Methods and Applications of Analysis.
Vol. 20, No. 4, pp. 383–404, December 2013

THE LANDSCAPE OF COMPLEX NETWORKS – CRITICAL NODES AND A HIERARCHICAL DECOMPOSITION[*]

WEINAN E[†], JIANFENG LU[‡], AND YUAN YAO[§]

Dedicated to Professor Stan Osher on the occasion of his 70*th birthday*

Abstract. Networks have recently emerged as a general tool for data representation in various fields. In the analysis of conformation transition networks in biomolecular dynamics (protein, RNA etc.), it is important to discover major transition bottlenecks which provides clues for drug design. Similarly in the analysis of social networks, it is helpful to identify nodes which act as bridges connecting different communities. Although there have been extensive studies on the community structure of networks, much less has been done about the connection between the communities. Inspired by the classical Morse theory, we introduce a new notion, critical nodes of functions on networks, based on the gradient flow of these functions. Critical nodes of different indices, together with their attraction basins, lead to a hierarchical decomposition of networks. This enables us to define a concise topological *landscape* of functions on networks. The usefulness of this new concept is illustrated by three examples: two social networks and one protein-ligand binding network. For the social networks, the index-0 critical nodes together with their attraction basins represent the different communities on the network; the index-1 and higher critical nodes play the role of bridges or hubs connecting the different communities. For the protein binding network, the index-0 critical nodes together with their basins explain the major metastable bound and misbound macrostates, while the index-1 and higher critical nodes represent the bottleneck between the misbound to the bound states. Computation of such critical nodes can be performed by a polynomial time algorithm based on recent developments in computational topology. In the non-degenerate case an almost linear time algorithm exists which is scalable for large scale network analysis.

Key words. Network, landscape, critical node, gradient flow, attraction basin, saddle, persistent homology, discrete Morse theory.

AMS subject classifications. 55U99, 68U05, 62-07, 62P35.

In recent years, networks have emerged as a general tool for representing data. A natural question arises then is how we can explore the structure of these networks in order to uncover the information hidden in these data. For example in a network representing the different metastable states of a ligand-protein complex, we would like to know the subsets of states that represent the same macro state as well as the states that represent the bottlenecks for the transition between the different macro states. This information is crucial for various applications such as drug design. For networks representing social circles, it is of interest to know the different sub-groups as well as the agents that act as bridges between the different groups. As we show in this paper, this kind of questions can be addressed by extending the notion of critical points from a manifold setting to the setting of networks.

There has been a large amount of recent work on detecting community structure in networks, *e.g.* [18, 17, 1, 21]. However relatively fewer results exist about exploring the

[*]Received December 31, 2012; accepted for publication April 23, 2013. W. E. was supported partially by ARO W911NF-11-10101. Y. Y. was supported partially by National Basic Research Program of China (973 Program 2012CB825501) and NSFC grant (61370004, 61071157).

[†]School of Mathematical Sciences and BICMR, Peking University, Beijing 100871, P. R. China; Department of Mathematics and Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544-1000 USA (weinan@math.princeton.edu).

[‡]Department of Mathematics, Physics, and Chemistry, Duke University, Box 90320, Durham, NC 27708-0320 USA (jianfeng@math.duke.edu).

[§]School of Mathematical Sciences, LMAM-LMEQF-LMPR, Peking University, Beijing 100871, P. R. China (yuany@math.pku.edu.cn).

connections between different communities, and this latter component may contain important information about the network. In this paper, we propose an approach for such a goal based on gradient flow type of dynamics associated with a function defined on the nodes of a network. Our approach is motivated by the Morse theory of manifolds [15] – by examining the critical points of a function on the manifold, one can extract information about the topology of the manifold.

This also allows us to define the "landscape" of the function on the network. The concept of landscape has been crucial in physics and chemistry in describing complex systems. For example in molecular dynamics, conformational changes are driven by the free energy of states, *e.g.* [25, 28], whence energy landscape decides metastable states, transition states and reactive pathways. In the world of complex networks, one often see flows of information or other quantities which are driven by the local gradients of a scalar function [23, 24]. For example traffic flow is driven by congestion function; heat flow is driven by temperature; internet browser's attention may be driven by the influence or relevance of webpages, possibly measured by quantities like PageRank etc; in social networks, gradient flow of some potential function may help identify communities as attraction basins of local minima [26]. Therefore it is helpful to embed community structure into the hierarchy of the complexes defined by the critical points of some functions and the associated basins. This will equip us with a concise description of the global structure of the network and help explain certain dynamical issues such as information diffusion and transition pathways.

Consider an undirected graph $G = (V, E)$ with a function defined on the node set $h : V \to \mathbb{R}$. The question we will attempt to address is: *given a function on its nodes, how can we endow the network with a landscape, so that one can distinguish critical nodes such as the local minima, local maxima, and saddles*?

There are several studies in the literature which may lead to critical nodes for graphs by extending Morse theory to discrete settings. In computational geometry one may embed the graph into a 2D-surface and then apply Morse theory for 2-manifolds [7]. However, such a surface embedding is not natural for general graphs in biological and social networks. Another candidate is discrete Morse theory [9], which studies functions defined on all faces of cell complexes and is therefore hard to use in the graph setting above. A related subject is the extension of the Poincare-Hopf theorem to the graph setting, *e.g.* in [12]. None of these gives a satisfactory answer to the question raised above.

In this paper we present a purely combinatorial approach which starts from a discrete gradient flow induced by the function on graph nodes. Such an approach does not need a surface embedding, and turns out to be closely related to persistent homology in computational topology [8, 6] and discrete Morse theory [9] without studying functions on high dimensional cells. In particular, given a function (often referred to as an energy function) on a network, we will define a discrete gradient flow associated with that function, as well as minimum energy paths between two disjoint sets of nodes. This allows us to define critical nodes or saddles. Roughly speaking, critical nodes are associated with minimum energy paths between node pairs: index-0 critical nodes are simply local minima; index-k critical nodes are the highest energy transition nodes of minimal energy paths connecting index-$(k-1)$ critical nodes. Note that for geometric random graphs whose nodes are random points in Euclidean spaces with a density function at the points, index-0 and index-1 critical nodes defined above capture density cluster trees which has been studied in statistics, *e.g.* [11, 28].

Such a critical node analysis, as we show by examples in social networks and biological networks, leads to a concise representation of networks while preserving some important structural properties. This approach provides us a hierarchical decomposition of networks into hypergraphs with hypernodes as attraction basins of critical nodes, a concise global visualization of networks adaptive to the landscape of a given function. In short, the local minima or maxima together with their attraction basins can be interpreted as communities or groups in networks; saddle points act as transition states between different critical points of lower indices. In particular, in social networks index-1 saddles act as hubs in connecting communities; in biomolecular dynamics, index-1 saddles play roles as intermediate or transition states connecting misfolded and native states. It is important to note that such an analysis in biomolecular dynamics does not rely on the Markovian assumption which requires simulation running long enough for convergence, whence can be applied to data analysis of conformational transition networks in more general settings with short or long simulations.

In algorithmic aspect, critical nodes in this paper can be computed at a polynomial time cost with an algorithm based on computational topology by monitoring topological changes over energy level sets. In particular in nondegenerate case we propose an almost linear algorithm. Therefore our approach is scalable for the analysis of large scale networks.

1. Landscape and critical nodes.

1.1. Discrete gradient flow. Let $h : V \to \mathbb{R}$ be a real valued function on vertex set V. Throughout this paper we assume that h is injective (one-to-one). Such functions are generic in the space of real functions on V. One may associate a *gradient flow* of h on the graph G, as a map $D_{h,0} : 2^V \to 2^V$ which maps a subset of vertices to its immediate neighbors with lower h values. More precisely, given $x \in V$, define the neighbor set of x with lower energy $\mathcal{N}^-(x) = \{y \in \mathcal{N}(x) : h(y) < h(x)\}$ and

$$D_{h,0}(\{x\}) = \begin{cases} \mathcal{N}^-(x), & \text{if } \mathcal{N}^-(x) \neq \emptyset; \\ \{x\}, & \text{otherwise.} \end{cases} \tag{1}$$

For any $X \subseteq V$, we define

$$D_{h,0}(X) = \bigcup_{x\in X} D_{h,0}(\{x\}). \tag{2}$$

Let $D_{h,0}^2 = D_{h,0} \circ D_{h,0}$, *etc.* We say that y is *reachable* from x, denoted by $x \succ y$ or $y \prec x$, if $y \in D_{h,0}^k(\{x\})$ for some $k \in \mathbb{N}$, *i.e.*, we can find an energy decreasing path from x to y.

Note that our construction of the gradient flow is related to, but different from the gradient network by [23, 24], in which each node is only connected to its neighbor with the lowest energy (*i.e.* the neighbor in the steepest descent direction). We also remark that the gradient flow can be viewed as a "zero temperature" limit of the stochastic gradient flow introduced in [26] in the study of network communities.

