


OPEN

Persistent homology of unweighted complex networks via discrete Morse theory

Harish Kannan¹, Emil Saucan^{2,3}, Indrava Roy¹ & Areejit Samal^{1,4} 

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.

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 data¹. 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 structures²) to imaging analysis (e.g., feature detection in 3D gray-scale images³) to biology (e.g., detection of breast cancer type with high survival rates⁴). 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 techniques¹. 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 theory⁵, 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 article⁶ 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 of 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 task¹⁴. Moreover, recent studies^{15–17} indicate that these higher-order correlations are not captured by usual network measures such as clustering coefficients. While a

¹The Institute of Mathematical Sciences (IMSc), Homi Bhabha National Institute (HBNI), Chennai, 600113, India.

²Department of Applied Mathematics, ORT Braude College, Karmiel, 2161002, Israel. ³Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, 3200003, Israel. ⁴Max Planck Institute for Mathematics in the Sciences, Leipzig, 04103, Germany. Correspondence and requests for materials should be addressed to I.R. (email: indrava@imsc.res.in) or A.S. (email: asamal@imsc.res.in)

Received: 17 January 2019

Accepted: 6 September 2019

Published online: 25 September 2019

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.

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 topology³¹, astrophysics². 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 (Fig. 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 Fig. 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 potency 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-world¹⁰ and scale-free¹¹ 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 simplicial 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.

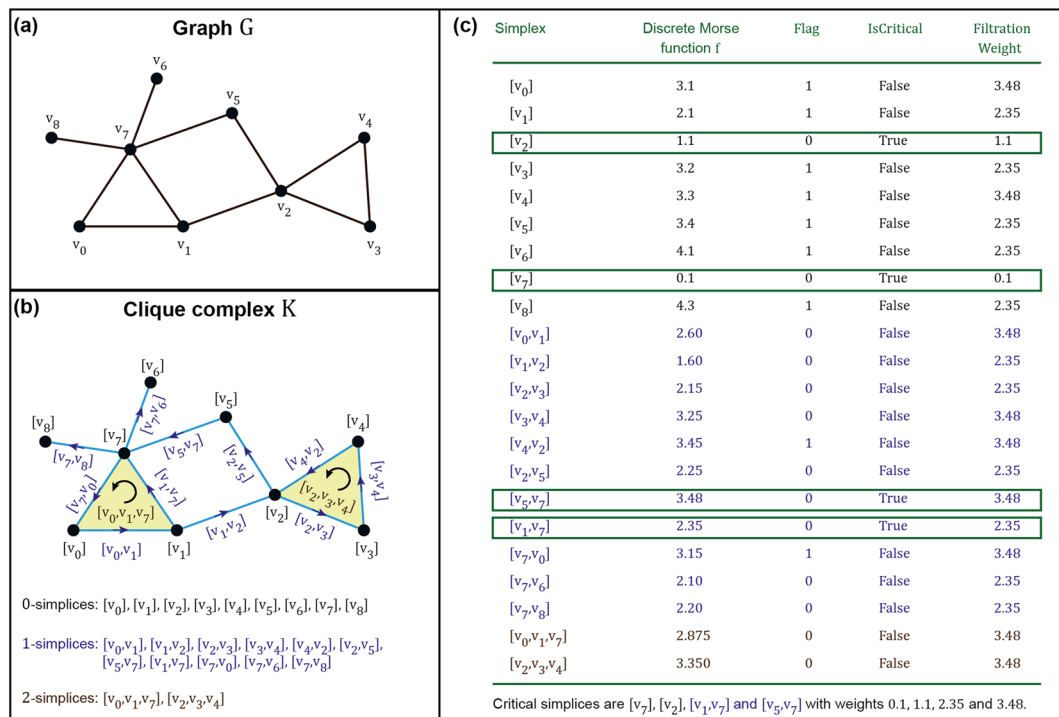


Figure 1. An illustration of the construction of a discrete Morse function f on a clique complex K corresponding to an unweighted and undirected graph G using our algorithm. (a) A simple example of an unweighted and undirected graph G containing 9 vertices and 11 edges. (b) The clique simplicial complex 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).

