Mathematical frameworks for quantitative network analysis

Cotiso Andrei Bura

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mathematics

Christian M. Reidys, Chair
Peter Haskell
Henning S. Mortveit
Daniel Orr

October 15, 2019
Blacksburg, Virginia

Keywords: graph theory, cores, topology, nerve homology, bi-structures

Mathematical frameworks for quantitative network analysis

Cotiso Andrei Bura

(ABSTRACT)

This thesis is comprised of three parts. The first part describes a novel framework for computing importance measures on graph vertices. The concept of a D-spectrum is introduced, based on vertex ranks within certain chains of nested sub-graphs. We show that the D-spectrum integrates the degree distribution and coreness information of the graph as two particular such chains. We prove that these spectra are realized as fixed points of certain monotone and contractive SDSs we call t-systems. Finally, we give a vertex deletion algorithm that efficiently computes D-spectra, and we illustrate their correlation with stochastic SIR-processes on real world networks. The second part deals with the topology of the intersection nerve for a bi-secondary structure, and its singular homology. A bi-secondary structure R, is a combinatorial object that can be viewed as a collection of cycles (loops) of certain at most tetravalent planar graphs. Bi-secondary structures arise naturally in the study of RNA riboswitches - molecules that have an MFE binary structural degeneracy. We prove that this loop nerve complex has a euclidean 3-space embedding characterized solely by $H_2(R)$, its second homology group. We show that this group is the only non-trivial one in the sequence and furthermore it is free abelian. The third part further describes the features of the loop nerve. We identify certain disjoint objects in the structure of R which we call crossing components (CC). These are non-trivial connected components of a graph that captures a particular non-planar embedding of R. We show that each CC contributes a unique generator to $H_2(R)$ and thus the total number of these crossing components in fact equals the rank of the second homology group.

Mathematical frameworks for quantitative network analysis

Cotiso Andrei Bura

(GENERAL AUDIENCE ABSTRACT)

This Thesis is divided into three parts. The first part describes a novel mathematical framework for decomposing a real world network into layers. A network is comprised of interconnected nodes and can model anything from transportation of goods to the way the internet is organized. Two key numbers describe the local and global features of a network: the number of neighbors, and the number of neighbors in a certain layer, a node has. Our work shows that there are other numbers in-between the two, that better characterize a node. We also give explicit means of computing them. Finally, we show that these numbers are connected to the way information spreads on the network, uncovering a relation between the network's structure and dynamics on said network. The last two parts of the thesis have a common theme and study the same mathematical object. In the first part of the two, we provide a new model for the way riboswtiches organize themselves. Riboswitches, are RNA molecules within a cell, that can take two mutually opposite conformations, depending on what function they need to perform within said cell. They are important from an evolutionary standpoint and are actively studied within that context, usually being modeled as networks. Our model captures the shapes of the two possible conformations, and encodes it within a mathematical object called a topological space. Once this is done, we prove that certain numbers that are attached to all topological spaces carry specific values for riboswitches. Namely, we show that the shapes of the two possible conformations for a riboswich are always characterized by a single integer. In the last part of the Thesis we identify what exactly in the structure of riboswitches contributes to this number being large or small. We prove that the more tangled the two conformations are, the larger the number. We can thus conclude that this number is directly proportional to how complex the riboswitch is.

Dedication

To the Paper & Pendragons,

Acknowledgments

I'd like to thank my advisor and chair, Christian Reidys, for the knowledge and wisdom imparted, his unerring support and his boundless patience when I endlessly inquired on matters of mathematics and life. I'd also like to thank my committee members for their willingness to guide me and the valuable advice they provided during this process. Finally, I'd like to thank the colleagues in my research group for the insightful conversations without which this work would, no doubt, be lacking.

Attributions

This small section contains the mandatory attributions to the papers in the Thesis.

For the first part:

Ricky X. F. Chen and Christian M. Reidys planned and performed this research. Ricky X. F. Chen also partly implemented the simulation. Andrei C. Bura implemented the simulation and performed the research. All authors discussed the results, wrote the paper and reviewed the manuscript.

For the second and third parts:

All authors planned and performed the research, discussed the results, wrote the papers and reviewed the manuscripts.

Contents

Introduction

Overview

This small chapter aims to provide context for the three parts (papers) of the Thesis as well as an overview of the work performed in each of them. This chapter is divided into two sections, thematically. The first section deals with a graph theoretical framework geared towards general network decomposition, and its scope is the first paper. The last two sections have a common theme, both dealing with a topological perspective on certain graphs, and so were collated in a single section with subsections dedicated to each of the two papers respectively.

Graph Theoretical Framework

From modelling the flow of information within social networks, to casting the atomic lattice structure of crystals, the usefulness of the notion of a network is undeniable. Most interaction structures in the physical world benefit from this abstraction. A fundamental, yet difficult to nail down question that people studying networks encounter is: "What is the most important node within a network?". As we shall see, the importance of a node is inextricably linked to the the inquirer's own definition of importance.

In the following, we will use the terms network and graph interchangeably. When we say graph within this contex, we will always mean a simple graph. Namely, a graph that is unweighted, undirected and free of loops or multiple edges. We will usually denote such an object by $G(V, E)$, where V is the set of vertices of the graph G, and $E \subset V \times V$ is its set of edges. For the sake of simplicity, although not necessary for the treatment in the following section, we will restrict our attention to connected graphs only. From a graph theoretical

standpoint, many established tools exist that capture, to varying degrees, the local or global structure of a graph. Among the most well studied ones are the degree sequence and the core decomposition. The degree sequence assigns to each vertex $v \in V$ a number $deg(v) \in \mathbb{N}$ called its degree - the number of vertices in G that share an edge with v. Note that this sequence captures purely local information about the vertex v. This restricts the knowledge about the structure of G than can be inferred from $deg(v)$ to that of the neighborhood of v. For the core number of a vertex v, we first construct a nested sequence of vertex induced sub-graphs that form a cover of G:

$$G = G_0 \geq G_2 \geq \ldots \geq G_{\Delta(G)} \ , \ \Delta(G) = \max_{v \in V} deg(v)$$

where G_k is the maximal sub-graph of G with the property that any vertex within G_k has at least k neighbors in G_k. We call these sub-graphs the cores of G. The core number of a vertex $v \in V$ is then the maximum k such that $v \in G_k$. We note that the core numbers of G provide global information about the structure of G, with v having a higher core number if it resides within a more connected (in terms of degree) community of vertices within G. The thing of note is that between these two numbers, one local and the other global, the change in information provided about the structure of G is rather abrupt. The goal of the first paper in this Thesis was to provide a unified framework for discussing both these numbers. We showed that the framework leads to intermediate structures that gradually "interpolate" the information from the local scale of the degree to the global scale of the cores. The spectrum obtained, carries information about G at various structure scales and is thus a more efficient predictor of a vertex v's importance than the degree of v or the coreness of v alone.

Overview of Part One: D-chain tomography of networks: a new structure spectrum and an application to the SIR process

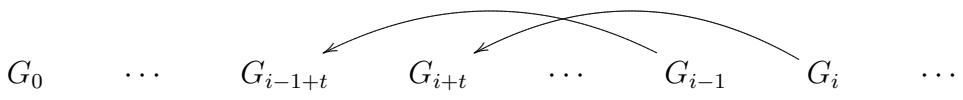
One of the key observations in this paper is the following: when we consider the i core of G, G_i, the definition of $v \in G_i$, is the requirement that v has at least i neighbors in G_i ,i.e., for a $v \in G_i$, in the chain of cores

$$G = G_0 \geq G_2 \geq \ldots \geq G_i \geq \ldots$$

v references its starting position when considering neighbors. The objects of study of this first paper are the chains obtained when the "referencing" is not the "core" that v belongs to when considering its neighbors. Namely, by introducing a parameter $t \in \mathbb{Z}$, we can define a new chain of sub-graphs that forms a cover of G. Let

$$L : G = G_0 \geq G_1 \geq \ldots \geq G_i \geq \ldots$$

be a chain of vertex induced sub-graphs of G. L is a D-chain of order $t \leq 0$ if for any i, and any $v \in G_i$, v has at least i neighbhors in G_j for $j = max\{0, i + t\}$. Thus the referencing for $t \leq 0$ fixed, is to the left of the initial position in the chain.



We further show that, under a certain definition of maximality, there is a unique maximal D-chain of order t, for every order $-\Delta(G) \leq t \leq 0$. This means that, to a fixed vertex $v \in V$, we can associate a vector of integer values comprised of the various modified "core" numbers for each of the values of the parameter t (i.e. the right most positions in L where v still appears). The vector obtained in this way we call the D-spectrum of v. Note that, for $t = 0$

the modified core number is in fact the standard coreness of v. This is since the referencing is now $i + t = i$ and, this is just the definition of the standard i core. For $t = -\Delta(G)$ the referencing is always G. As such, the modified core number obtained is the degree of the vertex v. Hence the first and last value of the D-spectrum for v, is the core number and the degree of v respectively.

The paper then moves on to methods of computing D-spectra for a given graph. We first introduce and discuss various properties of MC-systems. These are discrete sequential dynamics on a graph, where the vertex local functions are monotone and contractive. These two properties of the local functions imply a certain phase-space structure of the limit-cycles of these systems - namely, that all these limit cycles are in fact fixed points. It can be shown that if the original vertex states possess a linear order, then certain poset relations on these fixed points also arise. The main result of the following section is that if we were to interpret the modified core numbers for all the vertices in G, for $-\Delta(G) \leq t \leq 0$ fixed, as a state space vector, then this vector can be computed as the fixed point of a specific type of discrete sequential dynamical system we introduce, called a t-system. This system's underlying graph is G itself, while the local functions at each vertex are monotone. It can then be shown that, along the transient of the degree sequence of G, interpreted as a state vector for such a t-system, the local functions are also contractive. Hence, along this phase space transient, the particular t-system is in fact an MC-system. Levying the theory we previously developed on MC-systems we can then show that given a t-system, for $-\Delta(G) \leq t \leq 0$ fixed, its fixed point along the aforementioned transient is in fact the sequence of modified core numbers obtained from a maximal D-chain of order t.

We further prove that these modified core numbers can also be computed for all vertices simultaneously, via a vertex deletion algorithm which is an augmentation of the classical vertex deletion algorithm used for the computation of standard cores.

Finally, we examine the potency of using vertex specific D-spectra as a vertex importance

measure. To this end, we note that a vertex's importance is dependent on the definition of importance we desire to employ. As such, we choose as a reasonable definition, a vertex's spreading power within a discrete stochastic percolation process on G called an SIR process. We set the vertex to be examined, v, in an infected state, while all the other vertices are set to be susceptible. Then, at each time step, we allow the susceptible neighbors of infected vertices to become infected in a probabilistic fashion. The probability of infection of a neighboring vertex depends on a virality parameter that is fixed throughout the simulation. This probability also depends on the number of current infected neighbors the susceptible vertex in question has. At the end of each step, any previously infected vertices become recovered and are no longer infectious nor can they be subsequently infected. The proportion of vertices of G in the recovered state, once the simulation is completed and no more infection events occur, is the spreading power of the initial source vertex v. We average this spreading power for a vertex over many such simulations. A case can be made that these average spreading powers are a natural importance measure for the vertices in the graph G. This is since they model stochastic percolation along the graph G, and as such, are "coupled" to G's global structure. Finally, we conduct various tests to see how well correlated our D-spectrum importance measure is to this "natural" average spreading power measure. We find that our D-spectra outperform degree and core sequences as well as h-indices in terms of this correlation.

Topological Framework

RNA is a single stranded nucleic acid that self organizes into a variety of conformations and is of crucial importance to bio-processes within living organisms. This molecule's folding is mainly studied within the paradigm of secondary structures. These are planar arc diagrams,

drawn as non-crossing arcs in the upper half-plane, that represent the base pairing of the strand's nucleotides within a conformation. Such a structure also has a loop decomposition, each loop being a particular set of nucleotides that are part of, or subtended by, specific arcs in the diagram. A loop also corresponds to a boundary component of the secondary structure when said structure is considered as an orientable fat-graph (See Figure 1).

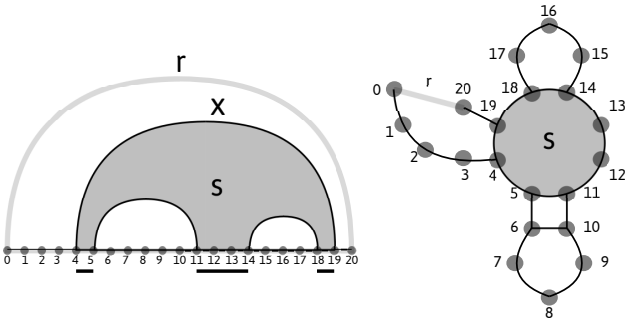


Figure 1: LHS: a secondary structure, S, and a distinguished loop $s = [4, 5] \cup [11, 14] \cup [18, 19]$. r and x are arcs. RHS: S represented as a planar RNA molecule.

As in this model any two loops within a given secondary structure intersect either trivially or in exactly two nucletides (vertices), this makes secondary structures amenable to recursions. Thus, dynamic programming routines can be employed for the computation of structure dependent parameters of the molecules the secondary structures model. For instance, the recursive loop-decomposition for secondary structures can facilitate the computation of the molecule's free energy.

Bi-secondary structures are pairs of such RNA secondary structures. A bi-secondary structure is represented by drawing its respective secondary structures in the upper and lower half-plane while the set of nucleotides remains the same. Such a bi-structure also has a loop decomposition, but in bi-secondary structures the intersection of loops is usually more complex.

The various ways the loops can intersect within a bi-structure is of crucial importance to current algorithmic work being done in bio-informatics and evolutionary optimization. To

better understand the complexity of such loop intersections, we cast this in the language of topology and study the associated topological space that captures these intersections.

Overview of Part Two: Loop Homology of Bi-secondary Structures

In this paper we construct an object $K(R)$, called a loop nerve. We do this by associating a d-dimensional simplex to a d+1-fold non-trivial loop intersection within the loop decomposition for a given bi-secondary structure (bi-structure for short) $R = (S, T)$, where S and T are the two secondary structures involved. Collecting all these simplices into a complex, we then study the simplicial homology of the associated topological space. The main result of this paper is that this space's homology sequence is trivial except for the second homology group, $H_2(R)$, which, in fact, can be show to be free abelian.

We begin from the simple observation that, the planarity of S and T requires that any five-fold intersections of the loops of R be trivial. As such, we can immediately conclude that $H_{d\geq 4} = 0$. The triviality of $H_3(R)$ follows from identifying certain "exposed" 2-faces for any 3-simplex in $K(R)$. We prove that any such 3-simplex possesses at least two 2-faces that are not faces of any other 3-simplex in $K(R)$. This allows us to use a marking scheme for any linear combination of 3-simplices in the $Ker(\partial_3)$, that then allows us to show that each one of their coefficients must be zero and hence this kernel is trivial. Since the $H_4(R) = 0$ the claim then follows. Incidentally, a somewhat similar marking argument is used to show that $H_2(R)$ is free abelian. The bulk of the difficulty resides in showing that $H_1(R) = 0$. To accomplish this an induction scheme is set up as follows: Firstly, for a fixed loop t of R, a graph denoted by Δ_t is shown to exist. This graph is defined from the 1-skeleta of the t-neighboring loops within $K(R)$ subject to certain constraints. Then, building R by inductively adding arcs to bi-structures for which we can assume by hypothesis that $H_2 = 0$, we show that the contributions to H_2 by adding t are also trivial. We do this by

sequentially processing a t contribution along Δ_t. The complexity of this argument resides in showing the existence of the loop specific Δ_t graphs. This is accomplished by recursively decomposing the bi-structure. We do this by removing certain distinguished arcs which result in simpler bi-structures for which we can inductively hypothesize that Δ_t graphs exist. We then observe what changes occur in the Δ_t graphs corresponding to the simpler structures, when the distinguished arcs are added back in. We then show that the newly obtained graphs resulting from these changes are Δ_t graphs for the more complicated bi-structures. The rank of H_2 is thus the only non-trivial discriminator, in terms of the loop nerve, when it comes to bi-structures.

Finally, we examined the rank of H_2 for known bi-structures in the form of naturally occurring riboswiches. We observe that, contrary to random secondary structure pairs, riboswitches are of rank one, while the random pairs tend to be of rank strictly higher than one.

Overview of Part Three: Loop Homology of Bi-secondary Structures II

While in the second part we prove that $H_2(R)$ is free abelian, we've yet to identify the combinatorial object within the diagram of the bi-structure R that contributes a generator to $H_2(R)$. In the third part we do precisely that.

We first introduce the notion of a crossing component (CC). This is a non-trivial connected component of a graph associated to R whose vertices are the arcs of R. An edge $(s \in S, t \in T)$ is drawn between the arcs s and t from the upper and lower half planes of the diagram of R respectively, if and only if, when flipping t to the upper half plane it crosses s. We denote the set of crossing components of R by $\chi(R)$. The main result of the third part is the proof that the rank of $H_2(R)$ is equal to the number of crossing components of R.

In order the prove this, we first require the notions of a decoration at a nucleotide and that

of a closure for a given CC. A decoration is a copy of a 2-simplex from $K(R)$, that is labeled by one of the nucleotides present in the loop intersection that corresponds to said 2-simplex. To such a decoration there corresponds a unique arc, either from S or T. A closure is a set of decorations corresponding to the arcs in the CC. We first prove that a given closure, simplicially glues into the triangulation of a 2-sphere within $K(R)$, and two closures correspond to distinct such spheres. Thus, for each closure there corresponds a unique, distinct generator in $H_2(R)$.

To show that the closures are solely responsible for the existence of all generators, we first distinguish the case when R has no nucleotides of degree four in its arc diagram. By construction, in this case, this means that no 3-simplices are present in $K(R)$. For this particular class of bi-structures, we then prove that there exits a decomposition of $K(R)$ into "irreducible" sub-complexes, of which all the aforementioned CC spheres are part of. We show that this decomposition's overall structure is "tree-like". This can provide the basis for recursively processing elements from $Ker(\partial_2)$ in order to show that each closure of a CC, i.e. each CC sphere, contributes exactly one generator to this group, and furthermore, that all generators are accounted for in this way. This allows us to conclude that, for R with no degree four nucleotides in its arc diagram, the rank $r(H_2(R)) = |\chi(R)|$.

Finally, to conclude the proof for the general case, i.e. when degree four nucleotides might be present, we show that there exists a particular injective mapping between such bi-structures and bi-structures with more nucleotides, same number of arcs, but no degree four nucleotides. We then show that the nerve complexes of a bi-structure and its image under this map, are homotopy equivalent as topological spaces and, as such, their homology groups are isomorphic. Furthermore, we prove that this degree four removal map does not introduce new crossing arcs and so preserves the number of crossing components. Thus, we can conclude that for an arbitrary R, $r(H_2(R)) = |\chi(R)|$.

Part One: D-chain tomography of networks: a new structure spectrum and an application to the SIR process

D-chain tomography of networks: a new structure spectrum and an application to the SIR process*

Ricky X. F. Chen†, Andrei C. Bura‡, and Christian M. Reidys§

Abstract. The analysis of the dynamics on complex networks is closely connected to structural features of the latter. In this context, features like, cores and node degrees have been studied ubiquitously. Here we introduce the D-spectrum of a network, a novel framework that is based on a collection of nested chains of subgraphs within the network and developed rigorously from the mathematical point of view. Each such chain gives rise to a ranking of nodes and, for a fixed node, the collection of these ranks provides us with its D-spectrum. Within this framework, cores and node degrees become rankings of two particular such chains, whence D-spectra integrate both concepts. As for computing the D-spectra, we present a node deletion algorithm, similar to that of k-cores and furthermore establish a connection between the D-spectra and fixed points of certain sequential dynamical systems. Finally, in order to show applicability we employ D-spectra in order to identify nodes of similar spreading power in the susceptible-infectious-recovered (SIR) model for a variety of real world networks. Simulation results show that D-spectra provide a meaningful augmentation of the concepts of cores and node degrees.

Key words. D-chain, Network structure, Spreading process, k-core, SIR model, Fixed point

AMS subject classifications. 05C70, 05C82, 93C55

1. Introduction.

Structural properties of complex networks are of central importance for understanding the formation principles of said networks and dynamics associated to them. Various network features have been studied, such as node degree [1, 2], path distance [3, 4], k-core decomposition [5, 6, 7], motif identification [8] and community identification [9, 10]. Spreading dynamics on networks, such as, for instance, information diffusion, knowledge dissemination, disease spreading, etc. [11, 12, 13, 14, 15, 16, 17] is ubiquitous and has been studied extensively. In the analysis particular focus has been put on the identification of nodes, that are the most effective spreaders [18, 19, 20, 21, 22]. Their localization is of key relevance for designing strategies to decelerate or stop the spread, for instance in infectious disease outbreaks, or accelerate the process in the case of knowledge dissemination.

At first glance, the most connected nodes (hubs) seem to be natural candidates for being "good" spreaders. However, Kitsak et al. [23] argue that the "location" of a node is more important than its degree, where said location is characterized by its core number [24]. These two perspectives differ significantly in that degree is a local feature while graph-cores are (potentially) extended subgraphs. Recently, h-index families were proposed as a measure, and it was shown that h-index outperforms both, degree as well as core-based measures in several cases [25]. In addition, a discussion of the integration of node degree, h-index as well as core

*Submitted to the editors DATE.

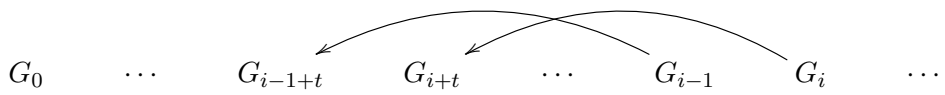
†Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, VA 22908, USA (chen.ricky1982@gmail.com).

‡Dept. of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA.

§Biocomplexity Institute and Initiative and Dept. of Mathematics, University of Virginia, Charlottesville, VA 22908, USA (duck@santafe.edu).

number was presented; a line of thought that can be also found implicitly in an earlier work of Montresor, Pellegrini and Miorandi [26].

In this paper we first present a new framework of characterizing network structure by introducing dendritic (D) spectrum of a network. The formal definition of D-chains and D-spectrum will be developed in Section 2 and here we shall present two motivations. The first originates from k-core of a network (graph) G: the k-core of G is the maximum subgraph where every vertex has degree at least k. From the definition of the k-core, the vertices contained in the k-core form a locally closed system in the sense that the interaction to the vertices outside of the k-core is not obvious and the i- and the j-cores are somewhat independent. [27] provides insight into possible drawbacks of this, where it was observed that those nodes contained in the k-core but that are weakly interacting with outside vertices are not good "spreaders". The second motivation comes from the work on protein interaction networks [28] where emphasis was put on the interaction of a particular substructure to its outside in order to quantify importance. Accordingly we may "modify" the concept of k-core introducing such interactions; that is, these modified i- and j-cores have some intertwined relation, depicted below:



As a result, we obtain D-spectra of nodes which integrate (parameterized by t), node degrees and core numbers as endpoints of a sequence, along which we have a transition from local to global information.

While the concept of D-spectrum is motivated by practical problems (e.g., spreading power and important substructures), it is not our primary objective to focus on a particular scenario in which D-spectra outperform existing approaches. Nevertheless, we shall discuss some applications. We stipulate rather, that there is a general benefit of bringing this new framework to the analysis tools of networks.

It is well-known that there is a node deletion algorithm to obtain the k-core of a graph. In Section 2, we shall show that there exists a node deletion algorithm for obtaining the D-spectra of nodes, exhibiting a substantially different property in Section 2.

In Section 3, we present an alternative approach for obtaining the D-spectra, computing certain fixed points of specific graph dynamical systems on the network which we call $[t]$-systems. In order to have a self-contained presentation of the second approach, we provide a brief investigation of MC systems, representing an abstraction of $[t]$-systems.

Finally, we show in Section 4 the applicability of the new framework, by employing D-spectra of nodes to characterize node similarity in the susceptible-infectious-recovered (SIR) model. We evaluate D-spectra for five distinct real world networks and show that they represent a meaningful augmentation of cores and node degrees.

2. D-chains and D-spectrum.

In this section, we introduce D-chains of networks. We shall use the notions graph and network as well as those of node and vertex, interchangeably.

2.1. D-chains of networks. Suppose G is a graph (without loops and multiple edges for simplicity). We write $H \leq G$ if H is a subgraph of G, and write $H < G$ if $H \leq G$ but $H \neq G$.

Let $L\colon G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_k$ be a chain of nonempty subgraphs of G, where $G_0 = G$, and G_i is a vertex-induced subgraph of G_{i-1} for $1 \leq i \leq k$. The chain L is called a *D-chain of order* $t \leq 0$ if for any $0 \leq i \leq k$, every vertex $v \in G_i$ has at least i neighbors in G_j where $j = \max\{0, i+t\}$ ($i+t$ could be negative.) We call the number k the *length* of the chain L and denote $|L| = k$.

Clearly, each graph G has a D-chain of order t for any non-positive integer t, since $G_0 = G$ is a D-chain of order t of length 0.

Lemma 2.1. *Let G be a graph. Suppose $t < t' \leq 0$. Then, any D-chain of order t' of G is a D-chain of order t of G.*

Proof. Let $L : G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_l$ be a D-chain of order t'. Then, by definition, for any i, every vertex in G_i has at least i neighbors in $G_{j'}$ ($j' = \max\{0, i+t'\}$). Note that $t < t' \leq 0 \Rightarrow i + t < i + t'$. Hence $G_{j'} \leq G_j$ where $j = \max\{0, i+t\}$. Thus, every vertex in G_i has at least i neighbors in G_j. Hence L is also a D-chain of order t of G. ∎

A D-chain of order t, $L\colon G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_k$, is called *maximal* if (i) there does not exist a D-chain L' with $|L'| > k$; and (ii) there does not exist a D-chain $L' : G'_0 \geq G'_1 \geq G'_2 \geq \cdots \geq G'_k$, where for some $1 \leq i \leq k$, $G_i < G'_i$. Clearly, for any G and non-positive integer t, there exists a unique maximal D-chain of order t, since the union of two D-chains of order t of maximum length is again a D-chain of order t of the same length.

Let $L\colon G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_k$ be the maximal D-chain of order t of G. Then L induces a *ranking* C_t of nodes, that is, $C_t(v) = i$ if and only if v is contained in G_i but not contained in G_{i+1}. Let $\Delta(G) = \max_{v \in V(G)}\{\deg(v)\}$, where $V(G)$ denotes the vertex set of G and $\deg(v)$ denotes the degree of v. We call the vector $\big(C_0(v), C_{-1}(v), \ldots, C_{-\Delta(G)}(v)\big)$ the D-spectrum of the vertex v. The collection of all maximal D-chains of G, or the D-spectra of all nodes, is called the D-spectrum of the network G.

It is easy to check that in the maximal D-chain of order $t = -\Delta(G)$, G_k is the maximal subgraph, where every vertex has degree at least k in G. Thus, the induced rank there for a vertex v is exactly the degree of v (in G). In the k-core decomposition of a graph G, the *core number* $C(v)$ of a vertex v is defined as, $C(v) = k$ if v is contained in the k-core but not $(k+1)$-core of G. Then, we have

Proposition 2.2. *Let G be a graph. Then, for any vertex $v \in V(G)$, we have the core number $C(v) = C_0(v)$.*

Proof. Let G_i be the i-core of G. Then, we have a chain $L : G_0 \geq G_1 \geq \cdots$. Note that the i-core of G is the largest subgraph of G where each node has degree at least i. This is equivalent to saying that every vertex in G_i has at least i neighbors in G_i. Thus L is a D-chain of G of order 0. Due to the maximality of the i-core, this order zero D-chain of G is maximal and hence unique, whence $C(v) = C_0(v)$. ∎

Accordingly, the two established rankings, core number and degree, become two particular entries in the D-spectrum.

For a general (not necessarily maximal) D-chain L of order t of G, we denote by $L(v) = i$ if $v \in V(G_i)$ and $v \notin V(G_{i+1})$ and call the number i the rank of v w.r.t. the chain L.

Proposition 2.3. *Let G be a graph and t be a non-positive integer. Let furthermore $L : G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_l$ be a D-chain of order t of G. Then, for any $v \in V(G)$, $L(v) \leq C_t(v)$.*

Proof. Let $C : G'_0 \geq G'_1 \geq G'_2 \geq \cdots \geq G'_k$ be the maximal D-chain of G of order t. Then $k \geq l$. First, we claim:

Claim. For any $0 \leq i \leq l$, $V(G_i) \subseteq V(G'_i)$.

Suppose the claim is not true. Then, there exists an index i such that $0 < i \leq l$, and a vertex $v \in V(G)$ such that $v \in G_i$ but $v \notin G'_i$. In this case, the chain $G'_0 \bigcup G_0 \geq \cdots \geq G'_i \bigcup G_i \geq G'_{i+1} \geq \cdots \geq G'_k$ is also a D-chain of G of oder t by definition. This however contradicts the maximality of C, whence the claim.

Now, if the rank of v w.r.t. the chain L is $L(v) = i$, then it must be that $v \in V(G_i)$ and $v \notin V(G_{i+1})$. The above claim then implies that $v \in V(G_i) \subset V(G'_i)$. But then the rank of v w.r.t the chain C must be at least i, that is $C_t(v) \geq i = L(v)$ and the proposition follows. ∎

Applying Proposition 2.3 and Lemma 2.1, we obtain

Corollary 2.4. *Let G be a graph. Suppose $t < t' \leq 0$. Then, $C_{t'}(G) \leq C_t(G)$, and for any vertex v of G, $C_{t'}(v) \leq C_t(v)$.*

Proof. Lemma 2.1 guarantees that the maximal D-chain of order t' is a D-chain of order t and then the rank relations follow from Proposition 2.3. ∎

Maximal D-chains are related as follows, where $G \to H$ denotes H being a subgraph of G:

$$\begin{array}{ccccccc}
 & \vdots & & \vdots & & \vdots & \\
 & \downarrow & & \downarrow & & \downarrow & \\
\cdots & \longrightarrow G_{i-1}^{t-1} & \longrightarrow & G_i^{t-1} & \longrightarrow & G_{i+1}^{t-1} & \longrightarrow \cdots \\
 & \downarrow & & \downarrow & & \downarrow & \\
\cdots & \longrightarrow G_{i-1}^{t} & \longrightarrow & G_i^{t} & \longrightarrow & G_{i+1}^{t} & \longrightarrow \cdots \\
 & \downarrow & & \downarrow & & \downarrow & \\
\cdots & \longrightarrow G_{i-1}^{t+1} & \longrightarrow & G_i^{t+1} & \longrightarrow & G_{i+1}^{t+1} & \longrightarrow \cdots \\
 & \downarrow & & \downarrow & & \downarrow & \\
 & \vdots & & \vdots & & \vdots &
\end{array}$$

In Figure 1 we display the maximal D-chains for a specific network, as well as their embeddings into the latter. We also display the D-spectra of all nodes of the network. The induced rank of a vertex is represented by its color.

2.2. Computing the D-spectra via a deletion algorithm. For fixed $k > 0$ and $t < 0$, suppose $i - mt = k$ for some $m \geq 0$ and $1 \leq i \leq -t$. Given a graph G, we will show that the following algorithm will produce the maximal D-chain of order t of G.

