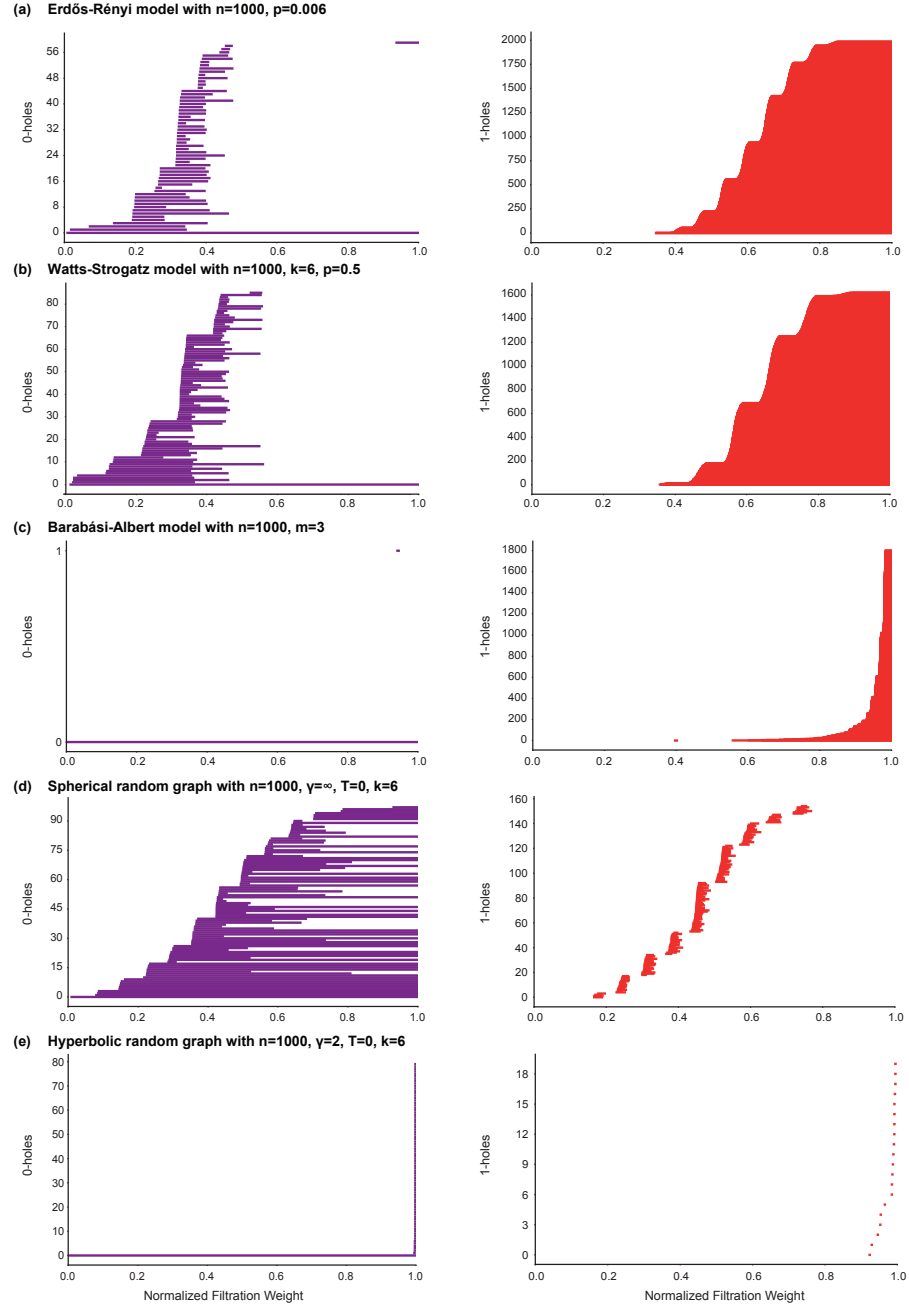# Supplementary Information (SI)
Figures

FIG. S1. Barcode diagrams for H_0 and H_1 in model networks with expected average degree 6.