Theory

Graphs and simplicial complexes. Consider a finite simple graph $G(\mathcal{V}, \mathcal{E})$ having vertex set $\mathcal{V} = \{v_0, v_1, \dots, v_n\}$ and the edge set \mathcal{E} . Note that a simple graph does not contain self-loops or multi-edges³⁴. 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 topology³⁶.

Figure 1 displays an example of the correspondence between a simple graph and its clique complex. The ordering of the vertex set $\{v_0, v_1, v_2, \dots, v_p\}$ 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 $[v_0, v_1, v_2, \dots, v_p]$ ³⁶. In Fig. 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:

$$\emptyset \subseteq K^0 \subseteq K^1 \subseteq \dots \subseteq K^q = K \quad (1)$$

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:

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

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:

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

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.

Discrete Morse theory. Recalling from the preceding 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 theory⁷⁻⁹. 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 Forman⁷⁻⁹ 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 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 $f: K \rightarrow \mathbb{R}$, for each simplex $\alpha^p \in K$, two sets U_α^f and V_α^f are defined as follows:

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

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

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 $f: K \rightarrow \mathbb{R}$ is a discrete Morse function^{8,9} if and only if for each simplex $\alpha^p \in K$:

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

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:

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

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 $\{w_k\}_{k=0, \dots, n}$ 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:

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

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 Forman⁹, 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 $\{w_k\}_{k=0, \dots, m}$ 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 $\{w_k\}_{k=0, \dots, m}$ is a subsequence of $\{w_k\}_{k=0, \dots, n}$ 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 $\{w_k\}_{k=0, \dots, m}$ of $\{w_k\}_{k=0, \dots, n}$, 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:

$$\emptyset \subseteq K(w_0) \subseteq K(w_1) \subseteq \dots \subseteq K(w_{c_{m-1}}) \subseteq K(w_m) \subseteq K. \tag{10}$$

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 algorithms 2 and 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 Fig. 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 Forman⁹ can be stated as follows.

- (i) For each $p = 0, 1, 2, \dots, d$ (where d is the dimension of K), $m_p \geq \beta_p$.
- (ii) $m_0 - m_1 + m_2 - \dots + (-1)^d m_d = \beta_0 - \beta_1 + \beta_2 - \dots + (-1)^d \beta_d$.

In other words, the above theorem gives a lower bound of the number of critical p -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 2.11⁹), 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 (Eq. 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:

$$W_{\infty}(X, Y) = \inf_{\eta: X \rightarrow Y} \sup_{x \in X} \left\| x - \eta(x) \right\|_{\infty}. \quad (11)$$

Similarly, given two persistence diagrams X and Y , the q -Wasserstein distance³⁸ between X and Y is defined as follows:

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

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 Erdős-Rényi (ER)⁴⁰, the Watts-Strogatz (WS)¹⁰, the Barabási-Albert (BA)¹¹ and the Hyperbolic Graph Generator (HGG)⁴¹. The ER model⁴⁰ 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 model¹⁰ 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 model¹¹ 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 Eq. 6. In the pseudocode of the algorithm 1, lines 2–6 initialize a variable $\text{Flag}[\alpha]$ for every simplex α in clique complex K with the value 0. We remark that the variable $\text{Flag}[\alpha]$ associated with a simplex α in K serves as a counter for the size of the set U_α defined in Eq. 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 Eq. 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 .

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.