Consider the maximal D-chain of order t of G, $G_0 \geq G_1 \geq \cdots \geq G_k \geq \cdots$: a vertex is contained in G_k iff at least k of its neighbors are contained in G_{k+t}, i.e. referencing the vertex degree within G_{k+t}, a predecessor in the chain. This referencing propagates down to G_i, after which we reference vertex degrees within G itself. Reversing this backtracking, the following vertex-deletion algorithm constructs the sequence $(G_i, G_{i-t}, \ldots, G_{i-mt} = G_k)$ starting from G as follows: it first constructs G_i, by deleting any vertices having G-degree less than i and

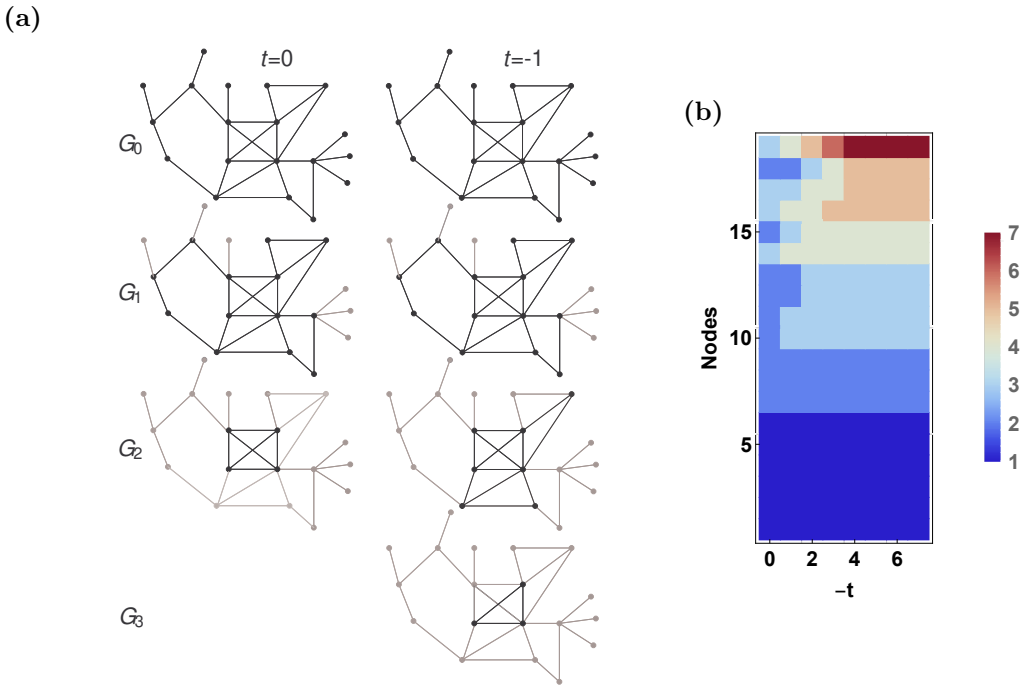


Figure 1: (a) the maximal D-chains of order $t = 0$ and $t = -1$ are highlighted in black. (b) the D-spectra of all vertices, where the induced ranks are indicated by colors.

Algorithm 2.1 A node deletion algorithm

```
H ← G
j ← 0
while j ≤ m do
    delete all nodes with degree smaller than i − jt in H
    H ← the resulting graph
    j ← j + 1
end while
return H
```

then constructs $G_{i-t} \leq G_i$ by deleting all vertices of G_i-degree less than $i - t$ (note $t < 0$). This continues inductively until it arrives at $G_k \leq G_{k+t}$, $k = m(-t) + i$. The formal proof is given below.

Theorem 2.5. *Let $G_0 \geq G_1 \geq \cdots \geq G_k \geq \cdots$ be the maximal D-chain of order t of the graph G. Then the graph H produced by Algorithm 2.1 equals G_k.*

Proof. For any $l > 0$, let H_l be the graph produced by Algorithm 2.1, by setting $k = l$. Firstly, for any $1 \leq i \leq -t$, the chain L_i: $H_0 = G \geq H_i \geq H_{i-t} \geq H_{i-2t} \geq \cdots$ satisfies by the construction of Algorithm 2.1 that any vertex v contained in H_{i-jt} has at least $i - jt$ neighbors in $H_{i-(j-1)t}$. Secondly, we claim that L_i is maximal: there does not exist a chain

L'_i: $H'_0 = G \geq H'_i \geq H'_{i-t} \geq H'_{i-2t} \geq \cdots$ satisfying the same degree properties such that $H'_{i-jt} > H_{i-jt}$ for some $j \geq 0$. To see this, suppose this were the case and let j' be the minimal such that $H'_{i-j't} > H_{i-j't}$. Then, there is a vertex u contained in $H'_{i-j't}$ but not $H_{i-j't}$. However, this is impossible, since, by assumption, we have $H'_{i-(j'-1)t} = H_{i-(j'-1)t}$ and the fact that u is deleted from $H_{i-(j'-1)t}$ is tantamount to u having degree less than $i - j't$ in $H_{i-(j'-1)t}$.

Finally, we claim that interlacing these $-t$ chains, that is, arriving at, $H_0, H_1, H_2, \ldots$, gives us the maximal D-chain of order t of G. It only remains to check the nesting property of adjacent graphs in the combined chain. Suppose this is not true. Then there exists a minimum r such that there is a vertex u such that $u \in H_{r+1}$ but $u \notin H_r$. Assume $r = i - jt$ whence $r+1 = (i+1) - jt$. Note that both chains L_i and L_{i+1} satisfy the degree property and $H_{i-j't} \geq H_{i+1-j't}$ for $0 \leq j' < j$. Thus, the resulting chain obtained by replacing H_{i-jt} with $H_{i-jt} \bigcup \{u\}$ in L_i still satisfies the degree property, contradicting the maximality just proved. This concludes the proof. ∎

Remark 2.6. Recall the node deletion algorithm for the conventional k-core decomposition: start with G, delete all vertices with degree smaller than k and the corresponding incident edges, this may cause new vertices of degree smaller than k to appear, delete them as well, iterate the process until every node in the remaining subgraph has degree at least k. This remaining subgraph is the k-core of G. Compared to our node deletion algorithm for D-chains, we observe a significant difference: in the deletion algorithm for the conventional k-cores, we do not know how many iterations are needed until the process stops. However, for the latter, there are only m iterations (depending on k).

3. Computing the D-spectra via $[t]$-systems.

In this section we present a different approach for computing the D-spectra, namely, as fixed points of certain discrete dynamical systems. To this end, let us briefly recapitulate some basic facts about such systems. A *discrete dynamical system* (or graph dynamical system) over a network involves the following ingredients [30, 29, 31, 32, 33, 34]: a network, a *local function* associated with each node of the network that specifies how the state of the node evolves and an *update schedule* that reflects when each individual node updates its state. Von Neumann's cellular automata (CA) are such systems. Given a network and local functions, the system dynamics is concerned with how the system state varies in time. Various classes of dynamical systems have been studied, e.g. linear, sequential systems [35, 36], monotone systems [37, 38], and threshold systems [39].

Let $G = (V, E)$ be a network with vertex set $V = \{1, 2, \ldots, n\}$ and edges in the set E. Suppose each node, i, has states contained in the finite set P. We associate a function f_i, that specifies how the vertex i updates its state, x_i. The update entails considering the states of the neighbors of i and i itself as arguments of f_i, whence we call f_i the local function at i. An infinite sequence $W = W_1 W_2 \cdots$, where $W_i \subseteq V$, is called a *fair update schedule*, if for any $k \geq 1$, and any $1 \leq i \leq n$, there exists $l > k$ such that $i \in W_l$. The system dynamics is being generated if nodes update their states using their respective local functions, following the order specified by a fair update schedule W. That is, suppose the initial system state at time $t = 0$ is $x^{(0)}$. For $j > 0$, the system state $x^{(j)}$ at time $t = j$ is obtained by the nodes contained in W_j updating their states by means of their local functions taking as arguments the states of their respective neighbors in $x^{(j-1)}$. The states of the nodes not in W_j remain

unchanged. Namely, the system state $x^{(i)} = (x_1^{(i)}, ..., x_n^{(i)})$ at time step i is obtained by the following:

$$x_j^{(i)} = \begin{cases} x_j^{(i-1)}, & v_j \notin W_i; \\ f_{v_j}(x^{(i-1)}), & v_j \in W_i. \end{cases}$$

We denote this dynamical system by $[G, f, W]$, and we write $[G, f, W]^{(j)}(x)$ for the system state at time $t = j$ assuming the system has initial state x at time $t = 0$. For a given dynamical system $[G, f, W]$, a system state x is said reaching a *fixed point* (or stable state) z if there exists $k \geq 0$ such that for any $j > k$, we have $[G, f, W]^{(j)}(x) = [G, f, W]^{(k)}(x) = z$. This also implies that z will not change by further updating any nodes in any order, and thus for any two fair update schedules W and W', we have $[G, f, W]^{(i)}(z) = [G, f, W']^{(i)}(z) = z$ for all $i \geq 0$. Namely, the set of fixed points is invariant when the update schedule varies.

3.1. MC systems. MC systems are a particular class of monotone dynamical systems. Suppose there is a linear order '$\leq$' on the set P. Let $P^q = \{(x_1, x_2, \ldots, x_q) \mid x_j \in P, 1 \leq j \leq q\}$. We extend the linear order on P to a partial order on P^q as follows: $(x_1, x_2, \ldots, x_q) \leq (y_1, y_2, \ldots, y_q)$ iff for all $1 \leq j \leq q$, $x_j \leq y_j$ in P. A function $g\colon P^q \to P$ is called *monotone* if for any $x \leq y$ in P^q, $g(x) \leq g(y)$ in P. A local function $f_i\colon (x_i, x_{k_1}, x_{k_2}, \ldots, x_{k_i}) \mapsto x_i'$ is called *contractive* if for any argument $(x_i, x_{k_1}, x_{k_2}, \ldots, x_{k_i}) \in P^{k_i+1}$, $x_i' \leq x_i$. For example, the Boolean functions 'AND' and 'OR' on $P^q = \{0, 1\}^q$ are monotone, under both assumptions that $0 < 1$ and that $1 < 0$. It is also easy to check that for f_{v_i} being the Boolean function 'AND', it is contractive for $0 < 1$ and for f_{v_i} being the Boolean function 'OR', it is contractive for $1 < 0$. A dynamical system in which local functions are monotone and contractive is called a *monotone-contractive* (MC) system.

Proposition 3.1. *Let $[G, f, W]$ be an MC system. Then, any system state $x \in P^n$ is reaching a fixed point.*

Proof. Since each local function f_{v_j} is contractive, every time the vertex v_j updates, its state will decrease (not necessarily strictly) w.r.t. the linear order on P, and this happens regardless of the states of other vertices. Thus $[G, f, W]^{(i+1)}(x) \leq [G, f, W]^{(i)}(x)$ by definition of the induced partial order on P^n. Hence, for any $i > 0$, if $[G, f, W]^{(i+1)}(x) \neq [G, f, W]^{(i)}(x)$, we have $[G, f, W]^{(i+1)}(x) < [G, f, W]^{(i)}(x)$. Denote by $S_x = \{[G, f, W]^{(i)}(x)\}_{i \geq 0}$, the sequence of iterates of the state x. Since P^n is finite and $S_x \subset P^n$, there can not exist an infinite, strictly decreasing subsequence of S_x. Therefore, $x \in P^n$ has to reach a fixed point $z \in P^n$. ■

Proposition 3.2. *Let $[G, f, W]$ be a monotone system (not necessarily contractive), and suppose a system state $x \in P^n$ is reaching a fixed point $z \in P^n$ with $z < x$. Then, any system state $y \in P^n$ such that $y \leq x$ but $y \not\leq z$ can not be a fixed point. (By $y \not\leq z$ we mean either $y \geq z$ or y is not comparable to z under the P^n partial order.)*

Proof. Since $y \not\leq z$, there exists at least one coordinate in z strictly smaller than the corresponding coordinate in $y = (y_1, y_2, \ldots, y_n)$. Since $[G, f, W]^{(0)}(x) = x \geq y$, the set $I = \{i : [G, f, W]^{(i)}(x) \geq y\} \neq \varnothing$.

Since x is reaching the fixed point z, and since there is at least one coordinate in z strictly smaller than the corresponding coordinate in y, the set I is finite. Let $k = \max\{i : i \in I\}$. Denote by $y' = (y_1', ..., y_n') = [G, f, W]^{(k)}(x)$ and by $y'' = (y_1'', ..., y_n'') = [G, f, W]^{(k+1)}(x)$.

Now, since k is the maximum element in I, there is at least one coordinate in y'' strictly

smaller than the corresponding coordinate in y. Without loss of generality, we can assume the coordinate holding the state of v_1 to be such a coordinate. Namely $y_1'' < y_1$, and so $v_1 \in W_{k+1}$. Then, by assumption, we have $y \leq y'$ and $f_{v_1}(y') = y_1'' < y_1$.

If y were a fixed point, then $f_{v_1}(y) = y_1$ holds. However, as f_{v_1} is monotone we also have $y \leq y' \Rightarrow f_{v_1}(y) = y_1 \leq y_1'' = f_{v_1}(y')$, a contradiction, as we previously established $y_1'' < y_1$. Therefore y cannot be a fixed point. ∎

Now we are ready to present a key property of MC systems:

Theorem 3.3. *For any two fair update schedules W and W', a system state $x \in P^n$ is reaching the same fixed point $z \in P^n$ in the MC systems $[G, f, W]$ and $[G, f, W']$. In addition, any state y such that $z \leq y \leq x$ is reaching the fixed point z.*

Proof. Based on Proposition 3.1, x is reaching a fixed point under any fair update schedule. Suppose x is reaching the fixed point $z \leq x$ under $[G, f, W]$ while reaching $z' \leq x$ under $[G, f, W']$. There are the following two cases: if $z' \not\leq z$, according to Proposition 3.2, z' can not be a fixed point; if $z' < z$, according to Proposition 3.2, z can not be a fixed point. Therefore, we must have $z = z'$, whence the first part of the theorem. ∎

Theorem 3.3 has several implications:

- Since we know that each state x will reach a unique stable state regardless of the update schedule, we can choose any update schedule to compute the stable state (some of them may be easier to implement). Accordingly, we shall not explicitly reference update schedules for the following analysis.
- Suppose we are given a state x and wish to compute its fixed point, z. Suppose further we can identify a state, y, that satisfies $z \leq y$. Then computing z via y may accelerate the computation.

We also believe that MC systems deserve future investigations due to their own independent interest.

3.2. $[t]$-systems. Particular MC systems, called $[t]$-systems, are employed in order to compute the D-spectra. Given a network G with n vertices $v_1, v_2, \ldots, v_n$, we will assume that each vertex has a state from the set $[n] = \{1, 2, \ldots, n\}$. Suppose the local function f_v associated to the vertex v returns the maximum integer k such that at least k of the neighbors of v in G have states at least $k + t$. Together with a specified fair update schedule W, we call the system a $[t]$-*system* and denote it by (G, f, W).

First, the f_v's are, by constriuction, monotone. However, they are in general not contractive. For example, $x_v = 1 \Rightarrow f_v(1, 2, 2) = 2 > 1$ for $t = 0$, whence f_v is not contractive.

However, Lemma 3.4 will provide a subset of system states such that the local functions f_v, when restricted to it, are both monotone and contractive.

Lemma 3.4. *Let t be fixed. Let $x = (\deg(v_1), \deg(v_2), \ldots, \deg(v_n))$, and let*

$$\mathcal{Q} = \bigcup_W \{z : z = [G, f, W]^{(i)}(x) \text{ for some } i \geq 0\},$$

where the union ranges over the $[t]$-systems $[G, f, W]$ for all possible fair update schedules W. Then, for any vertex v in G, the returned state for v after applying the local function f_v at v on any system state $x' \in \mathcal{Q}$ is smaller than its original state x'_v in x', i.e. f_v is contractive w.r.t. $x' \in \mathcal{Q}$.

Proof. For a fixed update schedule W, we shall prove by induction on $i \geq 0$ such that f_v is contractive w.r.t. $[G, f, W]^{(i)}(x)$ for all i and all v. We first check the case $i = 0$, that is, f_v is contractive w.r.t. x. This is clear as by definition of f_v the returned value k is smaller or equal to $\deg(v)$ for any v. Next, we suppose for any v, f_v is contractive w.r.t. $[G, f, W]^{(i)}(x)$ for $i \geq 0$. Then, this implies that $[G, f, W]^{(i+1)}(x) \leq [G, f, W]^{(i)}(x)$.

Next, for a vertex v, there are the following two cases:

(i) Suppose $v \notin W_{i+1}$. Then

$$v \notin W_{i+1} \Longrightarrow [G, f, W]^{(i)}(x)[v] = [G, f, W]^{(i+1)}(x)[v],$$

where $[G, f, W]^{(i)}(x)[v]$ stands for the state of the vertex v in $[G, f, W]^{(i)}(x)$. Now, by inductive assumption, we have contractivness of f_v in the i'th case, i.e.,

$$f_v([G, f, W]^{(i)}(x)) \leq [G, f, W]^{(i)}(x)[v].$$

And, by monotonicity of f_v, since $[G, f, W]^{(i+1)}(x) \leq [G, f, W]^{(i)}(x)$, we have

$$f_v([G, f, W]^{(i+1)}(x)) \leq f_v([G, f, W]^{(i)}(x)).$$

Combining the three relationships above, we arrive at

$$f_v([G, f, W]^{(i+1)}(x)) \leq [G, f, W]^{(i+1)}(x)[v].$$

Namely that f_v is contractive in the case $i + 1$.

(ii) Suppose $v \in W_{i+1}$. Then, by inductive assumption for the case i, it follows that

$$f_v([G, f, W]^{(i)}(x)) \leq [G, f, W]^{(i+1)}(x)[v]$$

Again, by monotonicity of f_v, since $[G, f, W]^{(i+1)}(x) \leq [G, f, W]^{(i)}(x)$, we have

$$f_v([G, f, W]^{(i+1)}(x)) \leq f_v([G, f, W]^{(i)}(x)) \leq [G, f, W]^{(i+1)}(x)[v],$$

meaning that f_v is again, contractive in the case $i + 1$. Therefore, we can conclude that f_v is contractive w.r.t. $[G, f, W]^{(i+1)}(x)$, and the lemma follows. ∎

Now, when restricted to the subset of system states $\mathcal{Q}$, the $[t]$-system (G, f, W) is an MC system for any t. Thus, starting with any system state in $\mathcal{Q}$, the system will converge to a stable state, which does not depend on the update schedule. In particular, we have the following proposition.

Proposition 3.5. *For the $[t]$-system on G, the state $\big(\deg(v_1), \deg(v_2), \ldots, \deg(v_n)\big)$ converges to a stable state $C^t = (C_{v_1,t}, C_{v_2,t}, \ldots, C_{v_n,t})$. In addition, if $t < t'$, $C^t \geq C^{t'}$.*

Proof. Since on the set Q, any $[t]$-system is an MC system, $x \in Q$ must reach a fixed point regardless of the update schedule, whence the first part of the proposition. For $t < t'$, first it is clear by definition that the returned value (for any v) by applying f_v with t is no less than that of t'. By induction using the monotonicity of the local functions, we can conclude $C_{v,t} \geq C_{v,t'}$, whence $C^t \geq C^{t'}$. ∎

3.3. The D-spectra via fixed points. Now we are ready to present our second approach of computing the D-spectrum of a network.

Theorem 3.6. *For the $[t]$-system on G, the state $x = \big(\deg(1), \deg(2), \ldots, \deg(n)\big)$ is reaching the stable state $C^t = (C_{1,t}, C_{2,t}, \ldots, C_{n,t})$, where*

i. $C_{i,t} = 0$ for any $1 \leq i \leq n$, if $t > 0$;

ii. $C_{i,t} = C_t(i)$ for any $1 \leq i \leq n$, if $t \leq 0$. In particular, if $t \leq -\Delta(G)$, $C_{i,t} = \deg(i)$ for any $1 \leq i \leq n$.

In addition, the state C^t is reaching the stable state C^{t+1} in the $[t+1]$-system on G.

Proof. Suppose first, $t > 0$ and suppose there exists some v such that $C_{v,t} > 0$. By definition, there is at least one neighbor u of v such that $C_{u,t} \geq C_{v,t} + t > 0$ holds. Iterating this argument, the node u has a neighbor with C_t-value at least $C_{u,t} + t$, effectively implying the existence of vertices with unbounded degrees, which is, given the fact that the nework is finite, impossible.

Secondly, let $t \leq 0$. We shall first prove that the state $z = \big(C_t(v_1), \ldots, C_t(v_n)\big)$ is a fixed point of the $[t]$-system, i.e. applying the local function f_v (with parameter t) to z will return $C_t(v)$ for any vertex v. Let $L\colon G_0 \geq G_1 \geq \cdots$ be the maximal D-chain of order t of G. Then, for any vertex v, by definition of $C_t(v) = i$, v belongs to G_i but not to G_{i+1}. This implies the following:

(i) there are at least i neighbors of v are contained in G_j ($j = \max\{0, i+t\}$). Note that for any u among these, we have $C_t(u) \geq i + t$ by definition. Thus, among the neighbors of v, there are at least i having values at least $i + t$ in z. This implies $f_v(z) \geq i = C_t(v)$;

(ii) there cannot be at least $i+1$ neighbors of v, that are contained in $G_{j'}$ ($j = \max\{0, i + 1 + t\}$), as otherwise, the chain $G_0 \geq \cdots \geq G_i \geq G_{i+1} \bigcup \{v\} \geq G_{i+2} \geq \cdots$ gives a D-chain of order t, which contradicts the maximality of L. Hence, among the neighbors of v, there can not be that at least $i + 1$ of them with values at least $i + 1 + t$ in z, whence $f_v(z) < i + 1$ holds.

(i) and (ii) establish $f_v(z) = i = C_t(v)$ and z is a fixed point.

We proceed with the proof by observing that, in case of $z = x$, we are done. By construction, we otherwise have $z < x$. Let y be the fixed point reached by x. In case of $y < z$ or y being incomparable to z, Proposition 3.2 guarantees that z cannot be a fixed point, which is a contradiction. Otherwise we have $z < y \leq x$. In this cases there exists a coordinate, which we shall index by v, that satisfies $y_v > C_t(v)$. Consider the sequence of subgraphs induced by the sequence of sets of vertices $S^0, S^1, \ldots$ which are inductively defined as follows: (i) $S^0 = \{v\}$; (ii) for $r > 0$, $S^r = \{u \mid y_u \geq y_w + t, w \in S^{r-1}, \text{and } u = w \text{ or } u \text{ is a neighbor of } w\}$. Clearly, by construction we have $S^{r-1} \subseteq S^r$, and by induction $y_u \geq y_v + rt$ for $u \in S^r$. If y is a fixed point, then we have: (a) there are at least y_v neighbors of v with values at least $y_v + t$ in y. These neighbors must be contained in S^1; (b) for $r \geq 0$, any $w \in S^r$, there are at least $y_w + t$ neighbors contained in S^{r+1}.

By abuse of notation, we will denote the subgraph induced by the set S^r-vertices as S^r. From (a), (b) and the fact that $y_u \geq y_v + rt$ for $u \in S^r$, we can conclude that for $r \geq 0$, any

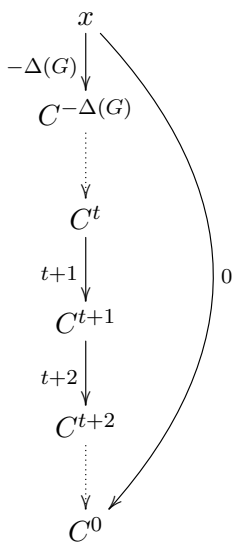
vertex in S^r has at least $y_v + rt$ neighbors in S^{r+1}. Then, the chain

$$\cdots \geq G_{y_v+2t} \bigcup S^2 \geq G_{y_v+2t+1} \bigcup S^1 \geq \cdots \geq G_{y_v+t} \bigcup S^1 \\ \geq G_{y_v+t+1} \bigcup S^0 \geq \cdots \geq G_{y_v-1} \bigcup S^0 \geq G_{y_v} \bigcup S^0 \geq G_{y_v+1} \geq \cdots$$

is a D-chain of order t of G, implying $C_t(v) \geq y_v$, which is a contradiction. Thus x cannot reach a fixed point y such that $z < y \leq x$, whence x is reaching the fixed point z as claimed.

In particular, if $t \leq -\Delta(G)$, we have $C_t(v) = \deg(v)$, whence x itself is a fixed point. Finally, Theorem 3.3 in turn implies that since $C^{t+1} \leq C^t \leq x$, C^t also converges to C^{t+1} in the $[t+1]$-system. ■

Remark. The fact that the state C^t converges to the stable state C^{t+1} for the $[t+1]$-system on G, as claimed in Theorem 3.6, guarantees essentially the same complexity for computing the D-spectra of all nodes as computing core numbers alone. That is, it is not necessary to initialize with the degree sequence.



4. Application to predicting similarity. In this section, we present some applications of our framework. We shall be interested in analyzing the connection between the D-spectra and the spreading power of nodes in the process such as disease outbreak or information spreading. Specifically, we will be using the SIR model to get the data on infection rates characterizing the spreading power of nodes. To begin, let us briefly review the SIR model and our simulation setup. The SIR process is a stochastic model for studying the spread of disease within a population. It works as follows: a population is modeled as a network, where each node represents an individual, while links (edges) between nodes represent their interaction relation. Each node can be in either of three states: susceptible (S), infected (I), and recovered (R). During the process, at each step, each infected node may infect each of its susceptible neighbors with a certain probability. At the subsequent step the infected node may become recovered with another probability. Once a node is in the state R, it will never infect other nodes and never become infected again. The process stops when there are no nodes in the state I. In a sense, an SIR process can be viewed as a stochastic discrete dynamical system, whereas the previously discussed discrete dynamical systems (e.g. MC systems) are deterministic (the local functions being deterministic).

Our SIR simulations are designed as follows: for each respective network, we initialize the process with exactly one node, the infected source, in the state I. We shall assume that the probability of an infected node becoming recovered in the next time step equals 1 and we assume one fixed transmission probability for all nodes throughout the simulation. For each infected source, we run 1000 simulations and for each of these we compute the ratio between the number of recovered nodes and the total number of nodes in the network. We refer to the average of these ratios as the infection rate of the node.

We execute this for each node in the network, for the nine transmission probabilities $h \cdot \beta$, where $h \in \{0.1, 0.5, 1, 1.5, 2, 4, 6, 8, 10\}$ and β being the epidemic threshold value of the network. The epidemic threshold value can be computed as $\beta = \frac{<k>}{<k^2>-<k>}$ [11, 12], where $< k >$ denotes the average node degree of the network and $< k^2 >$ is the average of the squares of the degrees. Accordingly, we obtain for each node nine distinct infection rates.

Kitsak et al. [23] shows that with respect to infection rates, nodes that are contained in the same core are generally more isotropic, than nodes having the same degree. In other words, in order to identify nodes of similar spreading power, core numbers are more suited than vertex degrees. In the following, we compare D-spectrum and core number.

The D-spectrum of a node is a vector whose coordinates are the ranks of the node in the respective D-chains of different orders. As a result, the Euclidian distance between D-spectra is a natural criterion for categorizing (possibly) similar nodes. In order to compare such a categorization to the one derived by restricting to core numbers, i.e. nodes having the same core number being considered similar, we study five networks proceeding as follows. We partition all nodes by means of the Euclidean distance between their D-spectra into an *a priori* specified number of clusters called D-blocks. (The clusters are derived by calling the standard function Findcluster in Mathematica 10.0.) We furthermore group the nodes according to their core numbers into clusters to which we refer to as C-blocks.

By construction, a D-block may contain nodes from multiple C-blocks and vice versa. The intersection of a C- and D-block is called an I-cell. In Figure 2, Figure 4 and Figure 6, the data for a row is from the nodes in the same C-block (and the row index is the actual core number of the nodes associated to the row), while the data for a column is from the nodes in the same D-block (and the column index has no particular meaning other than a label). Specifically, we compute the average infection rate of the nodes contained in the same I-cell and project the infection rate into a color of the corresponding cell in the 2D heatmaps of Figure 2, Figure 4 and Figure 6. Note that two clusters may have empty intersections, which is indicated in black. Clearly, if nodes with the same core number have similar spreading power, these cells on the same row necessarily have very similar (or isotropic) colors; if the colors on the same row vary across the spectrum, then nodes with that same core number do not have similar spreading power. The same logic holds for columns. From Figure 2, Figure 4 and Figure 6, we can observe that, colors on the same column are generally (statistically) more isotropic than those of the same row (although not strikingly pronounced). This means that nodes from the same D-block are more likely to have similar spreading power than nodes from the same C-block.

In Figure 3, Figure 5 and Figure 7, we compare the two clusterings from another point of view. For each C-block (resp. I-cell), we compute the dispersion (variance-to-mean ratio) of

infection rates of nodes within said block (resp. I-cell). Clearly, dispersion reflects similarity as well. The smaller the dispersion is, the more similar the underlying set of nodes are. We compute the average dispersion over I-cells from the same C-block too. Then, for each C-block, we obtain a number of points on the same vertical line in Figure 3, Figure 5 and Figure 7, where the x-coordinates of the points are the dispersion from the same C-block, the y-coordinate of the black point is the average dispersion over the I-cells from the same C-block, and the y-coordinate of a grey point corresponds to the dispersion from a certain I-cell. The red line indicates $x = y$. From Figure 3, Figure 5 and Figure 7, we can see that most of the points are below the red line, meaning the y-coordinates are generally smaller than the x-coordinates. That implies, for example, if we randomly pick two nodes from the same C-block, and randomly pick an I-cell from the same C-block and pick two nodes from the I-cell, the infection rates of the latter are closer than that of the former, i.e. refining C-blocks based on D-spectrum provides more accuracy for the clustering of similar nodes.

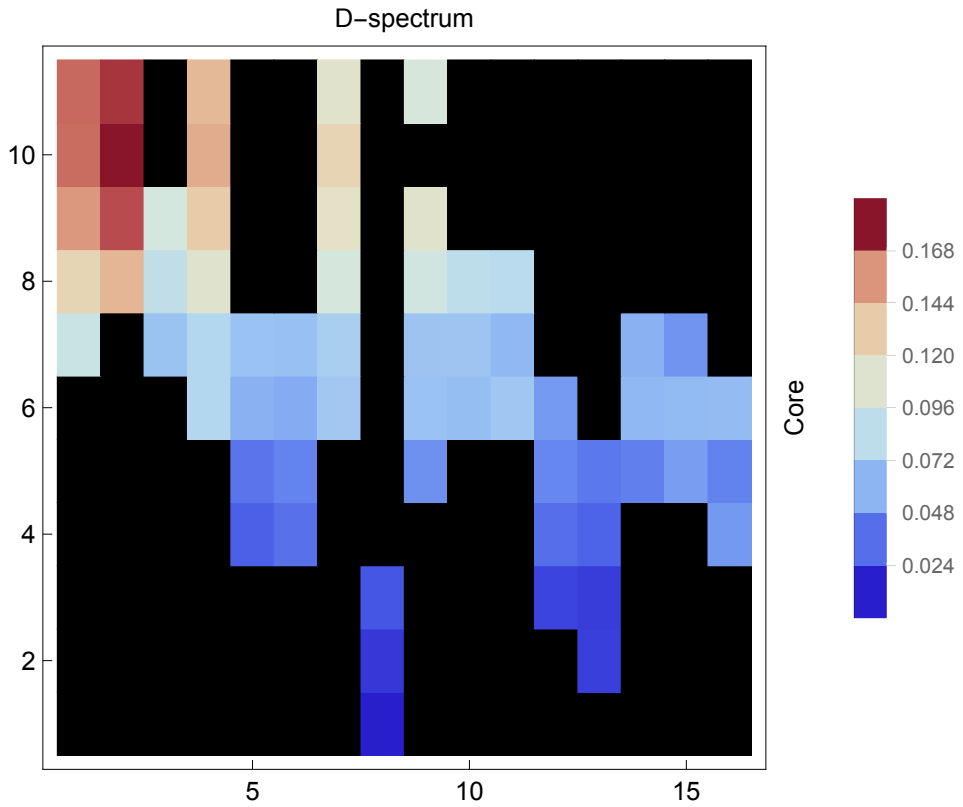


Figure 2: Email: comparason of core- and D-spectrum-clusterings by infection rate.