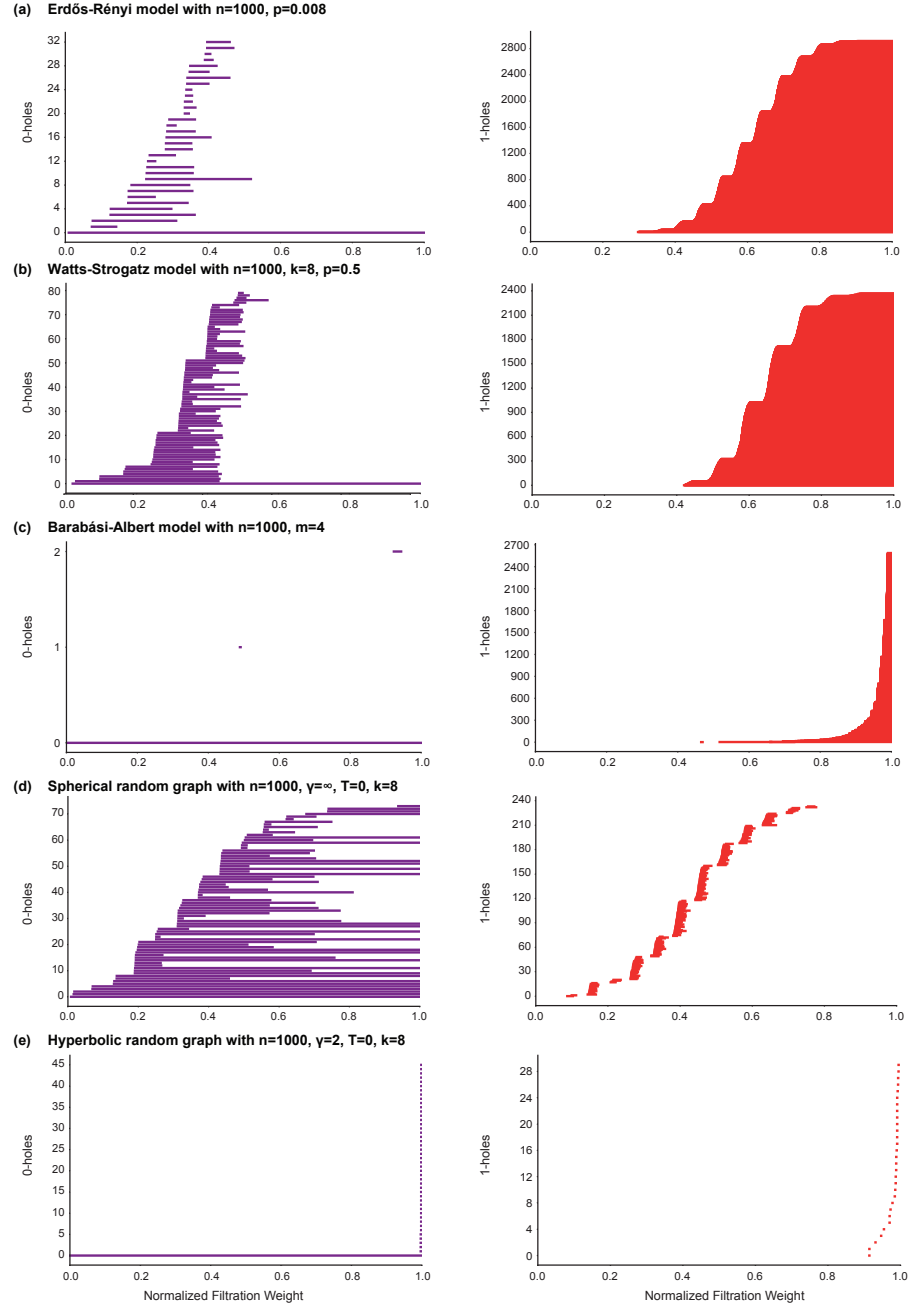
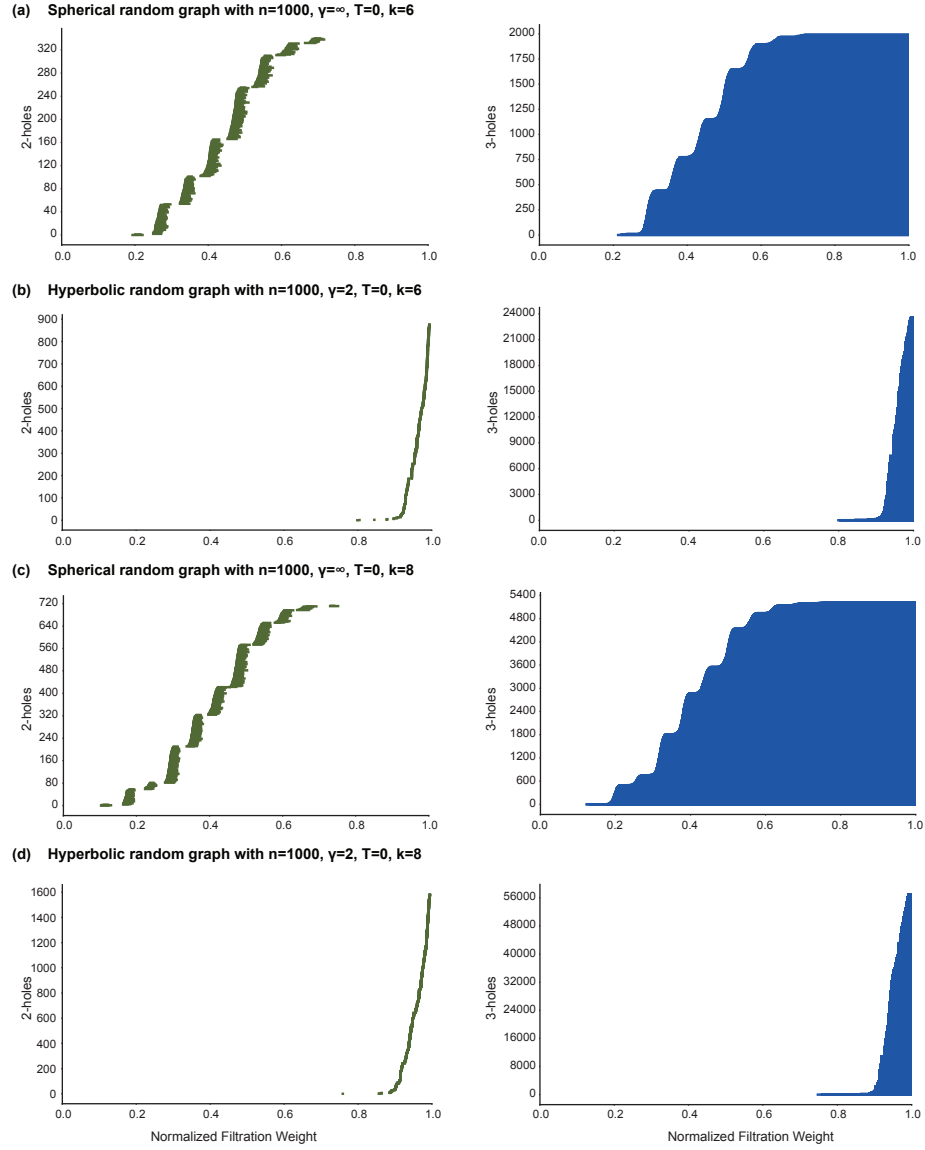FIG. S2. Barcode diagrams for H_0 and H_1 in model networks with expected average degree 8.