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

```

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

2:   for  $p = 0, \dots, d$  do                                     ▷ Initialize  $\text{Flag}$  variable associated with each simplex in  $K$ 
3:     for each  $p$ -simplex  $\alpha \in K$  do
4:        $\text{Flag}[\alpha] = 0$                                        ▷ The variable  $\text{Flag}$  keeps track of the size of the set  $U_\alpha$  for each simplex  $\alpha$ 
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  $p = 1, \dots, d$  do
11:    for each  $p$ -simplex  $\alpha \in K$  do                               ▷ Assign weights to  $p$ -simplices in  $K$  with  $p \geq 1$ 
12:      Let  $\text{Faces}[\ ]$  be an array of all  $(p-1)$ -dimensional faces of  $\alpha$ 
13:      Sort  $\text{Faces}[\ ]$  such that  $f(\text{Faces}[i]) \geq f(\text{Faces}[i+1])$  for each  $i \in \{0, 1, \dots, p-1\}$ 
14:      Let  $\gamma_0 = \text{Faces}[0]$ 
15:      Let  $\gamma_1 = \text{Faces}[1]$ 
16:      if  $\text{Flag}[\gamma_0] = 0$  and  $f(\gamma_0) > f(\gamma_1)$  then
17:         $f(\alpha) = (f(\gamma_0) + f(\gamma_1))/2$ 
18:         $\text{Flag}[\gamma_0] = 1$ 
19:      else
20:         $\varepsilon = \text{random}(0, 0.5)$                                ▷  $\varepsilon$  is generated randomly during runtime using a uniform distribution on  $(0, 0.5)$ 
21:         $f(\alpha) = f(\gamma_0) + \varepsilon$ 
22:      end if
23:    end for
24:  end for

25:  return  $f$ 

26: end function

```

Rationale for the choice of function on vertices. 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, $g: \mathcal{V} \rightarrow \mathbb{R}$, is as follows:

$$g(v) = \text{deg}_{\max} - \text{degree}(v) + \varepsilon \quad (13)$$

where $\text{degree}(v)$ 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 $(0, 0.5)$.

| 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 friendship | 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 |

Table 1. 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.

In the Theory section, we had highlighted the Theorem 2.11 by Forman⁹ 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 Eq. 10). This motivated our choice for the real-valued function (Eq. 13) which determines the weights of 0-simplices, and the rationale for this choice is as follows.

Ignoring the noise term ε in Eq. 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 Fig. 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 Fig. 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 Eq. 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 Fig. 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[v7]` and `Flag[v8]` 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[v8]` is set to 1 instead of `Flag[v7]`, 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 Eq. 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 $\text{degree}(v) = \text{degree}(w)$. The absence of a random noise ε in Eq. 13 forces $g(v) = g(w)$. Thus, irrespective of the state of $\text{Flag}[v]$ and $\text{Flag}[w]$, the *if condition* fails. This implies that the set $V_{e_{vw}}$ (See Eq. 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 Eq. 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 $\text{Flag}[v]$ and $\text{Flag}[w]$. 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 Eq. 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 Hard⁴⁹. 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 1). 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.

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 GUDHI⁵⁰, 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 Figs 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 (Fig. 3 and SI Figs S1 and S2). A standard result in algebraic topology³⁶ 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 indicates that the scale-free BA network has a strong tendency to maintain a low number of connected components during filtration (Fig. 3 and SI Figs S1 and 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 (Fig. 3 and SI Figs S1 and 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 (Fig. 3 and SI Figs S1 and 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 (Fig. 3 and SI Figs S1 and 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 1 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 1 and SI Table S2). Although, both spherical and hyperbolic networks exhibit a

