frontiers

Origin of Biomolecular Networks

Heeralal Janwa [1,*], **Steven E. Massey** [2], **Julian Velev** [3] **and Bud Mishra** [4]

[1] *Department of Mathematics, University of Puerto Rico, San Juan, PR 00931*
[2] *Department of Biology, University of Puerto Rico, San Juan, PR 00931*
[3] *Department of Physics, University of Puerto Rico, San Juan, PR 00931*
[4] *Departments of Computer Science, Mathematics and Cell Biology, Courant Institute and NYU School of Medicine, New York University , NY*

Correspondence*:
Bud Mishra
mishra@nyu.edu

ABSTRACT

Biomolecular networks have already found great utility in characterizing complex biological systems arising from pair-wise interactions amongst biomolecules. Here, we review how graph theoretical approaches can be applied not only for a better understanding of various proximate (mechanistic) relations, but also, ultimate (evolutionary) structures encoded in such networks. A central question deals with the evolutionary dynamics by which different topologies of biomolecular networks might have evolved, as well as the biological principles that can be hypothesized from a deeper understanding of the induced network dynamics. We emphasize the role of gene duplication in terms of signaling game theory, whereby sender and receiver gene players accrue benefit from gene duplication, leading to a preferential attachment mode of network growth. Information asymmetry between sender and receiver genes is hypothesized as a key driver of the resulting network topology. The study of the resulting dynamics suggests many mathematical/computational problems, the majority of which are intractable but yield to efficient approximation algorithms, when studied through an algebraic graph theoretic lens.

Keywords: Biomolecules, Regulation and Communication, Interaction (Binary) Relationship, Network Model, Network Analysis, Spectral analysis

1 GENESIS OF BIOMOLECULAR INTERACTIONS

1.1 Introduction and a Road Map

A range of complex phenotypes of biomolecular systems can be inferred from macromolecular interactions, represented using networks. Such biomolecular networks include gene (regulatory) networks (GRNs) Thompson et al. (2015), protein-protein interaction (PPI) networks Huang et al. (2017), protein and RNA neutral networks Schuster et al. (1994) Govindarajan and Goldstein (1997), metabolic networks McCloskey et al. (2013) and meta-metabolic networks (composite metabolic networks of communities) Yamada et al. (2011). Our major focus here will be on GRNs and PPI networks, but the principles outlined are also applicable to the other types of biomolecular networks.

The paper covers the following topics: (i) A brief introduction to biomolecular networks (a topic also covered by other accompanying articles); (ii) A compendium of known results in (algebraic and combinatorial) graph theory ; (iii) Algorithmic (and algebraic) complexity, arising in the study of evolution

of networks; (iv) Current state of the field and open problems. The list of open problems focuses largely on the following: How to devise efficient (algebraic) algorithms that can shed important lights on *game theoretic models of the evolution of biological interactions*, given that they are driven by information asymmetry (leading to duplications, complementation, pseudogenization, etc.). Some of these important mechanisms have been studied qualitatively elsewhere, albeit not mathematically rigorously.

1.2 Ohno's Evolution by Duplication

At the genetic level, the growth of a GRN or PPI network is driven by gene mutation: e.g., duplication, translocation, inversion, deletion, short indels, and point mutations, of which duplication plays an outsized role. Susumu Ohno coined the phrase evolution by duplication (EBD) to emphasize this evolutionary dynamic Ohno (1970). The classic view of molecular evolution is that gene families may expand and contract over evolutionary time due to gene duplication and deletion Demuth et al. (2018). Here, we wish to present a more complex view, by exploring how biomolecular networks may grow, contract, or alter their topology over time, from the relative dynamic contributions and interactions of their constituent genes and gene families. This evolution is ultimately driven by the process of gene duplication and deletion, which leads to node and edge addition, or removal, from a biomolecular network, respectively. Since such variations in the network alter the phenotypes, over which selection operates, the evolution of networks and their features ultimately captures the essence of Darwinian evolution.