We note that for clustering based on core numbers, the number of clusters is determined by the underlying network structure itself, i.e. the number of different values of the core numbers of the nodes. For instance, there are no canonical criteria to distribute two nodes with the same core number into different clusters, and there exist very large networks where nodes have a few different core numbers. In difference, D-spectra, as high-dimensional data, provide

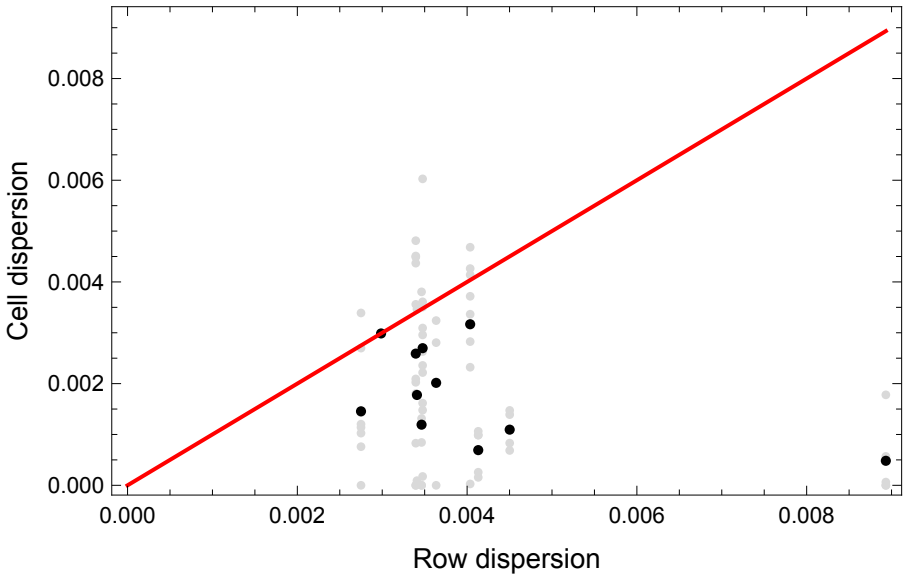


Figure 3: Email: comparason of core- and D-spectrum-clusterings by dispersion.

more freedom: the number of clusters can be determined as necessary based on criteria, as for instance, Euclidean distance. We consider this aspect to be an advantage of D-spectra. In our simulations, we tested three different values for the number of clusters based on D-spectra: 1, 1.5 and 2 times of the number of clusters based on core number.

The five networks studied here, include networks of social interaction, transportation, internet routing and communication and are listed below:

- *Email* network [41]: e-mail interchanges between members of the Univeristy Rovira i Virgili (Tarragona).
- *USAir* network [42]: the US air transportation network.
- *Jazz* network [43]: collaborations between jazz musicians.
- *PB* network [44]: US political blogs whose original links were directed, regarded as undirected edges.
- *Router* network [45]: a symmetrized snapshot of the structure of the Internet at the level of autonomous systems.

All networks have been converted to simple graphs by eliminating multi-edges or loops and considering directed edges as un-directed. Also, only the largest connected component is considered if the original network happens to be disconnected. The main network parameters are summarized in Table 1.

For the analysis of the three networks in Figures 2–7, the transmission probabilities were set to 1.5β, where β is their respective epidemic threshold value. In the Supplementary Materials, we extend the analysis of these networks incorporating the following additional two transmission probabilities: 1β and 2β. The analysis of the additional probabilities shows the robustness of the observation made in Figures 2–7 that the infection rates of nodes having the same core number, are generally more heterogeneous than that of nodes in the same

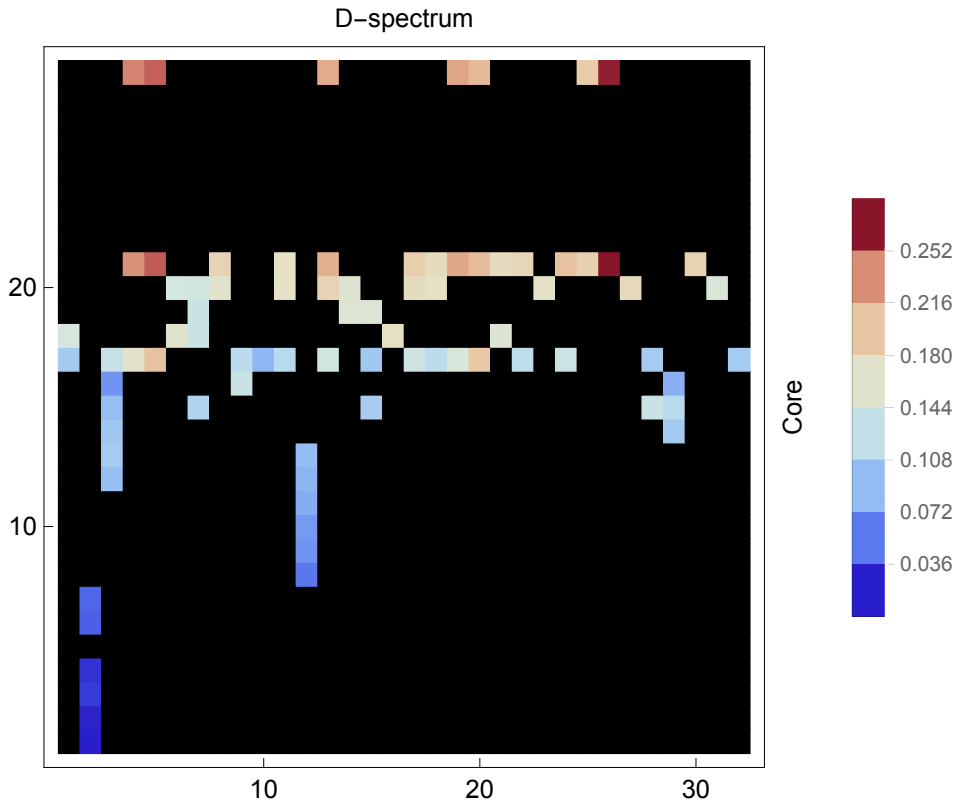


Figure 4: Jazz: comparason of core- and D-spectrum-clusterings by infection rate.

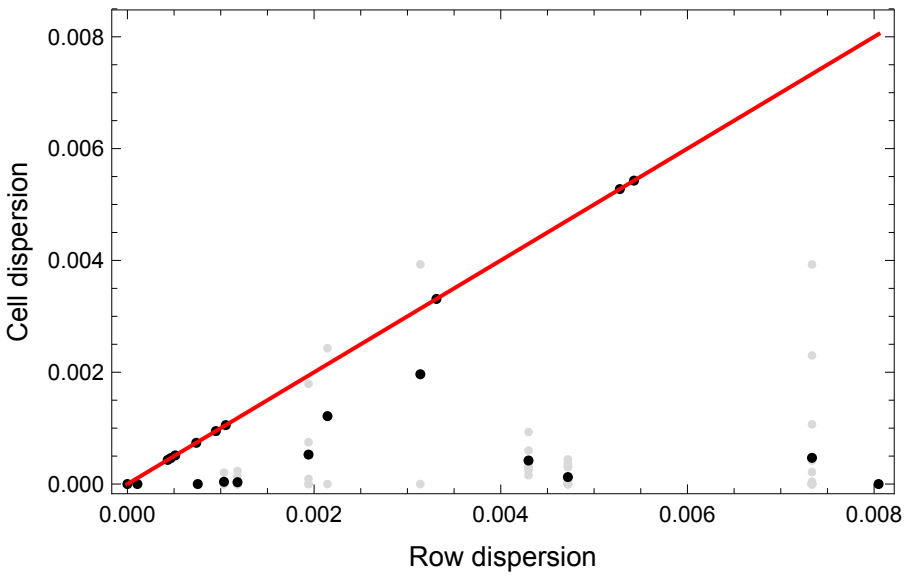


Figure 5: Jazz: comparason of core- and D-spectrum-clusterings by dispersion.

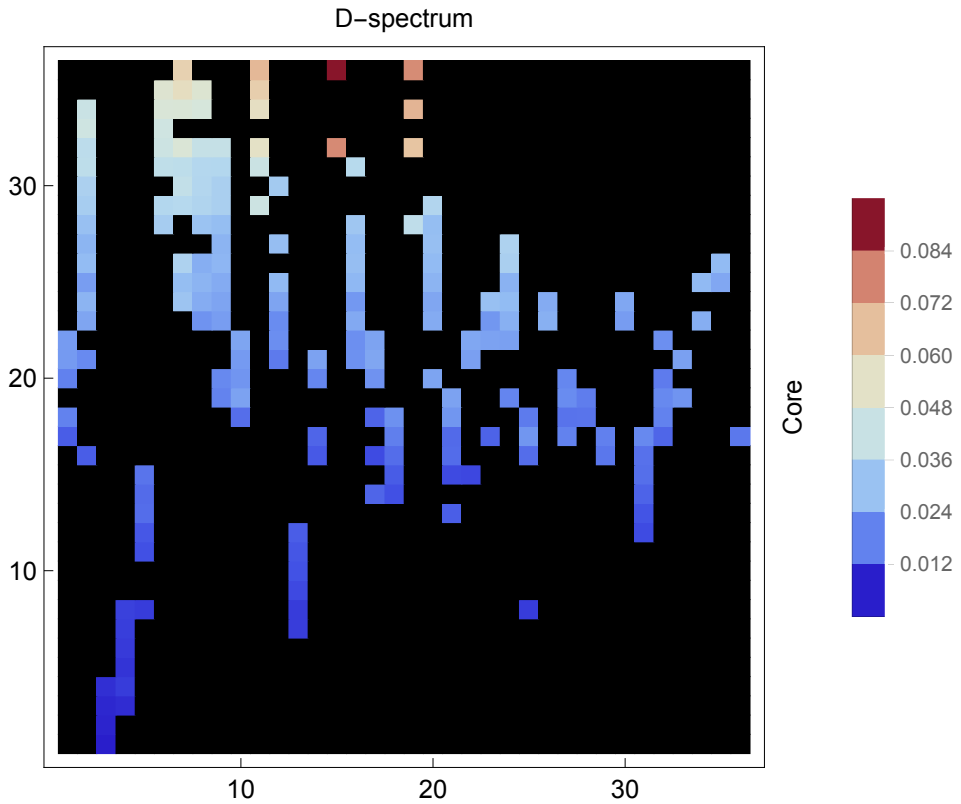


Figure 6: PB: comparason of core- and D-spectrum-clusterings by infection rate.

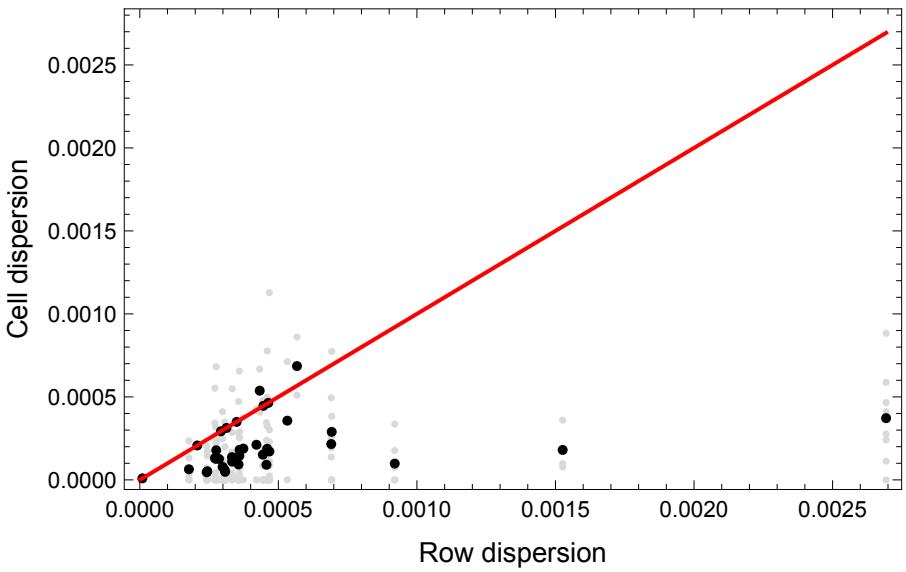


Figure 7: PB: comparason of core- and D-spectrum-clusterings by dispersion.

Network name	N	N_E	k_{max}	$< k >$	β
Email	1,133	5,451	71	9.62	0.056
Jazz	198	742	100	27.69	0.026
PB	1,222	16,714	351	27.35	0.012
Router	5,022	6,258	106	2.49	0.078
USAir	332	2,126	139	12.80	0.023

Table 1: Topological characteristics: here N denotes the number of nodes, N_E denotes the number of edges, k_{max} the maximum degree, $< k >$ the average degree and β the epidemic threshold.

D-spectrum block. See Figures S1–S13 in the Supplementary Materials. Accordingly, node partitions obtained via D-spectra provide a meaningful enhancement over categorizations obtained using conventional core numbers.

We next qualify the correlation between the spreading power of nodes observed in the SIR process and the D-spectra of nodes more directly. That is, we ask to what extent do nodes, categorized via D-spectra, exhibit isotropic spreading power in the SIR process. To this end, we firstly cluster the nodes according to their spreading power (i.e. the sequences of infection rates at the nine transmission probabilities) and secondly we cluster them w.r.t. their D-spectra. Then we inspect the mutual intersection of these clusters from the two approaches and plot the distribution of the sizes of the intersections in Figure 8. In Figure 8, each row represents a cluster based on spreading power while each column represents a cluster based on D-spectra, and the colors of the cells there represent the sizes of the intersections. Clearly, if two clusterings completely correlate with each other, then each row has exactly one non-empty intersection with the columns and vice versa, implying exactly one non-black cell in each row and each column. As a result, the fewer cells the mass of a row (resp. column) concentrates in, the more correlated the two clusterings are.

From Figure 8, we observe that generally speaking, the mass of rows and columns indeed respectively concentrate in a few cells. Accordingly, the clustering based on D-spectra exhibits a good correlation with the clustering based on spreading power, implying D-spectrum is a good candidate measure of detecting nodes of similar spreading power. The observed correlation is also robust w.r.t. different specified number of clusters, see Figures S14–S18 in the Supplementary Materials.

5. Discussion. In this paper, our primary objective is to present the D-spectrum framework for networks analysis. We motivate D-spectra as an enhancement of the framework of graph cores by introducing specific relations between vertices contained in specific "cores". As such D-spectra integrate the local (degree) with the global (core) information and fit well into existing approaches. We then systematically develop the concept of D-spectra from a rigorous mathematical perspective. We furthermore present two approaches for computing them: first a parametric deletion algorithm, reminiscent of the algorithm used for computing k-cores and secondly, computing D-spectra via certain fixed points of $[t]$-systems. Computing D-spectra of nodes has by construction the same time complexity as obtaining core numbers of nodes.

We then apply D-spectra to analyzing the SIR processes: we identify nodes of similar

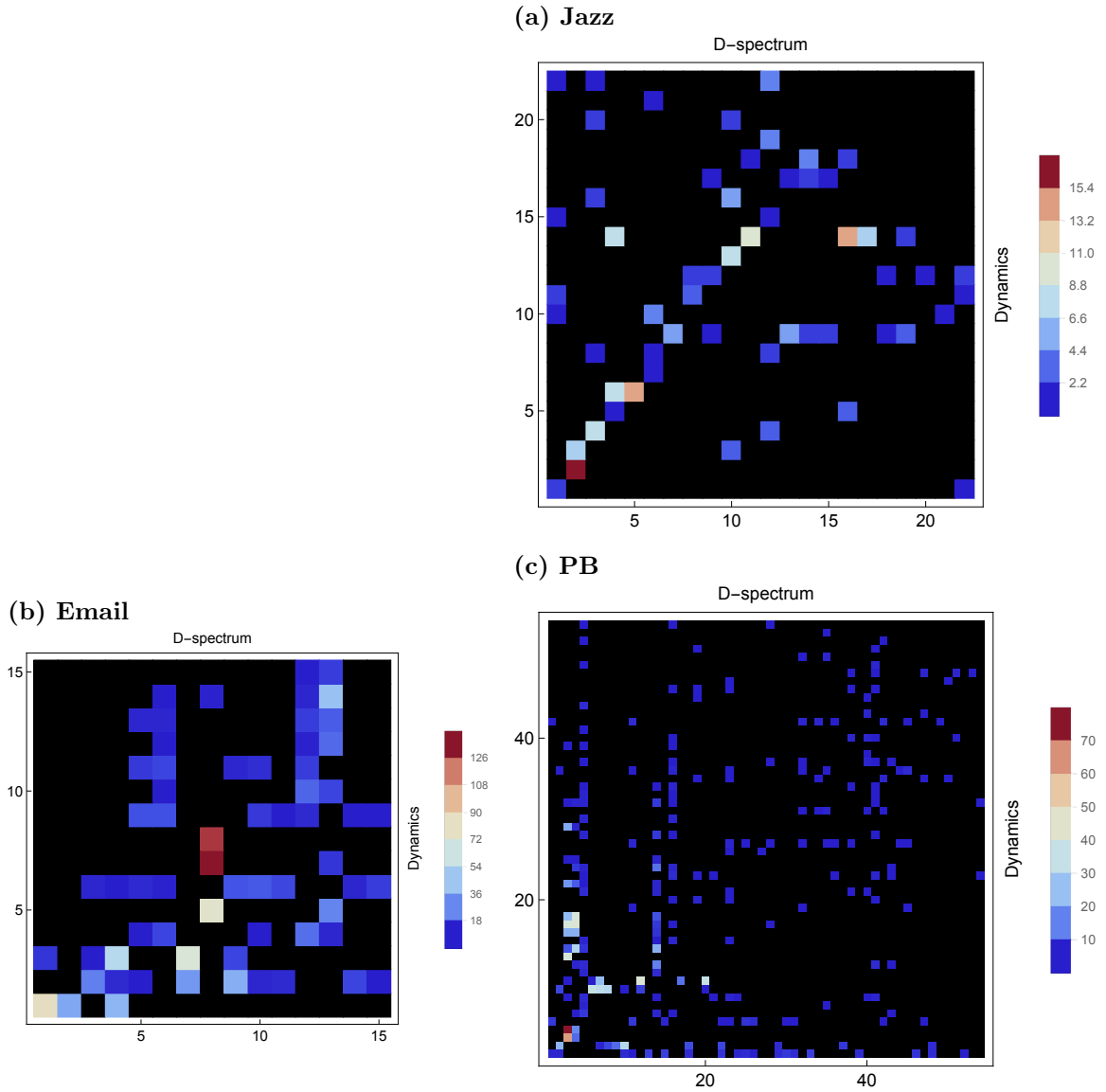


Figure 8: Partitions induced by D-spectra and spreading power are correlated. In (a), (b), (c), the color of any cell represents the size of the intersection between blocks of the partitions induced by spreading power and D-spectra, respectively. For all networks, we observe that there are only a few distinct cells for any row or column that contain almost all vertices. The extent of this concentration reflects how well D-spectra capture the spreading power of vertices.

spreading power based on D-spectra employing an approach, along the lines of the analysis based on core numbers or vertex degrees. We observe that nodes of similar spreading power exhibit similar D-spectra, and the latter similarity manifests using the natural Euclidean distance. The simulation results of a variety of networks imply that D-spectra lead to an

improvements compared to core numbers and vertex degree based approaches.

It is beyond the scope of this paper to provide a more extensive analysis of the applicability of D-spectra. It is intuitive to anticipate that D-spectra are a powerful analysis tool for processes that sensibly depend on the relations between different "cores".

The D-spectrum framework as a theoretical concept is far from being fully explored. For instance, are D-spectra useful in studying the long-standing graph isomorphism problem? That is, are two networks having the same D-spectrum (instead of a single ranking such as degree or core number) isomorphic? While it is easy to construct two non-isomorphic graphs that give the same degree sequence or core numbers, it is not easy to construct two non-isomorphic graphs that have the same D-spectrum. It is also well-known that the maximum core number that nodes of a graph can have provides an upper bound for the chromatic number of the graph. Similar connections between combinatorial observables and the D-spectra are interesting to explore.

In conclusion, we remark that the framework itself is not restricted to graphs, it can easily be extended to hypergraphs, weighted networks, and k-truss decompositions [48, 49].

Acknowledgments. We thank Stephen Eubank and Henning Mortveit for valuable discussions, Linyuan Lü and Qian-Ming Zhang for providing some data related to the used networks. We also thank the anonymous referees for valuable comments and suggestions which improved the presentation of the paper.

Contributions. R. X. F. C. and C. M. R. planned and performed this research. R. X. F. C. also partly implemented the simulation. A. C. B. implemented the simulation and performed the research. All authors discussed the results, wrote the paper and reviewed the manuscript.

Competing interests. The authors declare no competing financial interests.

Corresponding authors. Correspondence to R. X. F. Chen (chen.ricky1982@gmail.com) or C. M. Reidys (duck@santafe.edu).

REFERENCES

[1] A.-L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science 286 (1999), pp. 509–512.
[2] R. Albert, H. Jeong and A.-L. Barabasi, *Error and attack tolerance of complex networks*, Nature 406 (2000), pp. 378–382.
[3] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature 393 (1998), pp. 440–442.
[4] L. C. Freeman, *A set of measures of centrality based on betweenness*, Sociometry 40 (1977), pp. 35–41.
[5] S. N. Dorogovtsev, A. V. Goltsev and J. F. F. Mendes, *K-core organization of complex networks*, Phys. Rev. Lett. 96 (2006), 040601.
[6] S. B. Seidman, *Network structure and minimum degree*, Social Networks 5 (1983), pp. 269–287.
[7] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt and E. Shir, *A model of Internet topology using k-shell decomposition*, Proc. Natl Acad. Sci. USA 104 (2007), pp. 11150–11154.
[8] U. Alon, *Network motifs: theory and experimental approaches*, Nat. Rev. Genet. 8 (2007), pp. 450–461.
[9] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E 69 (2004), 026113.
[10] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Phys. Rev. E 74 (2006), 036104.

[11] C. Castellano and R. Pastor-Satorras, *Thresholds for epidemic spreading in networks*, Phys. Rev. Lett. 105 (2010), 218701.
[12] M. E. J. Newman, *Spread of epidemic disease on networks*, Phys. Rev. E 66 (2002), 016128.
[13] R. Pastor-Satorras and A. Vespignani, *Epidemic spreading in scale-free networks*, Phys. Rev. Lett. 86 (2001), 3200.
[14] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*, Princeton Univ. Press, 2008.
[15] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.
[16] S. Eubank et al, *Modelling disease outbreaks in realistic urban social networks*, Nature 429 (2004), pp. 180–184.
[17] E. M. Rogers, *Diffusion of Innovation* 4th edn, Free Press, 1995.
[18] R. Pastor-Satorras and A. Vespignani, *Immunization of complex networks*, Phys. Rev. E 65 (2002), 036104.
[19] P. Wang, J. Lu and X. Yu, *Identification of important nodes in directed biological networks: a network motif approach*, PLoS ONE 9 (2014), e106132.
[20] F. Morone and H. A. Makse, *Influence maximization in complex networks through optimal percolation*, Nature 524 (2015), pp. 65–68.
[21] R. M. Anderson, R. M. May and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control*, Oxford Science Publications, 1992.
[22] S. Aral and D. Walker, *Identifying Influential and Susceptible Members of Social Networks*, Science 337 (2012), pp. 337–341.
[23] M. Kitsak et al, *Identification of influential spreaders in complex networks*, Nat. Phys. 6 (2010), pp. 888–893.
[24] B. Bollobás, *Graph Theory and Combinatorics: Proceedings of the Cambridge Combinatorial Conference in Honor of P. Erdös* Vol. 35, Academic, 1984.
[25] L. Lü, T. Zhou, Q.-M. Zhang and H.E. Stanley, *The h-index of a network node and its relation to degree and coreness*, Nat. Commun. 7 (2016), 10168.
[26] A. Montresor, F. D. Pellegrini and D. Miorandi, *Distributed k-core decomposition*, IEEE Trans. Parallel Distrib. Syst. 24 (2013), pp. 288–300.
[27] Y. Liu, M. Tang, T. Zhou and Y. Do, *Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition*, Sci. Rep. 5 (2015), 9602.
[28] F. Luo et al., *Modular organization of protein interaction networks*, Bioinformatics 23 (2007), pp. 207–214.
[29] J. Von Neumann, *Theory of Self-Reproducing Automata*, University of Illinois Press, Chicago, 1966.
[30] S. A. Kauffman, *Metabolic stability and epigenesis in randomly constructed genetic nets*, J. Theor. Biol. 22 (1969), pp. 437–467.
[31] S. Wolfram, *Cellular Automata and Complexity*, Addison-Wesley, New York, 1994.
[32] C. L. Barrett and C. M. Reidys, *Elements of a theory of simulation I*, Appl. Math. Comput. 98 (1999), pp. 241–259.
[33] C. L. Barrett, H. S. Mortveit and C. M. Reidys, *Elements of a theory of simulation II: sequential dynamic systems*, Appl. Math. Comput. 107 (2000), pp. 121–136.
[34] H. S. Mortveit and C. M. Reidys, *An Introduction to Sequential Dynamic Systems*, Springer, 2008.
[35] B. Elspas, *The theory of autonomous linear sequential networks*, IRE Transactions on Circuit Theory 6 (1959), pp. 45–60.
[36] R. X. F. Chen and C. M. Reidys, *Linear sequential dynamical systems, incidence algebras, and Möbius functions*, Linear Algebra Appl. 553 (2018), pp. 270–291.
[37] R. X. F. Chen, H. S. Mortveit and C. M. Reidys, *Dependence of update schedules of monotone sequential Boolean networks*, submitted.
[38] H. Daniels and M. Velikova, *Monotone and Partially Monotone Neural Networks*, IEEE Trans. Neural Netw. 21 (2010), pp. 906–917.
[39] E. Goles, *Comportement oscillatoire d'une famille d'automates cellulaires non uniformes*, Thèse IMAG, Grenoble, 1980.
[40] J. E. Hirsch, *An index to quantify an individual's scientific research output*, Proc. Natl Acad. Sci. USA 102 (2005), pp. 16569–16572.
[41] R. Guimerà et al., *Self-similar community structure in a network of human interactions*, Phys. Rev. E

68 (2003), 065103.
[42] V. Batageli and A. Mrvar, *Pajek Datasets*, Available at http://vlado.fmf.uni-lj.si/pub/networks/data/2007.
[43] P. Gleiser and L. Danon, *Community structure in Jazz*, Adv. Complex Syst. 6 (2003), 565.
[44] L. A. Adamic and N. Glance, in: Proceedings of the 3rd International Workshop on Link Discovery, ACM 2004, pp. 36–43.
[45] N. Spring, R. Mahajan, D. Wetherall and T. Anderson, *Measuring ISP topologies with Rocketfuel*, IEEE/ACM Trans. Networking 12 (2004), pp. 2–16.
[46] Email Dataset, Available at http://www-levich.engr.ccny.cuny.edu/webpage/hmakse/software-and-data.
[47] The Internet Movie Database, Available at http://www.imdb.com.
[48] J. Cohen, *Trusses: Cohesive subgraphs for social network analysis*, 2008.
[49] J. Wang and J. Cheng, *Truss decomposition in massive networks*, Proceedings of the VLDB Endowment 5 (2012), pp. 812–823.