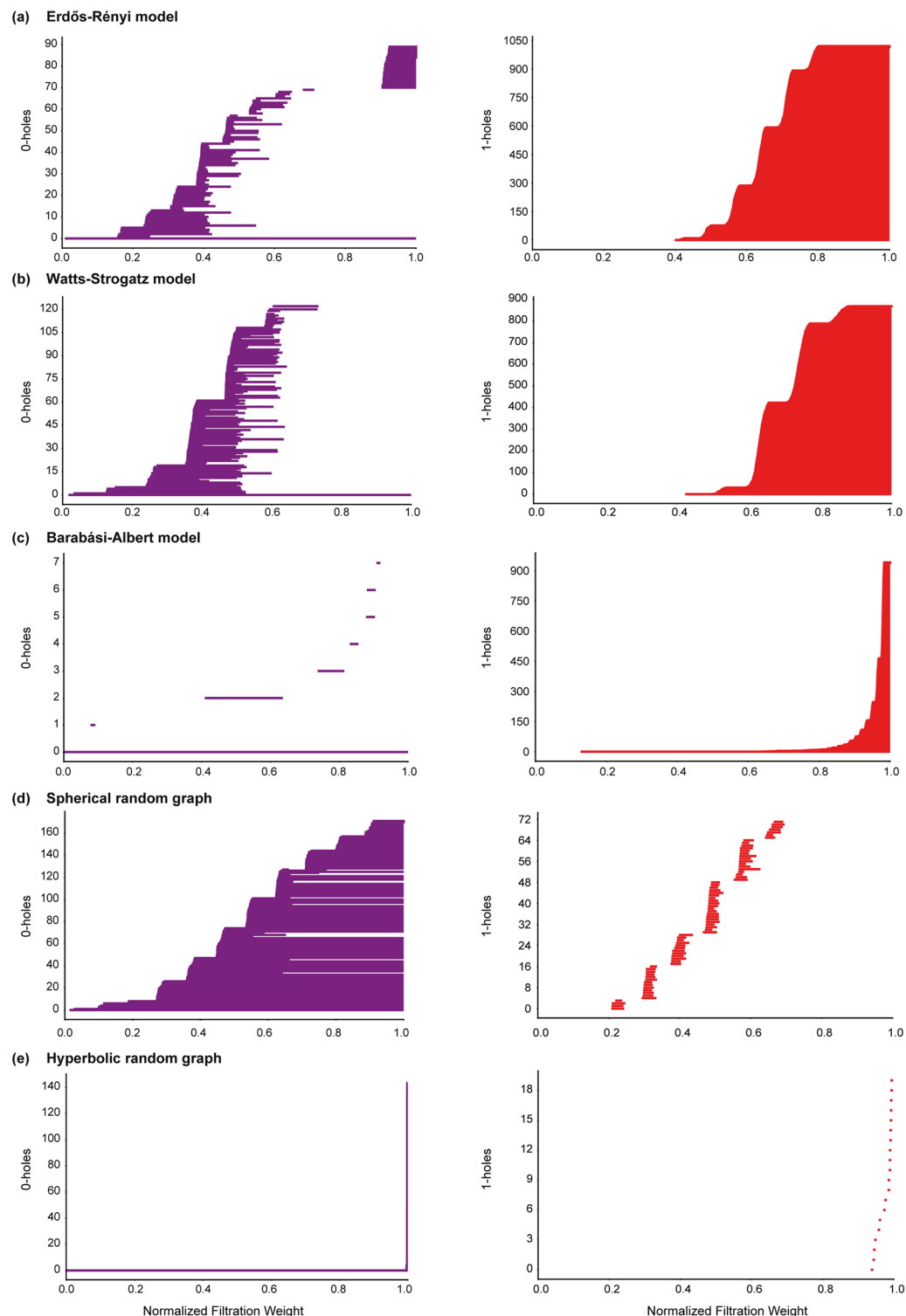


Figure 3. Barcode diagrams for H_0 and H_1 in model networks. (a) ER model with $n = 1000$ and $p = 0.004$. (b) WS model with $n = 1000$, $k = 4$ and $p = 0.5$. (c) BA model with $n = 1000$ and $m = 2$. (d) Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$. (e) Hyperbolic random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = 2$.

higher number of connected components, they differ from each other with respect to the evolution of these connected components during filtration (Fig. 3 and SI Figs S1 and 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 (Fig. 3 and SI Figs S1 and S2). This is in contrast with the behavior of the H_0