1.2. Local minima. The *local minima* of h are those vertices whose h value is no larger than the values of its neighbors.

$$\mathcal{C}_0 = \{x \mid h(x) \leq h(y),\ \forall\, y \in \mathcal{N}(x)\}. \tag{3}$$

In other words, the set of local minima are precisely the maximal vertex set of *fixed points* of the gradient flow $D_{h,0}$.

Given a local minimum $x \in V$, its *attraction basin* is defined to be:

$$\mathcal{A}_0(x) = \{y \mid D^{\infty}_{h,0}(\{y\}) = \{x\}\}. \tag{4}$$

These are the points that reach the local minimum x but not any other local minima.

Boundary or separatrix consists of those nodes which can reach more than one local minimum following the gradient flow

$$\mathcal{B}_0 = \{x \mid |D^{\infty}_{h,0}(\{x\})| > 1\}. \tag{5}$$

It is clear by definition that we have the non-overlapping decomposition

$$V = \mathcal{B}_0 \bigcup \bigcup_{x \in C_0} \mathcal{A}_0(x). \tag{6}$$

1.3. Index-1 critical nodes. Our next task is to classify the nodes in $\mathcal{B}_0$. We do so according to their role in the pathways connecting the different local minima. In particular, *index*-1 *critical nodes (saddles)* are defined as the maxima on *local minimum energy paths* connecting different local minima.

Clearly such a definition relies on the notion of *local minimal energy paths*, which depends on the topology of the path space. Given two local minima, we examine all the paths connecting them. If a path γ_1 can be deformed by the gradient flow to another path γ_2, we say that γ_1 is *deformable* to γ_2. The *local minimum energy paths* are paths which cannot be deformed by the gradient flow.

To be more precise, given two points $a, b \in V$, we define a *path* from a to b as $\gamma = (w_0 \cdots w_n)$ such that $w_0 = a$, $w_n = b$, and $w_{i+1} \in \mathcal{N}(w_i)$ for $i = 0, \cdots, n-1$. We denote the collection of paths from a to b as $\mathscr{P}_{a,b}$.

We note the following elementary lemma, whose proof is obvious.

Lemma 1. *Let $x \succ y$, we can then find a path $\gamma = (w_0 \cdots w_n)$ from x to y such that $h(w_i) > h(w_{i+1})$ for $i = 0, 1, \cdots, n-1$.*

Given two paths $\gamma_1, \gamma_2 \in \mathscr{P}_{a,b}$, we say γ_1 is *deformable to* γ_2, if there is a map $F : \gamma_1 \to 2^{\gamma_2}$, such that

- (reaching) every node in γ_1 reaches some nodes in γ_2, *i.e.* for any $x \in \gamma_1$, $F(x)$ is not empty and for each $y \in F(x) \subset \gamma_2$, $y \prec x$;
- (onto) every node in γ_2 is reachable from γ_1, *i.e.* for any $y \in \gamma_2$, there exists $x \in \gamma_1$, so that $y \in F(x)$, or equivalently,

$$\gamma_2 = \bigcup_{x \in \gamma_1} F(x).$$

Let a, b be two local minima. We call a path $\gamma \in \mathscr{P}_{a,b}$ *local minimum energy path*, if it is not deformable to any other path in $\mathscr{P}_{a,b}$.

We define the energy of a path the maximal energy traversed by the path, *i.e.* $h(\gamma) = \max_{y \in \gamma} h(y)$. From the definition, if γ_2 is deformable to γ_1, we have $h(\gamma_2) \geq h(\gamma_1)$, so in terms of energy barrier, γ_1 is a more preferable path than γ_2.

Given a local minimum energy path, we call the node of maximal energy on the path an *index*-1 *critical node*. The set of all index-1 critical nodes is denoted by $\mathcal{C}_1$. We will also call local minima *index*-0 *critical nodes*, and hence the notation $\mathcal{C}_0$.

The following fact gives a characterization of index-1 critical nodes. The proof can be found in Appendix A.

Proposition 2 (Classification of index-1 critical nodes). *All local minima in $\mathcal{B}_0$ are index-*1 *critical nodes. The other index-*1 *critical nodes will reach one of the local minima in $\mathcal{B}_0$ by the gradient flow.*

We call the index-1 critical nodes that are also local minima in $\mathcal{B}_0$ the nondegenerate index-1 critical nodes, the set of which will be denoted as $\overline{\mathcal{C}}_1$. The other index-1 critical nodes are called degenerate. Not every index-1 critical node is a local minimum in $\mathcal{B}_0$, for example in some cluster trees (see Figure 1). In density cluster trees studied in statistics [11, 28], networks are geometric random graphs in Euclidean spaces where the node set consists of sample points and edges are given if two points are within certain neighborhood, density function on sample points is studied and the index-0 and index-1 critical nodes thus lead to a clustering tree.

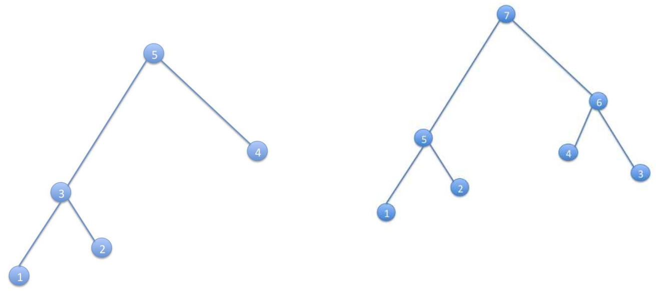


Fig. 1. *Left: an example of degenerate index-*1 *critical node, where node 5 on top of the tree is a degenerate index-*1 *saddle while nodes 3 is a nongenerate index-*1 *saddle. Right: an example of both degenerate and non-degenerate critical node, where node 7 on top of the tree is a degenerate index-1 saddle as it lies on the minimum energy path connecting local minima 1 (or 2) and 3 (or 4), and as well a non-degenerate index-2 saddle as it is on the minimum energy path linking index-1 saddles 5 and 6.*

1.4. Higher index critical nodes. The procedure presented above can be extended to define higher index critical nodes.

To define index-2 critical nodes, we consider the subgraph with nodes in $\mathcal{B}_0$ and edges restricted on this subset, denoted by $G_1 = (V_1 = \mathcal{B}_0, E_1)$. The gradient flow $D_{h,1} : 2^{\mathcal{B}_0} \to 2^{\mathcal{B}_0}$ on $\mathcal{B}_0$ is defined similarly as for $D_{h,0}$. We define the attraction basins for $x \in \mathcal{C}_1$ as

$$(7) \qquad \mathcal{A}_1(x) := \{y \in \mathcal{B}_0 \mid D^{\infty}_{h,1}(\{y\}) = \{x\}\}.$$

Note that for any nondegenerate index-1 critical node, the attraction basin is nonempty. While for a degenerate index-1 critical node, the attraction basin is an empty set. This explains the notion "degenerate" for the critical nodes that are not local minima in $\mathcal{B}_0$.

We define the boundary set as

$$(8) \qquad \mathcal{B}_1 = \{x \in \mathcal{B}_0 \mid |D^{\infty}_{h,1}(\{x\})| > 1\}.$$

As shown in Proposition 2, all local minima on $\mathcal{B}_0$ are in $\mathcal{C}_1$. Therefore, we have the decomposition

$$(9) \qquad \mathcal{B}_0 = \mathcal{B}_1 \bigcup \bigcup_{x \in \mathcal{C}_1} \mathcal{A}_1(x) = \mathcal{B}_1 \bigcup \bigcup_{x \in \overline{\mathcal{C}}_1} \mathcal{A}_1(x).$$

Analogously, we define *index-2 critical nodes* as the maxima on local minimum energy paths connecting different nondegenerate index-1 critical nodes. It is clear that index-2 critical nodes, if exist, must be in $\mathcal{B}_1$.

We remark that under our definition, a degenerate index-1 critical node can also be an index-2 critical node, as shown in Figure 1. This ambiguity is actually quite natural from the network point of view, as these points play multiple roles in the structure of the network. The degenerate index-1 critical node can lie either in the basin of a nondegenerate critical node or link together two different nondegenerate critical nodes.

Higher index critical nodes can be defined recursively through further decomposition of $\mathcal{B}_1$. Classification for high index critical points can be done following similar arguments as above. Combining these, we obtain:

Theorem 1 (Node Decomposition). *V admits the following decomposition*

$$V = \mathcal{B}_0 \bigcup \bigcup_{x \in \mathcal{C}_0} \mathcal{A}_0(x)$$

where

$$\mathcal{B}_{k-1} = \mathcal{B}_k \bigcup \bigcup_{x \in \overline{\mathcal{C}}_k} \mathcal{A}_k(x).$$

Here $\mathcal{A}_k$ is the attraction basin of local minima restricted on the $k-1$-th boundary set $\mathcal{B}_{k-1}$ and $\overline{\mathcal{C}}_k$ is the set of nondegenerate index-k critical nodes.

The theorem gives us a hierarchical representation of the network associated to the energy landscape. It actually leads to a hypergraph representation whose hypernodes are made up of critical nodes with their attraction basins.