Recently, we introduced a signaling games perspective of biochemistry and molecular evolution Massey and Mishra (2018). There, we focused on interactions between biological macromolecules, which may be described using the framework of sender-receiver signaling games, where an expressed macromolecule such as a protein or RNA, constitutes a signal sent on behalf of a sender agent (e.g., gene). The signal comprises the three-dimensional conformation and physicochemical properties of the macromolecule. A receiver agent (e.g., a gene product, another macromolecule) may then bind to the signal macromolecule, which produces an action (such as an enzymatic reaction). The action produces utility for the participating agents, sender and receiver, and thereby – albeit indirectly – a change in overall fitness of the genome (in evolutionary game theory, utility and fitness are treated as analogous). When there is common interest, the utility is expected to benefit both sender and receiver and their selection, thus driving Darwinian evolution.

Replicator dynamics allow the signaling game to be couched in evolutionary terms Taylor and Jonker (1978). Replicator dynamics arise from the increased replication of players with higher utility (fitness). Thus, if a gene has a strategy that results in increased utility, then it will increase in frequency in a population. For a sender gene this would entail sending a signal that results in an increase in utility, while for a receiver gene this would entail undertaking an action that likewise results in an increase in utility. As already suggested, these dynamics represent a process analogous to Darwinian (adaptive) evolution or positive selection.

Biomolecular signaling games are sustained by information asymmetry between sender and receiver and so their interactions can be represented using directed graphs. Information asymmetry arises because the receiver is uninformed regarding the identity of the sender gene: it must rely on the expressed signal macromolecule to determine the identity of the sender gene. But, this strategy may be open to deception. Most biomolecular signaling games in the cell are between sender and receiver genes which have perfect common interest. This is so, because they are *cellularized*, chromosome replication is synchronized and so the genes replicate in concert. Such games are termed 'Lewis' signaling games, and rely on honest signaling from sender to the receiver Lewis (1969). A biomolecular signaling game is illustrated in Figure 1, part (1).