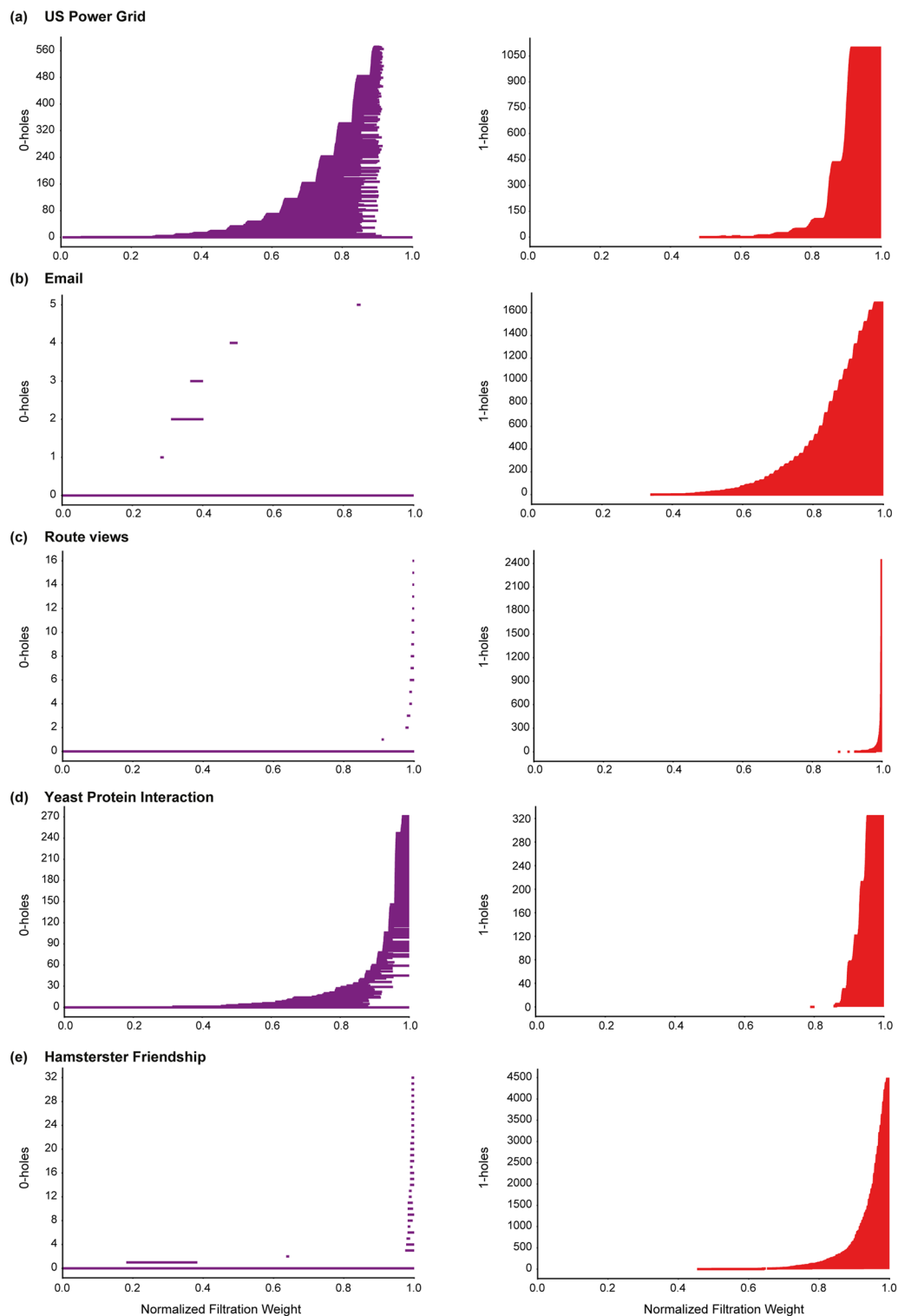


Figure 4. Barcode diagrams for H_0 and H_1 in real networks. **(a)** US Power Grid. **(b)** Email communication. **(c)** Route views. **(d)** Yeast protein interaction. **(e)** Hamsterster friendship.

barcode diagram of the spherical model which exhibits a more distributed evolution of connected components during filtration (Fig. 3 and SI Figs S1 and 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 1 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). 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