2. Finding critical nodes using persistent homology.

2.1. Persistent homology algorithm.

The landscape introduced above can be naturally formulated in terms of a flooding procedure, from low to high values of the height function $h : V \to \mathbb{R}$. Flooding starts from local minima, followed by the attraction basins. Once the relevant index-1 saddle is passed, basins of local minima are merged together. This procedure then continues on to critical points of higher indices.

More precisely, this procedure can be described in terms of persistent homology. Persistent homology, firstly proposed by [8] and developed afterwards largely in [10, 6, 2], is an algebraic tool for computing the Betti numbers and homology groups of a simplicial complex when its faces are added sequentially. To work with persistent homology, we extend the graph G into a simplicial complex up to dimension 2, and also define a filtration which consists of such simplicial complexes, in a spirit close to [7] for PL-manifolds.

An abstract simplicial complex Σ_V is a collection of subsets of V, which is closed under deletion or inclusion, i.e. if $\sigma \in \Sigma_V$, then $\tau \in \Sigma_V$ for any $\tau \subset \sigma$.

We define *the flooding complex of network G associated with the function h*, $\Sigma_{G,h} \subseteq 2^V$ as follows:

- 0-simplex: the vertex set V;
- 1-simplex: the vertex pairs $\{x, y : h(x) \le h(y)\}$ that $x \prec y$, *i.e.*, $x \in D^k_{h,0}(\{y\})$ for some k;

- 2-simplex: collections of triangles $\{x, y, z : h(x) \leq h(y) \leq h(z)\}$, such that $x \prec y$ and $y \prec z$.

One can similarly extend the definition above to general k-simplex. However for our purpose it suffices to define up to dimension 2 simplices.

A filtration of flooding complex $\Sigma_{G,h}$ is a nested family $\mathcal{F}_t \subseteq \Sigma_{G,h}$ with $\mathcal{F}_{t-1} \subset \mathcal{F}_t$ which respects the order of deletion or inclusion in $\Sigma_{G,h}$, i.e. if $\sigma \in \mathcal{F}_t$ and $\tau \subset \sigma$ then $\tau \in \mathcal{F}_t$.

Assume that $h : V \to \mathbb{R}$ is injective or one-to-one, which is generically the case. By taking the maximum over vertices, one can extend h from the vertex set to simplicies, and thus to the simplicial complex $\Sigma_{G,h}$. For a simplex $\sigma \in \Sigma_{G,h}$ let $h(\sigma) = \max\{h(i) : i \in \sigma\}$. This implies that a face's h-value is always no more than that of its associated simplex, i.e. $\sigma \subset \tau \Rightarrow h(\sigma) \leq h(\tau)$.

A filtration $(\mathcal{F}_t : t \in \mathbb{N})$ respecting the order of h can be defined in the following way:

1. $\mathcal{F}_0 = \emptyset$;
2. $\#\{\sigma \in \mathcal{F}_{t+1} \backslash \mathcal{F}_t : \dim(\sigma) = 0\} = 1$, *i.e.* there is precisely one node being added into the filtration for each step;
3. $h(\mathcal{F}_t) < h(\mathcal{F}_{t+1})$, where $h(\mathcal{F}_t) = \max\{h(\sigma) : \sigma \in \mathcal{F}_t\}$, *i.e.* when a node is added into the filtration, all the simplices of the same energy are added into the filtration simultaneously.

Note that under this construction, $\mathcal{F}_1$ consists of the global minimum of h.

In this construction, we consider the filtration corresponding to the flooding procedure from low to high h values. The change of Betti numbers identifies the index-0 and index-1 critical nodes. Once the filtration is defined, persistent homology computes the Betti numbers of the simplicial complex in $\mathcal{F}_t$ for each $t \in \mathbb{Z}$, and draws the barcodes of Betti number versus the t or h values, *e.g.* using JPLEX toolbox[1]. The proof of the following theorem is in Appendix A.

THEOREM 2. *Consider the filtration $(\mathcal{F}_t)$. For all $t \in \mathbb{N}$, $\mathcal{F}_{t+1} \backslash \mathcal{F}_t$ contains an index-*0 *critical node if and only if β_0 increases from $\mathcal{F}_t$ to $\mathcal{F}_{t+1}$; $\mathcal{F}_{t+1} \backslash \mathcal{F}_t$ contains an index-*1 *critical node if and only if either β_0 decreases or β_1 increases from $\mathcal{F}_t$ to $\mathcal{F}_{t+1}$.*

Figure 3 illustrates how such a persistent homology algorithm works for the Zachary's Karate Club network. Detailed information about this network can be found in Section 3.1.

To find higher index saddles, we restrict on the subgraph $G_k = (V_k, E_k)$ where $V_k = \mathcal{B}_{k-1} = V \backslash \cup_{0 \leq i \leq k-1} \cup_{x \in \mathcal{C}_i} \mathcal{A}_i(x)$ and E_k consists of edges restricted on V_k. We can analogously construct the filtration corresponds to the flooding procedure $(\mathcal{F}_{k,t}, t \in \mathbb{N})$ on the subgraph G_k. Similar identification holds for higher index saddles.

THEOREM 3. *Consider the filtration $(\mathcal{F}_{k,t})$ on subgraph G_k for $k \geq 2$. For all $t \in \mathbb{N}$ such that $\mathcal{F}_{k,t+1} \backslash \mathcal{F}_{k,t}$ contains an index-k critical node if either β_0 decreases or β_1 increases from $\mathcal{F}_{k,t}$ to $\mathcal{F}_{k,t+1}$.*

Clearly our characterization of high order critical nodes above only exploits simplicial complex up to dimension 2, whose persistent homology computation is recently improved to be of complexity $O(m^{2.376})$ [16] with $m = O(n^3)$ the total number of

[1] `http://comptop.stanford.edu/programs/`

simplices and n the number of nodes. Such a complexity does not suffer the curse of dimensionality as the computation of high order Betti numbers in general.

2.2. Efficient search of nondegenerate Saddles.

Algorithm 1 Fast search of nondegenerate critical nodes

```
Sort the nodes according to $h$ in increasing order;
Set $G_0 = G$;
for $k = 0, \dots, n$ do
  for $x \in V_k$ in an increasing order of $h$ do
    Find neighbors of $x$ with lower energy, $\mathcal{N}_k^-(x) = \{y \in \mathcal{N}(x) \cap V_k \mid h(y) < h(x)\}$;
    if $\mathcal{N}_k^-(x) = \emptyset$ then
      Add $x$ to $\overline{\mathcal{C}}_k$ and set the color of $x$ as its node index;
    else
      if $\mathcal{N}_k^-(x)$ contains a single color then
        Set the color of $x$ as the single color;
      else
        Leave the color of $x$ as blank;
      end if
    end if
  end for
  return :
    (1) local minima $\overline{\mathcal{C}}_k$ as nondegenerate critical nodes;
    (2) attraction basins $\mathcal{A}_k(x_0)$ ($x_0 \in \overline{\mathcal{C}}_k$) as color components;
    (3) boundary $\mathcal{B}_k$ as the blank nodes;
  Set $G_{k+1} = (\mathcal{B}_k, E_{k+1})$ where $E_{k+1}$ are edges restricted on $\mathcal{B}_k$;
end for
```

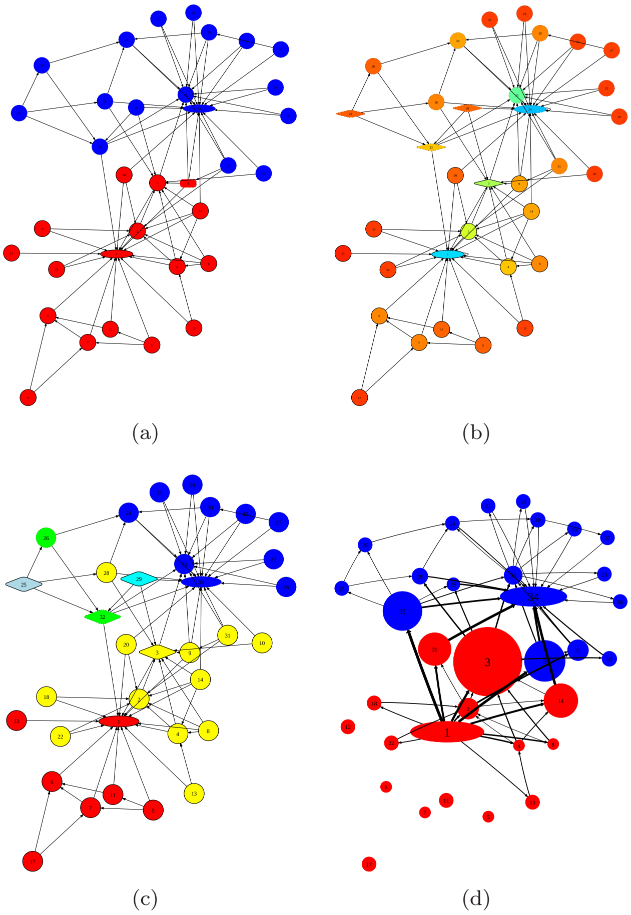
As we know from Proposition 2 that nondegenerate critical nodes are actually local minimum in sub-graphs G_k, this leads to an efficient algorithm for finding nondegenerate critical nodes. In fact, all the examples shown in this paper have only nondegenerate critical nodes and thus can be found efficiently using this algorithm.