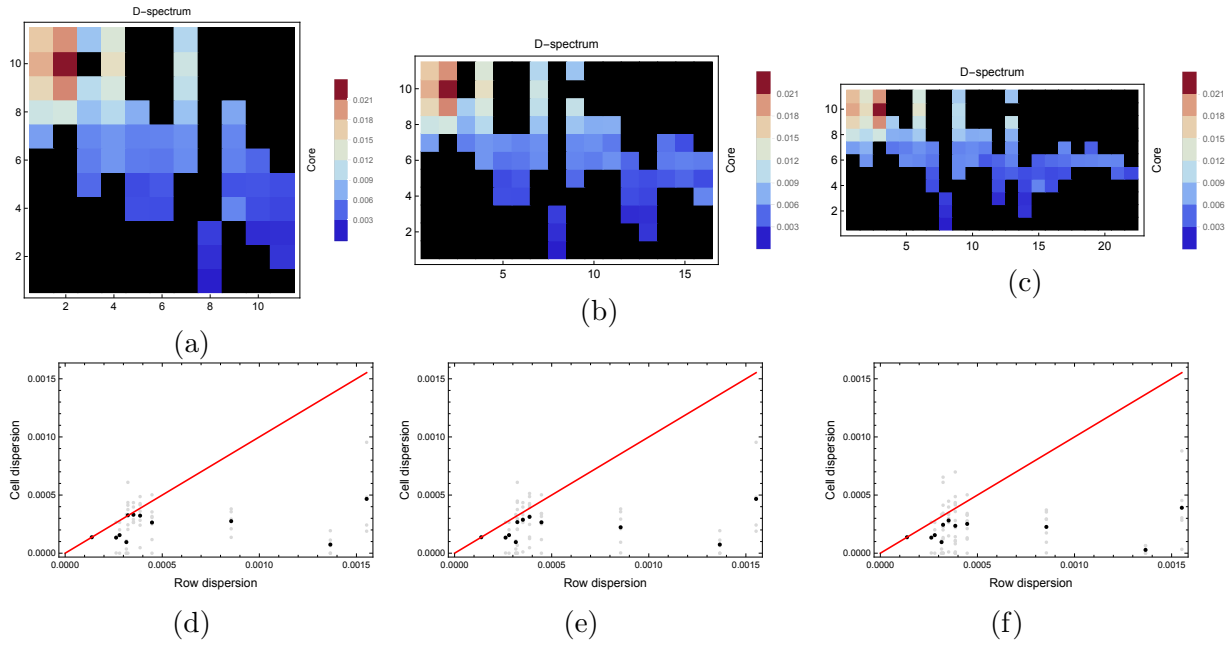
D-chain tomography of networks: a new structure spectrum and an application to the SIR process

SUPPLEMENTARY MATERIALS

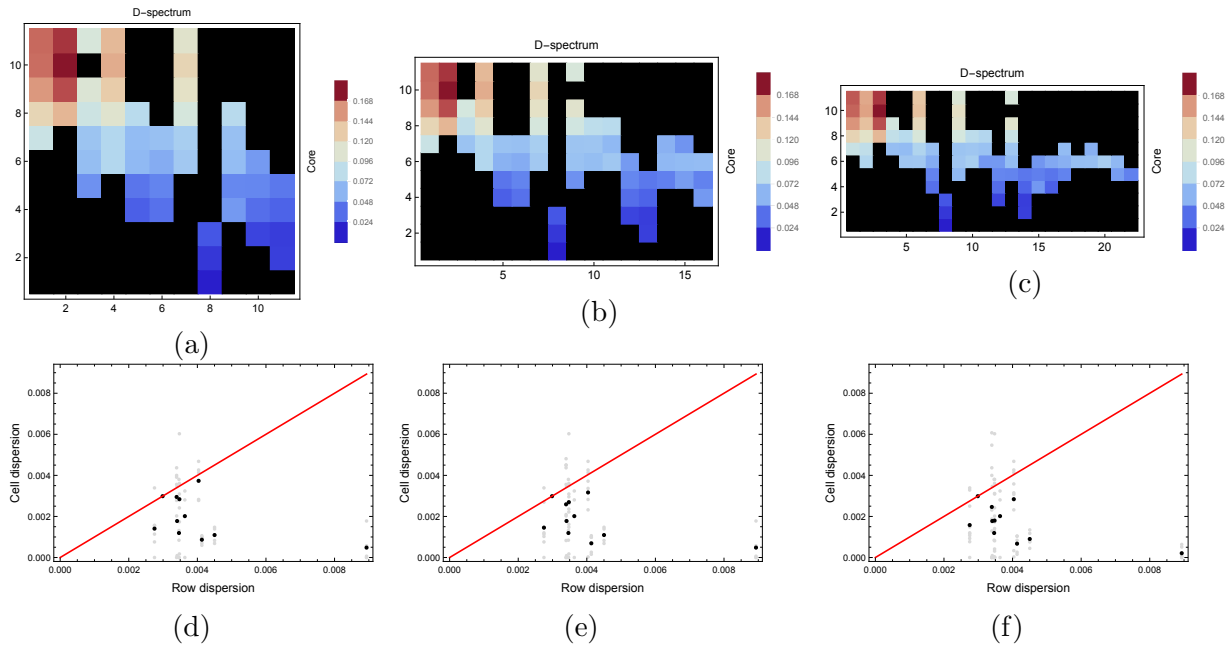
Ricky X. F. Chen, Andrei C. Bura, Christian M. Reidys

A More evaluation results on the SIR process

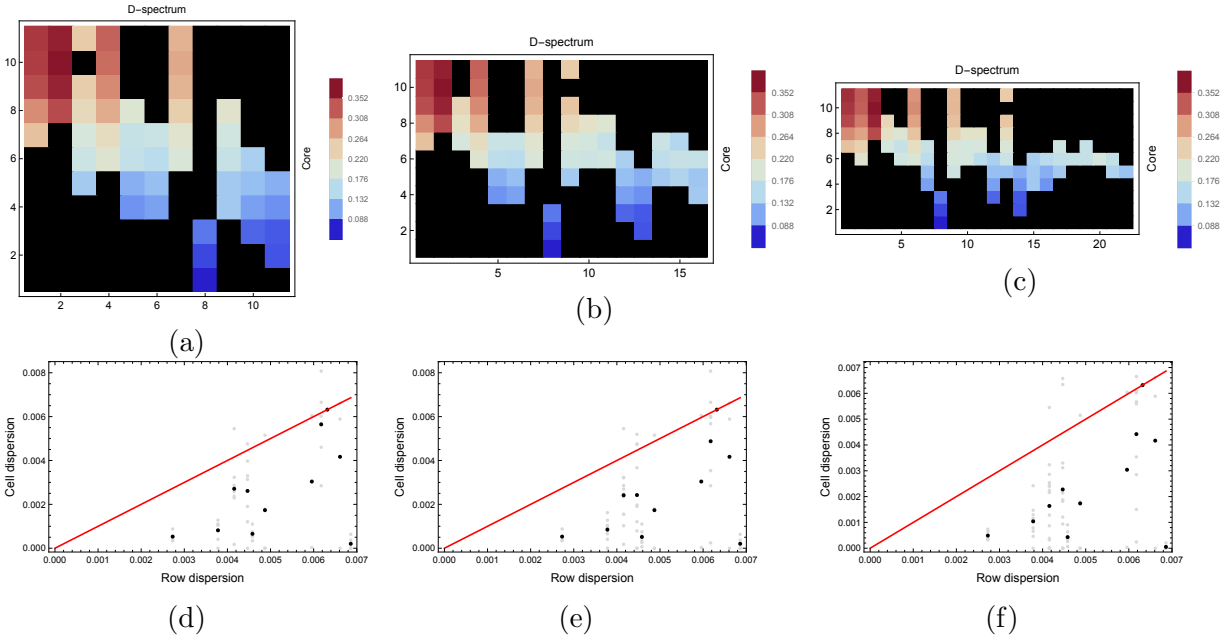
In Figures S1–S13, we provide further evaluation results on comparing clusterings based on core numbers and D-spectra.



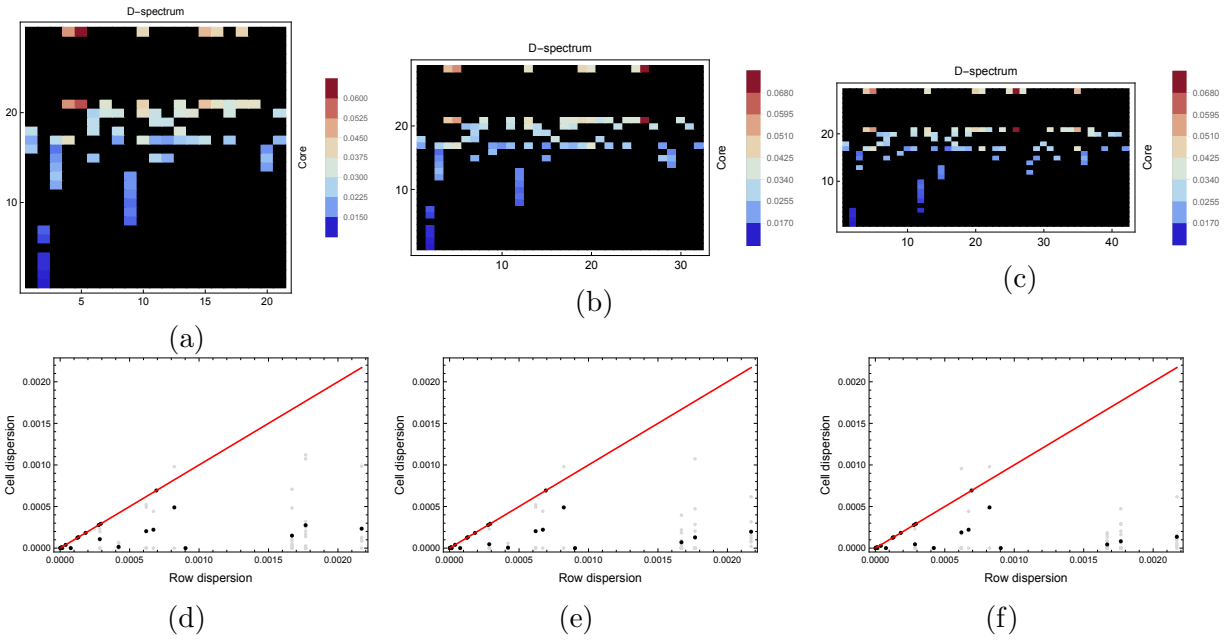
Supplemental Figure S1: Email network with β and different number of clusters from the D-spectra.



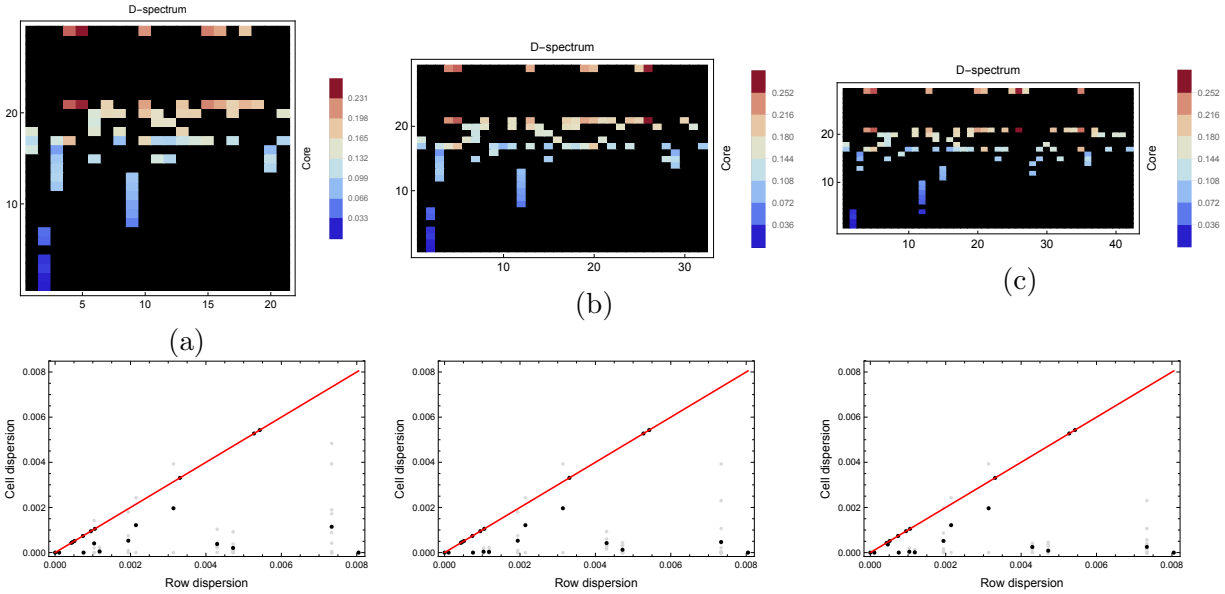
Supplemental Figure S2: Email network with 1.5β and different number of clusters from the D-spectra.



Supplemental Figure S3: Email network with 2β and different number of clusters from the D-spectra.

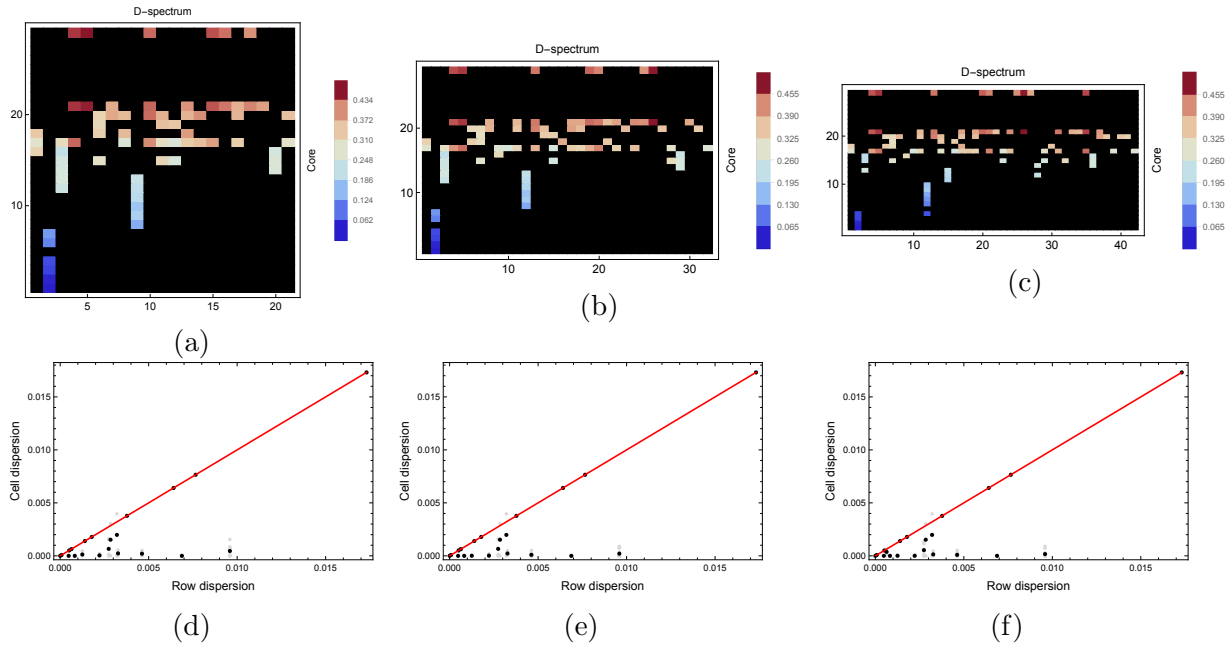


Supplemental Figure S4: Jazz network with β and different number of clusters from the D-spectra.

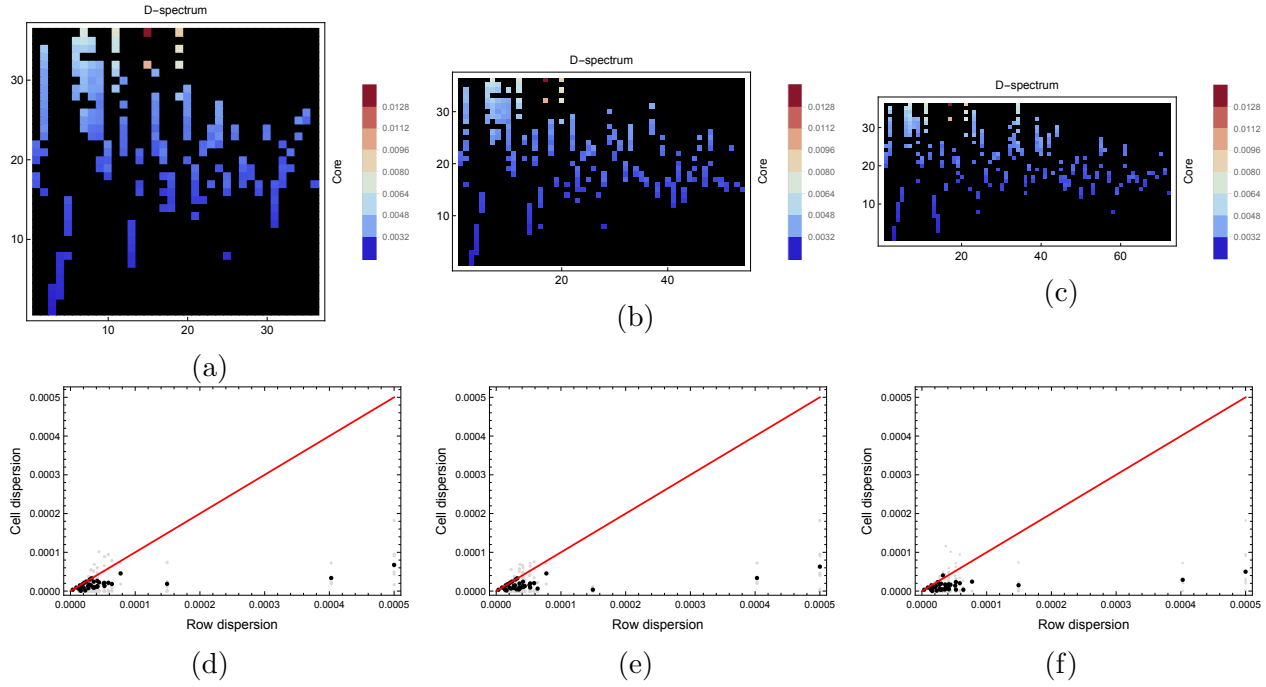


(a) (b) (c) (d) (e) (f)

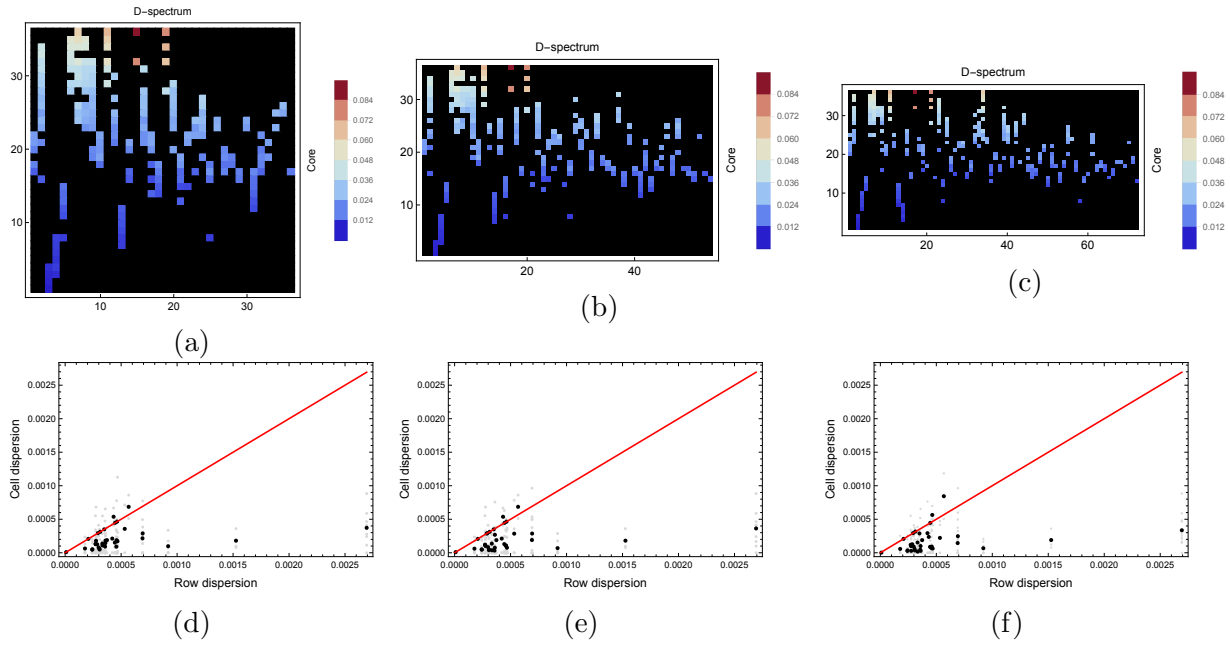
Supplemental Figure S5: Jazz network with 1.5β and different number of clusters from the D-spectra.



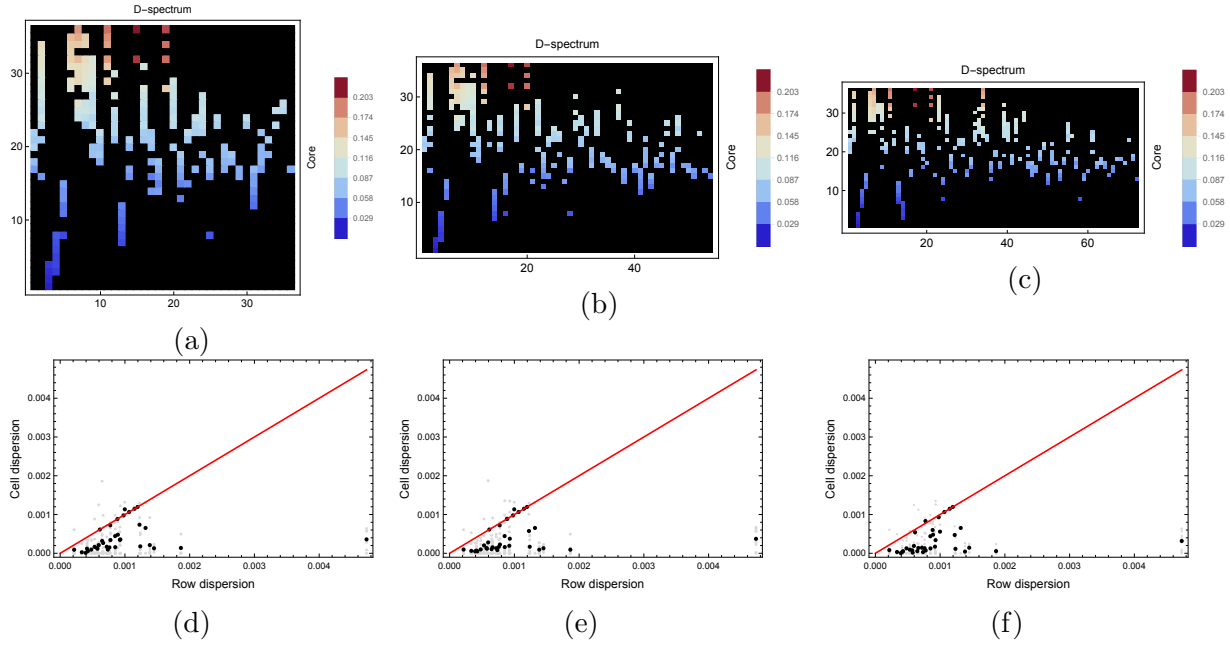
Supplemental Figure S6: Jazz network with 2β and different number of clusters from the D-spectra.



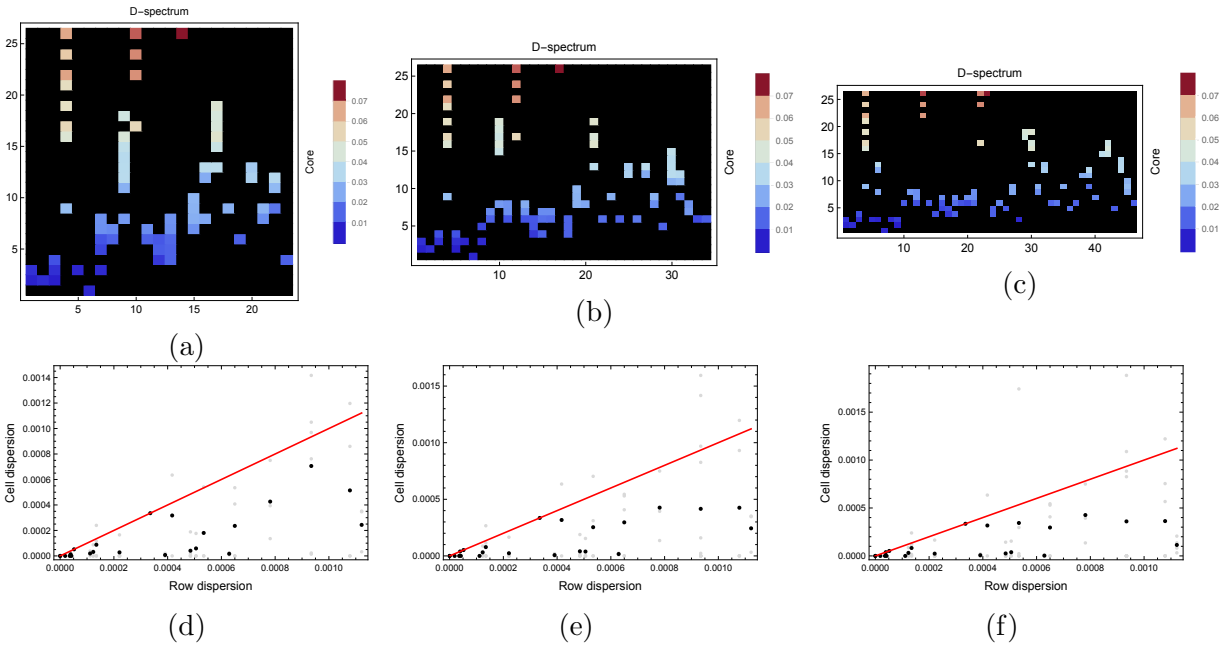
Supplemental Figure S7: PB network with β and different number of clusters from the D-spectra.



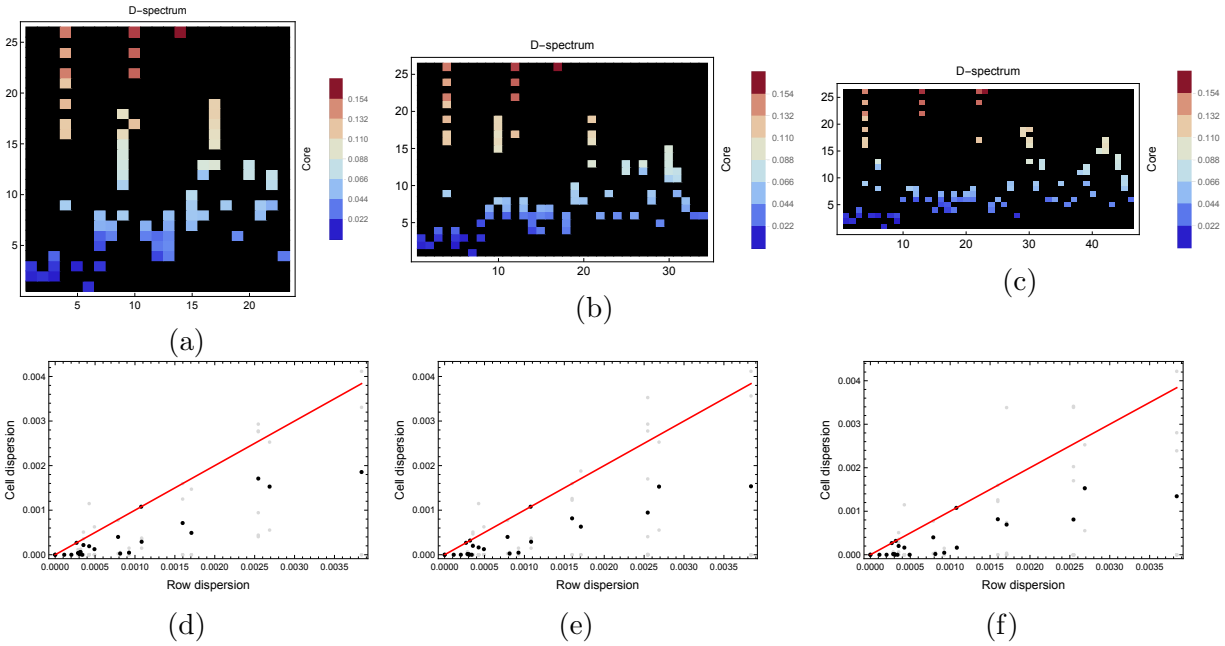
Supplemental Figure S8: PB network with 1.5β and different number of clusters from the D-spectra.



Supplemental Figure S9: PB network with 2β and different number of clusters from the D-spectra.

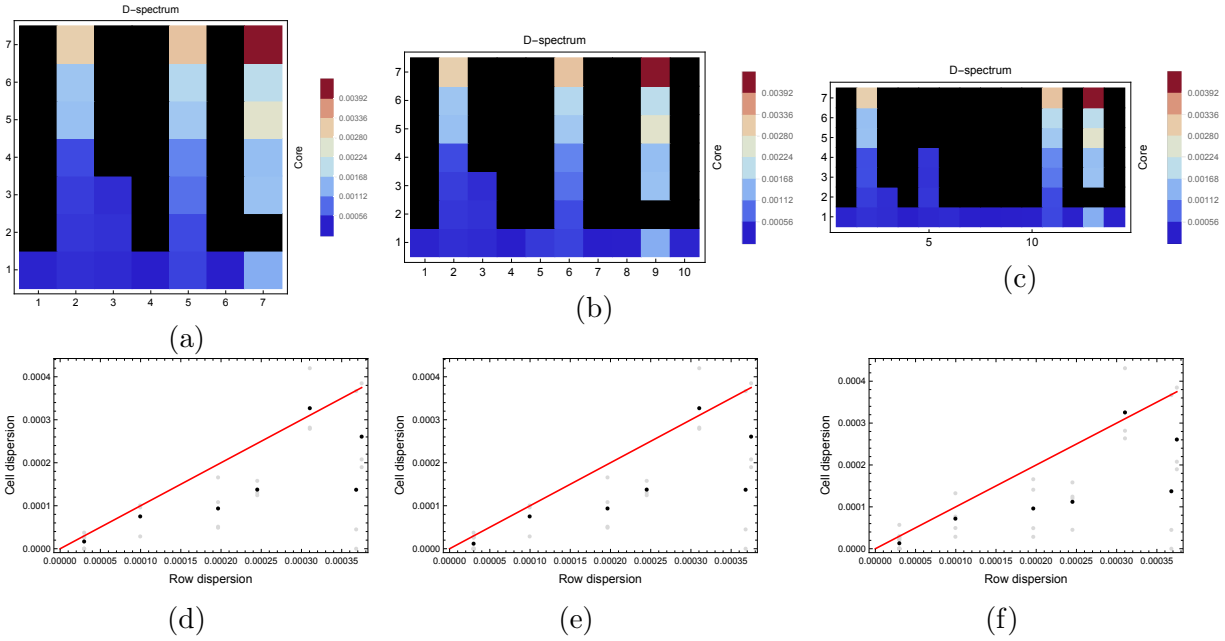


Supplemental Figure S10: USAir network with 1.5β and different number of clusters from the D-spectra.

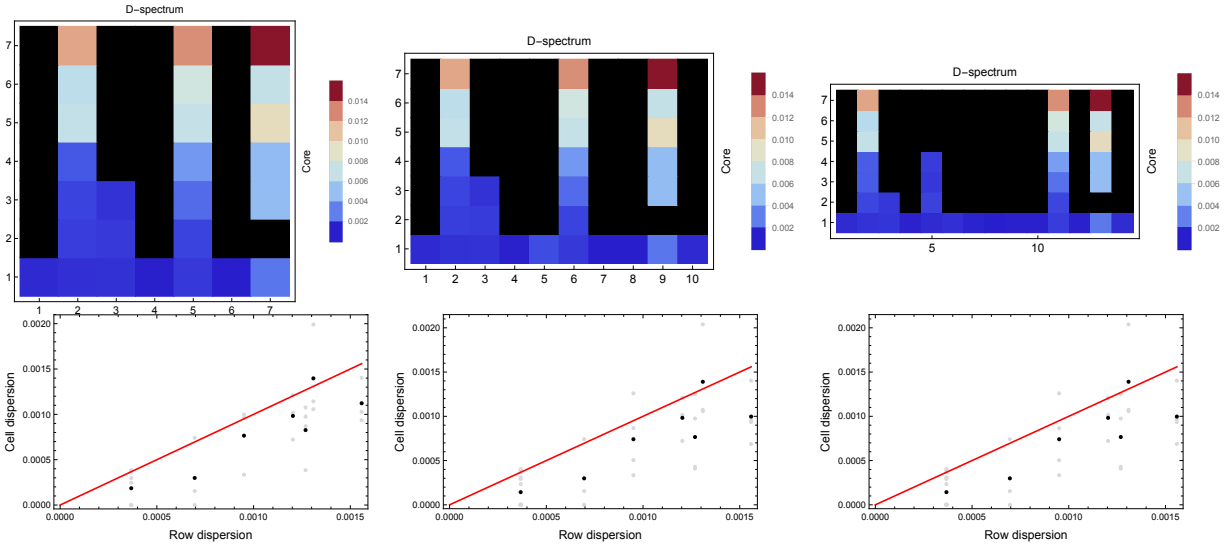


Supplemental Figure S11: USAir network with 2β and different number of clusters from the D-spectra.

Here we have adopted 7β and 8β for the network Router, because the infection rates around β are too low. Moreover, it seems that Router always behaves differently in



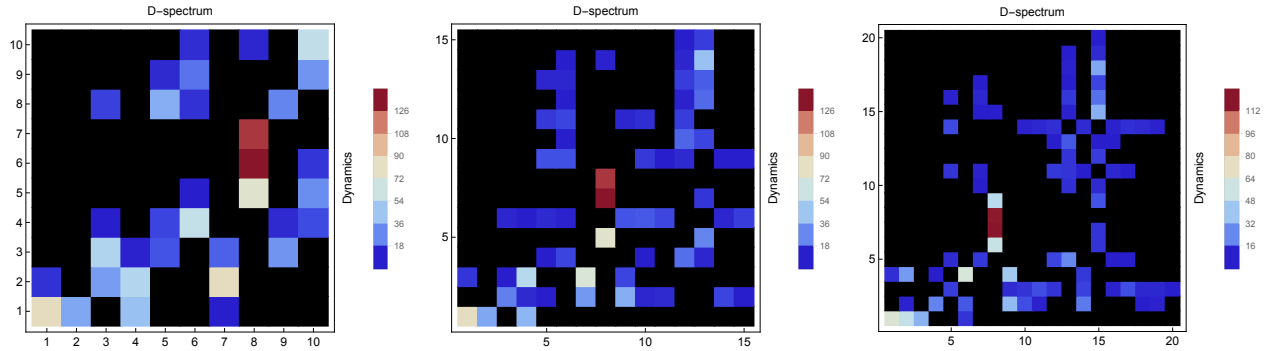
Supplemental Figure S12: Router network with 7β and different number of clusters from the D-spectra.