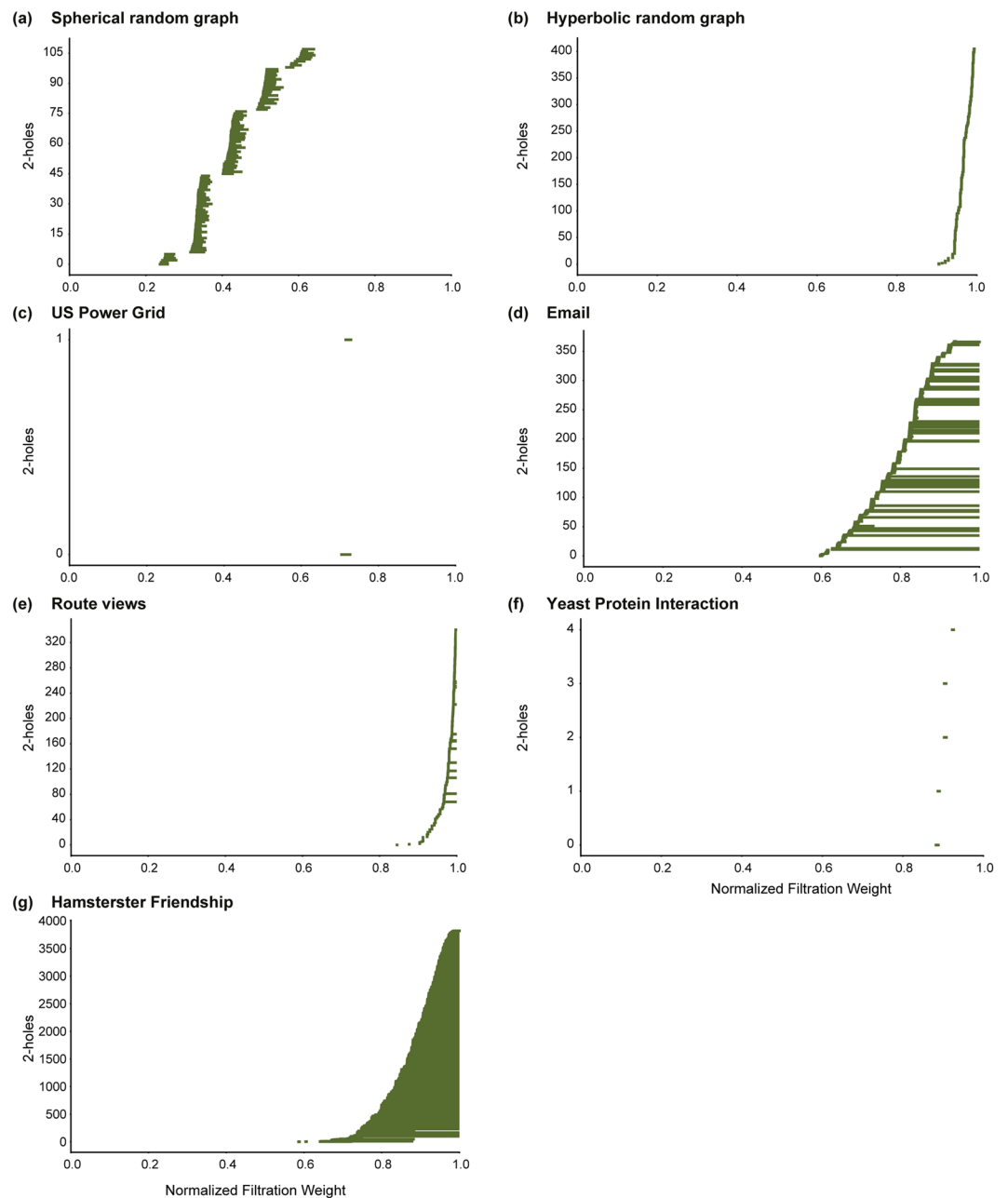


Figure 5. Barcode diagrams for H_2 in model and real networks. **(a)** Spherical random graphs produced from HGG model with $n = 1000, T = 0, k = 4$ and $\gamma = \infty$. **(b)** Hyperbolic random graphs produced from HGG model with $n = 1000, T = 0, k = 4$ and $\gamma = 2$. **(c)** US Power Grid. **(d)** Email communication. **(e)** Route views. **(f)** Yeast protein interaction. **(g)** Hamsterster friendship.

of 3-simplices successively fill in the 2-holes (Fig. 5 and SI Fig. S3). The H_3 barcode diagrams of spherical and hyperbolic networks (See SI Figs S3 and 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 (Figs 3 and 5 and SI Figs S1–S4) suggests that the different models can be distinguished based on their persistent homology. In Theory section, we had 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