Given an injective function h on the vertices, we obtain the local mimina and nondegenerate index-k saddles using Algorithm 1.

The bottleneck in this algorithm is in finding the attraction basins of local minima, whose complexity can be $O(nd)$ where n is the number of vertices and d is the maximum degree a node has. The total complexity is $O(Knd)$ where K is the maximum index of critical points. The algorithm is much faster than the previous algorithm for finding all critical nodes.

3. Examples.

3.1. Zachary's karate club network. Zachary's karate club network [27] consists of 34 nodes, representing 34 members in a karate club with node 1 being the instructor and node 34 being the president (Figure 2). An edge between two nodes means that the two members join some common activities beyond the normal club classes and meetings. Conflicts broke out between the instructor and the president when the instructor sought to raise the fee and the president opposed the proposal. The club eventually split into two, one formed by the president (blue nodes in Figure 2(a)) and another one led by the instructor (red nodes in Figure 2(a)). A lot of information about this fission can be disclosed by looking at the graph structure of this social network.



(a) (b)

(c) (d)

Fig. 2. *Landscape of Karate Club. (a) The fission of Karate Club into two new clubs marked by red and blue colors [27]: the coach is node 1 and the president is node 34, where the box node joined the red club (coach) instead of the blue due to his necessity to finish the course. (b) A gradient flow on edges, node colors from blue to red indicate the energy from low to high, four nodes in diamond shape to-be-disclosed soon as critical nodes. (c) Node decomposition with each color component representing a critical node with its basin: two local minima are in oval shape in which node 1 has basin in red and node 34 in blue; two index-1 saddles are in diamond shape in which node 3 has basin in yellow and node 32 in green; two index-2 saddles, node 25 in light blue and node 29 in cyan. (d) A transition path analysis with source node 1 and target node 34. Committor function with thresholding probability* 0.5 *is used to divide all the nodes into two communities, one with node 1 in red and the other with node 34 in blue. Node size is in proportion to transition current connecting two communities through the node. Effective reactive currents from node 1 to node 34 are drawn with arrows on edges, whose width is determined from effective reactive current with a threshold greater than* 0.001. *It can be seen that index-1 saddles (3, 32) host a majority of transition currents.*

Let d_i be the degree of node i, and define $\overline{h}_i = -\log d_i$. To avoid the same degree between two nodes in neighbor, a small enough random perturbation is added such that $h_i = \overline{h}_i + \epsilon_i$ is injective. Figure 2(b) shows the gradient flow of h. The arrows on the edges point from low degree nodes to high degree ones. Note that nodes 24 and 25 both have degree 3, hence a small random perturbation is added resulting in the arrow from 25 to 26. The same is done for nodes 5 and 11.

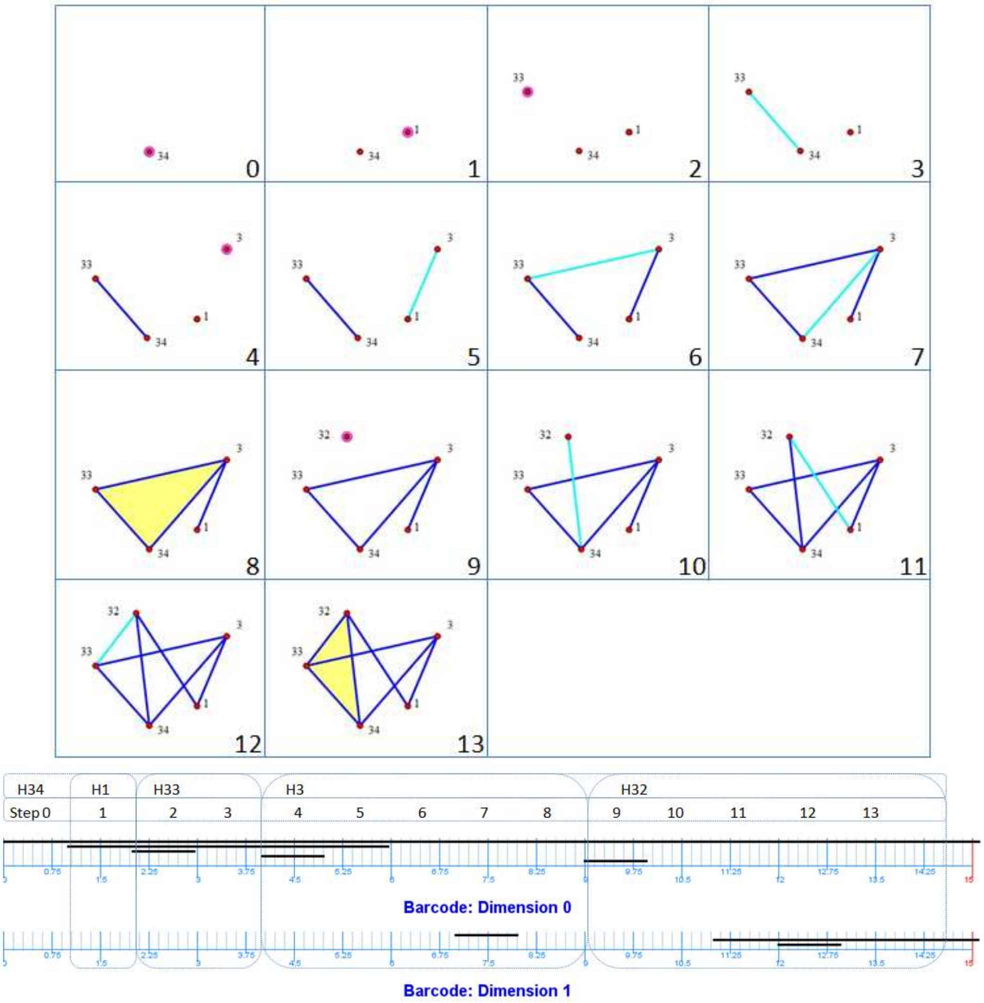


FIG. 3. *An illustration of finding index-0 (node 1 and 34) and index-1 (node 3 and 32) critical nodes in Karate Club network using the persistent homology algorithm. Top: a flooding complex is created with a filtration adapted to the order of energy function $H : V \to R$ defined in Section 3.1. At each step from 0 to 13, a cell (node, edge, and triangle) is created such that, (1) node emergence follows the increasing order of H, (2) an edge appears after its associated nodes are created, and (3) an triangle appears after its associated edges are created. Bottom: persistent Betti numbers (β_0 and β_1) is plotted as barcodes with step numbers and energy values. Node 34 with the lowest energy is added at step $t = 0$ which creates a connected component (β_0) which never disappears. Node 1 with the second smallest energy is added at step $t = 1$ which creates a new connected component (β_0) disappeared after index-1 saddle 3 and its associated cells (edges and triangles) are all added during step $t \in [4, 8]$. As the second index-1 saddle, node 32 and its associated cells are added during step $t \in [9, 13]$ which creates a loop (β_1).*

Figure 2(c) shows the node decomposition for Karate club network with each color component for a critical node and its attraction basin. Two local minima, nodes 1 and 34, are in oval shape together with their attraction basins marked in red and blue, respectively. Two index-1 saddles, nodes 3 and 32, are yellow and green diamond nodes, whose basins are in yellow (nodes 3) and green (node 32) correspondingly.

Node 3 is the lowest energy node connecting the local minima nodes 1 and 34 via a minimum energy path $\gamma_1 = (1, 3, 33, 34)$. Node 32 links the two local minima by another local minimum energy path, $\gamma_2 = (1, 32, 34)$. Two index-2 saddles, nodes 25 (in light blue diamond) and 29 (in cyan diamond), which connect two index-1 saddles via two non-deformable minimal energy paths $(3, 29, 32)$ and $(3, 28, 25, 32)$. Figure 2(d) further depicts a transition path analysis (see [4, 5]) of a Markov chain induced by a random walk on the undirected graph. On each node, a random walker jumps to its neighbors with equal probability. Transition path analysis (see Appendix B) of such a random walk from local minimum node 1 to node 34 shows two index-1 saddles capture most of transition currents.

Figure 3 shows the barcodes by persistent homology algorithm for the computation of index-0 and index-1 critical nodes of this network. Similar computation can be carried over to index-2 critical nodes.

Although Zachary's original paper [27] does not disclose detailed information about the nodes beyond the coach and the president, from the analysis above, one can see that index-1 and index-2 critical nodes play important roles on characterizing information diffusion pathways. For example, index-1 saddle 3 is the most popular node connecting the basins of the coach (node 1) and the president (node 34), and information released from one of the basins mostly probably pass node 3 to reach the other. Therefore if node 3 is off or blocked for such kind of information transition, another index-1 saddle 32 will play the same role. Similarly index-2 saddles 25 and 29 are most popular nodes connecting basins of index-1 saddles. In this way, one reaches an hierarchical decomposition of the network.