Supplemental Figure S13: Router network with 8β and different number of clusters from the D-spectra.

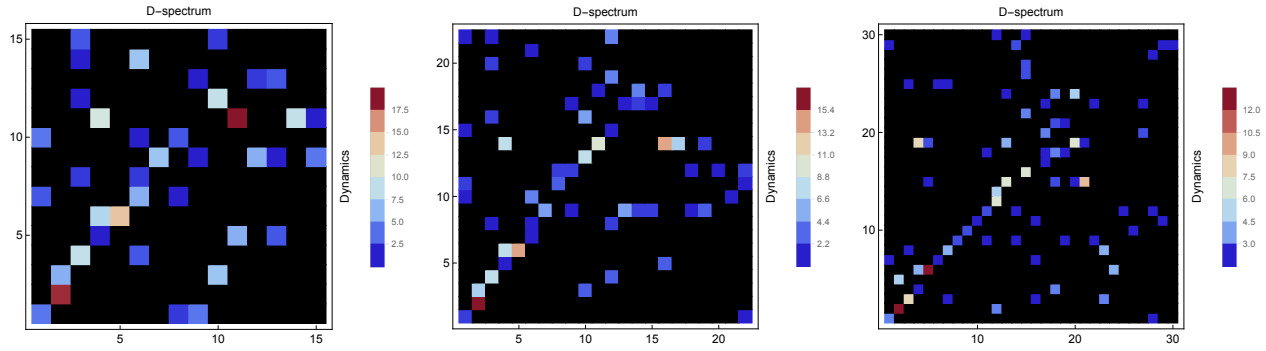
these kinds of studies (see also Lü et al. [25]), which may imply a very different network structure than other networks.

In Figures S14–S18, we provide further evaluation results on comparing clusterings based on D-spectra and spreading power. Since the infection rates at low transmission probabilities and high transmission probabilities are of different orders, say 10^{-3} and 10^{-1}, respectively, the part of high transmission probabilities will dominate the Euclidean distance if computing directly based on the absolute infection rates. In order to avoid this possible domination issue, we normalize the infection rate of each probability by the greatest infection rate (among all nodes) for that probability, before we apply the grouping function.



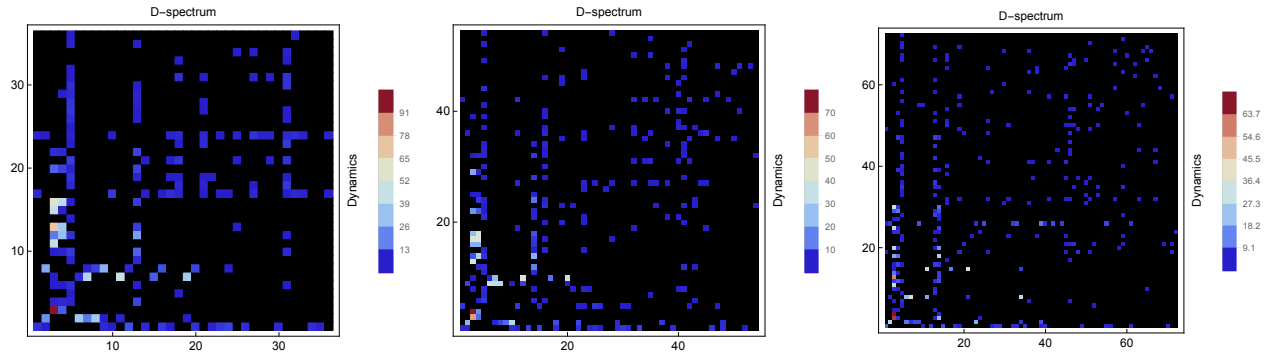
(a) (b) (c)

Supplemental Figure S14: Email network: dynamics vs D-spectra based clusterings into different number of clusters.



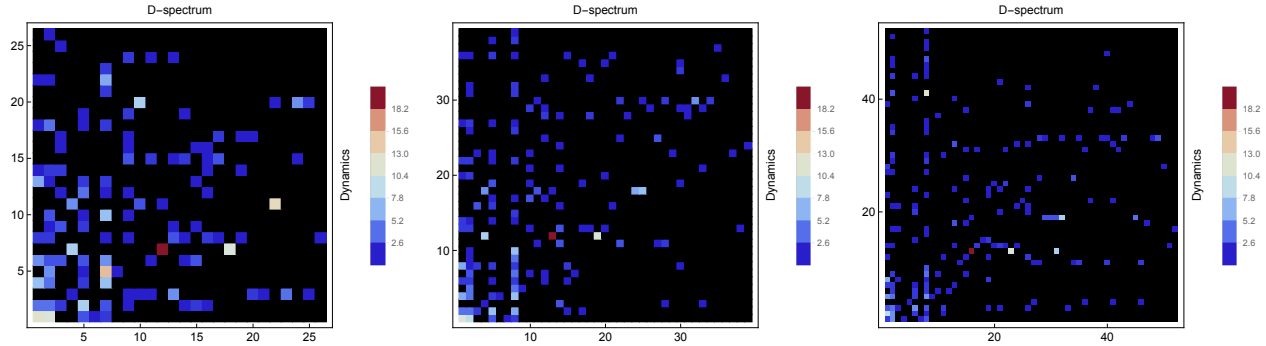
(a) (b) (c)

Supplemental Figure S15: Jazz network: dynamics vs D-spectra based clusterings into different number of clusters.



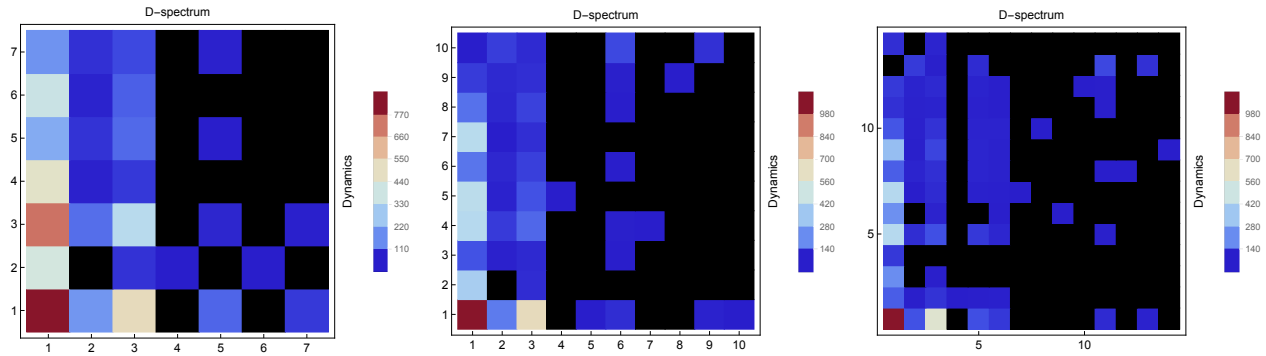
(a) (b) (c)

Supplemental Figure S16: PB network: dynamics vs D-spectra based clusterings into different number of clusters.



(a) (b) (c)

Supplemental Figure S17: USAir network: dynamics vs D-spectra based clusterings into different number of clusters.



(a) (b) (c)

Supplemental Figure S18: Router network: dynamics vs D-spectra based clusterings into different number of clusters. (Again, Router seems not behave nicely.)

Part Two: Loop Homology of Bi-secondary Structures

Loop homology of bi-secondary Structures

Qijun He[a,*], Andrei C. Bura[b,c], Christian M. Reidys[a,d]

[a] *Biocomplexity Institute and Initiative, University of Virginia, 995 Research Park Boulevard, Charlottesville, VA 22911*
[b] *Department of Mathematics, Virginia Tech, 225 Stanger Street, Blacksburg, VA 24061-1026*
[c] *Biocomplexity Institute of Virginia Tech, 1015 Life Sciences Circle Blacksburg, VA 24061*
[d] *Department of Mathematics, University of Virginia, 141 Cabell Dr, Charlottesville, VA 22903*

Abstract

In this paper we compute the loop homology of bi-secondary structures. Bi-secondary structures were introduced by Haslinger and Stadler and are pairs of RNA secondary structures, i.e. diagrams having non-crossing arcs in the upper half-plane. A bi-secondary structure is represented by drawing its respective seconday structures in the upper and lower half-plane. An RNA secondary structure has a loop decomposition, where a loop corresponds to a boundary component, regarding the secondary structure as an orientable fatgraph. The loop-decomposition of secondary structures facilitates the computation of its free energy and any two loops intersect either trivially or in exactly two vertices. In bi-secondary structures the intersection of loops is more complex and is of importance in current algorithmic work in bio-informatics and evolutionary optimization. We shall construct a simplicial complex capturing the intersections of loops and compute its homology. We prove that only the zeroth and second homology groups are nontrivial and furthermore show that the second homology group is free. Finally, we provide evidence that the generators of the second homology group have a bio-physical interpretation: they correspond to pairs of mutually exclusive substructures.

Keywords: RNA, bi-secondary structure, loop, nerve, simplicial homology.

1. Introduction

RNA sequences are single stranded nucleic acids that, in difference to DNA, can form a plethora of structural conformations. Over the last several decades, researchers have discovered an increasing number of important roles for RNA [1]. The folded structure of RNA is critically important to its function [2] and has

*Corresponding author
Email addresses: qhe196@gmail.com (Qijun He), anbur12@vt.edu (Andrei C. Bura), duck@santafe.edu (Christian M. Reidys)

been extensively studied at the coarse grained level of base pairing interactions. This leads to the notion of RNA secondary structures [3], that represent particular contact matrices and do not take into account the embedding in 3-space [4].

The thermodynamic stability of a secondary structure is characterized by its free energy, and is computed by summing the energy contribution of its loops [5, 6]. Prediction of the minimum free energy (i.e. the most stable) secondary structure for a given sequence, is an important problem at the most basic biological level [7].

The first mfe-folding algorithms for RNA secondary structures are due to [8, 4, 9]. Waterman studied the loop decomposition and the recursive construction of secondary structures and derived the first dynamic programming (DP) folding routines for secondary structures [10]. The DP routine facilitates polynomial time folding algorithms [11, 12, 13] and partition function calculation [14]. In [15], Haslinger and Stadler extended the notion of secondary structures to bi-secondary structures in order to study pseudoknotted structures, RNA structures exhibiting cross serial interactions [16]. Bi-secondary structures play furthermore a central role for studying sequences that can realize two, oftentimes mutually exclusive, conformations, in the context of evolutionary transitions [17] and in the study of RNA riboswitches, i.e. sequences that exhibit two stable configurations [18].

The partition function of structures w.r.t. a fixed sequence has a dual: the partition function of sequences compatible with a fixed structure [19]. Partition function and Boltzmann sampling have a variety of applications in sequence design [20, 21], extracting structural semantics [22] and to analyze mutational robustness [23].

RNA structures, viewed as abstract diagrams or trees, have been studied in enumerative combinatorics [10, 24, 25, 15], algebraic combinatorics [26], matrix-models [27, 28] and topology [29, 30, 31].

In [24], a bijection between linear trees and secondary structures was constructed. This facilitated beautiful, explicit formulae for the number of secondary structures on n vertices, having exactly k arcs.

Jin *et al.* [26] enumerate k-non-crossing RNA structures, based on the bijection given by Chen *et al.* [32], between k-non-crossing partial matchings and walks in $\mathbb{Z}^{k-1}$ which remain in the interior of the Weyl-chamber C_0. The bijection between oscillating tableaux and matchings originated from Stanley [33] and was generalized by Sundaram [34].

Penner and Waterman connected RNA structures with topology by studying the space of RNA secondary structures. They proved that the geometrical realizations of the associated complex of secondary structures is a sphere [35]. In [29], Bon *et al.* presented a topological classification of secondary structures, based on matrix models.

In the course of computing the Euler characteristics of the Moduli space of a curve, [36], Harer and Zagier computed the generating function of the number of linear chord diagrams of genus g with n chords. Based on this line of work, Andersen *et al.* [28], enumerated the number of chord diagrams of fixed

genus with specified numbers of backbones and chords. Such an enumeration of chord diagrams provides the number of secondary structures of a given genus as well as the number of cells in Riemann's moduli spaces for bordered surfaces. This was done by using Hermitian matrix model techniques and topological recursions along the lines of [37]. Employing an augmented version of the topological recursion on unicellular maps of Chapuy [38], Huang *et al* [31] derived explicit expressions for the coefficients of the generating polynomial of topological shapes of RNA structures and the generating function of RNA structures of genus g. This lead to uniform sampling algorithms for structures of fixed topological genus as well as a natural way to resolve crossings in pseudoknotted structures [39].

Bi-secondary structures emerge naturally in the context of evolutionary transitions, since they are closely connected to sets of sequences, that are *simultaneously* compatible with two structures [40]. This paper is motivated by the dynamical programming (DP) routine of Huang [41], that is based on sub-problems associated with sets of loops. The sub-problems were constructed incrementally by adding one loop at a time, where subsequently added nucleotides affect the energy calculation if they appear in multiple loops. This naturally leads one to consider intersections of loops and eventually to introduce the *nerve* of loops as a simplicial complex.

In this paper, we study the homology of bi-secondary structures [15]. We show that for any bi-secondary structure R, we have only two nontrivial homology groups, $H_0(R)$ and $H_2(R)$. The key to establish $H_1(R) = 0$ is to establish in Lemma 6 the existence of certain, spanning, sub 1-skeleta, whose existence follows from an inductive argument over the arcs of one of the secondary structures. These skeleta give rise to specific trees, which in turn allow one to systematically process elements of $\mathrm{Ker}(\partial_1)$. We show that $H_0(R) \cong \mathbb{Z}$ and $H_2(R) \cong \bigoplus_{k=1}^{r} \mathbb{Z}$, introducing the rank of $H_2(R)$ as a new invariant of the bi-secondary structures. We show that $H_2(R)$ is free by showing that it is a subgroup of a free group, whose freeness in turn is a consequence of Lemma 4 which guarantees the existence of exposed faces of 3-simplices. We then discuss the new invariant, observing that all RNA riboswitch sequences in data-bases exhibit $\mathrm{rank}(H_2(R)) = 1$, seldomly assumed by random secondary structure pairs and provide an outlook on future work.

2. Some basic facts

We shall begin by defining loops in an RNA secondary structure and then present results on its loop decomposition.

An RNA diagram S over $[n]$, is a vertex-labeled graph whose vertices are drawn on the horizontal axis and labeled by $[n] = \{1, \ldots, n\}$. An *arc* (i, j), is an ordered pair of vertices, which represents the base pairing between the i-th and j-th nucleotides in the RNA structure. Furthermore, each vertex can be paired with at most one other vertex, and the arc that connects them is drawn in the upper half-plane. We introduce two "formal" vertices associated with positions

0 and $n+1$, respectively, closing any diagram by the arc $(0, n+1)$, called the rainbow. The set $[0, n+1]$ is called the diagram's *backbone*, see Figure 1.

Let S be an RNA diagram over $[n]$. Two arcs (i, j) and (p, q) are called *crossing* if and only if $i < p < j < q$. S is called a *secondary structure* if it does not contain any crossing arcs. The arcs of S can be endowed with a partial order as follows: $(k, l) \prec_S (i, j) \iff i < k < l < j$. We denote this by $(S, \prec_S)$ and call it the arc poset of S. Finally, an *interval* $[i, j]$ on the backbone is the set of vertices $\{i, i+1, \ldots, j-1, j\}$.

Let S be a secondary structure over $[n]$. A *loop* s in S is a subset of vertices, represented as a disjoint union of a sequence of contiguous blocks on the backbone of S, $s = \dot{\bigcup}_{i=1}^{k}[a_i, b_i]$, such that (a_1, b_k) and (b_i, a_{i+1}), for $1 \leq i \leq k-1$, are arcs and such that any other interval-vertices are unpaired. Let α_s denote the unique, maximal arc (a_1, b_k) of the loop.

In this paper we shall identify a secondary structure with its set of loops. Let S be a secondary structure over $[n]$ and $s = \dot{\bigcup}_{i=1}^{k}[a_i, b_i]$ a loop in S, then
(1) each unpaired vertex is contained in exactly one loop,
(2) (a_1, b_k) is maximal w.r.t. $\prec_S$ among all arcs contained in s, i.e. there is a bijection between arcs and loops, mapping each loop to its maximal arc,
(3) the Hasse diagram of the S arc-poset is a rooted tree $\mathrm{Tr}(S)$, having the rainbow arc as root,
(4) each non-rainbow arc appears in exactly two loops.

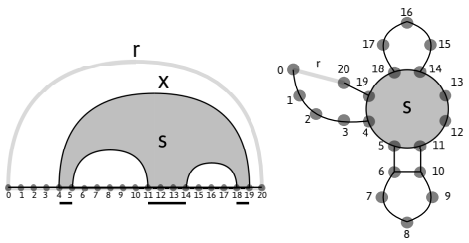


Figure 1: LHS: a secondary structure, S, and a distinguished loop $s = [4,5] \cup [11,14] \cup [18,19]$. r is the rainbow arc and $\alpha_s = x$. RHS: S represented as a planar RNA molecule.

Proposition 1. *Let s, s' and s'' be three distinct loops in a secondary structure S. Then*
(1) $s \cap s' \cap s'' = \varnothing$, (2) $s \cap s' \neq \varnothing$ implies $|s \cap s'| = 2$.

Proof. Vertices of S are either paired or unpaired. In the latter case, they are contained in exactly one loop. In the former, by construction, they are endpoints of arcs and contained in exactly two distinct loops. Hence, no vertex can be contained in three distinct loops. In case of $s \cap s' \neq \varnothing$ the loops intersect in the endpoints of exactly one arc, which is maximal for exactly of of them, whence $|s \cap s'| = 2$. □

In this section we introduce bi-secondary structures and their nerves. To this end we introduce the nerve over a finite collection of sets:

Let $X = \{x_0, x_1, \ldots, x_m\}$ be a collection of finite sets. We call $Y = \{x_{i_0}, \ldots, x_{i_d}\} \subseteq X$ a *d-simplex* of X iff $\bigcap_{k=0}^{d} x_{i_k} \neq \varnothing$. We set $\Omega(Y) = \bigcap_{k=0}^{d} x_{i_k}$ and refer to $\omega(Y) = |\Omega(Y)| \neq 0$ as the *weight* of Y. Let $K_d(X)$ be the set of all d-simplices of X, then the *nerve* of X is

$$K(X) = \dot{\bigcup}_{d=0}^{\infty} K_d(X) \subseteq 2^X.$$

A d'-simplex $Y' \in K(X)$ is called a d'-face of Y if $d' < d$ and $Y' \subseteq Y$. By construction, $K(X)$ is an abstract simplicial complex. Let S be a secondary structure over $[n]$. The geometric realization of $K(S)$, the nerve over the set of loops of S, is a tree. By means of the correspondence between arcs and loops, this tree of loops is isomorphic to $\mathrm{Tr}(S)$.

Definition 1. *Given two secondary structures S and T over $[n]$, we refer to the pair $R = (S, T)$ as a bi-secondary structure. Let $S \cup T$ be the loop set of R and $K(R) = \dot{\bigcup}_{d=0}^{\infty} K_d(R)$ its nerve of loops.*

We represent the diagram of a bi-secondary structure $R = (S, T)$ with the arcs of S in the upper half plane while the arcs of T reside in the lower half plane. Let $R = (S, T)$ be a bi-secondary structure with loop nerve $K(R)$. A 1-simplex $Y = \{r_{i_0}, r_{i_1}\} \in K_1(R)$ is called *pure* if r_{i_0} and r_{i_1} are loops in the same secondary structure and *mixed*, otherwise.

Suppose Y is a pure 1-simplex in $K(R)$, then by Proposition 1 we have $\omega(Y) = 2$, see Figure 2.

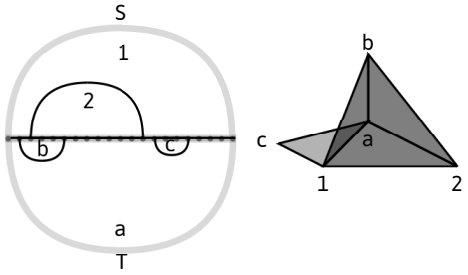


Figure 2: LHS: a bi-secondary structure $R = (S, T)$. RHS: the geometric realization of its loop nerve, $K(R)$. the 1-simplices $\{c, 1\}$ and $\{1, 2\}$ are mixed and pure, respectively.

Lemma 1. *Let $R = (S, T)$ be a bi-secondary structure with nerve $K(R)$. For any $Y \in K_2(R)$, exactly one of its three 1-faces is pure, the other two being mixed. Furthermore, we have $\omega(Y) \leq 2$.*

Proof. Let $Y = \{r'_0, r'_1, r'_2\} \in K_2(R)$ be a 2-simplex of $K(R)$. By Proposition 1, $\cap_{i=0,1,2} r'_i \neq \varnothing$ implies that not all three loops can be from the same structure. W.l.o.g. suppose $r'_0, r'_1 \in S$ and $r'_2 \in T$. Certainly $Z = \{r'_0, r'_1\}$ is a pure 1-face of Y and two other 1-faces of Y are by construction mixed since they contain $r'_3 \in T$. For any 1-face Z' of Y, we have $\omega(Y) \leq \omega(Z')$ and Proposition 1 guarantees $\omega(Z) = 2$, whence the lemma. □

Lemma 2. *Let $R = (S, T)$ be a bi-secondary structure with nerve $K(R)$ and let $Y = \{r_0, r_1, r_2, r_3\} \in K_3(R)$ be a 3-simplex. Then we have*
(a) $Y = \{s_0, s_1, t_0, t_1\}$, where $s_0, s_1 \in S$ and $t_0, t_1 \in T$,
(b) Y has exactly two pure 1-faces, $\{s_0, s_1\}$ and $\{t_0, t_1\}$,
(c) $\omega(Y) \leq 2$.

Proof. Any 2-simplex in the loop nerve is of the form $\{s, t, t'\}$ or $\{t, s, s'\}$ and Y has the 2-faces $\{r_0, r_1, r_2\}$, $\{r_1, r_2, r_3\}$, $\{r_0, r_2, r_3\}$ and $\{r_0, r_1, r_3\}$. In view of the 2-simplex $\{r_0, r_1, r_2\}$, we can, w.l.o.g. set $s_0 = r_0$ $s_1 = r_1$ and $t_0 = r_2$. The 2-simplex $\{r_0, r_1, r_3\}$ then implies that r_3 is a T-loop, whence we can set $t_1 = r_3$ and (a) follows. Assertion (b) follows immediately from (a). Finally, (c), follows from $\cap_{i=0}^{3} r_i \subset s_0 \cap s_1$ and $|s_0 \cap s_1| = 2$. □

Lemma 3. *Let $R = (S, T)$ be a bi-secondary structure and let $K(R)$ be its loop nerve. Then any pure 1-simplex, P, appears as the 1-face of at most two distinct 3-simplices.*

Proof. W.l.o.g. we may assume $P = \{s_0, s_1\}$, for some $s_0, s_1 \in S$. If P is a 1-face of a 3-simplex Y then by Lemma 2, $Y = \{s_0, s_1, t_0, t_1\}$ for some $t_0, t_1 \in T$. For any such 3-simplex we have $\Omega(Y) \subset s_0 \cap s_1$, where $s_0 \cap s_1 = \{x, y\}$. Similarly $t_0 \cap t_1 = \{a, b\}$ and $\Omega(Y) \subset \{a, b\}$. In the case of $\{a, b\} = \{x, y\}$, $\{s_0, s_1\}$ is contained exclusively in the 3-simplex $\{s_0, s_1, t_0, t_1\}$. Otherwise, we obtain two 3-simplices $Y_x = \{s_0, s_1, t_0^x, t_1^x\}$ and $Y_y = \{s_0, s_1, t_0^y, t_1^y\}$ and in view of $\{x, y\} \cap t_0^x \cap t_1^x = \{x\}$ and $\{x, y\} \cap t_0^y \cap t_1^y = \{y\}$, both, Y_x and Y_y contain P, see Figure 3. □

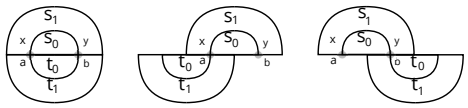


Figure 3: LHS: the case $\{a, b\} = \{x, y\}$. Center: the case of $Y_x = \{s_0, s_1, t_0^x, t_1^x\}$. RHS: the case of $Y_y = \{s_0, s_1, t_0^y, t_1^y\}$.

Definition 2. *Let $K(X) = \dot{\bigcup}_{d=0}^{\infty} K_d(X)$ be an abstract simplicial complex and let $Y \in K_d(X)$ be a d-simplex. Let Y' be a $(d-1)$-face of Y. We say Y' is* Y-exposed *if and only if no other d-simplices of K contain Y' as a $(d-1)$-face.*

Lemma 4. *Let $R = (S, T)$ be a bi-secondary structure with loop-nerve $K(R)$. Then any $Y \in K_3(R)$ contains at least two Y-exposed 2-faces.*

Proof. By Lemma 2, any 3-simplex, Y, is of the form $Y = \{s_0, s_1, t_0, t_1\}$ and has exactly two pure 1-faces, $P_1 = \{s_0, s_1\}$ and $P_2 = \{t_0, t_1\}$. We shall use P_1 to construct at least one specific, exposed 2-face of Y. For P_1, $W_1 = \{s_0, s_1, t_0\}$ and $W_2 = \{s_0, s_1, t_1\}$ are the only two distinct 2-faces, that contain P_1 as a pure 1-face. In $K(R)$, Y is the unique 3-simplex that contains both W_1 and W_2 as 2-faces. It thus remains to show that there cannot exist two distinct 3-simplices Y_1 and Y_2 having W_1 and W_2 as a 2-face, respectively. If this were the case,

Y, Y_1, Y_2 were, by construction, three distinct 3-simplices having P_1 as a pure 1-face, which, in view of Lemma 3, is impossible. Thus, either W_1 or W_2 is exposed in Y. We can argue analogously for P_2 and the lemma follows. □

3. Homology

In this section we consider the chain complex over the loop nerve $K(R)$ and compute its homology. We will show that only the second homology group $H_2(R)$ is nontrivial and that $H_2(R)$ is free. This produces a new invariant for bi-secondary structures, that provides insight into RNA riboswitch sequences, i.e. where a single sequence switches, depending on context, between mutually exclusive structures.

Suppose we are given a bi-secondary structure $R = (S, T)$ and let $(T, \prec_T)$ and $(S, \prec_S)$ be the posets of arcs on the secondary structures T and S respectively. $\prec_S$ and $\prec_T$ allow us to endow R with the poset-structure:

$$(R, \prec_R) = (T, \prec_T) \oplus (S, \prec_S),$$

where $R = T \cup S$ and $\prec_R$ is given by

$$r_1 \prec_R r_2 \Leftrightarrow \begin{cases} r_1, r_2 \in T \text{ and } r_1 \prec_T r_2 \\ r_1, r_2 \in S \text{ and } r_1 \prec_S r_2 \\ r_1 \in S,\ r_2 \in T \end{cases}$$

Let us next choose a linear extension of $(R, \prec_R)$, $(R, \leq)$, to which we refer to as the simplicial order of the loop nerve. Any d-simplex, $Y \in K_d(R)$ becomes then the unique d-tuple $Y = (r_0, r_1, \ldots, r_d)$ where $r_0 \leq r_1 \leq \cdots \leq r_d$.

Let $R = (S, T)$ be a bi-secondary structure with loop nerve $K(R)$. Let $C_d(R)$ be the simplicial chain group of dimension d of $K(R)$. Let $Y = (r_0, r_1, \ldots, r_d) \in C_d(R)$ and $\partial_d : C_d(R) \longrightarrow C_{d-1}(R)$ be the boundary map given by

$$\partial_d(Y) = \sum_{i=0}^{d} (-1)^i \, (r_0, \ldots, r_{i-1}, r_{i+1}, \ldots, r_d) \, .$$

Let furthermore $H_d(R) = \mathrm{Ker}(\partial_d)/\mathrm{Im}(\partial_{d+1})$ be the d'th homology group of the loop nerve of R. In the following we shall show

Theorem 1. *The loop-nerve of a bi-secondary structure, R, has only the following nontrivial homology groups*

$$\begin{aligned} H_0(R) &= \mathbb{Z} \\ H_2(R) &= \bigoplus_{k=1}^{r} \mathbb{Z}. \end{aligned}$$

Let us begin proving Theorem 1 by first noting

Lemma 5. $H_0(R) \cong \mathbb{Z}$.

Proof. By construction, the 1-skeleton of $K(R)$ contains the two rooted trees associated to S and T, respectively. Their respective root-loops are connected by a 1-simplex as both rainbows share the vertices 0 and $n+1$. Thus any loop is path connected to a rainbow loop implying that any loop is, modulo boundaries, equivalent to a rainbow loop. Hence the assertion follows. □

Let $t \in T$ be a loop, we set

$$S(t) = \{s \in S \mid \{s, t\} \in K_1(R)\}$$

$$T(t) = \{t' \in T \mid t' \prec_T t, \nexists\, t'' \in T \text{ s.t. } \alpha_{t'} \prec_T \alpha_{t''} \prec_T \alpha_t, \{t, t'\} \in K_1(R)\},$$

the sets of S and T neighbors of t, respectively. Let $R(t) = S(t) \cup T(t)$ and let $\mathrm{Gr}(t)$ be the vertex induced sub-graph of the 1-skeleton in the geometric realization of $K(R)$, whose vertices are the loops in $R(t)$. By construction, $\mathrm{Gr}(t)$, does not contain the loop t as a vertex.

Let $R = (S, T)$ be a bi-secondary structure with loop nerve $K(R)$ and let $t \in T$ be a loop. A connected, spanning sub-graph, $G(t) \leq \mathrm{Gr}(t)$, in which each edge satisfies

$$\{r_a, r_b\} \in G(t) \quad \Longrightarrow \quad \{r_a, r_b, t\} \in K_2(R),$$

is called a Δ_t-graph and we refer to its edges as Δ_t-edges.

Theorem 2. *Let $R = (S, T)$ be a bi-secondary structure and $K(R)$ be its loop nerve, then $H_1(R) = 0$.*

Proof. We shall inductively build T, arc by arc, from bottom to top and from left to right.
For the induction basis assume $T = \varnothing$, then, by construction, $K(R) = K(S)$ and the geometric realization of its nerve is a tree, with edges between loops $p, q \in S$, whenever p directly covers q w.r.t. $\prec_S$. Hence $H_1(R) = 0$ and the induction basis is established.