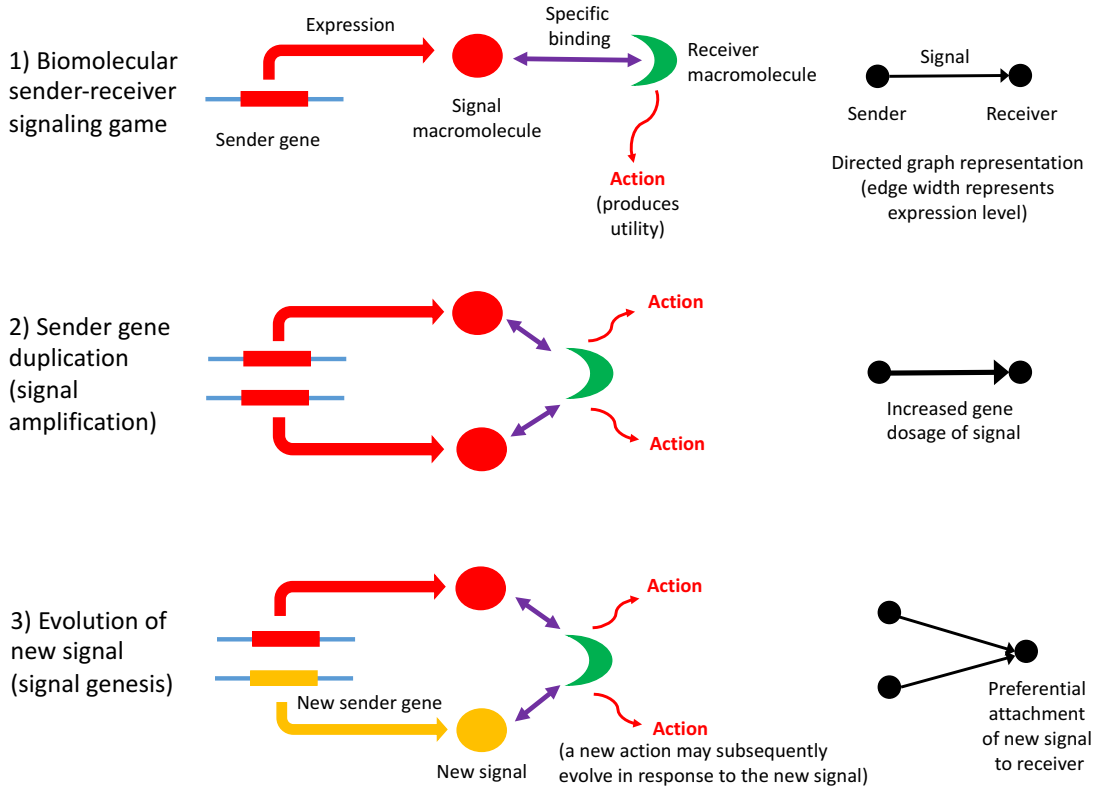

Figure 1. *The influence of information asymmetry on growth of a PPI network*. Interactions between macromolecules are envisaged as a biomolecular signaling game whereby a sender gene expresses a macromolecule, the signal, that then binds specifically to a receiver macromolecule, which then undergoes an action (such as an enzymatic reaction, or conformational change), which produces utility (fitness). The signal consists of the three-dimensional conformation and physicochemical properties of the macromolecule (1). The sender gene may undergo duplication, which has a dosage effect on the expressed macromolecule, resulting in signal amplification (2). This mechanism is expected to lower the Shapley value of the gene players in the genome, as the signal is partially redundant and so inefficient. Subsequently, the sender gene duplicate may acquire a new function (evolve a new signal) although the majority would be expected to undergo pseudogenization (3). Both these scenarios represent the re-establishment of a Nash equilibrium. If a new signal macromolecule evolves, it is likely to bind to the same receiver macromolecule initially. This preferential attachment arises because gene duplicates have a tendency to bind to their interaction partner initially, and then subsequently undergo interaction turnover Zhang et al. (2005), and is illustrated to the right of the figure. A key problem is how a new action by the receiver arises as the result of the evolution of a new signal; the new action may co-evolve with the new signal, or may be necessary first before a new signal can evolve. The latter would imply that receiver gene duplication and action genesis facilitates the evolution of new signals and sender genes (an exception would be when there is a conflict of interest; here the sender is more likely to make the first move in evolving a novel deceptive signal, and then the receiver would respond with a better discriminative recognition mechanism). This key, and novel aspect of gene duplication might be deciphered via consideration of the topology of directed graph representations of biomolecular interactions as sender-receiver signaling games. Refinements to the illustrated scheme include situations where the original signal protein binds to a variety of receiver proteins, or where the gene that codes for the receiver protein undergoes duplication (Figure 2).

However, situations may arise where a sender has a conflict of interest with the receiver. This kind of misalignment can occur when a sender gene is selfish, and would prefer to replicate itself at the expense of the rest of the genome. Such genes are termed ‘selfish elements,’ and come in a variety of forms, all marked by decoupled replication from the rest of the genome Burt and Trivers (2006). In a signaling game, when there is a conflict of interest between sender and receiver, then the sender is expected to adopt some degree of deceptive signaling Crawford and Sobel (1982). Consistent with this prediction, there are a range of examples of selfish elements that utilize molecular deception Massey and Mishra (2018).

Gene duplication is a fundamental evolutionary driver of organismal complexity Lespinet et al. (2002). The first step in the process of duplication of a sender gene may be viewed as one of signal enhancement. Because gene duplication results in gene dosage effects, it also results in amplification of the signal, the expressed gene product. This strategy can be viewed as lowering the overall utility of the genome, given that there is a cost involved in producing excessive signal. It is, thus, expected to lower the Shapley value Shapley (1969) of the gene players that cooperate within the genome. This conflict is usually resolved when the duplicated gene becomes pseudogenized, the usual fate of gene duplicates Innan and Kondrashov (2010).