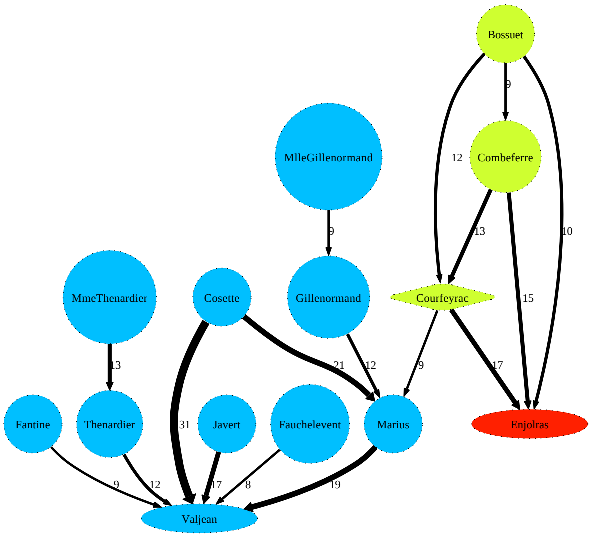


FIG. 4. *Landscape of a subnet of The Les Misérables Network. Edges are left with weights larger than* 7. *Two local minima, Valjean and Enjoras as well as an index-*1 *saddle, Courfeyrac, are identified.*

3.2. The social network of Les Misérables. The social network of Les Misérables, collected by Knuth [13], consists of 77 main characters in the novel by Victor Hugo. The edge weight w_{ij} record the number of co-occurrence of two characters i and j in the same scene. Thus it is a weighted graph where $h_i = -\log \sum_{j\sim i} w_{ij}$

as the negative logarithmic weighted degree. The original network exhibits a single local (global) minimum, Valjean, who is the central character as the whole novel was written around his experience.

However, dropping those edges whose weights are no more than a threshold value (7 here), there appears a subnetwork which is closely associated with the Paris uprising on the 5th and 6th of June 1832, see Figure 4. The subnetwork consists of two local minima, Enjoras and Valjean, the former being the leader of the revolutionary students called *Friends of the ABC*, the Abaissé. Led by Enjolras, its other principal members are Courfeyrac, Combeferre, and Laigle (nicknamed Bossuet) et al., who fought and died in the insurrection. Among them is an index-1 saddle, Courfeyrac, a law student and often seen as the heart of the revolutionary student group, who introduced various people to the Friends of ABC including Marius. Marius, a descend of the Gillenormands, though badly injured in the battle, was saved by the main character Valjean when the barricade fell and married to Cosette, the adopted daughter of Valjean. The landscape of this subnetwork highlights these events in the novel, where the index-1 saddle captures a hub connecting different events in the novel.

3.3. LAO protein binding transition network. This application examines the binding of Lysine-, Arginine-, Ornithine-binding (LAO) protein to its ligand, recently studied in [22]. The critical node analysis provides us a concise summary of global structure of networks while preserving important pathways, which enables us to reach a more thorough description than previous approximate analysis.

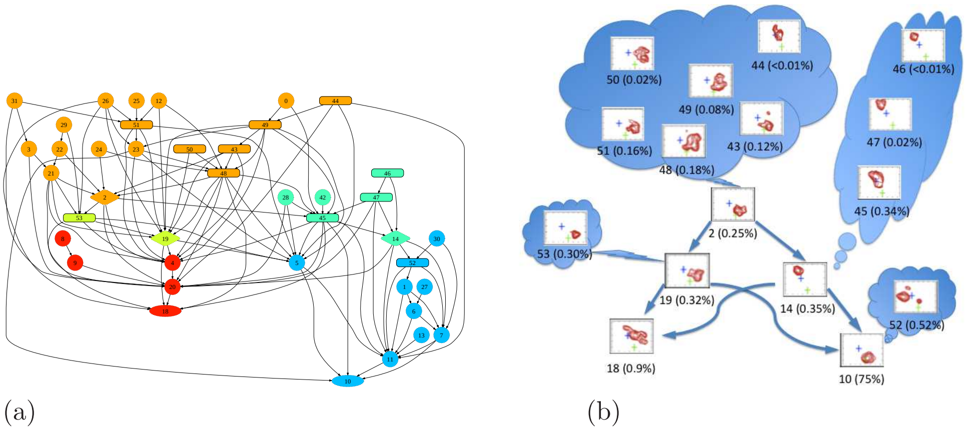


FIG. 5. *The landscape of the LAO-protein binding network. (a) The node decomposition is represented by color components. Local minima are represented as ovals, index-*1 *and index-*2 *saddles are shown in diamonds, circular nodes are regular nodes, and rectangular nodes are solvated states* $\{43, 44, \dots, 53\}$. *(b) A concise landscape is drawn with five critical nodes and eleven solvated states with their associated critical nodes. The structure of each state is represented by a free energy plot on the plane of the twist angle and opening angle, together with its percentage of population. In both figures, arrows between nodes indicate the gradient descent direction of the free energy.*

One major challenge in molecular dynamics simulations is the temporal gap between simulation time step and biologically experimental time scales. Recently Markov State Models are built to overcome this hurdle [20]. In this approach a conformational transition network is constructed from molecular dynamics simulation data, with a Markov chain which can be used to reproduce long term behavior in simulations. Free-energy landscape with some appropriately chosen continuous reaction coordinates has been very successful in providing human visualization and quantitative predictions, which is however not clear how to be analyzed in a network setting [3]. The following example shows that our landscape analysis provides a useful tool for the free-energy landscape on conformational transition networks.

In [22] a Markov state model is constructed which captures some long term ($\gg$ 6 ns) dynamical behavior of the LAO-protein binding. Now we examine the transition network constructed in [22] as a weighted directed graph $G = (V, E, W)$, where V consists of 54 nodes, each representing a metastable state whose detailed information can be found in Appendix C, an edge $(i, j) \in E$ if transitions from node i to j are observed in simulations with 6 ns delays (the implied time scale for approximate Markovian behavior), and the number of transitions is recorded as the weight w_{ij}. Among all the states, eleven of them ($\{43, 44, \ldots, 54\}$) are solvated or unbound states, and the binding state is node 10.

Let $p_{ij} = w_{ij}/\sum_j w_{ij}$ be the transition probability from state i to state j. This defines a Markov chain with a unique stationary distribution π. The detailed balance condition of this physical system implies that $\pi_i p_{ij} \approx \pi_j p_{ji}$ up to some sampling error. We threshold this graph to an undirected graph by keeping those edges $\{i, j\}$ such that $\dfrac{w_{ij} + w_{ji}}{2} > 30$, *i.e.* average count number is larger than 30. One reason for doing this is that small numbers of transitions may be heavily influenced by the noise caused by the way of counting the transition. Note that the mean transition count is about 120, and the qualitative behavior reported below shows certain stability under the variation of the threshold value.

The free-energy function is $h(i) = -\log \pi(i)$ where π is the stationary distribution of metastable states. Application of the method above gives rise to a landscape shown in Figure 5. Isolated states are dropped in this picture. Colors in this picture illustrate the node decomposition according to Theorem 1, where each color component represents the attraction basin of a critical node. Below we shall discuss structural properties of these nodes. A complete picture of structural information for all 54 states can be found in Appendix C.

The landscape exhibits two local minima, the bound state 10 and the misbound state 18. Bound state 10 is also the global minimum with the largest population of 74.9%, in contrast to the misbound state 18 with only 0.9% population. In the bound basin colored in light blue, there are two encounter states 11 (population 13.5%) and 5 (population 1.15%) where the ligand is in or close to the binding site and conformations in this state have a small (positive or negative) twisting but large opening angles. In the misbound state 18, the ligand interacts with the protein outside the binding site and close to the hinge region of two domains of the protein. State 18, together with state 4, 8, 9, and 20, forms a misbound basin marked in red. In these states, the ligand interacts with the protein from a distance to the binding site. State 8 and 9 exhibit similar structural properties with a negative twisting angle and a fixed distance to the binding site (about 10Å), while state 4, 18, and 20 exhibit similar but a different type of structures.

There are two index-1 saddles (node 14 of population 0.35% and 19 of population 0.32%) as metastable intermediate states connecting the bound and the misbound basins. At these saddles ligand interacts with the protein in different ways. In state 19, the ligand interacts with the protein from one twist direction (positive) and the protein is almost closed. In contrast, in state 14 the ligand approaches the protein from the opposite twist direction (negative) and the protein is still quite open. The index-2 saddle 2 (0.25%) is a high energy misbound state, where the ligand interacts with the protein from positive twist angles and the protein is closed. These saddles play different roles in binding pathways as shown in Figure 5 (b).

All the solvated states in the basins of the saddle 14 and bound state 10, *i.e.* node $\{45, 46, 47\}$ (total population 0.37% with second largest solvated state 45) and 52 (largest solvated state of 0.52%), are featured with negative twist angles, with the difference that those in the basin of saddle 14 has larger opening angles than those in the basin of state 52. This suggests that when the ligand approaches the protein from a negative twist angle, the system will easily form an encounter complex and reach the bound state, via saddle 14 or directly. In this case, the misbound state might only trap the conformations with large opening angle.