For the induction step, the induction hypothesis stipulates $H_1(S, T) = 0$. We shall show that $H_1(R') = 0$, where $R' = (S, T')$ and T' is obtained from T by adding the arc α_t, the maximal arc of the newly added loop t. We have the following scenario

$$\begin{array}{ccccccc} C_2(R') & \longrightarrow & C_1(R') & \longrightarrow & C_0(R') & \longrightarrow & 0 \\ \uparrow & & \uparrow & & \uparrow & & \\ C_2(R) & \longrightarrow & C_1(R) & \longrightarrow & C_0(R) & \longrightarrow & 0 \end{array} \tag{1}$$

where the vertical and horizontal maps are the natural embeddings and boundary homomorphisms, respectively.
Claim 1.

$$\mathrm{Ker}(\partial_1^{R'}) \subseteq \mathrm{Ker}(\partial_1^{R}) \oplus \mathrm{Im}(\partial_2^{R'}).$$

To prove the claim, we consider $\tau_0 \in C_1(R')$:

$$\tau_0 = \sum_{e_i \in K_1(R)} n_i e_i + \sum_{e_j = \{r,t\}, r \in R(t)} n_j e_j,$$

distinguishing any edges, that contain t, in the second term. The idea is to now process the edges containing t in a systematic way. To this end we first claim

Claim 2. Let $R = (S, T)$ be a bi-secondary structure with nerve $K(R)$ and let t be a T-loop, then, there exists a Δ_t-graph, $G(t)$.

We shall give the proof of Claim 2 by means of Lemma 6, below.

Given a Δ_t-graph, any of its vertices can be employed as the root of a spanning $G(t)$-sub-tree and we select the $\leq$-maximum $G(t)$-vertex as root. Let $A(t)$ denote this rooted tree. Any vertex, $r \in R(t)$, appearing in an edge $\{r, t\}$, occurs in $A(t)$ and any two $A(t)$-neighbors, $\{r_1, r_2\}$ are in the boundary of the 2-simplex $\{r_1, r_2, t\}$.

We examine now all $R(t)$-vertices in the following systematic way: starting with $A(t)$-leaves, pick r_0 and its unique, immediate, $A(t)$-ancestor, r_1. We then have either

Case 1*:* $r_0 \leq r_1$.

Then (r_0, r_1) is a simplex and using that $\{r_0, r_1\}$ is a Δ_t-edge, we are guaranteed that $\{r_0, r_1, t\}$ is a 2 simplex and

$$\partial_2(r_0, r_1, t) = (r_1, t) - (r_0, t) + (r_0, r_1).$$

We have a closer look at the sum of simplices $n_0(r_0, t) + n_1(r_1, t)$,

$$\begin{aligned} n_0(r_0, t) + n_1(r_1, t) &= n_0(r_0, t) + n_1(r_1, t) \pm n_0(r_0, r_1) \pm n_0(r_1, t) \\ &= -n_0[(r_1, t) - (r_0, t) + (r_0, r_1)] + (n_0 + n_1)(r_1, t) + n_0(r_0, r_1) \\ &= -n_0\partial_2((r_0, r_1, t)) + (n_0 + n_1)(r_1, t) + n_0(r_0, r_1). \end{aligned}$$

This produces on the RHS a boundary, a new term $(r_0, r_1) \in C_1(R)$, a modified coefficient for the simplex (r_1, t) and the term (r_0, t) has become part of a boundary.

Case 2*:* $r_1 \leq r_0$.

Here (r_1, r_0) is a simplex and

$$\partial_2(r_1, r_0, t) = (r_0, t) - (r_1, t) + (r_1, r_0).$$

Furthermore,

$$\begin{aligned} n_0(r_0, t) + n_1(r_1, t) &= n_0(r_0, t) + n_1(r_1, t) \pm n_0(r_0, r_1) \pm n_0(r_1, t) \\ &= n_0[(r_0, t) - (r_1, t) + (r_0, r_1)] + (n_0 + n_1)(r_1, t) - n_0(r_0, r_1) \\ &= n_0\partial_2((r_1, r_0, t)) + (n_0 + n_1)(r_1, t) - n_0(r_0, r_1). \end{aligned}$$

On the RHS we, again, have a boundary, a new term $(r_0, r_1) \in C_1(R)$, a modified coefficient for the simplex (r_1, t) and the term (r_0, t) has become part of a boundary.

Iterating this procedure, we step by step transform simplices $\{r, t\}$ into boundaries, working along the tree $A(t)$, from the leaves to the root. This finally produces the following expression for τ_0

$$\tau_0 = \epsilon_0 + n_k(r_k, t) + \tau_k,$$

where $\epsilon_0 \in \mathrm{Im}(\partial_2^{R'})$, i.e. ϵ_0 is a boundary, r_k is the root of $A(t)$ and $\tau_k \in C_1(R)$. At this point we cannot proceed transforming (r_k, t) into a boundary and shall argue as follows: suppose $\tau_0 \in \mathrm{Ker}(\partial_1^{R'})$. Then

$$\partial_1^{R'}(\tau_0) = \partial_1^{R'}(\epsilon_0) + n_k t - n_k r_k + \partial_1^{R'}(\tau_k).$$

Since $\epsilon_0 \in \mathrm{Im}(\partial_2^{R'})$ we certainly have $\partial_1^{R'}(\epsilon_0) = 0$. By construction of the Δ_t-graph $G(t)$, the 0-simplex $\{t\}$ does not appear in $\partial_1^{R'}(\tau_k)$, from which we conclude $n_k = 0$. As a result we have $n_k r_k = 0$ and since $\partial_1^{R'}(\tau_k) = \partial_1^{R}(\tau_k)$, we have $0 = \partial_1^{R'}(\tau_0) = \partial_1^{R}(\tau_k)$, and as a result

$$\tau_k \in \mathrm{Ker}(\partial_1^{R}).$$

The induction hypothesis guarantees $H_1(R) = 0$, i.e. $\mathrm{Ker}(\partial_1^R) = \mathrm{Im}(\partial_2^R)$. Hence $\tau_k \in \mathrm{Im}(\partial_2^R)$, which in view of diagram (1) implies $\tau_0 \in \mathrm{Im}(\partial_2^{R'})$ and we have proved $\mathrm{Ker}(\partial_1^{R'}) = \mathrm{Im}(\partial_2^{R'})$. □

It remains to show the proof of Claim 2. To this end, let $r \in R$ be a loop with $\alpha_r = (a, b)$ and denote $b(r) = b(\alpha_r) = a$ and $e(r) = e(\alpha_r) = b$.

Let $s = \dot{\bigcup}_{i=1}^{k}[a_i, b_i]$ be a loop in a given secondary structure S. We refer to the intervals $g_0(s) = [0, a_1]$, $g_i(s) = [b_i, a_{i+1}]$ for $1 \le i \le k-1$ and $g_k(s) = [b_k, n+1]$, as the gaps of the loop s. We call $g_0(s)$ and $g_k(s)$ exterior gaps and the rest interior gaps.

Claim 2 now follows from

Lemma 6. *Let $R = (S, T)$ be a bi-secondary structure with loop-nerve, $K(R)$, and let $t \in T$ be a loop, then, there exists a Δ_t-graph, $G(t)$.*

Proof. Let $S(t)$ and $T(t)$ be the S and T neighbors of t respectively. We prove the lemma by induction on N, the number of non-rainbow arcs in S. To this end, let us first consider the induction base case $N = 0$.
As there are no arcs other than the rainbow, α_r, we have $S(t) = \{r\}$. By construction, $b(r) \in g_0(t)$ and $e(r) \in g_k(t)$, the exterior t-gaps. We make the Ansatz

$$G(t) = \mathrm{Star}(r) = (R(t), \{\{r, t'\} | t' \in T(t)\}).$$

By construction, $\mathrm{Star}(r)$ is a connected spanning sub-graph of $\mathrm{Gr}(t)$. Furthermore, $\forall t' \in T(t)$ we have $b(r) < b(t) < b(t') < e(t') < e(t) < e(r)$. Hence

$$r \cap t \cap t' = \{b(t'), e(t')\} \neq \varnothing$$

and as a result $\{r, t, t'\} \in K_2(R(t))$. Thus, any edge $\{r, t'\} \in E(r) = \{\{r, t'\} | t' \in T(t)\}$ is a Δ_t-edge and $\mathrm{Star}(r)$ is a Δ_t-graph, establishing the induction basis, see Figure 4.

Let next $\overline{S}$ denote a secondary structure, having $N-1 \ge 0$ arcs. By induction hypothesis, for any such bi-secondary structure, $\overline{R} = (\overline{S}, T)$ and $t \in T$, a Δ_t-graph exists. We will denote such a graph by $\overline{G}(t)$.

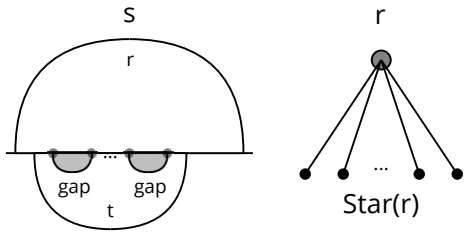


Figure 4: LHS: $S(t) = \{r\}$, RHS: $G(t) = Star(r)$.

We shall prove the existence of a Δ_t-graph as follows: first we identify and then remove a distinguished non-rainbow arc $x \in S$. This gives us the bi-secondary structure $\overline{R} = (\overline{S}, T)$, for which the induction hypothesis applies, i.e. a Δ_t-graph $\overline{G}(t)$ exists. We then reinsert the arc x and inspect how to obtain $G(t)$ from $\overline{G}(t)$.

Let $\mathrm{Exp}_t(S)$ be the set of non-rainbow S-arcs, x, having at least one t-exposed endpoint, i.e. either $b(x)$ or $e(x)$ are contained in t.

Case 1*:* $\mathrm{Exp}_t(S) \neq \varnothing$.

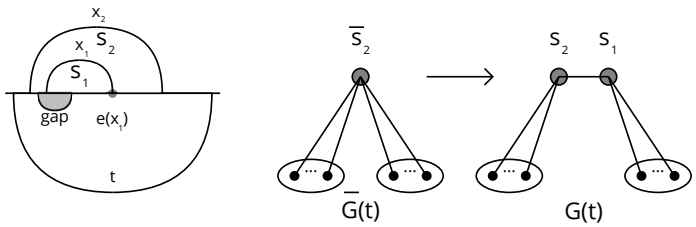


Figure 5: LHS: $x_1 \in \mathrm{Exp}_t(S)$. RHS: the effect of reintroducing x_1, passing from $\overline{G}(t)$ to $G(t)$.

Select $x_1 \in \mathrm{Exp}_t(S)$. Let $s_1 \in S(t)$ be the loop such that $x_1 = \alpha_{s_1}$ and let x_2 be the arc, that directly covers x_1 w.r.t. $\prec_S$. Let $s_2 \in S$ be the loop such that $\alpha_{s_2} = x_2$. W.l.o.g. we may assume that $e(x_1) \in t$. Clearly, $s_2 \in S(t)$ since $e(s_1) \in s_2 \cap t$, see Figure 5.

x_1-removal produces the secondary structure $\overline{S}$ and $\overline{R} = (\overline{S}, T)$, for which the induction hypothesis applies. Let $\overline{s_2} \in \overline{S}$ be such that $\alpha_{\overline{s_2}} = x_2$. Then $\overline{s_2} \in \overline{S}(t)$ since, in absence of x_1, $e(x_1) \in \overline{s_2} \cap t$. Hence $\overline{s_2}$ is a vertex in $\overline{G}(t) = (\overline{R}(t), \overline{E})$.

Reinserting x_1 into $\overline{R}$ splits $\overline{s_2}$ into the two S-loops s_1 and s_2, see Figure 5. We make the Ansatz

$$G(t) = ((\overline{R}(t) \setminus \{\overline{s_2}\}) \cup \{s_1, s_2\}, E),$$

where

$$\begin{aligned} E \quad = \quad & (\overline{E} \setminus \{\{\overline{s_2}, r'\} \mid r' \in \overline{R}(t)\}) \cup \{\{s_1, s_2\}\} \cup \\ & \{\{s_1, r'\} \mid r' \in R(t) \setminus \{s_1, s_2\},\ s_1 \cap r' \cap t \neq \varnothing\} \cup \\ & \{\{s_2, r'\} \mid r' \in R(t) \setminus \{s_1, s_2\},\ s_2 \cap r' \cap t \neq \varnothing\}. \end{aligned}$$

Since $\overline{s_2} = s_1 \cup s_2$ as sets, and $\overline{R}(t) \setminus \{\overline{s_2}\} = R(t) \setminus \{s_1, s_2\}$, we have

$$\{r' \in \overline{R}(t) \mid \{r', \overline{s_2}\} \in \overline{E}\} = \{r' \in \overline{R}(t) \setminus \{\overline{s_2}\} \mid \{r', s_1\} \in E \text{ or } \{r', s_2\} \in E\}.$$

Accordingly, any $\overline{R}(t)$-vertex connected in $\overline{G}(t)$ to $\overline{s_2}$ is, when considered in $R(t)$, connected to either s_1 or s_2. In view of $\{e(x_1)\} \subset (s_1 \cap\ s_2 \cap t)$, we can conclude that s_1 and s_2 are connected by a Δ_t-edge. This guarantees that $G(t)$ is a connected spanning sub-graph of $\mathrm{Gr}(t)$.

Case 2: $\mathrm{Exp}_t(S) = \varnothing$.

Having no arcs with exposed endpoints, for any loop $s \in S$, there exist t-gaps, containing $b(s)$ and $e(s)$. Suppose first, there exists an arc x having both endpoints in the same gap, see Figure 6. The associated loop, s, having $\alpha_s = x$, is not contained in $S(t)$. Upon inspection

$$\overline{S}(t) = S(t) \text{ and } G(t) = \overline{G}(t),$$

Hence the induction hypothesis directly implies the existence of $G(t)$.

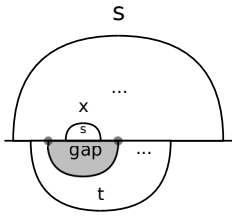


Figure 6: The case where $b(x)$ and $e(x)$ are contained in the same t-gap.

It thus remains to discuss S-arcs, whose endpoints belong to distinct t-gaps, see Figure 7. We shall distinguish the following two scenarios:

(a) $(S \setminus S(t)) \setminus \{r\} \neq \varnothing$*, where* α_r *is the rainbow arc.*

We shall show that the removal of an arc $x = \alpha_s, s \in (S \setminus S(t)) \setminus \{r\}$, will not affect $G(t)$, aside from relabeling of a single vertex. Since $s_1 \notin S(t)$ we have $\overline{s_2} \notin \overline{S}(t)$ if and only if $s_2 \notin S(t)$. In this case we set $G(t) = \overline{G}(t)$ and the assertion is directly implied by the induction hypothesis. In case of $\overline{s_2} \in \overline{S}(t)$, $G(t)$ is obtained from $\overline{G}(t)$ by relabeling $\overline{s_2}$ to s_2 exhibiting no other changes, see Figure 7:

$$G(t) = ((\overline{R}(t) \setminus \{\overline{s_2}\}) \cup \{s_2\}, E),$$

where

$$E = (\overline{E} \setminus \{\{\overline{s_2}, r'\} \in \overline{E} \mid r' \in \overline{R}(t)\}) \cup \{\{s_2, r'\} \mid r' \in \overline{R}(t), \{\overline{s_2}, r'\} \in \overline{E}\}$$

and $G(t)$ is consequently a Δ_t-graph.

(b) $(S \setminus S(t)) \setminus \{r\} = \varnothing$*, where* α_r *is the rainbow arc.*

We then have either $S \setminus S(t) = \varnothing$ or $S \setminus S(t) = \{r\}$. In the latter case we select x to be an arc, corresponding to a loop s_1, that is immediately covered by α_r. Let $s_2 = r$. Since $r \notin S(t)$ we make the Ansatz

$$G(t) = ((\overline{R}(t) \setminus \{\overline{s_2}\}) \cup \{s_1\}, E),$$

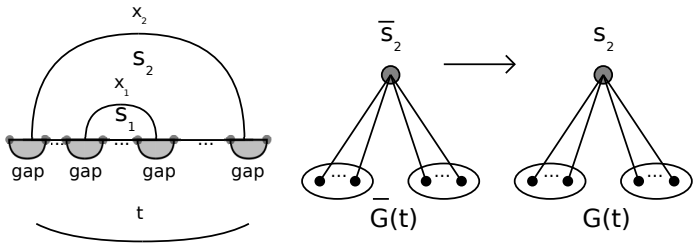


Figure 7: LHS: all S-arcs having their endpoints in distinct t-gaps. RHS: (a), $\overline{s_2} \in \overline{S}(t)$, $G(t)$ is obtained by a relabelling of $\overline{G}(t)$.

where

$$E = (\overline{E} \setminus \{\{\overline{s_2}, r'\} \in \overline{E} \mid r' \in \overline{R}(t)\}) \cup \{\{s_1, r'\} \mid r' \in \overline{R}(t), \{\overline{s_2}, r'\} \in \overline{E}\}.$$

Accordingly, $G(t)$ is obtained from $\overline{G}(t)$ by relabeling $\overline{s_2}$ by s_1 and $G(t)$ is a Δ_t-graph, see Figure 8.

It remains to analyze $S \setminus S(t) = \varnothing$, i.e. all S-arcs are contained in $S(t)$, where we recall we reduced the analysis to arcs whose endpoints belong to different t-gaps.

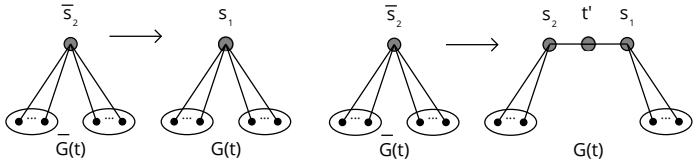


Figure 8: LHS: the case $S \setminus S(t) = \{r\}$. RHS: the case $S \setminus S(t) = \varnothing$.

Suppose now all S-loops are contained in $S(t)$. Consider the set of all minimal arcs of S w.r.t. $\prec_S$. We claim there exists one such minimal arc, call it α_{s_1}, such that its immediate cover w.r.t. $\prec_S$, call it α_{s_2}, is such that s_2 contains at least one of the endpoints of one of the t-gaps that contain one of the endpoints of α_{s_1}. To show this we observe that if all t-gaps would have their endpoints inside loops corresponding to $\prec_S$-minimal arcs, then at least one arc that immediately covers such minimal arcs would not correspond to a loop in $S(t)$. Hence, there must be a loop s_1 with α_{s_1} minimal w.r.t. $\prec_S$ and an arc α_{s_2} that immediately covers α_{s_1}, such that s_2 contains one of the endpoints of a gap that contains $b(s_1)$ or $e(s_1)$.

Let us denote this gap by h. W.l.o.g. we can assume $e(s_1) \in h$, see Figure 9. Then, the minimality of s_1 guarantees that s_1 contains the other endpoint of the gap h. We shall now remove $x_1 = \alpha_{s_1}$.

We consider the loop t' associated to h and note that $s_1 \cap t' \cap t \neq \varnothing$ as well as $s_2 \cap t' \cap t \neq \varnothing$. Accordingly, t' connects s_1, s_2 in $R(t)$ by means of Δ_t-edges, see Figure 8 and we immediately obtain that

$$G(t) = ((\overline{R}(t) \setminus \{\overline{s_2}\}) \cup \{s_1, s_2\}, E),$$

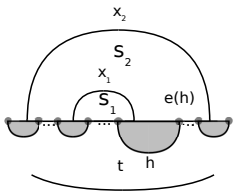


Figure 9: $x_1 = \alpha_{s_1}$ is minimal w.r.t. $\prec_S$.

where

$$\begin{aligned} E \quad = \quad & (\overline{E} \setminus \{\{\overline{s_2}, r'\} \in \overline{E} \mid r' \in \overline{R}(t)\}) \cup \\ & \{\{s_1, r'\} \mid r' \in R(t), s_1 \cap r' \cap t \neq \varnothing\} \cup \\ & \{\{s_2, r'\} \mid r' \in R(t), s_2 \cap r' \cap t \neq \varnothing\}, \end{aligned}$$

is a Δ_t-graph for $R(t)$. This concludes the proof of the induction step and the lemma follows. □

Next we compute $H_2(R)$,

Theorem 3. *For any bi-secondary structure, $R = (S, T)$, with loop nerve $K(R)$, we have*

$$H_2(R) \cong \bigoplus_{i=1}^{k} \mathbb{Z},$$

i.e. $H_2(R)$ is free of finite rank.

Proof. Claim 1. $\mathrm{Im}(\partial_3) \cong C_3(R)$, *i.e.* $\mathrm{Im}(\partial_3)$ *is a free Abelian group and freely generated by* $P = \{\partial_3(Y_i) \mid Y_i \text{ is a 3-simplex}\}$.

Claim 1 is a consequence of two facts: (a) $C_4(R) = 0$ and (b) $H_3(R) = 0$, both of which we prove below. It is obtained as follows: $C_4(R) = 0$ guarantees $\mathrm{Im}(\partial_4) = 0$, which in view of $0 = H_3(R) = \mathrm{Ker}(\partial_3)/\mathrm{Im}(\partial_4)$ implies $\mathrm{Ker}(\partial_3) = 0$. This in turn implies that ∂_3 is an embedding, i.e. $\mathrm{Im}(\partial_3) \cong C_3(R)$, whence $\mathrm{Im}(\partial_3)$ is a free Abelian group. P certainly generates $\mathrm{Im}(\partial_3)$ and a $\mathbb{Z}$-linear combination

$$\sum_j \lambda_j \partial_3(Y_j) = \partial_3(\sum_j \lambda_j Y_j) = 0$$

means that $\sum_j \lambda_j Y_j \in \mathrm{Ker}(\partial_3)$. Since the latter is trivial we arrive at

$$\sum_j \lambda_j Y_j = 0$$

which implies $\lambda_j = 0$, for any j appearing in this sum. This shows that the P-elements are $\mathbb{Z}$-linear independent.

Let $Y_i \in K_3(R)$, $0 \leq i \leq k$ denote the generators of $C_3(R)$. Lemma 4 guarantees that each 3-simplex Y has at least two Y-exposed 2-faces. Hence, to

each generator $Y_i \in K_3(R)$ there correspond at least two generators of the free group $C_2(R)$ that appear as terms only in the image $\partial_3(Y_i)$. Let us write

$$\partial_3(Y_i) = \sum_{u(i)} U_{u(i)} + \sum_{c(i)} C_{c(i)},$$

distinguishing exposed, signed and covered, signed 2-faces of Y_i, i.e. we consider the sign, induced by the boundary map, to be part of $U_{u(i)}$ and $C_{c(i)}$, respectively. In particular, for any $u(i)$, there exists an unique r, such that we have either $U_{u(i)} = +Z_r$ or $U_{u(i)} = -Z_r$ where Z_r is a generator of $C_2(R)$.

Claim 2. *$C_2(R)/\mathrm{Im}(\partial_3)$ is free.*

We consider $\overline{C_2(R)} = C_2(R)/\mathrm{Im}(\partial_3)$ as a $\mathbb{Z}$-module and suppose X is a torsion element of order n in $\overline{C_2(R)}$. Then we can represent X as

$$X = \sum_r \lambda_r Z_r + \mathrm{Im}(\partial_3),$$

where, w.l.o.g. we assume that all $\lambda_r \neq 0$. Since X is a torsion element, we have $nX = 0$ in $\overline{C_2(R)}$, i.e.

$$\begin{aligned} n(\sum_r \lambda_r Z_r) &= \sum_{i=1}^{k} \alpha_i \partial_3(Y_i) \\ &= \sum_{i=1} \alpha_i (\sum_{u(i)} U_{u(i)}) + \sum_{i=1} \alpha_i (\sum_{c(i)} C_{c(i)}) \end{aligned}$$

where $\lambda_r, \alpha_i \in \mathbb{Z}$ are unique nonzero integer coefficients. Clearly, each unique signed 2-face, $U_{u(i)}$ of the RHS corresponds to a unique generator $Z_{r(U_{u(i)})}$ and hence, irrespective of the particular choice of the $u(i)$ and the sign of $U_{u(i)}$, we obtain for any i of the sum on the RHS

$$n\lambda_{r(U_{u(i)})} = \alpha_i,$$

only depending on the index i. This is an equation in $\mathbb{Z}$ and hence implies $\alpha_i \equiv 0 \mod n$. Accordingly, we derive

$$(\sum_r \lambda_r Z_r) = \sum_{i=1}^{k} \frac{\alpha_i}{n} \partial_3(Y_i),$$

which means $X \in \mathrm{Im}(\partial_3)$, since $\frac{\alpha_i}{n} \in \mathbb{Z}$, i.e. $X = 0$. By transposition we have proved that $X \neq 0$ in $\overline{C_2(R)}$ implies for any $n \in \mathbb{N}$, $nX \neq 0$ in $\overline{C_2(R)}$, whence $\overline{C_2(R)}$ is free and Claim 2 is proved.

As a result, $\mathrm{Ker}(\partial_2)/\mathrm{Im}(\partial_3)$ is, as a subgroup of the free group $\overline{C_2(R)}$, itself free and the theorem is proved. □

It remains to show $C_d(R) = 0$ for $d \geq 4$ and $H_3(R) = 0$.

Lemma 7. *Let $R = (S, T)$ be a bi-secondary structure with loop-nerve $K(R)$ and let $d \geq 4$, then $K_d(R) = \varnothing$, $C_d(R) = 0$ and $H_d(R) = 0$.*

Proof. For any $Y = \{r_{i_0}, \ldots, r_{i_d}\} \in K_d(R)$ for some $d \geq 4$ we have $\Omega(Y) = \bigcap_{k=0}^{d} r_{i_k} \neq \varnothing$. Since $d \geq 4$, $|Y| \geq 5$, whence at least three loops $r'_0, r'_1, r'_2 \in Y$ are contained in the same secondary structure, which is a contradiction to Proposition 1, which stipulates that three loops of one secondary structure intersect only trivially. □

Next we show that $H_3(R) = 0$.

Theorem 4. *Let $R = (S, T)$ be a bi-secondary structure with loop nerve $K(R)$, then $H_3(R) = 0$.*

Proof. Consider

$$X = \sum_{Y_i \in K_3(R)} n_i Y_i \in C_3(R) \in \mathrm{Ker}(\partial_3),$$

then

$$\begin{aligned} \partial_3(X) &= \sum_{Y_i \in K_3(R)} n_i \partial_3(Y_i) \\ &= \sum_i n_i (\sum_{u(i)} U_{u(i)}) + \sum_i n_i (\sum_{c(i)} C_{c(i)}), \end{aligned}$$

where the $U_{u(i)}$ and $C_{c(i)}$ are the signed exposed and covered 2-faces of Y_i, respectively. Since we have at least two unique exposed 2-faces and, by assumption, $\partial_3(X) = 0$, we conclude that for any i of $X = \sum_{Y_i \in K_3(R)} n_i Y_i$ we have $n_i = 0$. Therefore $X = 0$ and $\mathrm{Ker}(\partial_3)$ contains no nontrivial elements, whence $H_3(R) = 0$. □

4. Discussion

In the previous section, we showed that $H_2(R)$ is non-trivial and free, leading to a novel observable for the pair of secondary structures (S, T), namely the rank of $H_2(R = (S, T))$, $r(H_2(R))$. We shall see that the generators of $r(H_2(R))$ represent key information about the "switching sequence" [42], a segment of the sequence, that engages w.r.t. each respective structure in a distinct, mutually exclusive fashion.

It is well known from experimental work that native riboswitch pairs, ncRNAs, exhibit two distinct, mutually exclusive, stable secondary structures [43]. We analyzed all nine riboswitch sequences contained in the Swispot database [44] and observed that $r(H_2(R)) = 1$. In Figure 10 we illustrate the connection between a $H_2(R)$-generator and pairs of mutually exclusive substructures. The ranks, $r(H_2(R))$, for uniformly sampled structures pairs, are displayed in Figure 11, showing that 6.7% of the uniform random pairs exhibit $r(H_2(R)) = 1$.

As for future work, the complexity analysis and optimal scheduling problems arising from the work of Huang [41] suggest to consider a graded version of the homologies, developed here. Let $t \geq 1$ be an integer and $R = (S, T)$ be a

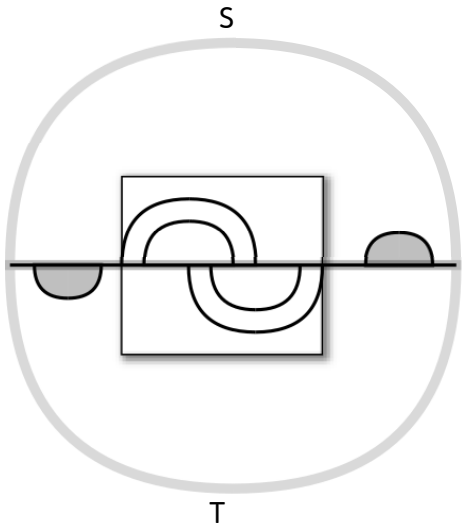


Figure 10: $H_2(R)$-generators and mutually exclusive structure pairs: the two helices (boxed) are mutually exclusive, while the two substructures (shaded) are not. The former two, together with the two rainbows correspond to a generator of $H_2(R)$.

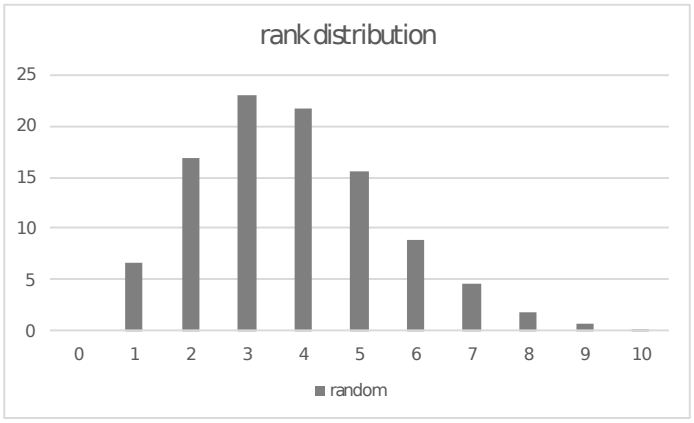


Figure 11: $r(H_2(R))$ for uniformly random structure pairs: $r(H_2(R))$ (x-axis) and the relative frequencies (y-axis).

bi-secondary structure. We set $K_d^t(R) = \{Y \in K_d(R) | \omega(Y) \geq t\}$, the set of d-simplices of weight at least t. We define $C_d^t(R)$ to be the free abelian group generated by $K_d^t(R)$, i.e. the chain group of rank d and weight at least t. It is easy to see then that $C_d(R) = C_d^1(R)$ for all $d \geq 0$ and that $C_d^{t+1}(R) \leq C_d^t(R)$ for all $t \geq 1$ and all $d \geq 0$. We can naturally define boundary operators for these groups in terms of restrictions of our original boundary maps as $\partial_d^t : C_d^t(R) \to C_{d-1}^t(R)$ with $\partial_d^t = \partial_d|_{C_d^t(R)}$. As such, we obtain a t-parametric sequence of nerves $\{K^t(R)\}_{t\geq 1}$ each of which gives rise to its t-labelled homology sequence. Tracking the persistence of homology group generators across the newly obtained homological t-spectrum gives rise to a more granular analysis of the structure of the complete nerve [45]. This analysis represents a version of persistent homology, pioneered by Edelsbrunner and by Gunnar Carlson [46, 47, 48] and is of central importance for designing an optimal loop-removal schedule in [41].

We extend the homology analysis to planar interaction structures [49]. Due to the fact that the physical $5' - 3'$ distance for RNA strands is in general very small [50], the formation of an interaction structure is connected to the alignment of two discs, each representing the respective, circular backbone. That is, interpreting the two circles corresponding to two interacting secondary structures S_1, S_2 to be $\partial(D(0,1))$, the boundaries of unit disks in $\mathbb{C}$. These boundaries contain distinguished points that correspond to the paired vertices in the secondary structures. The connection between interaction structure and disc-alignment leads one to consider one disc being acted upon by Möbius transforms. This action is well defined since the Möbius maps of the disc map the boundary to itself and, being holomorphic, cannot introduce crossings. Different alignments are then captured by these automorphisms and give rise to a spectrum of homologies, as introduced in this paper.