Subsequent to duplication, the gene duplicates will sometimes diverge in function, although the exact mechanism remains to be elucidated Innan and Kondrashov (2010). This process represents signal divergence if the gene is a sender gene, and action divergence if the gene codes for a receiver macromolecule. The genesis of a new sender gene with a new signal may then promote evolution of a novel action by the receiver macromolecule, potentially facilitating duplication of the receiver gene itself. Likewise, the duplication of a receiver gene may facilitate the diversification of macromolecular signals that interact with the two duplicated receiver macromolecules. The process modifies the GRN or PPI network in a non-obvious manner and it deviates considerably from the way evolution of random graphs is usually treated, following Erdös and Rényi, discussed in more detail in Section 3 Erdös and Rényi (1959).

Signal and action genesis via gene duplication may have features in common with a Pólya's urn model of signal genesis Alexander et al. (2012) (Pólya's urn models are statistical models that involve sampling with replacement influenced by the identity of the sampled element. These models can lead to a '*rich get richer*' effect, of which 'preferential attachment' is an example, discussed in more detail in Subsection 3.2). In this model, reinforcement of signals (similar to reinforcement learning) may promote the invention of new synonyms. These considerations may provide parallels for how signals originate elsewhere, not dissimilar to how new words in a language can arise from existing words by a process of derivation Cotterell et al. (2017). Mechanistic commonalities in the process of signal genesis in these diverse systems as exhibited in GRNs remain to be explored. These models hint at a possibly new, but universal form of "preferential attachment" that drives the variations in biomolecular networks as well as the selectivity in Darwinian evolution.

1.3 Network Topology, Evolution by Duplication, and Preferential Attachments

Consequently, the topology of gene networks is non-deterministic and yet not memoryless, since it must encode layers of ripples produced earlier via the dynamics of gene duplication (paralogs and orthologs), as amplified during the network's history. Just as physicists infer the theories of origin of universe from the cosmic background radiation, we expect to enrich our understanding of the origin of machinery of life (e.g., codon evolution, evolution of multicellularity, evolution of sex etc.) from a rigorous analysis of the signaling games and their equilibria, which has rippled through the extant biomolecular networks. Taking this analogy further, we observe that the ripples in gravitational waves have been proposed to reflect the existence of parallel universes, whose presence created asymmetries in the initial conditions, giving rise to filamentary structures in the visible universe Hawking and Hertog (2018) This comparison is inspired by the notion of a 'protein big bang' from a single (or handful of) ur-protein(s) in the first complex life forms, evolving by gene duplication into the extant 'protein universe,' hinting at the information asymmetries fossilized in the GRN and PPI networks. Dokholyan et al. (2002).

Likewise, we point out that information asymmetry in macromolecular sender-receiver interactions may point to evolutionary paths that might have been abandoned unexplored; which may suggest new

engineering approaches needed by synthetic biology, or in drug discovery, or immuno-therapy. Note that during the process of evolution of signaling, gene duplication and deletion contribute to a certain degree of non-determinism and "conventionality" to the Nash equilibria that stabilize and manifest as non-trivial anisotropies in gene network topology.

In summary, the process of gene duplication, tempered by signal and action genesis can be thought of as a driver of preferential attachment in shaping the topology of gene networks, in which information asymmetry between senders and receivers is expected to play an indelible role. Figure 1 illustrates a basic mechanism whereby signal genesis may lead to preferential attachment during the growth of a PPI network. Topological features expected to hint at this process include: (i) the degree distribution, (ii) hierarchicity, (iii) assortativity and many others; they require powerful statistical and algebraic tools – covered in the later sections, where it is assumed that genome evolution is a complex process involving diverse groups of mutations such as insertions, deletions, conversions, duplications, transpositions, translocations and recombinations, and that it is further affected by selective constraints and effective population size and other factors such as the environment. With recent understanding of large scale cellular networks (regulatory, metabolic, protein-protein interactions) one must now aim at investigation between the evolutionary rates of a gene mutations and its effects on the network topology using mathematical models and analytics: see Wagner (1994). For instance, combining sequence analysis in a single genome and its close relatives, one can infer the rate and tempo of the evolutionary dynamics acting on the genome, and the resulting effects on the network's algebraic structures. We provide an example of how evolution by duplication leads to a preferential attachment mode of gene network growth in Figure 2, using the duplication of the p53 gene, and its paralogs p63 and p73 – all transcription factors regulating pathways involved in related phenotypes of somatic or developmental surveillance and interacting with similar family of genes (e.g., MDM2 or MDMX), as illustration [1].