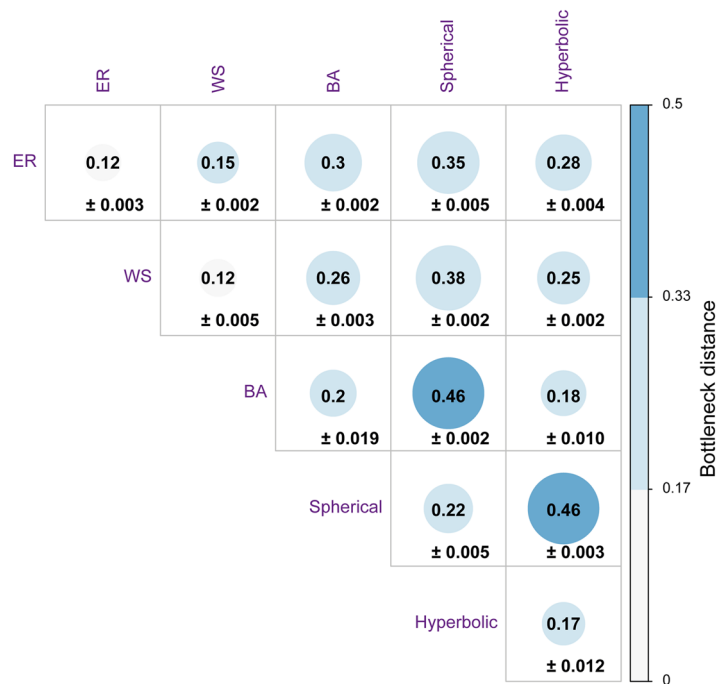


Figure 6. Bottleneck distance between persistence diagrams of model networks, namely, ER model with $n = 1000$ and $p = 0.004$, WS model with $n = 1000$, $k = 4$ and $p = 0.5$, BA model with $n = 1000$ and $m = 2$, Spherical random graphs produced from HGG model with $n = 1000$, $T = 0$, $k = 4$ and $\gamma = \infty$, 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.

and other parameters of the model. In Fig. 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 network with the same number of vertices and similar average degree (Fig. 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 (Fig. 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 (Fig. 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 Fig. 6 remain consistent if the 1-Wasserstein or 2-Wasserstein distance is employed in place of the bottleneck distance (data not shown).

Table 1 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 networks. In Table 2 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.

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

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 simplicies. The value of μ ranges from 0 to 1. $\mu = 1$ indicates that the discrete Morse function achieves exactly the minimum number of critical simplicies thereby corresponding to the most optimal situation, while $\mu = 0$ indicates that all simplicies are critical thereby corresponding to the least optimal

| Network | Optimality 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 interaction | 0.942157 |
| Hamsterster friendship | 0.793236 |
| Euro road | 0.840678 |
| Human protein interaction | 0.957924 |

Table 2. The table lists the value of optimality indicator μ for various model and real networks analyzed here. Eq. 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.

case. Note that the value of μ for a particular network increases linearly with a decrease in the number of critical simplices. In Table 2, 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 2, 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 Figs 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 Figs 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 Fig. 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 2). 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 Figs 4, 5 and SI Fig. 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 (Fig. 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 (Fig. 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 (Fig. 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 (Fig. 5). In contrast, from the H_2 and H_3 barcode diagrams (Fig. 5, SI Fig. 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 2 lists the value of the optimality indicator μ (See Eq. 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 1 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.

Data Availability

All data generated or analysed during this study are included in this article or is available upon request from the corresponding author.

References

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. & Nigmatov, 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. Erdős, 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., Giral, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Physical Review E* **68**, 065103 (2003).
48. Kunegis, J. Konec: The Koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web companion*, 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).

Acknowledgements

We thank Amritanshu Prasad for fruitful discussions. A.S. would like to acknowledge support from the Max Planck Society, Germany, through the award of a Max Planck Partner Group in Mathematical Biology, and I.R. from the Science and Engineering Research Board (SERB) of the Department of Science and Technology (DST) India through the award of a MATRICS grant [MTR/2017/000835]. We thank the anonymous reviewers for their comments which have helped improve this manuscript.

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.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-50202-3>.

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019