5. Declarations of interest

None.

6. Acknowledgments

We gratefully acknowledge the comments from Fenix Huang. Many thanks to Thomas Li, Ricky Chen and Reza Rezazadegan for discussions.

References

[1] J. E. Darnell, RNA: life's indispensable molecule, Cold Spring Harbor Laboratory Press New York, 2011.

[2] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, A. Zamir, Structure of a ribonucleic acid, Science (1965) 1462–1465.

[3] D. Thirumalai, N. Lee, S. A. Woodson, D. Klimov, Early events in rna folding, Annual review of physical chemistry 52 (1) (2001) 751–762.

[4] J. R. Fresco, B. M. Alberts, P. Doty, et al., Some molecular details of the secondary structure of ribonucleic acid., Nature 188 (1960) 98–101.

[5] J. Gralla, D. M. Crothers, Free energy of imperfect nucleic acid helices: Ii. small hairpin loops, Journal of molecular biology 73 (4) (1973) 497–511.

[6] D. H. Turner, D. H. Mathews, Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure, Nucleic acids research 38 (suppl_1) (2009) D280–D282.

[7] J. M. Pipas, J. E. McMAHON, Method for predicting rna secondary structure, Proceedings of the National Academy of Sciences 72 (6) (1975) 2017–2021.

[8] C. Delisi, D. M. Crothers, Prediction of rna secondary structure, Proceedings of the National Academy of Sciences 68 (11) (1971) 2682–2685.

[9] I. Tinoco, O. C. Uhlenbeck, M. D. Levine, Estimation of secondary structure in ribonucleic acids, Nature 230 (5293) (1971) 362.

[10] M. S. Waterman, Secondary structure of single-stranded nucleic acids, Adv. math. suppl. studies 1 (1978) 167–212.

[11] R. Nussinov, A. B. Jacobson, Fast algorithm for predicting the secondary structure of single-stranded rna, Proceedings of the National Academy of Sciences 77 (11) (1980) 6309–6313.

[12] M. S. Waterman, T. F. Smith, Rapid dynamic programming algorithms for rna secondary structure, Advances in Applied Mathematics 7 (4) (1986) 455–464.

[13] M. Zuker, D. Sankoff, Rna secondary structures and their prediction, Bulletin of mathematical biology 46 (4) (1984) 591–621.

[14] J. S. McCaskill, The equilibrium partition function and base pair binding probabilities for rna secondary structure, Biopolymers: Original Research on Biomolecules 29 (6-7) (1990) 1105–1119.

[15] C. Haslinger, P. F. Stadler, Rna structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties, Bulletin of mathematical biology 61 (3) (1999) 437–467.

[16] M. Taufer, A. Licon, R. Araiza, D. Mireles, F. Van Batenburg, A. P. Gultyaev, M.-Y. Leung, Pseudobase++: an extension of pseudobase for easy searching, formatting and visualization of pseudoknots, Nucleic acids research 37 (suppl_1) (2008) D127–D135.

[17] C. M. Reidys, Random induced subgraphs of generalizedn-cubes, Advances in Applied Mathematics 19 (3) (1997) 360–377.

[18] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, M. Zehl, Design of multistable rna molecules, Rna 7 (2) (2001) 254–265.

[19] A. Busch, R. Backofen, Info-rna—a fast approach to inverse rna folding, Bioinformatics 22 (15) (2006) 1823–1831.

[20] A. Levin, M. Lis, Y. Ponty, C. W. O'donnell, S. Devadas, B. Berger, J. Waldispühl, A global sampling approach to designing and reengineering rna secondary structures, Nucleic acids research 40 (20) (2012) 10041–10052.

[21] C. Barrett, Q. He, F. W. Huang, C. M. Reidys, An efficient dual sampling algorithm with hamming distance filtration, Journal of Computational Biology 25 (11) (2018) 1179–1192.

[22] C. Barrett, F. W. Huang, C. M. Reidys, Sequence–structure relations of biopolymers, Bioinformatics 33 (3) (2017) 382–389.

[23] Q. He, F. W. Huang, C. Barrett, C. M. Reidys, Genetic robustness of let-7 mirna sequence-structure pairs, arXiv preprint arXiv:1801.05056.

[24] W. R. Schmitt, M. S. Waterman, Linear trees and rna secondary structure, Discrete Applied Mathematics 51 (3) (1994) 317–323.

[25] I. L. Hofacker, P. Schuster, P. F. Stadler, Combinatorics of rna secondary structures, Discrete Applied Mathematics 88 (1-3) (1998) 207–237.

[26] E. Y. Jin, J. Qin, C. M. Reidys, Combinatorics of rna structures with pseudoknots, Bulletin of mathematical biology 70 (1) (2008) 45–67.

[27] H. Orland, A. Zee, Rna folding and large n matrix theory, Nuclear Physics B 620 (3) (2002) 456–476.

[28] J. E. Andersen, L. O. Chekhov, R. Penner, C. M. Reidys, P. Sułkowski, Topological recursion for chord diagrams, rna complexes, and cells in moduli spaces, Nuclear Physics B 866 (3) (2013) 414–443.

[29] M. Bon, G. Vernizzi, H. Orland, A. Zee, Topological classification of rna structures, Journal of molecular biology 379 (4) (2008) 900–911.

[30] J. E. Andersen, R. C. Penner, C. M. Reidys, M. S. Waterman, Topological classification and enumeration of rna structures by genus, Journal of mathematical biology 67 (5) (2013) 1261–1278.

[31] F. W. Huang, C. M. Reidys, Shapes of topological rna structures, Mathematical biosciences 270 (2015) 57–65.

[32] W. Chen, E. Deng, R. Du, R. Stanley, C. Yan, Crossings and nestings of matchings and partitions, Transactions of the American Mathematical Society 359 (4) (2007) 1555–1575.

[33] R. P. Stanley, Enumerative combinatorics, wadsworth publ, Co., Belmont, CA.

[34] S. Sundaram, The cauchy identity for sp (2n).

[35] R. Penner, M. S. Waterman, Spaces of rna secondary structures, Advances in Mathematics 101 (1) (1993) 31–49.

[36] J. Harer, D. Zagier, The euler characteristic of the moduli space of curves, Inventiones mathematicae 85 (3) (1986) 457–485.

[37] J. Ambjørn, L. Chekhov, C. F. Kristjansen, Y. Makeenko, Matrix model calculations beyond the spherical limit, Nuclear Physics B 404 (1-2) (1993) 127–172.

[38] G. Chapuy, A new combinatorial identity for unicellular maps, via a direct bijective approach, Advances in Applied Mathematics 47 (4) (2011) 874–893.

[39] F. W. Huang, C. M. Reidys, Topological language for rna, Mathematical biosciences 282 (2016) 109–120.

[40] C. M. Reidys, Neutral networks of rna secondary structures.

[41] F. W. Huang, Personal Communication.

[42] R. R. Breaker, Riboswitches and the rna world, Cold Spring Harbor perspectives in biology 4 (2) (2012) a003566.

[43] A. Serganov, E. Nudler, A decade of riboswitches, Cell 152 (1-2) (2013) 17–24.

[44] M. Barsacchi, E. M. Novoa, M. Kellis, A. Bechini, Swispot: modeling riboswitches by spotting out switching sequences, Bioinformatics 32 (21) (2016) 3252–3259.

[45] A. Zomorodian, G. Carlsson, Computing persistent homology, Discrete & Computational Geometry 33 (2) (2005) 249–274.

[46] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, Discrete & Computational Geometry 28 (4) (2002) 511–533.

[47] G. Carlsson, Topology and data, Bull. Amer. Math. Soc. 46 (2009) 255–308.

[48] G. Carlsson, A. Zomorodian, A. COllins, L. J. Guibas, Persistence barcodes for shapes, International Journal of Shape Modeling 11 (02) (2005) 149–187.

[49] J. E. Andersen, F. W. Huang, R. C. Penner, C. M. Reidys, Topology of rna-rna interaction structures, Journal of Computational Biology 19 (7) (2012) 928–943.

[50] A. M. Yoffe, P. Prinsen, W. M. Gelbart, A. Ben-Shaul, The ends of a large rna molecule are necessarily close, Nucleic acids research 39 (1) (2010) 292–299.

Part Three: Loop Homology of Bi-secondary Structures II

Loop Homology of Bi-secondary Structures II

Andrei C. Bura[a,b], Qijun He[c,*], Christian M. Reidys[c,d]

[a] *Department of Mathematics, Virginia Tech, 225 Stanger Street, Blacksburg, VA 24061-1026*
[b] *Biocomplexity Institute of Virginia Tech, 1015 Life Sciences Circle Blacksburg, VA 24061*
[c] *Biocomplexity Institute and Initiative, University of Virginia, 995 Research Park Boulevard, Charlottesville, VA 22911*
[d] *Department of Mathematics, University of Virginia, 141 Cabell Dr, Charlottesville, VA 22903*

Abstract

In this paper we further describe the features of the topological space $K(R)$ obtained from the loop nerve of R, for $R = (S, T)$ a bi-secondary structure. We will first identify certain distinct combinatorial structures in the arc diagram of R which we will call crossing components. The main theorem of this paper shows that the total number of these crossing components equals the rank of $H_2(R)$, the second homology group of the loop nerve.

Keywords: RNA, bi-secondary structure, loop, nerve, simplicial homology.

1. Introduction

In Part Two: Loop Homology of Bi-secodary Structures, we proved that $H_2(R)$ is free abelian. However, we've yet to identify the combinatorial object within the diagram of the bi-structure R that contributes a generator to $H_2(R)$. In the the following, we will identify the precise sub-structures of a given bi-secondary structure R, that when considered within the loop nerve $K(R)$, correspond to sub-complexes that triangulate 2-spheres. These sub-structures we will call crossing components (CCs). We will show that there is a bijective correspondence between any minimal generating set of $H_2(R)$ and the set of CCs of R and thus, the number of CCs equals the rank of $H_2(R)$.

2. Secondary and Bi-Secondary Structures

Definition 1. *An RNA diagram S over $[n]$, is a vertex-labeled graph whose vertices are drawn on the horizontal axis and labeled by $[n] = \{1, \ldots, n\}$. An* arc *$\mu = (i, j), i < j$, is an ordered pair of vertices, which represents the base*

*Corresponding author
Email addresses: anbur12@vt.edu (Andrei C. Bura), qhe196@gmail.com (Qijun He), duck@santafe.edu (Christian M. Reidys)

pairing between the i-th and j-th nucleotides in the RNA structure. We denote by $b(\mu) = i$ and $e(\mu) = j$ the start and endpoints of an arc $\mu \in S$. Furthermore, each vertex can be paired with at most one other vertex, and the arc that connects them is drawn in the upper half-plane. We introduce two "formal" vertices associated with positions 0 and $n+1$, respectively, closing any diagram by the arc $(0, n+1)$, called the rainbow. The set $[0, n+1]$ is called the diagram's backbone.

Definition 2. *Let S be an RNA diagram over $[n]$. Two arcs (i,j) and (p,q) are called* crossing *if and only if $i < p < j < q$. S is called a* secondary structure *if it does not contain any crossing arcs. The arcs of S can be endowed with a partial order as follows: $(k,l) \prec_S (i,j) \iff i < k < l < j$. We denote this by $(S, \prec_S)$ and call it the arc poset of S. Finally, an* interval $[i,j]$ *on the backbone is the set of vertices $\{i, i+1, \ldots, j-1, j\}$.*

Definition 3. *Let S be a secondary structure over $[n]$. A* loop *s in S is a subset of vertices, represented as a disjoint union of a sequence of intervals on the backbone of S, $s = \dot{\bigcup}_{i=1}^{k}[a_i, b_i]$, such that (a_1, b_k) and (b_i, a_{i+1}), for $1 \leq i \leq k-1$, are arcs and such that any other interval-vertices are unpaired. Let α_s denote the unique, maximal arc (a_1, b_k) of the loop s.*

In this paper we shall identify a secondary structure with its set of loops.

Remark 1. *Let S be a secondary structure over $[n]$ and $s = \dot{\bigcup}_{i=1}^{k}[a_i, b_i]$ a loop in S, then*

- *Each unpaired vertex is contained in exactly one loop.*
- *The arc (a_1, b_k) is maximal w.r.t. $\prec_S$ among all arcs contained in s, i.e. there is a bijection between arcs and loops, mapping each loop to its maximal arc.*
- *The Hasse diagram of the S arc-poset is a rooted tree $Tr(S)$, having the rainbow arc as the root.*
- *Each non-rainbow arc appears in exactly two loops.*

Let $X = \{x_0, x_1, \ldots, x_m\}$ be a collection of finite sets. We call $Y = \{x_{i_0}, \ldots, x_{i_d}\} \subseteq X$ a *d-simplex* of X iff $\bigcap_{k=0}^{d} x_{i_k} \neq \varnothing$. We set $\Omega(Y) = \bigcap_{k=0}^{d} x_{i_k}$ and denote by $\omega(Y) = |\Omega(Y)| \neq 0$. Let $K_d(X)$ be the set of all d-simplices of X, then the *nerve* of X is

$$K(X) = \dot{\bigcup}_{d=0}^{\infty} K_d(X) \subseteq 2^X.$$

A d'-simplex $Y' \in K(X)$ is called a d'-face of Y if $d' < d$ and $Y' \subseteq Y$. By construction, $K(X)$ is an abstract simplicial complex.

Let S be a secondary structure over $[n]$. The geometric realization of $K(S)$, the nerve over the set of loops of S, is a tree.

Definition 4. *Given two secondary structures S and T over $[n]$, we refer to the pair $R = (S, T)$ as a bi-secondary structure. Let $S \cup T$ be the loop set of R and $K(R) = \dot{\bigcup}_{d=0}^{\infty} K_d(R)$ be its nerve of loops.*

We represent the arc diagram of a bi-secondary structure $R = (S, T)$ with the arcs of S in the upper half plane while the arcs of T reside in the lower half plane.

Let $R = (S, T)$ be a bi-secondary structure with loop nerve $K(R)$. A 1-simplex $Y = \{r_{i_0}, r_{i_1}\} \in K_1(R)$ is called *pure* if r_{i_0} and r_{i_1} are loops in the same secondary structure and *mixed*, otherwise. Any 2-simplex $Y \in K_2(R)$ had exactly one pure edge and two mixed edges as its 1-faces (See Part Two: Loop Homology of Bi-secodary Structures).

Definition 5. *Given $R = (S, T)$, a bi-secondary structure on $[n]$, R is called a* simple *bi-secondary structure, if any nucleotide $q \in \{1, \ldots, n\}$ has degree at most three in the arc diagram of R.*

3. Decorations and Closures

Definition 6. *Let $R = (S, T)$ be a simple bi-secondary structure on $[n]$. Let $q \in \{1, \ldots, n\}$ be a nucleotide of degree exactly three in the diagram of R. Then, by definition of $K(R)$, q will correspond to exactly one 2-simplex $Y \in K_2(R)$, namely, the 2-simplex (triangle) for which $q \in \Omega(Y)$ holds. We denote the pair (q, Y) by Y_q and call it a* decoration *of the diagram of R at q. We denote by $K_2(R)^*$ the set of all possible decorations of R.*

Remark 2. *Since $1 \leq \omega(Y) \leq 2$ for any $Y \in K_2(R)$ there exist at most two, and at least one decoration $(\star, Y)$ in $K_2(R)^*$.*

Definition 7. *Let $Y = [x, y, z] \in K_2(R)$ for R a simple bi-secondary structure. W.l.o.g. we may assume that $[x, y]$ is the pure edge of Y. Note then that $x \leq y \leq z$ in terms of the simplicial ordering on $K(R)$(See Part Two: Loop Homology of Bi-secodary Structures). Then, for any decoration $Y_q \in K_2(R)^*$ we have $q = b(\alpha_x)$ or $q = e(\alpha_x)$. Hence, to each $Y \in K_2(R)$ and any decoration $Y_q \in K_2(R)$ there corresponds a unique arc $\gamma(Y) = \alpha_x$ such that either $b(\gamma(Y)) = q$ or $e(\gamma(Y)) = q$. We call this arc the* pure arc *of Y. When convenient we refer to this arc as the pure arc of a decoration, provided that said decoration is of the form $Y_\star = (\star, Y)$. By faces (i.e. edges or vertices) of a decoration $Y_\star = (\star, Y)$ we mean the respective faces of Y (See Figure 1).*

Definition 8. *Let $R = (S, T)$ be a bi-secondary structure with loop set $R = S \cup T$. We define the arc* line graph *of R to be $G(R) = (R, E)$ where*

$$E \ni e = (s \in S, t \in T) \Leftrightarrow$$

$$b(\alpha_s) < b(\alpha_t) < e(\alpha_s) < e(\alpha_t) \text{ or } b(\alpha_t) < b(\alpha_s) < e(\alpha_t) < e(\alpha_s)$$

i.e. the arc α_s intersects the arc α_t if we were to flip α_t to the upper half plane. In this case we say the two arcs α_s and α_t are crossing.

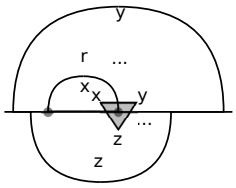


Figure 1: A decoration $Y_{e(r)} = [x, y, z]_{e(r)}$, its pure arc $\gamma(Y) = r = \alpha_x$. Its pure edge is $[x, y]$.

Definition 9. *We call the set of arcs X, associated to a connected component of the $G(R)$ arc line graph, an* irreducible component *(IC) of R. When the component X is non-trivial, i.e. $|X| > 1$, we call X a* crossing component *(CC) of R. We denote the set of all CCs of R by $\chi(R)$. When convenient, and when no possibility of confusion exists, we will also identify X with the set of loops whose unique maximal arcs are the elements of X.*

Remark 3. *By definition, an IC is either a non-crossing arc in R, or a CC in R and any two ICs are disjoint. Hence, the arc set of any bi-secondary structure R is paritioned by the ICs of R.*

Definition 10. *Let X be an IC of a simple bi-secondary structure $R = (S, T)$. We call*

$$C(X) = \{\ Y_\delta \in K_2(R)^* | \gamma(Y) \in X\}$$

the closure of X.

Remark 4. *Suppose X is a non-empty CC for a simple bi-secondary structure $R = (S, T)$. Then, by definition, there must exists at least two crossing arcs $\alpha_v, \alpha_w \in X$. Hence, there exist at least four decorations $Y_{b(\alpha_v)}$, $Y'_{e(\alpha_v)}$, $Y'''_{b(\alpha_w)}$, $Y''''_{b(\alpha_w)}$ in $C(X)$.*

Lemma 1. *Let X be a CC of a simple bi-secondary structure $R = (S, T)$. Then, for all $Y_p, Y'_q \in C(X)$ we have $Y = Y' \implies p = q$. I.e. the closure of a crossing component does not contain two decorations that have the same underlying 2-simplex.*

Proof. We will prove the contra-positive $p \neq q \implies Y \neq Y'$ for any pair of decorations $Y_p, Y'_q \in C(X)$. Suppose then by absurd that $Y' = Y = [x, y, z] \in K_2(R)$. W.l.o.g. we can assume that $[x, y]$ is the pure edge of Y. Since $p \neq q$, it must be then that $\{p, q\} = \{b(\gamma(Y)), e(\gamma(Y))\}]$. However since $Y' = Y$, we must have by the definition of decorations that $p \in z \ni q$. Hence $b(\gamma(Y)) \in z \ni e(\gamma(Y))$, i.e. both of the endpoints of the arc $\gamma(Y)$ belong to the same loop z. However, this means that $\gamma(Y)$ does not cross any other arc in R. A contradiction since $\gamma(Y) \in X$ and so must cross at least one other arc in R. So our assumption that $Y' = Y$ must be false. Thus if $p \neq q$, we must have $Y' \neq Y$, and so the lemma follows. □

Lemma 2. *Let X, X' be two distinct CCs of the simple bi-secondary structure $R = (S, T)$ i.e. $X \cap X' = \varnothing$. Then $C(X) \cap C(X') = \varnothing$.*

Proof. Suppose $Y_p \in C(X) \cap C(X')$. Then we must have $\gamma(Y) \in X$ and $\gamma(Y) \in X'$ hence $X \cap X' \neq \varnothing$ a contradiction. The lemma thus follows. □

4. Closures and Spheres

Definition 11. *Let $R = (S, T)$ be a simple bi-secondary structure and let $C(X)$ be the closure of a CC X of R. Let $N(X) = \{\delta | Y_\delta \in C(X)\}$ be the set of nucleotides that index the decorations in the closure of the CC X of R. We can introduce a cyclical ordering on $N(X)$ by letting $p \in N(X)$ precede $q \in N(X)$ if q is the smallest nucleotide such that $p < q$. Furthermore we set* $\max[N(X)]$ *to precede* $\min[N(X)]$*. This cyclical order induces a cyclical order on $C(X)$ where $Y_p \in C(X)$ precedes $Y'_q \in C(X)$ if p precedes q in $N(X)$. By virtue of Lemma 1 this order is well defined. We call this ordered set the orbit of $C(X)$.*

Lemma 3. *Let $R = (S, T)$ be a simple bi-secondary structure and let $C(X)$ be the closure of a CC X of R. Then, for any $Y_p \in C(X)$ and any 1-face $[u, v]$ of Y, there exists $Y'_q \in C(X)$, $Y_p \neq Y'_q$, with $Y_p \cap Y'_q = [u, v]$. I.e. any decoration (triangle) in $C(X)$ is glued along all of its 1-faces (edges) to decorations still in $C(X)$. Furthermore, the only decorations in $C(X)$ that have as a face the edge $[u, v]$ are Y_p and Y'_q.*

Proof. W.l.o.g. we can consider $Y_p = [x, y, z]_p \in C(X)$ to be a decoration at p with the pure edge of Y being $[x, y]$ for $x, y \in S$. We thus have $\alpha_x = \gamma(Y) \in X$. Furthermore we can assume that $p = b(\gamma(Y)) = b(\alpha_x)$.

For each edge of Y_p we would like to identify another decoration $Y'_q \in C(X)$ such that Y'_q shares that edge with Y_p.

Firstly, clearly Y_p shares the pure edge $[x, y]$ with the decoration $Y'_{e(\gamma(Y))} \in C(X)$ (See Figure 2). Since $[x, y]$ is a pure edge, it can only appear as a 1-face in two 2-simplices of $K_2(R)$. By construction, these are Y and Y'. Hence Y'_q is the unique decoration in $C(X)$ that shares the 1-face $[x, y]$ with Y_p.

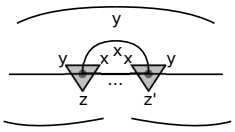


Figure 2: The decoration $Y_p = [x, y, z]_p$ with $p = b(\alpha_x)$, is glued along its pure edge $[x, y]$ to the decoration $Y'_q = [x, y, z']_q$ with $q = e(\alpha_x)$. Note that in this case $\gamma(Y) = \gamma(Y') = \alpha_x$.

Consider now a mixed edge of Y_p. We claim that this edge is present in the decoration Y'_q that: is the predecessor OR that succeeds Y_p in the the orbit of $C(x)$. Suppose we consider the edge is $[x, z] \subseteq Y_p$ (the argument for the other choice being similar to the following). Let Y'_q succeed Y_p in the orbit. Note that, by definition, we must then have that q is the closest (minimal)

nucleotide to p (w.r.t. the cyclic ordering on $N(X)$). To show that $[x,z] \subseteq Y'_q$ it suffices to note that if r would be a nucleotide at which we would have a decoration Y''_r, and said nucleotide would be in between p and q then, we must have $\forall Y''_r \in K_2(R)^* \implies Y''_r \notin C(X)$. Otherwise r would violate the minimality of q. Thus $p < b(\gamma(Y'')) < e(\gamma(Y'')) < q$. Hence we must have $[x,z] \subseteq Y'_q$ (See Figure 3).

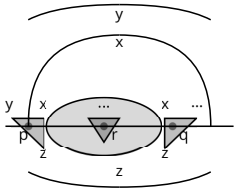


Figure 3: The decoration $Y_p = [x,y,z]_p$ with $p = b(\alpha_x)$, is glued along its mixed edge $[x,z]$ to the decoration $Y'_q = [x,y,z']_q$. By minimality of Y'_q we must have that for any decoration Y''_r with $p < r < q$, $Y''_r \notin C(x)$.

Now, to show that Y'_q is the only other decoration in $C(X)$ that contains the face $[x,z]$ we argue as follows:

Suppose there exists another decoration $Y''_r \in C(X)$, $r \neq q$ that also contains the face $[x,z]$. Then by lemma 1 we must have $Y'' \neq Y'$ and so $\gamma(Y') \neq \gamma(Y'')$. Note that $\alpha_z \neq \gamma(Y')$, since if that were the case, we would have to have $p < r < q$ which would contradict the minimality of q. Thus it must be that $q = b(\gamma(Y')) < e(\gamma(Y')) < r$ (See Figure 4).

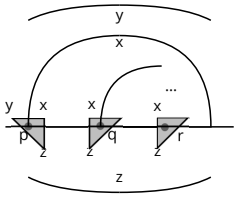


Figure 4: The decoration $Y_p = [x,y,z]_p$ with $p = b(\alpha_x)$, is glued along its mixed edge $[x,z]$ to the decoration $Y'_q = [x,y,z']_q$. By minimality of q we must have that for any decoration $Y''_r \in C(x)$ with $[x,z] \subseteq Y''_r$, $q = b(\gamma(Y')) < e(\gamma(Y')) < r < e(\alpha_x)$.

Since $\gamma(Y') \in X$, there must exists a sequence τ of pairwise crossing arcs that terminates with α_x, i.e. a path between $\gamma(Y')$ and α_x in the X-vertex induced arc line sub-graph of R. Note that for such arcs w in this sequence τ, we cannot have $b(w) < b(x) < e(w)$ otherwise the edge $[x,z] \subseteq Y_p$ would have to contain the loop corresponding to w in its labeling. Hence τ must connect $\gamma(Y')$ to α_x through an arc w' such that $b(w') < e(x) < e(w')$.

However, since $e(\gamma(Y')) < b(\gamma(Y'')) < e(\gamma(Y'')) < e(\alpha_x)$, then either $b(w'') < r < e(w'')$ for some w'' in τ, or $\gamma(Y'')$ must belong to τ. In the first case, the label of the edge $[x,z] \subseteq Y''_r$ would have to contain the loop corresponding to w''. But since $[x,z]$ is fixed, so is its labeling, and hence a contradiction

arises. We examine the second case where we suppose $\gamma(Y'')$ belongs to τ which connects $\gamma(Y')$ and α_x. In this case, either $r = b(\gamma(Y''))$ or $r = e(\gamma(Y''))\}$. Suppose $r = b(\gamma(Y''))$. Since there is a sub-sequence $\tau_L \subseteq \tau$ of arcs connecting q to r there must exist a w'' in this sub-sequence such that $b(w'') < r < e(w'')$. But then, again, the label of the edge $[x, z] \subseteq Y''_r$ would have to contain the loop corresponding to w'', and since $[x, z]$ is fixed, so is its labeling, and we reach another contradiction. Finally, if we suppose $r = e(\gamma(Y''))$, there is a sub-sequence $\tau_R \subseteq \tau$ of arcs connecting r to $e(\alpha_x)$. Thus, there must exist a w'' in this sub-sequence such that $b(w'') < r < e(w'')$. This again precludes the edge $[x, z] \subseteq Y''_r$ having the correct labeling and so the final contradiction arises.

Hence, there does not exists another decoration $Y''_r \in C(x)$ with $[x, z]$ as a face. As mentioned previously, a similar set of arguments hold for the edge $[x, z] \subseteq Y_p$, and thus the lemma follows. □

Lemma 4. *Let $R = (S, T)$ be a simple bi-secondary structure and let $C(X)$ be the closure of a CC X of R. There exists a Euclidean 3-space embedding of $C(X)$ that is homeomorphic to a 2-sphere.*

Proof. By Lemma 3 and Lemma 1 we can conclude that there exists a Euclidean 3-space embedding of $C(X)$ that is a closed surface. It suffices to show that this surface is a sphere. To this end we argue as follows: Let P be the triangulated annular region obtained by the pairwise consecutive gluing of the decorations in $C(X)$ following the orbit, only along edges that are mixed (See Figure 5).

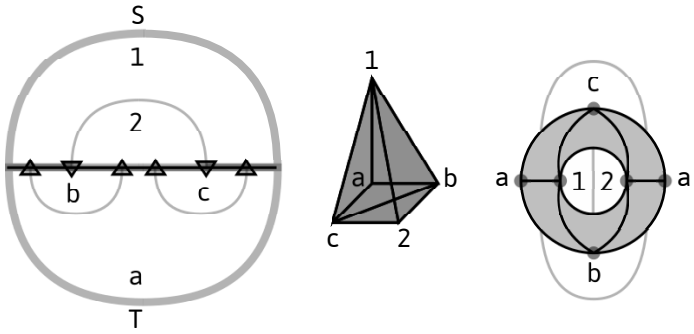


Figure 5: LHS: a bi-secondary structure with one CC, $X = \{\alpha_b, \alpha_c, \alpha_2\}$, and the CC's closure in terms of corresponding decorations. MS: The closure as a triangulation of a 2-sphere in $K(R)$. RHS: the triangulation of the annular region P with the gluing arcs corresponding to the arcs in the CC. Note that the gluing is performed in a cyclical fashion along the orbit of $C(X)$.

We draw a "gluing" arc between two pure edges in P if they are to be glued. It suffices to show that these arcs can be embedded in $\mathbb{R}^2 \setminus P$ without crossing. The $\mathbb{R}^2 \setminus P$ embedding is given by the fact that, as mentioned in the proof of Lemma 3, pure edges of a decoration at the endpoint of a given gluing arc will be glued to pure edges of a decoration at the other endpoint of the gluing arc.

Hence, the "gluing" arcs are actually the pure arcs themselves. Furthermore the pure-arcs corresponding to the inside boundary of P will be arcs from the secondary structure S while those corresponding to the outside boundary of P correspond to arcs in T. Since $R = (S, T)$ is a bi-secondary the pure arcs will thus have a planar embedding into $\mathbb{R}^2 \setminus P$ by virtue of the planarity of $R = (S, T)$. Hence the lemma follows. □

Remark 5. *Lemma 4 and Lemma 2 allow us to immediately conclude that*

$$|\chi(R)| \leq r(H_2(R))$$

in the case where $R = (S, T)$ is a simple bi-secondary structure. This prompts the natural question as to whether or not we actually have strict equality in the above relation. As we shall see in the following, that will indeed be the case.