Note that these abstract models generate refutable hypotheses that need experimental verification and support from mechanistic explanations. However, unfortunately, the biochemical processes involved in the hypothesized preferential attachment dynamics are not fully understood. For example, the duplication processes are often driven by Non-Homologous End Joining (NHEJ), a pathway that repairs double-strand breaks in DNA. To guide repair, NHEJ typically uses short homologous DNA sequences called microhomologies, which are often present in single-stranded overhangs on the ends of double-strand breaks Chang et al. (2017). When the overhangs are perfectly compatible, NHEJ usually repairs the break accurately. However, imprecise repair can lead to inappropriate NHEJ resulting in translocations, duplications and rearrangements Rodgers and McVey (2016), which add to variations that are random but not memoryless. Perhaps some of such hypotheses may need to be carefully examined using cancer genome data such as TCGA, and models of tumor progression. This analysis may also explain efficacy of certain therapeutic interventions in cancer as well as their failures via drug and immuno resistance.

2 NETWORK ANALYSIS

In this section, we discuss fundamentals of graphs, a mathematical formalism used in the study of biomolecular networks, as well as other related important topics. Consider a set of entities, denoted V and a set of binary relations between the entities $E \subseteq V \times V$. When V denotes biomolecules and E denotes interactions between them (e.g., regulations, proximity, synteny, etc.), the resulting graph represents a biomolecular network. One important advantage of graphs is that they have intuitive graphical

[1] A mutation in MDM affects all p53, p63 and p73 allowing utility tradeoffs between fecundity (through increased embryonic lethality) and cancer risks (through reduced somatic surveillance) in a population.