On the other hand, the protein-ligand misbinding typically happens when the ligand approaches the protein from a positive twist angle. In fact, all the solvated states where the protein has positive twisting angles, *i.e.* $\{43, 48, 49, 50, 51\}$ and 53 (the third largest solvated state of 0.30%), lie in the basins of the misbound saddle 2 and saddle 19, respectively. The majority of misbound events occur when the ligand interacts with the protein from positive twist angles, *i.e.* states $\{18, 20, 4\}$ (total population 1.68%). The only exception is a small solvated state 44 (population $< 0.01\%$) in the basin of the misbound saddle 2, which has a negative twist angle but from small to large opening angles. Figure 5 (a) further shows that state 44 connects to both misbound state 20 and encounter state 11. This fact confirms the observation above that negative twisting may lead to binding, but when the protein is largely open, a negative twist angle may lead to misbinding. As a further evidence for the latter, in a minority of misbound basin states (states $\{8, 9\}$ of total population 0.28%) the protein has a negative twist angle and is largely open.

The analysis above can be further improved by quantitative studies using the transition path theory. For a quantitative analysis on the roles of index-1 saddles, we conduct two kinds of transition path analysis using transition path theory (see [4, 5] or Appendix B). First, we study reactive currents from the misbound state 18 to the bound state 10. This analysis shows that a majority of flux passes through the saddle 19. Therefore once the ligand and protein fall in the misbound state 18, the major pathway to escape and enter the bound state is via saddle 19.

The other analysis, as was also did in [22], studies transition paths from the eleven solvated states marked from 43 to 53 to the bound state 10. In particular, we investigate reactive currents from each of the solvated states to the bounded state, respectively. The results are summarized as follows. A large part of these details has been ignored in [22], since they only examined 10 transition pathways, ignoring the others.

1. Solvated state 52 lies in the basin of bound state 10, whence misbound state 18 has little influence on its pathway.
2. Solvated state 53 only passes through index-1 critical node 19 to enter the bound state 10, which is heavily influenced by the misbound state 18.
3. Solvated states $\{45, 46, 47\}$ lie in the basin of index-1 critical node 14 and enter the bound state 10 directly or via 14. They are not much influenced by the misbound state 18.
4. Other solvated states are in the basin of index-2 critical node 2. Transition path analysis further shows that misbound state 18 has a stronger influence on them than those in the basin of 14. In particular state 50 is mostly influenced with near 50% of transition currents trapped by the misbound state 18.

In summary, the misbound state 18 only affects some of the pathways from solvated states to the bound state. If the ligand interacts with the protein from a negative twist angle and the protein is moderately open, binding will typically happen. The misbound state mostly traps those conformations when the ligand interacts with the protein from a positive twist angle, or from a negative twist but the protein has a large opening angle. In particular, if we can design some mutation to disrupt the stability of index-1 saddles 14 (negatively twisting and largely open) and/or 19 (positively twisting and closed), we may be able to make the binding much more difficult. Finally we note that the critical node analysis can be further applied to the conformational transition networks beyond the Markovian time scale, which is important in applications as Markovian behavior only holds approximately in certain range of time scale.

4. Discussion and conclusion. We have introduced a notion of critical points for network which can be used to reduce a complex network to a coarse-grained representation while preserving structural properties associated with functional gradient flows. Examples have shown that the information obtained this way is of great value in capturing global structure and dynamics of the network, such as diffusive or reactive pathways. Moreover, the critical point analysis leads to a hierarchical decomposition which may enable us to perform multiscale analysis of complex networks. These perspectives will be systematically pursued in the future.

An interesting question is the stability of these objects against noise. To answer this question, one has to clarify the source of noise. There are two types of noise one should consider in landscape analysis of networks – one associated with the energy function h and the other associated with the network structure. The former can be dealt with traditional persistent homology denoising, where critical nodes with shallow basins can be merged with their saddles. The latter is however more challenging as there are no systematic studies yet on perturbation or bootstrap of networks. In the examples above, we used edge thresholding on the Les Misérables and the protein binding networks, which is equivalent to modeling such networks as a superposition of a signal graph and some noise as Erdös-Rényi type random graphs. In such cases one would like to develop statistics based on critical nodes, such as Fréchet means. In a summary, it will be one of our future direction to study critical points of random functions on random graphs.

Appendix A. Proofs.

Proof of Proposition 1. We show first that every local minimum in $\mathcal{B}_0$ must be an index-1 critical node. Let x be a local minimum in $\mathcal{B}_0$. Then x reaches at least two local minima, say $y_1, y_2 \in \mathcal{C}_0$. Consider the subgraph with node set

$$S = \big(\{x\} \cup \mathcal{A}(y_1) \cup \mathcal{A}(y_2)\big) \cap \{y \mid h(y) \leq h(x)\}.$$

Clearly, S is connected and x is the unique maximum node in S. By the definition of the attraction basin, the set $S\backslash\{x\}$ is not connected.

Since S is connected, it contains at least a path from y_1 to y_2. Let γ be the local minimal energy path from y_1 to y_2 in the subgraph S. As $S\backslash\{x\}$ is not connected, γ must pass x, so that $h(\gamma) = h(x)$.

We now show by contradiction that γ is also a local minimal energy path in the original graph V. Suppose we can find another path from y_1 to y_2, called $\widetilde{\gamma}$, so that γ is deformable to $\widetilde{\gamma}$. For any $z \in \widetilde{\gamma}$, we have $h(z) \leq h(x)$. Consider the set $\widetilde{\gamma} \cap \mathcal{B}_0$, which is non-empty. We distinguish two cases:

a) $\widetilde{\gamma} \cap \mathcal{B}_0 = \{x\}$. Then, $\widetilde{\gamma}\backslash\{x\} \subset \mathcal{A}(y_1) \cup \mathcal{A}(y_2)$, so that $\widetilde{\gamma} \subset S$. By construction of γ, we have $\widetilde{\gamma} = \gamma$;

b) If there exists $z \in \widetilde{\gamma} \cap \mathcal{B}_0$ and $z \neq x$, we have some point $x' \in \gamma$ that $z \prec x'$. It is easy to see that x' must be x, since other points on γ are in attraction basins of y_1 and y_2. Using Lemma 1, there exists a path $\gamma_1 = (w_0 \cdots w_n)$ from z to x ordered in energy increase. In particular, consider the point w_{n-1}, we have $z \prec w_{n-1}$ so that $w_{n-1} \in \mathcal{B}_0$. Moreover, $w_{n-1} \in \mathcal{N}(x)$ and $h(w_{n-1}) < h(x)$. This contradicts with the fact that x is a local minimizer in $\mathcal{B}_0$.

Therefore, γ is a local minimal energy path, and x is an index-1 critical node.

Let $z \in \mathcal{C}_1$ which is not a local minimum in $\mathcal{B}_0$. Then, z must reach a local minimum x in $\mathcal{B}_0$ by the gradient flow. By the first part of the proposition, $x \in \mathcal{C}_1$. The proposition is proved. □

Proof of Theorem 2. (Necessity). We first show that index-0 and index-1 critical nodes, when added into the filtration, will change Betti numbers in the way above.

For index-0 critical nodes, they are local minima of graph G. When a local minima is added into the filtration, it must create a new connected component which increases the 0-th Betti number, β_0.

Index-1 saddles will play a more complicated role. We have two situations

- if an index-1 saddle lies on top of a global minimal energy path, it will decrease β_0 upon being added;
- if an index-1 saddle lies on top of a local minimal energy path other than the global one, it will increase β_1 upon being added.

Given a pair of index-0 critical nodes $y_1, y_2 \in \mathcal{C}_0$, among all local minimal energy paths connecting them (if exist), there must be a global minimal energy path γ_0, so that $h(\gamma_0)$ is less than any other local minimal energy paths between y_1 and y_2. We denote the maximal node of the global minimal energy path as x. Such x is an index-1 critical node. When x is added into the filtration, the 0-th Betti number β_0 will decrease as x connects two components contains y_1 and y_2 respectively.

For the other local minimal energy paths connecting y_1 and y_2, the associated index-1 critical nodes will increase the first Betti number β_1 when added into the filtration. Indeed, let z be such an index-1 critical node. Thus z is a maximum of a local minimum energy path γ_1 such that $h(\gamma_1) = h(z) > h(x) = h(\gamma_0)$. γ_1 is not deformable to the global minimal energy path γ_0 between y_1 and y_2. Then two paths γ_0 and γ_1 forms a loop, and hence the first Betti number β_1 increases when z is added into the filtration.

(Sufficiency). We show next that no other nodes when added into the filtration will change the first two Betti numbers in the same way.

For any node x which lies in the attraction basin of a local minima $\mathcal{A}_0(x_0)$ for some $x_0 \neq x$, x reaches x_0 by gradient flow. For any edge $\{x, x'\} \in E$ with $x' \in \mathcal{A}_0(x_0)$, x' reaches x_0 and thus the triangle $\{x, x', x_0\}$ is included in the simplicial complex. This implies that $\mathcal{A}_0(x_0)$ is contractible (star-shape), whence no node in $\mathcal{A}_0(x_0)$ other than local minimum x_0 will change Betti numbers.