5. The Tree of Irreducible Components

Definition 12. *Let $R = (S, T)$ be a simple bi-secondary structure. Let X and X' be two distinct ICs of R. Then we say X is* nested by *X' which we denote by $X \ll X'$, if and only if there exists an arc $\epsilon' \in X'$, such that for all $\epsilon \in X$, we have $\epsilon \prec_S \epsilon'$ or $\epsilon \prec_T \epsilon'$ (when ϵ is flipped to the same side as ϵ').*

Remark 6. *Clearly, the $\ll$ relation defines a poset structure on the the set of ICs of R. As a result, a bi-secondary structure can be constructed from ICs via nesting and concatenation. Hence, each IC has a unique cover (parent) w.r.t the $\ll$ poset order. Furthermore, letting $\langle C(X)\rangle$ denote the sub-simplicial complex of $K(R)$ generated by $\{Y | \gamma(Y) \in X\}$, Lemma 4 shows that when X is a CC, $\langle C(X)\rangle$ is homeomorphic to a 2-sphere. We now show that when X is a trivial IC, i.e., X contains only 1 arc, $\langle C(X)\rangle$ is a single 2-simplex (triangle).*

Lemma 5. *Let X be a trivial IC of a simple bi-secondary structure $R = (S, T)$. Then $\langle C(X)\rangle$ is a 2-simplex.*

Proof. W.l.o.g, we can assume $X = \{\mu\}$, where $\mu \in S$. Let ϵ be the cover of μ w.r.t. $\prec_S$. Let β be the cover of μ w.r.t. $\prec_T$ (when μ is flipped to the T side of the diagram). Since $\mu \in S$ does not cross an arc in T we must w.l.o.g. have $b(\beta) < b(\mu) < e(\mu) < e(\beta)$. Let $Y = [s_0, s_1, t]$ with $\alpha_{s_0} = \mu, \alpha_{s_1} = \epsilon, \alpha_t = \beta$. Then $Y_{b(\mu)} = Y_{e(\mu)}$ and so we must have $C(X) = \{Y_{b(\mu)}, Y_{e(\mu)}\} = \{Y\}$ (See Figure 6).

Hence, the lemma follows. □

Let now X be an IC of a simple bi-secondary structure $R = (S, T)$. We say ϵ is the minimal S-arc that nests X and β is the minimal T-arc that nests X if and only if $\forall \mu \in X, \epsilon \prec_S \mu \prec_T \beta$ (when μ is flipped to the S and T sides of the diagram respectively). Two such arcs always exist and are unique, since R is a bi-secodary structure. We observe the following:

Letting $p = \min N(X)$, w.l.o.g., we can assume p is an end point of an arc ϵ' in S. Note then that $b(\epsilon) < p < e(\epsilon)$ and similarly, $b(\beta) < p < e(\beta)$. Let

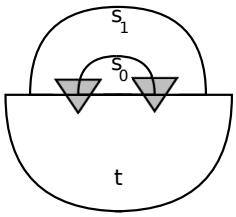


Figure 6: Here $\mu = \alpha_{s_0}$. The decorations at $b(\alpha_{s_0})$ and $e(\alpha_{s_0})$ come from the same 2-simplex $Y = [s_0, s_1, t]$.

$Y = \{s_0, s_1, t\}$ with $\alpha_{s_0} = \epsilon', \alpha_{s_1} = \epsilon, \alpha_t = \beta$. Then we must have $Y_p \in C(X)$ (See Figure 7). We can thus make the following defintion:

Definition 13. *The* 1*-simplex* $\{s, t\} \in \langle C(X) \rangle$ *with* $\alpha_s = \epsilon, \alpha_t = \beta$ *as defined above, is called the* up (mixed) edge *of* $\langle C(X) \rangle$*. All other mixed* 1*-simplices of* $\langle C(X) \rangle$ *are called* down (mixed) edges *of* $\langle C(X) \rangle$*.*

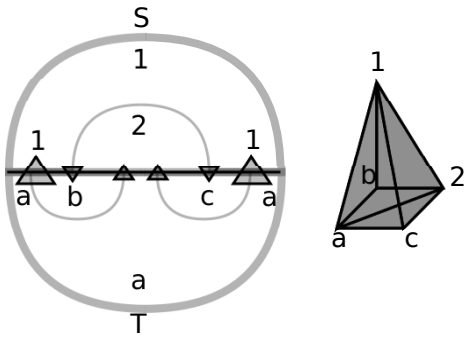


Figure 7: LHS: a bi-secondary structure with a single CC $X = \{\alpha_b, \alpha_c, \alpha_2\}$. The 1-simplex $[1, a]$ is the up-edge of the CC. RHS: The closure of the LHS's CC.

Lemma 6. *Let* X *be an IC of a simple bi-secondary structure* $R = (S, T)$ *and let* $\{s, t\}$ *be the up edge of* $\langle C(X) \rangle$*, where* $s \in S$*,* $t \in T$ *and* $\alpha_s = \epsilon$*,* $\alpha_t = \beta$*. Then* X'*, the cover of* X *under the* $\ll$ *poset order, is the unique IC such that* $\langle C(X') \rangle$ *contains* $\{s, t\}$ *as a down edge.*

Proof. We distinguish two cases depending on whether or not ϵ and β cross.

Case 1: ϵ and β cross and so must be contained in the same IC. In this case, by definition of $\ll$, both ϵ and β must be contained in X', since ϵ and β are the minimal S-arc and T-arc respectively that both nest X. Let p' be the largest nucleotide in $N(X')$ that is smaller than the smallest nucleotide $p \in N(X)$. The decoration $Y_{p'} \in C(X')$ thus contains $[s, t]$ as a mixed edge for $\alpha_s = \epsilon$ and $\alpha_t = \beta$. Suppose that $[s, t]$ is the mixed up edge of $\langle C(X') \rangle$. Then, no arc from the set $\{\epsilon$ and $\beta\}$ can be contained in X'. Since in this case both ϵ and β are

contained in X', $[s,t]$ is a down mixed edge of $\langle C(X')\rangle$. Furthermore, since X' is the unique IC that contains ϵ and β, X' is the unique IC such that $\langle C(X')\rangle$ contains $[s,t]$ as a down mixed edge.

Case 2: ϵ and β are non-crossing and so must be contained in different ICs. In this case, ϵ and β must be nested within one another. W.l.o.g. we can assume $b(\beta) < b(\epsilon) < e(\epsilon) < e(\beta)$, i.e. ϵ is nested by β. Then, by definition of $\ll$, ϵ is contained in X'. However, since ϵ and β are contained in different ICs, β must then be the minimal T-arc that also nests X'. Let now p' be the largest nucleotide in $N(X')$ that is smaller than the smallest nucleotide $p \in N(X)$. The decoration $Y_{p'} \in C(X')$ thus contains $[s,t]$ as a mixed edge for $\alpha_s = \epsilon$ and $\alpha_t = \beta$. Recall that if $[s,t]$ is the mixed up edge, then no arc from the set $\{\epsilon$ and $\beta\}$ can be contained in X'. Since ϵ is contained in X', $[s,t]$ must again be a down mixed edge of $\langle C(X')\rangle$. On the other hand, let X'' be the IC that contains β. Since ϵ is nested by β, $[s,t]$ can not be a 1-face of $\langle C(X'')\rangle$ (See Figure 8).

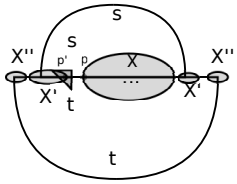


Figure 8: Here $\epsilon = \alpha_s$, $\beta = \alpha_t$ and $X'' \ll X' \ll X$.

Since X' is the unique IC that contains ϵ, X' is the unique IC such that $\langle C(X')\rangle$ contains $[s,t]$ as a down mixed edge. The lemma then follows. □

Remark 7. *We have revealed that a tree-like structure on the set* $\{\langle C(X)\rangle\}_{X=IC}$ *is inherited from the poset order* $\ll$ *over the set of all ICs in the following sense: We construct a graph* Γ *over* $\{\langle C(X)\rangle\}_{X=IC}$ *considered as vertices, where* $\langle C(X)\rangle$ *is connected via an edge to* $\langle C(X')\rangle$ *if the unique up edge of* $\langle C(X)\rangle$ *is a down edge of* $\langle C(X')\rangle$ *(See Figure 9). Lemma 6, together with the fact that* $H_1(R) = 0$ *(See Part Two: Loop Homology of Bi-secodary Structures)), establishes that* Γ *is in essence, a tree.*

6. Crossing Components and Homology Ranks for simple Bi-structures

Definition 14. *Let* $K = \dot{\bigcup}_{d=0}^{\infty} K_d$ *be an abstract simplicial complex and let* $Y \in K_d$ *be a* d*-simplex. Let* Y' *be a* k*-face of* Y*, where* $k < d$*. We say* Y' *is* Y-exposed *if and only if any simplex of* K *that contains* Y' *as a* k*-face must be a face of* Y*.*

Lemma 7. *Let* $R = (S,T)$ *be a simple bi-secondary structure. For any* 2*-simplex* $Y \in K_2(R)$*, if* $\gamma(Y)$ *is not contained in any CC of* R*, then the pure edge of* Y *is* Y*-exposed.*

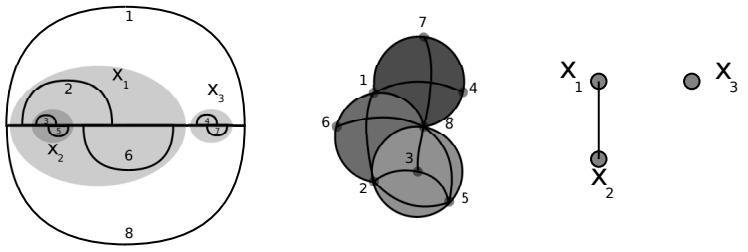


Figure 9: LHS: a bi-secondary R structure with tree CCs, $X_1 = \{\alpha_2, \alpha_6\}, X_2\{\alpha_3, \alpha_5\}$ and $X_3 = \{\alpha_4, \alpha 7\}$. MS: $K(R)$ obtained by gluing $\langle C(X_1)\rangle$, $\langle C(X_2)\rangle$ and $\langle C(X_3)\rangle$. RHS: the poset $(\chi(R), \ll)$. Note that only X_1 and X_2 and connected by an edge in the poset diagram of $\ll$ since $X_2 \ll X_1$, while $\langle C(X_1)\rangle$ and $\langle C(X_3)\rangle$ share the mixed up edge $[1, 8]$ and so must be connected by an edge in Γ.

Proof. W.l.o.g., let us assume $Y = [s_0, s_1, t_0]$ with pure arc $\gamma(Y) = \alpha_{s_0}$. Since R is simple, $b(\gamma(Y))$ and $e(\gamma(Y))$ are unpaired nucleotides in the T secondary structure. Hence, each of $b(\gamma(Y))$ and $e(\gamma(Y))$ are contained in exactly one loop in T. Furthermore, as $\gamma(Y)$ does not cross any arc in T, then for any arc $z \in T$ we must have that $[b(z) < b(\gamma(Y)) < e(z)] \Leftrightarrow [b(z) < e(\gamma(Y)) < e(z)]$. Therefore, $b(\gamma(Y))$ and $e(\gamma(Y))$ are contained in the same loop in T, namely, t_0. Since t_0 is the unique loop in T that has nonempty mutual intersection with s_0 and s_1, $Y = [s_0, s_1, t_0]$ is the unique 2-simplex in $K(R)$ that contains $[s_0, s_1]$ as an edge. Thus $[s_0, s_1]$ is Y-exposed and the lemma follows. □

Theorem 1. *Let $R = (S, T)$ be a simple bi-secondary structure. Let $r(H_2(R))$ denote the rank of the second homology group of $K(R)$. Then*

$$r(H_2(R)) = |\chi(R)|.$$

Proof. The basic idea behind this proof is to recursively decompose $Ker(\partial_2)$, following the tree-like structure of $K(R)$ such that each CC will contribute exactly one basis vector to $Ker(\partial_2)$.

Since R is a simple bi-secondary structure, $K_3(R) = \varnothing$. Therefore $Im(\partial_3) = 0$ and thus $H_2(R) \cong Ker(\partial_2)$. Let us consider $\tau \in Ker(\partial_2)$ where

$$\tau = \sum_{Y \in K_2(R)} n_Y Y.$$

Note that for each Y, its corresponding pure arc $\gamma(Y)$ is either crossing or non-crossing. Furthermore, if $\gamma(Y)$ is crossing, then it must be contained in exactly one of the CCs by definition. Assume $|\chi(R)| = k$ and let $X_1, X_2, \ldots, X_k$ be the CCs of R. We can further decompose τ into the following sum

$$\tau = \sum_{\gamma(Y)\ non-crossing} n_Y Y + \sum_{j=1}^{k} \sum_{\gamma(Y^j) \in X_j} n_{Y^j} Y^j.$$

Since $\tau \in Ker(\partial_2)$, we have

$$\partial_2(\tau) = \sum_{\gamma(Y)\ non-crossing} n_Y \partial_2(Y) + \sum_{j=1}^{k} \partial_2\Big(\sum_{\gamma(Y^j)\in X_j} n_{Y^j} Y^j\Big)$$

$$= \sum_{\gamma(Y)\ non-crossing} n_Y \overline{Z^P} + \sum_{\gamma(Y)\ non-crossing} n_Y(\overline{Z_1^M} + \overline{Z_2^M}) +$$

$$+ \sum_{j=1}^{k} \partial_2\Big(\sum_{\gamma(Y^j)\in X_j} n_{Y^j} Y^j\Big) = 0,$$

where $\overline{Z^P}$ and $\overline{Z_{1,2}^M}$ are the signed pure 1-faces and the mixed 1-faces of Y respectively, such that $\gamma(Y)$ is non-crossing. By Lemma 7, we know that for all non-crossing arcs $\gamma(Y)$, $\overline{Z^P}$ is exposed. Thus the coefficient of $\overline{Z^P}$ in $\partial_2(\tau)$ is n_Y. Since $\partial_2(\tau) = 0$, we must have $n_Y = 0$. Thus, in the expression of $\partial_2(\tau)$, the sum over non-crossing arcs disappears.

Next, we will focus on the term $j = 1$ in the expression of $\partial_2(\tau)$, namely $\partial_2(\sum_{\gamma(Y^1)\in X_1} n_{Y^1} Y^1)$, where X_1 is a CC that is minimal w.r.t. $\ll$ among all other CCs of R (i.e. X_1 does not nest any other CC of R). We will rewrite this term as a linear combination of 1-faces of $\langle C(X_1)\rangle$ while further partitioning said linear combination based on the types of 1-faces in $\langle C(X_1)\rangle$, namely, pure, down mixed and up mixed

$$\partial_2\Big(\sum_{\gamma(Y^1)\in X_1} n_{Y^1} Y^1\Big) = \sum_{Z^P\in\langle C(X_1)\rangle\ pure} m_{Z^P} Z^P +$$

$$+ \sum_{Z^D\in\langle C(X_1)\rangle\ down\ mixed} m_{Z^D} Z^D + m_{Z^U} Z^U.$$

The first sum is taken over all pure 1-faces Z_p of $\langle C(X_1)\rangle$. The second sum is taken over all down mixed 1-faces of $\langle C(X_1)\rangle$. The last term corresponds to the unique up mixed edge of $\langle C(X_1)\rangle$.

Let us examine the first sum. Note that each pure edge of $K_1(R)$ corresponds to a unique arc in R, namely, the pure arc of any decoration that contains said pure edge (see Definition 7). By Remark 6 we can conclude that for any Z^P, X_1 is the unique IC such that $\langle C(X_1)\rangle$ contains Z^P as a pure edge. Therefore, the coefficient of Z^P in $\partial_2(\tau)$ is m_{Z^P}. Since $\partial_2(\tau) = 0$, we must have $m_{Z^P} = 0$. Hence, the first sum in the decomposition of $\partial_2(\sum_{\gamma(Y^1)\in X_1} n_{Y^1} Y^1)$ disappears.

Now, for the second sum, since X_1 is a CC that does not nest any other CC in R, by Lemma 6, each Z^D is either: the up edge of some Y^{Z^D} where $\gamma(Y^{Z^D})$ is non-crossing, OR it is not contained in any other $\langle C(X')\rangle$ for X' another CC of R. We can then conclude that the coefficient of Y^{Z^D} in τ must be zero, since if $\gamma(Y^{Z^D})$ is non-crossing then its coefficient in τ must be 0 by the argument above regarding the first sum. Therefore, regardless, the coefficient of Z^D in $\partial_2(\tau)$ is m_{Z^D}. Since $\partial_2(\tau) = 0$, we must have $m_{Z^D} = 0$. Hence, the second sum in the decomposition of $\partial_2(\sum_{\gamma(Y^1)\in X_1} n_{Y^1} Y^1)$ disappears.

We can thus conclude that

$$\partial_2(\sum_{\gamma(Y^1)\in X_1} n_{Y^1}Y^1) = m_{Z^U}Z^U.$$

Note however that

$$0 = \partial_1(\partial_2(\sum_{\gamma(Y^1)\in X_1} n_{Y^1}Y^1)) = m_{Z^U}\partial_1(Z^U).$$

Since $K(R)$ is a simplicial complex, each of its 1-faces contains two distinct 0-faces. Therefore, $\partial_1(Z^U) \neq 0$. As a result, we must have $m_{Z^U} = 0$. Hence we can conclude that if $\tau \in Ker(\partial_2)$ then

$$\partial_2(\sum_{\gamma(Y^1)\in X_1} n_{Y^1}Y^1) = 0.$$

We now apply the above argument recursively, from bottom to top, following the $\ll$ poset order on the CCs of R. Thus, for each CC $X_j \in R$, we will eventually have

$$\partial_2(\sum_{\gamma(Y^j)\in X_j} n_{Y^j}Y^j) = 0.$$

Since for each CC $X_j \in R$, $\langle C(X_j)\rangle$ is a triangulation of a 2-sphere, by [1],

$$H_2(\langle C(X_j)\rangle) \cong \mathbb{Z}.$$

Thus, there exists $V_j = \sum_{\gamma(Y^j)\in X_j} v_{Y^j}Y^j$, such that $\sum_{\gamma(Y^j)\in X_j} n_{Y^j}Y^j$ can be uniquely represented as l_jV_j, for some $l_j \in \mathbb{Z}$. Furthermore, By Lemma 2, all closures $C(X_j)$ for X_j a CC of R are disjoint. Thus $\{V_j\}_{1\leq j\leq k}$ are linearly independent. Therefore, any $\tau \in Ker(\partial_2)$ can be uniquely represented as $\tau = \sum_{j=1}^{k} l_jV_j$. As a result, we have

$$H_2(R) \cong Ker(\partial_2) \cong \mathbb{Z}^k = \mathbb{Z}^{|\chi(R)|} \implies r(H_2(R)) = |\chi(R)|,$$

and the theorem follows. □

7. Scoops, Splits and Homology Ranks for arbitrary Bi-structures

Lemma 8. *Let $R = (S, T)$ be a bi-secondary structure. For any 3-simplex W in $K_3(R)$, there exists one mixed edge $Z \in K_1(R)$ that is W-exposed.*

Proof. Let $W = [s_0, s_1, t_0, t_1] \in K_3(R)$, with $s_0 \leq s_1 \leq t_0 \leq t_1$ (in terms of the simplicial ordering on $K(R)$). Since $s_0 \cap s_1 \cap t_0 \cap t_1 \neq \varnothing$, α_{s_0} and α_{t_0} must share at least one endpoint. W.l.o.g., we distinguish the following two cases (See Figure 10):
Case 1: $b(\alpha_{s_0}) < e(\alpha_{s_0}) = b(\alpha_{t_0}) < e(\alpha_{t_0})$.
In this case, we have $s_0 \cap t_0 = s_0 \cap s_1 \cap t_0 \cap t_1 = \{e(\alpha_{s_0})\}$. Suppose there exists

another 2-simplex (triangle) that contains the 1-simplex (edge) $[s_0, t_0]$. Namely, suppose there exists $x \in R$, with $s_{0,1} \neq x \neq t_{0,1}$, and such that $s_0 \cap t_0 \cap x \neq \varnothing$. Then

$$\varnothing \neq s_0 \cap t_0 \cap x = s_0 \cap s_1 \cap t_0 \cap t_1 \cap x \implies \begin{cases} s_0 \cap s_1 \cap x \neq \varnothing, x \in S \\ t_0 \cap t_1 \cap x \neq \varnothing, x \in T \end{cases}.$$

Either case this yields a contradiction, since three loops of the same secondary structure intersect trivially (See Part Two: Loop Homology of Bi-secodary Structures).Thus, it must be the case that $Z = [s_0, t_0]$ is W-exposed.

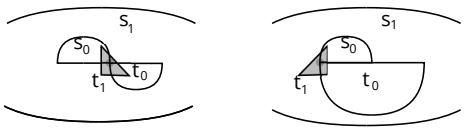


Figure 10: LHS: Case 1, $[s_0, t_0]$ is the mixed exposed W-edge. RHS: Case 2, $[s_0, t_1]$ is the mixed exposed W-edge.

Case 2: $b(\alpha_{s_0}) = b(\alpha_{t_0}) < e(\alpha_{s_0}) < e(\alpha_{t_0})$.
In this case, we have $s_0 \cap t_1 = s_0 \cap s_1 \cap t_0 \cap t_1 = \{b(\alpha_{s_0})\}$. By a similar argument as in Case 1, we conclude that $Z = [s_0, t_1]$ is W-exposed. The arguments for the remaining cases can be obtained by symmetry from the ones above and are similar to them. The lemma then follows. □

Let $R(S, T)$ be a bi-secondary structure over $[n]$ and let

$$P = \{p \in \{1, \ldots, n\} | deg(p) = 4 \text{ in the arc diagram of } R\}.$$

The two arcs that meet at p determine four mutually intersecting loops s_0, s_1, t_0, t_1 which contribute a unique 3-simplex $W \in K_3(R)$ to the simplicial complex $K(R)$ (See Figure 11).

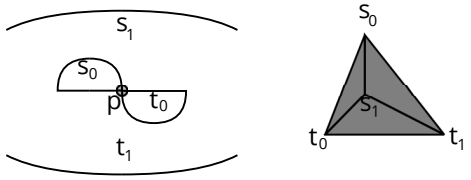


Figure 11: LHS: $s_0 \cap s_1 \cap t_0 \cap t_1 = \{p\}$. RHS: the 3-simplex $W = [s_0, s_1, t_0, t_1]$.

Lemma 8 guarantees that, among the 1-faces of the simplex W, at least one of them, call it $Z \in K_1(R)$, is W-exposed. W.l.o.g. we can assume that $Z = [s_0, t_0]$.

Definition 15. *Let $\mathcal{R}_p$ be a retraction*

$$\mathcal{R}_p : K(R) \longrightarrow \overline{K(R)}$$

where $\overline{K(R)} = \dot{\bigcup}_{d=0}^{\infty} \overline{K_d(R)}$ *is the induced topological space of the simplicial complex obtained by removing the* 1*-simplex* Z *and all subsequent higher dimensional simplices of* $K(R)$ *that have* Z *as a face. Namely,*

$$\overline{K_0(R)} = K_0(R), \overline{K_1(R)} = K_1(R) \setminus \{Z\},$$

$$\overline{K_2(R)} = K_2(R) \setminus \{[s_0, s_1, t_0], [s_0, t_0, t_1]\},$$

$$\overline{K_3(R)} = K_3(R) \setminus \{W\}, \overline{K_d(R)} = K_d(R) \text{ for all } d \geq 4.$$

We call $\mathcal{R}_p$ *the scoop of* R *at* p *(See Figure 12).*

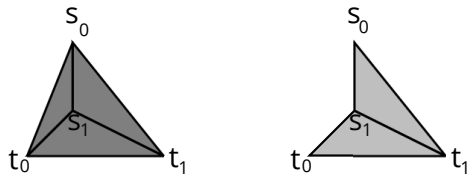


Figure 12: LHS: (before the scoop) the 3-simplex $W = [s_0, s_1, t_0, t_1]$. RHS: (after the scoop) removing the 1-simplex $Z = [s_1, t_0]$ and all higher dimensional simplices that contain it as a face, we are left with the two 2-simplices $[s_0, s_1, t_1]$ and $[s_0, t_0, t_1]$.

Remark 8. *For each* $p \in P$, $\mathcal{R}_p$ *induces a deformation retraction along the 1-face* Z*, of the* 3*-simplex of* $K(R)$ *(see Figure 13). Since a deformation retraction is a homotopy equivalence of the two spaces* $K(R)$ *and* $\overline{K(K)}$*, by [1], we can immediately conclude that*

$$H_2(\circ_{p \in P} \mathcal{R}_p(K(R))) \cong H_2(R).$$

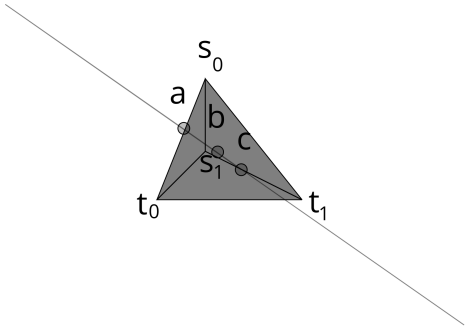


Figure 13: The deformation retraction D of the 3-simplex W, induced by the scoop R_p. We embed W in a euclidean 3-space and fix c the midpoint of the edge $[s_1, t_1]$. Then $D : W \times [0,1] \longrightarrow \langle\{[s_1, t_0, t_1], [s_0, s_1, t_1]\}\rangle$ with $D(a, t) = b$ when $b = ct + (1 - t)a$ is the linear combination along the ray between c and a.

Definition 16. *Let $\mathcal{S}_p$ be a mapping that takes the bi-secondary structure R over $[n]$ to the bi-secondary structure R' over $[n+1]$ by splitting the nucleotide p into two adjacent nucleotides q_1, q_2 such that the arcs in R that have one endpoint at p now have endpoints at q_1 and q_2 respectively and do not cross. We call $\mathcal{S}_p$ a split of R at p (See Figure 14).*

Remark 9. *For each $p \in P$, it is immediately clear that such a mapping $\mathcal{S}_p$ always exists.*

Lemma 9. *Let $R(S,T)$ be a bi-secondary structure over $[n]$ and let P be defined as above. Furthermore let $p \in P$ be fixed. Then,*

$$K(S_p(R)) \cong R_p(K(R)).$$

I.e. the simplicial complex of R split at p, is homeomorphic as a topological space to the scoop of R at p.

Proof. Let $W = [s_0, s_1, t_0, t_1] \in K_3(R)$, with $s_0 \leq s_1 \leq t_0 \leq t_1$ (in terms of the simplicial ordering on $K(R)$) be the 3-simplex determined by the two arcs that meet at p. Since $\{p\} \subseteq s_0 \cap s_1 \cap t_0 \cap t_1$, α_{s_0} and α_{t_0} must share at least one endpoint. W.l.o.g., we distinguish the following two cases (See Figure 10):

Case 1: $b(\alpha_{s_0}) < e(\alpha_{s_0}) = b(\alpha_{t_0}) < e(\alpha_{t_0})$.
In this case, after splitting R at p, we obtain $b(\alpha_{\overline{s_0}}) < e(\alpha_{\overline{s_0}}) < b(\alpha_{\overline{t_0}}) < e(\alpha_{\overline{t_0}})$ with the new loops $\overline{s_0} = (s_0 \setminus \{p\}) \cup \{q_1\}$, $\overline{t_0} = (t_0 \setminus \{p\}) \cup \{q_2\}$, $\overline{s_1} = (s_1 \setminus \{p\}) \cup \{q_1, q_2\}$ and finally $\overline{t_1} = (t_1 \setminus \{p\}) \cup \{q_1, q_2\}$.
Note that, $\overline{s_1} \cap x \neq \varnothing \Leftrightarrow s_1 \cap x \neq \varnothing, \forall x \in R$ and $\overline{t_1} \cap x \neq \varnothing \Leftrightarrow t_1 \cap x \neq \varnothing, \forall x \in R$. Also, $\overline{s_0} \cap x \neq \varnothing \Leftrightarrow s_0 \cap x \neq \varnothing, \forall x \in R \setminus \{t_0\}$ and $\overline{t_0} \cap x \neq \varnothing \Leftrightarrow t_0 \cap x \neq \varnothing, \forall x \in R \setminus \{s_0\}$.Finally, $\overline{s_0} \cap \overline{t_0} = \varnothing$. Thus, in this case we must have $K(S_p(R)) \cong R_p(K(R))$.(See Figure 14).

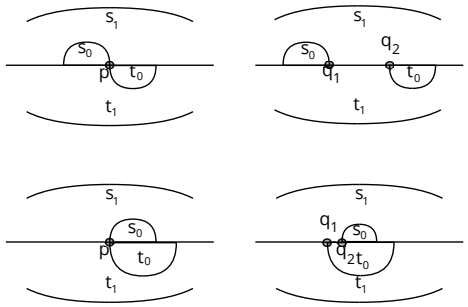


Figure 14: LHS: before the split. RHS: after the split. TOP: Case 1 split. BOTTOM: Case 2 split.

Case 2: $b(\alpha_{s_0}) = b(\alpha_{t_0}) < e(\alpha_{s_0}) < e(\alpha_{t_0})$.
In this case, after splitting R at p, we obtain $b(\alpha_{\overline{t_0}}) < b(\alpha_{\overline{s_0}}) < e(\alpha_{\overline{s_0}}) < e(\alpha_{\overline{t_0}})$ with the new loops $\overline{s_0} = (s_0 \setminus \{p\}) \cup \{q_2\}$, $\overline{t_0} = (t_0 \setminus \{p\}) \cup \{q_1\}$, $\overline{s_1} = (s_1 \setminus \{p\}) \cup \{q_1, q_2\}$ and finally $\overline{t_1} = (t_1 \setminus \{p\}) \cup \{q_1\}$.

Note that, $\overline{s_1} \cap x \neq \varnothing \Leftrightarrow s_1 \cap x \neq \varnothing, \forall x \in R$ and $\overline{t_1} \cap x \neq \varnothing \Leftrightarrow t_1 \cap x \neq \varnothing, \forall x \in R \setminus \{s_0\}$. Also, $\overline{s_0} \cap x \neq \varnothing \Leftrightarrow s_0 \cap x \neq \varnothing, \forall x \in R \setminus \{t_1\}$ and $\overline{t_0} \cap x \neq \varnothing \Leftrightarrow t_0 \cap x \neq \varnothing, \forall x \in R$. Finally, $\overline{s_0} \cap \overline{t_1} = \varnothing$. Hence in this case as well, we must have $K(S_p(R)) \cong R_p(K(R))$.

The arguments for the remaining cases can be obtained by symmetry from the ones above and the lemma then follows. □

Finally, we are in the position to prove the main result of this paper.

Theorem 2. *Let $R = (S, T)$ be an arbitrary bi-secondary structure. Then*

$$r(H_2(R)) = |\chi(R)|.$$

Proof. Denote by $R' = \circ_{p \in P} S_P(R)$ the bi-secondary structure obtained by sequential splits of R at all nucleotides $p \in P$ where P is defined as above. By Lemma 9 we must have that

$$K(R') \cong \circ_{p \in P} R_p(K(R)).$$

From this homeomorphism we obtain

$$H_2(K(R')) \cong H_2(\circ_{p \in P} R_p(K(R))).$$

By Remark 8

$$H_2(\circ_{p \in P} \mathcal{R}_p(K(R))) \cong H_2(R).$$

Hence $H_2(R) \cong H_2(R')$. Now R' is simple since each nucleotide of degree four in the arc diagram of R has been split into two nucleotides each of degree three in the arc diagram of R'. Thus, by Theorem 1, we have that $r(H_2(R')) = |\chi(R')|$. Finally, since each split introduces no new crossing arcs in R', the number of crossing components is conserved under splitting. Hence, we must have that $|\chi(R')| = |\chi(R)|$. Thus

$$r(H_2(R)) = r(H_2(R')) = |\chi(R')| = |\chi(R)|$$

and the theorem follows. □

8. Declarations of interest

None.

9. Acknowledgments

We gratefully acknowledge the comments from Fenix Huang. Many thanks to Thomas Li, Ricky Chen and Reza Rezazadegan for discussions.

References

[1] A. Hatcher, Algebraic topology, Tsinghua University Press, 2005.