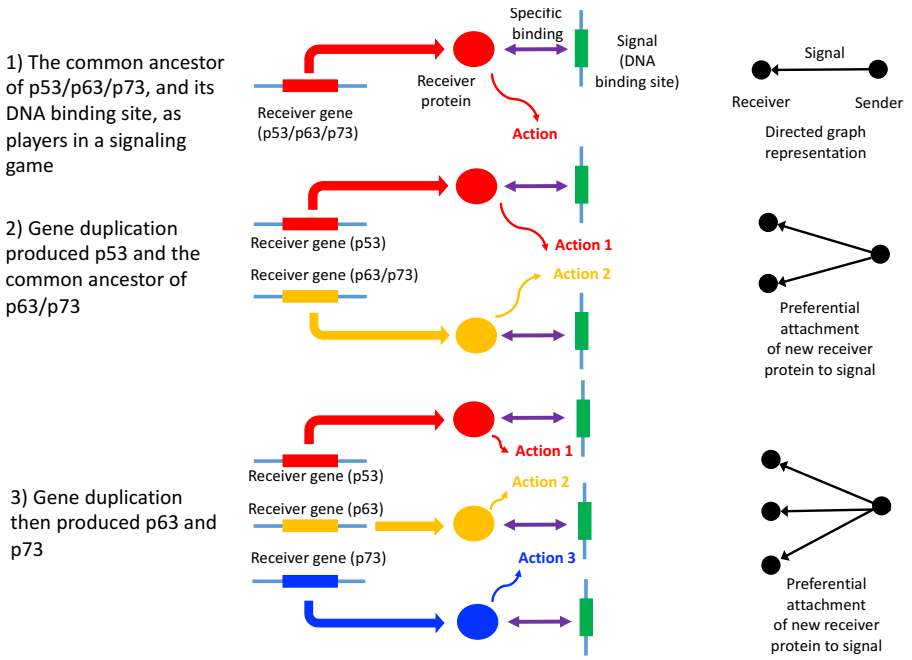

Figure 2. *Gene duplication of p53, p63 and p73 as a signaling game, and GRN growth.* An illustrative example of a signaling games view of network growth is provided by the paralogs p53, p63 and p73, which code for transcription factors, p53 being of critical importance in many cancers Joerger and Fersht (2006). Here, p53 and the common ancestor of p63/p73 duplicated (2), followed by the duplication and divergence of p63 and p73 Lu et al. (2009) Belyi et al. (2010) (3). The signal is the DNA binding site, while the receivers are the p53, p63 and p73 proteins (here the sender is the protein coding gene downstream of the DNA binding site). The receiver protein undergoes an action upon binding to the DNA binding site (the signal), which consists of the recruitment of additional transcription factors, and contribution to the assembly of the transcription initiation complex Nogales et al. (2017). The gene products of p53, p63 and p73 mostly bind to the same DNA binding sites Smeenk et al. (2008), thus each signal (and ultimately sender gene) has acquired two new binding partners, in addition to the original interaction with the gene product of the common ancestor of p53/p63/p73. This is a form of preferential attachment, which should influence network topology as the number of genes increase by duplication, as illustrated to the right of the figure. The signaling games perspective allows us to better understand scenarios where there is a conflict of interest between the genome, and a selfish entity such as a selfish element, a cancer or a virus. When there is a conflict of interest, a deceptive signal is expected to be emitted by the sender Crawford and Sobel (1982) (the selfish entity). Here, the DNA binding site of the selfish entity will mimic that of canonical DNA binding sites associated with normal cellular function, 'tricking' a transcription factor to bind to it, and altering the transcription of the sender gene (or alternatively abolishing transcription factor binding). Examples include *cis*-regulatory mutations in cancer Poulos et al. (2015)

representation. Such networks evolve over time with additions and deletions to the sets V and E. In order to create a bridge to algebraic approaches, we extend the standard combinatorial definition by endowing it with additional maps.

Formally, a graph is a pair of sets $G = (V, E)$ where V are the vertices (nodes, points) and $E \subseteq V \times V$ are the edges (arcs) respectively. When E is a set of unordered pair of vertices the graph is said to be undirected or simple. In a directed graph $G = (V, E,\ o, t)$, E consists of an ordered set of vertex pairs, i.e. for each edge $e \in E$, $e \to (o\,(e)\,, t\,(e))$ where $o\,(e)$ is called the origin of the edge e and $t\,(e)$ is called the terminus of the edge e [Serre (1980) and Biggs (1993)]. A graph is weighted if there is a map (weighting function, $w : E \to R_+$) assigning to each edge a positive real-valued weight.

If $G = (V, E, \cdot, \cdot)$ and $G' = (V', E', \cdot, \cdot)$ are two graphs such that $V' \subseteq V$ and $E' \subseteq E \cap (V' \times V')$, then $G' \subseteq G$, G' is a *subgraph* of G. If $E' = E \cap (V' \times V')$ (E' contains every edge in $e \in E$ with $o(e), t(e) \in V'$) then G' is an *induced subgraph* of G. G' and G are *isomorphic* ($G' \equiv G$) if there is a bijection $f : V' \to V$ with $(u, v) \in E' \iff (f(u), f(v)) \in E, \forall u, v \in V'$.

2.1 Topological Properties

Network properties are governed by its topology, such as degree distribution, clustering coefficients, motifs, assortativity, etc. Comprehensive treatments can be found in Thulasiraman et al. (2015); Loscalzo and Barabási (2016), and for more in-depth treatment regarding biomedical networks in Loscalzo et al. (2017).