It remains to show that any node in boundary $\mathcal{B}_0 \backslash \mathcal{C}_1$ will not change Betti numbers in the same way. Any such node $z \in \mathcal{B}_0$ must reach at least two local minima, say a and b. Then by Lemma 1 there is a path $\gamma = (a = w_0, \ldots, z = w_k, \ldots, b = w_l)$ for some $l \in \mathbb{N}$ such that $h(w_s) < h(w_{s+1})$ for $s \leq k-1$ and $h(w_s) > h(w_{s+1})$ for $s > k$. Moreover $z \notin \mathcal{C}_1$ implies that γ is deformable to a local minimal energy path $\pi = (a = v_0, \ldots, b = v_m)$ between the same end nodes, for some $m \in \mathbb{N}$. z can not decreases number of connected components as the path π, which appears first in the filtration, already connects a and b.

Now we show that the path γ will not create a loop either. Let $\pi_t = c \in \mathcal{C}_1$ be the maximal node on π. We must have $c \prec z$. To see this, as γ is deformable to π, there is a node $c' = w_{k'} \in \gamma$ which reaches $c \in \pi$. We may assume $c' \neq z$ ($k' \neq k$) since otherwise we are done. Then, by the construction of the path γ, we have $c' \prec z$, and hence $c \prec z$.

Note that both z and c reach both local minima a and b, node w_i with $i < k$ ($i > k$) reaches a (b, respectively), and node v_i with $i < t$ ($i > t$) reaches a (b, respectively). These will create a set of triangles such that γ is homotopy equivalent to π, *i.e.* loop-free. □

Proof of Theorem 3. The proof is analogous to that of Theorem 2. □

Appendix B. Transition path theory. The energy landscape gives us a global picture for the different attraction basins on the network. To understand the dynamics between the different basins, the transition path theory (TPT) provides a natural tool.

The transition path theory was originally introduced in the context of continuous-time Markov process on continuous state space [4] and discrete state space [14], see [5] for a review. Another description of discrete transition path theory for molecular dynamics can be also found in [19]. Here we adapt the theory to the setting of discrete time Markov chain with transition probability matrix P. We assume reversibility in the following presentation, the extension to non-reversible Markov chain is straightforward.

Given two sets A and B in the state space V, the transition path theory tells how these transitions between the two sets happen (mechanism, rates, etc.). If we view A as a reactant state and B as a product state, then one transition from A to B is a reaction event. The reactve trajectories are those part of the equilibrium trajectory that the system is going from A to B. To make the notion more precise, define the ordered family of times $\{n_j^A, n_j^B\}$ such that

$$\begin{aligned}&X_{n_j^A} \in A, \quad , X_{n_j^B} \in B,\\ &X_n \in V\backslash(A \cup B), \quad \forall n, n_j^A < n < n_j^B.\end{aligned}$$

Hence, a reaction happens from time n_j^A to time n_j^B.

Definition B-3. *Given any equilibrium trajectory $\{X_n\}$, we call each portion of the trajectory of between n_j^A and n_j^B a AB-*reactive trajectory*. We call the time during which the reaction occurs the* reactive times

$$R = \bigcup_{j \in \mathbb{Z}} (n_j^A, n_j^B). \tag{10}$$

The central object in transition path theory is the committor function. Its value at x gives the probability that a trajectory starting from x will hit the set B first than A, *i.e.*, the success rate of the transition at x. Given two sets A and B in the state space, q satisfies the equation

$$\begin{cases} \sum_{y \in V} p_{xy} q(y) - q(x) = 0, & x \notin A \cup B; \\ q(x) = 0, & x \in A; \\ q(x) = 1, & x \in B, \end{cases} \tag{11}$$

The committor function provides natural decomposition of the graph. If $q(x)$ is less than 0.5, x is more likely to reach A first than B; so that $\{x \mid q(x) < 0.5\}$ gives the set of points that are more attached to set A.

Once the committor function is given, the statistical properties of the reaction trajectories between A and B can be quantified. We state several propositions characterizing transition mechanism from A to B. The proof of them is an easy adaptation of [4, 14] and will be omitted.

Proposition B-4 (Probability distribution of reactive trajectories). *The probability distribution of reactive trajectories*

$$\pi_R(x) = \mathbb{P}(X_n = x, n \in R) \tag{12}$$

is given by

$$\pi_R(x) = \pi(x)q(x)(1 - q(x)). \tag{13}$$

The distribution π_R gives the equilibrium probability that a reactive trajectory visits x. It provides information about the proportion of time the reactive trajectories spend in state x along the way from A to B.

Proposition B-5 (Reactive current from A to B). *The reactive current from A to B, defined by*

$$J(xy) = \mathbb{P}(X_n = x, X_{n+1} = y, \{n, n+1\} \subset R), \tag{14}$$

is given by

$$J(xy) = \begin{cases} \pi(x)(1 - q(x))p_{xy}q(y), & x \neq y; \\ 0, & \textit{otherwise}. \end{cases} \tag{15}$$

The reactive current $J(xy)$ gives the average rate the reactive trajectories jump from state x to y. From the reactive current, we may define the effective reactive current on an edge and transition current through a node which characterizes the importance of an edge and a node in the transition from A to B, respectively.

Definition B-6. *The* effective current *of an edge xy is defined as*

$$J^+(xy) = \max(J(xy) - J(yx), 0). \tag{16}$$

The transition current *through a node $x \in V$ is defined as*

$$T(x) = \left\{ \begin{array}{ll} \sum_{y \in V} J^+(xy), & x \in A \\ \sum_{z \in V} J^+(zx), & x \in B \\ \sum_{y \in V} J^+(xy) = \sum_{z \in V} J^+(zx), & x \notin A \cup B \end{array} \right. \tag{17}$$

In applications one often examines partial transition current through a node connecting two communities $V^- = \{x : q(x) < 0.5\}$ and $V^+ = \{x : q(x) \geq 0.5\}$, *e.g.* $\sum_{y \in V^+} J^+(xy)$ for $x \in V^-$, which shows relative importance of the node in bridging communities.

The reaction rate ν, defined as the number of transitions from A to B happened in a unit time interval, can be obtained from adding up the probability current flowing out of the reactant state. This is stated by the next proposition.

Proposition B-7 (Reaction rate). *The reaction rate is given by*

$$\nu = \sum_{x\in A, y\in V} J(xy) = \sum_{x\in V, y\in B} J(xy). \tag{18}$$

Finally, the committor functions also give information about the time proportion that an equilibrium trajectory comes from A (the trajectory hits A last rather than B).

Proposition B-8. *The proportion of time that the trajectory comes from A (resp. from B) is given by*

$$\rho^A = \sum_{x\in V} \pi(x)q(x), \quad \rho^B = \sum_{x\in V} \pi(x)(1-q(x)). \tag{19}$$

Appendix C. Supplementary figures.

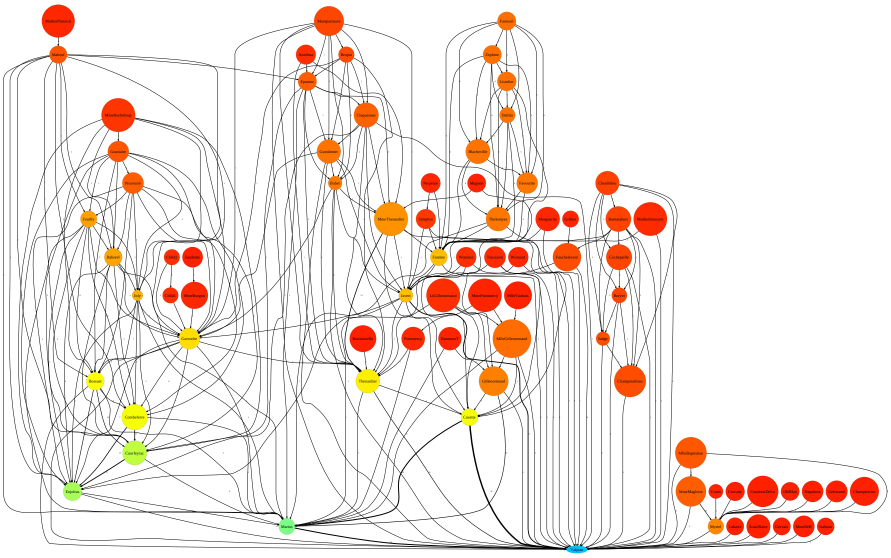


Fig. 6. *The Les Misérables Network. The whole network has 77 nodes as main characters in Victor Hugo's novel, Les Misérables, where Valjean is the only local minimum (global minimum) as the novel is written around his experience.*

The first figure is the whole co-appearance network of 77 main characters in the novel, Les Misèrables, by Victor Hugo [13]. It is an undirected weighted graph with edge weights as the number of co-appearances for a pair of characters. Without thresholding this network contains one local minimum, Valjean. However a thresholding with edge weight greater than 7 gives rise to the subnetwork in the main text.