FIG. S3. Barcode diagrams for H_2 and H_3 in model networks with expected average degree 6 and 8. (a) Spherical random graphs with expected average degr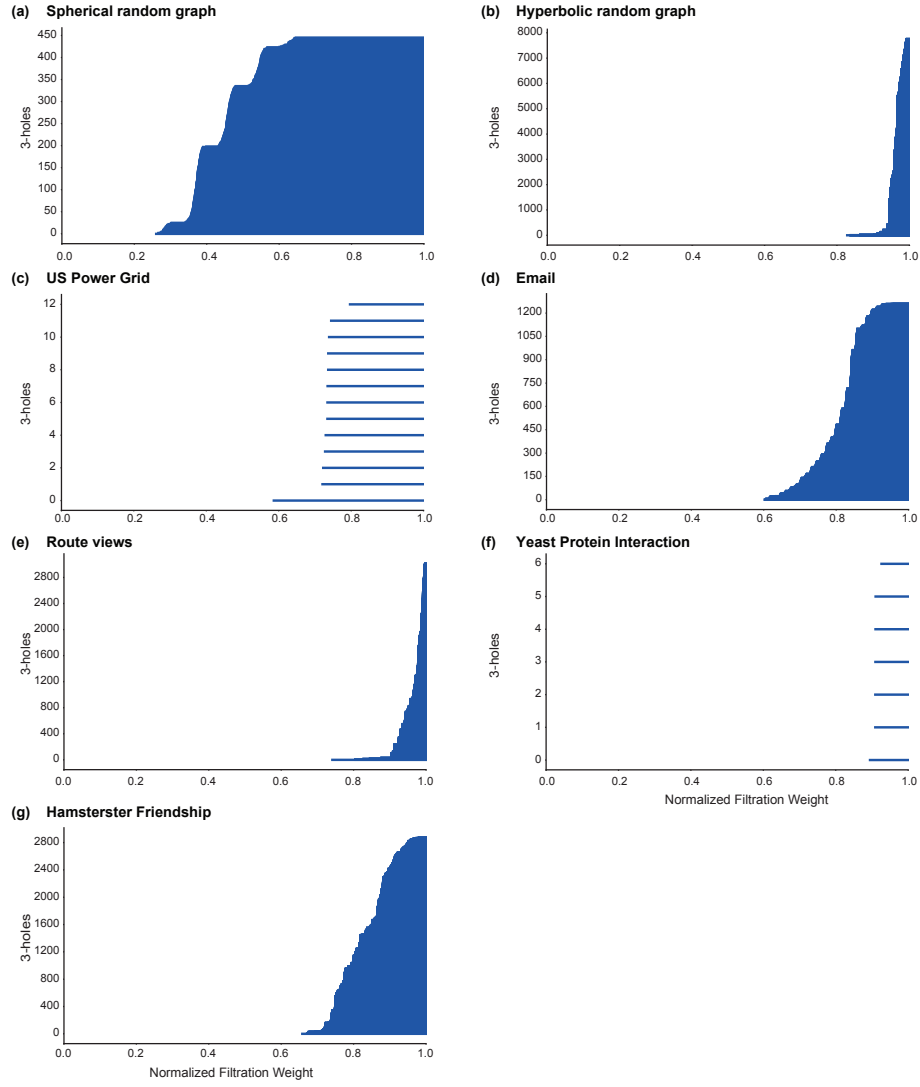ee 6. (b) Hyperbolic random graphs with expected average degree 6. (c) Spherical random graphs with expected average degree 8. (d) Hyperbolic random graphs with expected average degree 8.

FIG. S4. Barcode diagrams for H_3 in model and real networks. (a) Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$. (b) Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$. (c) US Power Grid. (d) Email communication. (e) Route views. (f) Yeast protein interaction. (g) Hamsterster friendship.
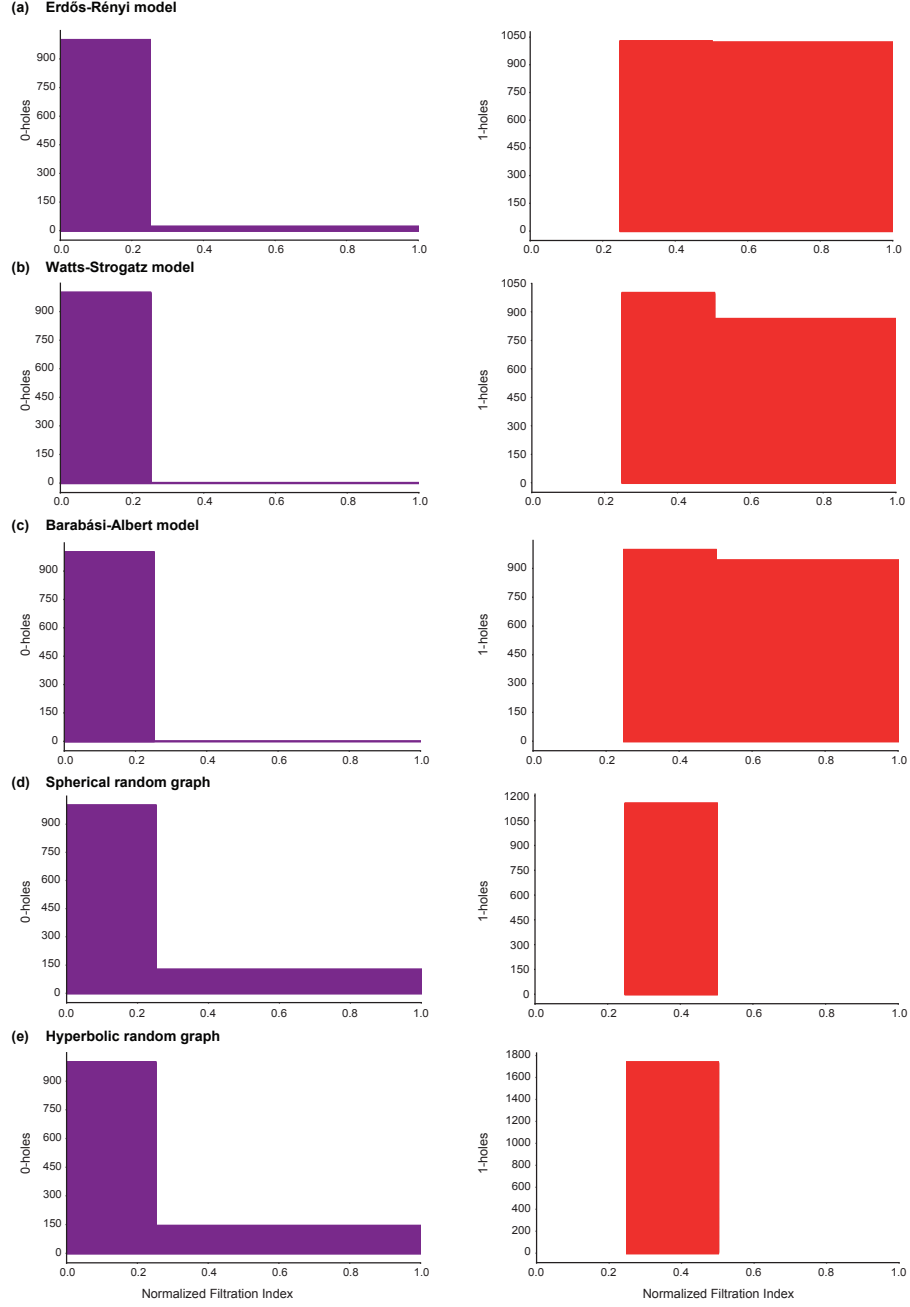
FIG. S5. Barcode diagrams for H_0 and H_1 in model networks with expected average degree 4 obtained using the dimensional filtration scheme used in Horak *et al.* We restrict our investigation to the three-dimensional clique complex while computing these barcode diagrams for model networks using the dimensional filtration scheme of Horak *et al.* In this figure, we normalize the filtration index to be in the range 0 to 1, and p-holes with normalized filtration index 1 indicate that they never die. (a) ER model with $n = 1000$ and $p = 0.004$. (b) WS model with $n = 1000$, $k = 4$ and $p = 0.5$. (c) BA model with $n = 1000$ and $m = 2$. (d) Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$. (e) Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$.