Degree Distribution

The degree of a vertex v, $\deg(v)$, is the number of edges that connect the vertex with other vertices. In other words, the degree is the number of immediate neighbors of a vertex. In directed graphs in-degree and out-degree of a vertex can be defined as the number of incoming and outgoing edges respectively. Let n_k be the number of vertices of degree k and $|V| = N$, the total number of vertices in the graph and $|E| = M$, the total number of edges in the graph. Note that $\sum_k n_k = N$ and $\sum kn_k = \sum_{v \in V} \deg(v) = 2|E| = 2M$. The degree distribution is the fraction of vertices of degree k, $P(k) = n_k/N$, and two isomorphic networks will have same degree distributions (though not necessarily the converse). Thus, the degree distributions can tell a great deal about the structure of a family of networks. For example, if the degree distribution is singly peaked, following the Poisson (or its Gaussian approximation) distributions, the majority of the nodes can be described by the average degree $\langle k \rangle = \sum_k kP(k) = 2M/N$. The graph is said to be *sparse*, if $\langle k \rangle = o(\log N)$ (or $M = o(N \log N)$). Biomolecular networks are usually sparse, which can be fruitfully exploited in their algorithmic analysis. We can talk of *typical* nodes of the networks as being those that have degree distribution as those within 1 to 2 standard deviations from the average, while, with probability decreasing exponentially, it is possible to find nodes with degree much different from the average. While power-law degree distributions follow a completely different pattern: they are *fat-tailed*; the majority of the nodes have only few neighbors, while many nodes have relatively large number of neighbors. The highly-connected nodes are known as *hubs*.

Distance Metrics

One of the most fundamental metrics is the *distance* on a graph. First we define a *walk* of length m in a graph G from a vertex u to v as a finite alternating sequence of vertices and edges $\langle v_0, e_1, v_1, e_2, \ldots, e_m, v_m \rangle$, such that $o(e_i) = v_{i-1}$ and $t(e_i) = v_i$, for $0 < i \leq m$, such that $u = v_0$ and $v = v_m$. Then the number of edges traversed in the shortest walk joining u to v is called the *distance* in G between u and v denoted by $d(u, v)$. If there is a walk from u to itself, then we say that the set of vertices (respectively edges) form a cycle. The smallest number of m edges in a walk from u to itself is called a cycle of length m. The girth $g(G)$, is the shortest cycle in G. A walk whose vertices are distinct is called a (simple) *path*.

The concept of a walk allows us to define other properties of the graph. A graph $G = (V, E, o, e)$ is said to be *connected*, if any two vertices are the extremities of at least one walk. The maximally connected subgraphs are called the *connected components* of G. A giant component is a connected component containing a significant fraction of the nodes. The maximum value of the distance function in a connected graph is called the *diameter* of the graph. Frequently real life networks have small diameter and are said

to exhibit *small world phenomenon*. For many biomolecular networks the average distance between two nodes depends logarithmically on the number of vertices in the graph.

Additionally, a *complete graph* G is the undirected graph, in which each vertex is a neighbor of all other vertices; $\deg(v) = N - 1, \forall v \in V$; or equivalently, each distinct pair of vertices are connected (or are adjacent) by a unique edge. G is then denoted as K_N. A *clique* in an undirected graph is a subset of vertices such that its induced subgraph is complete. Additional combinatorial invariants of graphs useful in the analysis of networks can be defined (see Supplementary material for details).

Expanding Constants

Let $G = (V, E, \cdot, \cdot)$ be an undirected graph. Then for all $F \subset V$, the *boundary* ∂F is the set of edges connecting F to $V \setminus F$. The *expanding constant*, or *isoperimetric constant* of X is defined as,

$$h(X) = \min_{\emptyset \neq F \subset V} \frac{|\partial F|}{\min\{|F|, |V \setminus F|\}}.$$

For molecular network, then, the invariant $h(X)$ measures the quality of the network with respect to the flow of information within it, (e.g., via chemical reactions, or signaling). Larger $h(X)$ implies better expansion, faster mixing, faster partitioning, and many other related properties that may give the network a selective advantage.