The second figure contains a list of structural information on 54 metastable states. It contains a typical crystal structure in each state, and some free energy plots on

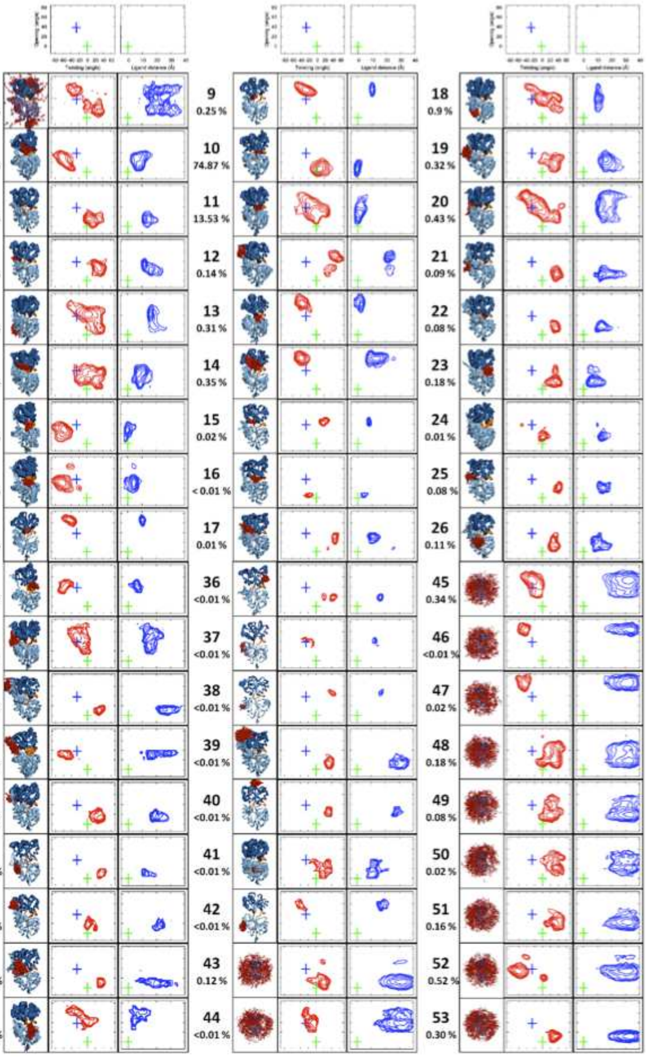


FIG. 7. *(Courtesy by Xuhui Huang) Three types of pictures for each of the 54 metastable states: on the left is the crystal structure of a representative conformation in each state, on the right are free energy plots of the protein opening angle versus twisting angle (O, T) (red), as well as the distance between the ligand and the binding site versus the opening angle (L,O) (blue). The green and blue crosses correspond to X-ray structures of the bound (PDB ID: 1LAF) and apo (PDB ID: 2LAO) conformations respectively.*

certain reaction coordinates. From these pictures one can read various structural properties of critical nodes in LAO-protein binding transition network discussed in the main text. More information about this system can be found in [22].

The third figure shows the ranking of transition currents out of misbound state 18 over eleven transition pathways. The experiment selects each of the eleven solvated states $\{43, \dots, 53\}$ as the source set and the misbound state 10 as the common target set. In each of the eleven experiments, relative transition current out of state 18 divided by total transition current from the source, is recorded and plotted in a descending order.

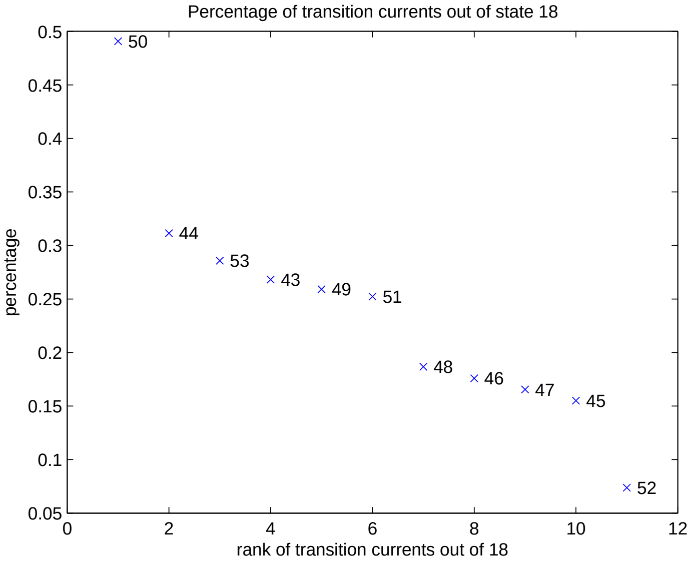


FIG. 8. *Transition Currents out of misbound state 18, with source set from each of solvated states* $\{43, \dots, 53\}$ *and target set as bound state* 10.

REFERENCES

[1] P. J. BICKEL AND A. CHEN, *A nonparametric view of network models and newman-girvan and other modularities*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 21068–21073.

[2] G. CARLSSON, *Topology and data*, Bulletin of the American Mathematical Society, 46:2 (2009), pp. 255–308.

[3] J. D. CHODERA AND V. S. PANDE, *The social network (of protein conformations)*, Proceedings of the National Academy of Sciences, 108:32 (2011), pp. 12969–12970.

[4] W. E AND E. VANDEN-EIJNDEN, *Towards a theory of transition paths*, J. Stat. Phys., 123 (2006), pp. 503–523.

[5] W. E AND E. VANDEN-EIJNDEN, *Transition-path theory and path-finding algorithms for the study of rare events*, Annual Review of Physical Chemistry, 61 (2010), pp. 391–420.

[6] H. EDELSBRUNNER AND J. HARER, *Persistent homology: a survey*, Contemporary Mathematics (2008), pp. 1–26.

[7] H. EDELSBRUNNER, J. HARER, AND A. ZOMORODIAN, *Hierarchical morse-smale complexes for piecewise linear 2-manifolds*, Discrete and Computational Geometry, 30:1 (2003), pp. 87–107.

[8] H. EDELSBRUNNER, D. LETSCHER, AND A. ZOMORODIAN, *Topological persistence and simplification*, Discrete and Computational Geometry, 28:4 (2002), pp. 511–533.

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

[10] R. GHRIST, *Barcodes: the persistent topology of data*, Bulletin of the American Mathematical Society, 45:1 (2007), pp. 61–75.

[11] J. A. HARTIGAN, *Consistency of single linkage for high-density clusters*, J. Amer. Statist. Assoc., 76 (1981), pp. 388–394.

[12] O. KNILL, *A graph theoretical poincare-hopf theorem*, 2012, preprint, arXiv:1201.1162.

[13] D. E. KNUTH, *The stanford graphbase: A platform for combinatorial computing*, Addison-Wesley, 1993.

[14] P. METZNER, C. SCHÜTTE, AND E. VANDEN-EIJNDEN, *Transition path theory for markov jump processes*, Multiscale Model. Simul., 7 (2009), 1192.

[15] J. MILNOR, *Morse theory*, Princeton University Press, 1963.

[16] N. MILOSAVLJEVIC, D. MOROZOV, AND P. SKRABA, *Zigzag persistent homology in matrix multiplication time*, Proceedings of the 27th Annual Symposium on Computational Geometry (SoCG'11) (2011), pp. 216–225.

[17] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Phys. Rev. E, 74:3 (2006), 036104.

[18] ———, *Modularity and community structure in networks*, Proc. Natl. Acad. Sci. USA, 103 (2006), pp. 8577–8582.

[19] F. Noè, C. Schütte, E. Vanden−Eijnden, L. Reich, and T. R. Weikl, *Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations*, Proceedings of the National Academy of Sciences of the United States of America, 106:45 (2009), pp. 19011–19016.

[20] V. S. Pande, K. Beauchamp, and G. R. Bowman, *Everything you wanted to know about markov state models but were afraid to ask*, Methods, 52:1 (2010), pp. 99–105.

[21] K. Rohe, S. Chatterjee, and B. Yu, *Spectral clustering and the high-dimensional stochastic block model*, Annals of Statistics, 39:4 (2011), pp. 1878–1915.

[22] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado, and X. Huang, *A role for both conformational selection and induced fit in ligand binding by the lao protein*, PLoS Computational Biology, 7:5 (2011), e1002054.

[23] Z. Toroczkai and K. E. Bassler, *Jamming is limited in scale-free systems*, Nature, 428 (2004), 55455.

[24] Z. Toroczkai, B. Kozma, K. E. Bassler, N. W. Hengartner, and G. Korniss, *Gradient networks*, Journal of Physics A: Mathematical and Theoretical, 41 (2008), 155103.

[25] D. J. Wales, *Energy landscapes*, Cambridge University Press, 2003.

[26] B. Yang, J. M. Liu, and J. F. Feng, *On the spectral characterization and scalable mining of network communities*, IEEE Transactions on Knowledge and Data Engineering, 24 (2012), pp. 326–337.

[27] W. W. Zachary, *An information flow model for conflict and fission in small groups*, Journal of Anthropological Research, 33:4 (1977), pp. 452–473.

[28] Q. Zhou and W. H. Wong, *Reconstructing the energy landscape of a distribution from monte carlo samples*, The Annals of Applied Statistics, 2:4 (2008), pp. 1307–1331.