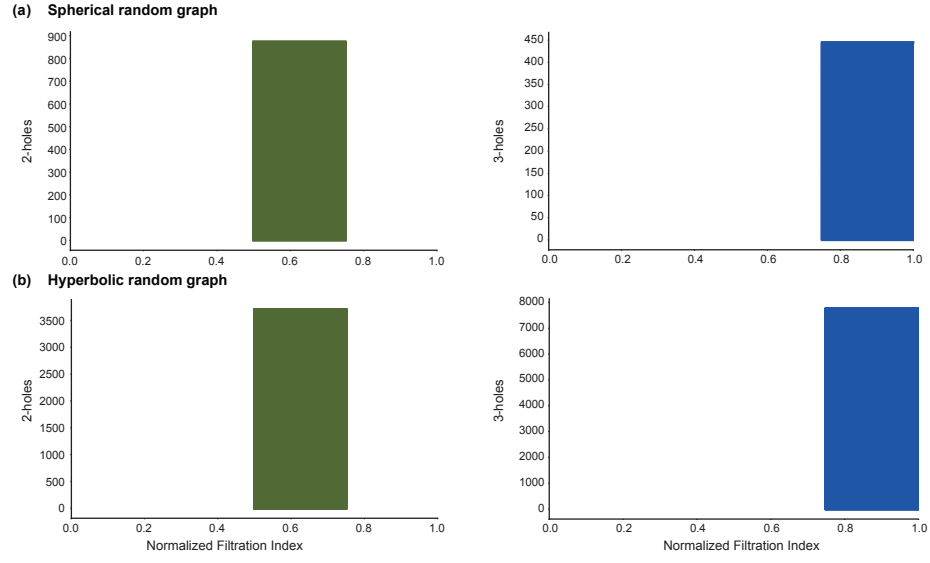
FIG. S6. Barcode diagrams for H_2 and H_3 in model networks with expected average degree 4 obtained using the dimensional filtration scheme used in Horak *et al.* (a) Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$. (b) Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$.
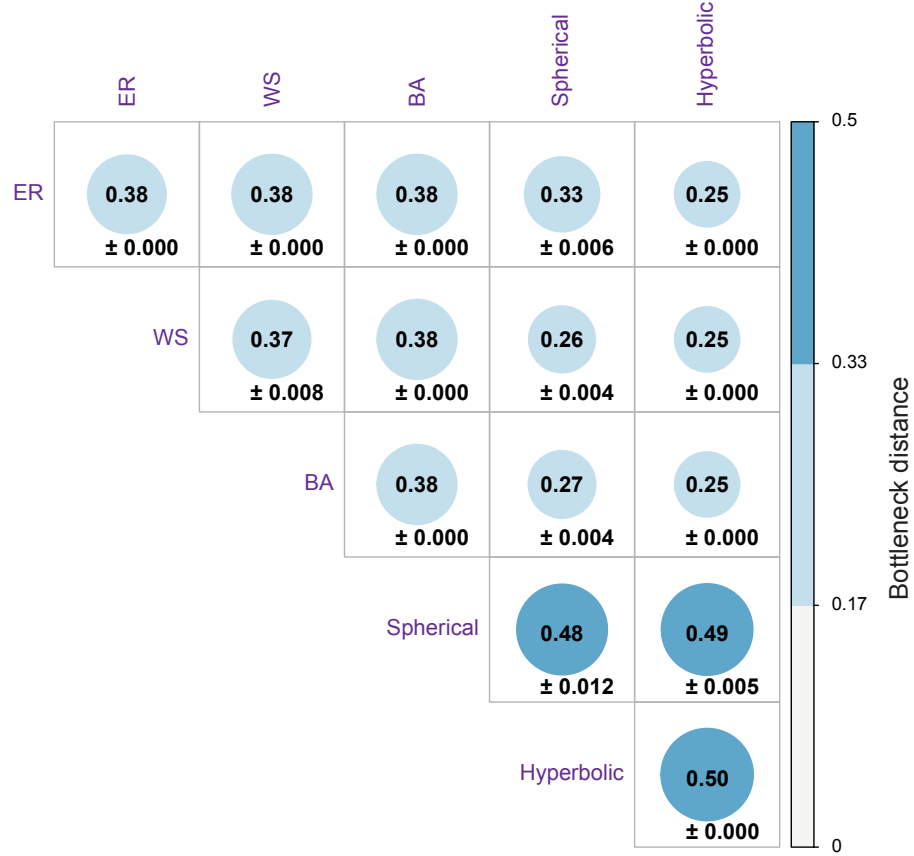
FIG. S7. Bottleneck distance between persistence diagrams obtained using the dimensional filtration scheme used in Horak *et al.* for model networks, namely, ER model with $n = 1000$ and $p = 0.004$, WS model with $n = 1000$, $k = 4$ and $p = 0.5$, BA model with $n = 1000$ and $m = 2$, Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$, and Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$. For each of the five model networks, 10 random samples are generated by fixing the number of vertices n and other parameters of the model. We report the distance (rounded to two decimal places) between two different models as the average of the distance between each of the possible pairs of the 10 sample networks corresponding to the two models along with the standard error.