Using various combinatorial algorithms devised for the study and analysis of biomolecular networks, one may compute $h(X)$ to determine their complexity. However precise characterization of $h(X)$ itself is an intractable (i.e., NP-complete) problem. Isoperimetric inequalities give bounds on $h(X)$ in terms of a related algebraic invariant, $\gamma(X)$ – called its *spectral gap*, determination of which has complexity $O(|V|)^c$, where c is at most 3; furthermore, $c = 1$ for many sparse graphs. We give isoperimetric bounds and results applicable to biomolecular networks in the Supplement, where we also introduce local Cheeger constant. We also introduce algebraic invariants in Section 2.2.

Clustering and Clustering Coefficients

Biological networks are modular, forming communities and hierarchies, likely to have been sculpted by EBD (Evolution by Duplication). To study these local structures in network science, one may perform *community* analysis, which aims to identify a group of nodes that have a higher probability of connecting to each other than to nodes from other communities (see for exmple Pellegrini (2019)). Various notions such as k-cliques, k-clubs and k-clans have been developed to detect communities, but they are ultimately closely connected to the problem of finding cliques and consequently, do not generally lend themselves to any reasonable algorithm other than brute-force enumeration. However, even detecting communities approximately may prove valuable for general evolutionary studies, since in these biological networks communities determine how specific biological functions are encoded in cellular networks – and thus subjected to Darwinian selective pressure, since these players are likely to have formed communities in the first place to carry out specific cellular functions. (see Hartwell, Hartwell et al. (1999)). Figure 4 highlights significant evidence that communities play important role human disease networks (see Loscalzo et al. (2017)).

Usually a simpler approach is commonly employed and deals with the problem of *clustering* in a graph, which seeks to partition the graph into disjoint subgraphs such that nodes in each such subgraph are "closer" to the other nodes in the same subgraph, while they are "farther" from the nodes of other subgraphs. Hierarchical clustering algorithms have been developed to uncover communities (approximately)

in polynomial time and depend upon the *similarity matrix* (x_{ij}), where the entry x_{ij} equals the distance between node i and node j. Among the classical algorithms are included those by Ravasz and by Girvan and Newman Girvan and Newman (2002). Other related algorithms include those for random-walk betweenness and network centrality.

The *local clustering coefficient* captures the degree to which the neighbors of a given node link to each other. In general, for undirected graphs, the *local clustering coefficient* C_i of node i with degree k_i is defined as

$$C_i := \frac{L_i}{k_i(k_i-1)/2}$$

where the numerator L_i is the actual number of connections between k_i immediate neighbors of i, and the denominator is the number of connections if the neighbors formed a complete graph (i.e. a clique). Note that an undirected complete graph K_{k_i} of k_i nodes has $k_i(k_i-1)/2$ edges. Thus, a fully clustered node will have $C_i = 1$ and for completely isolated node $C_i = 0$. We can define the *(average) clustering coefficient* of the whole network with N nodes as

$$\langle C \rangle = \frac{1}{N}\sum C_i.$$

The clustering coefficients can be used to characterize a network's *modularity*, as discussed later (in Section 3) in details.

Subgraphs and Motifs

Biomolecular networks have been found to contain network *motifs*, representing elementary interaction patterns between small subgraphs that occur substantially more often than as predicted by a completely random network of similar size and connectivity. The presence of such motifs is usually explained by an evolutionary process that can quickly create (usually by a variation involving duplication) or eliminate (usually by a selection process that favors pseudogenization and complementation) regulatory interactions in a fast evolutionary time scale – relative to the rate at which individual genes mutate. It is usually hypothesized that the underlying evolutionary processes are convergent. Thus efficient algorithms to detect such motifs are important in the analysis of biomolecular networks. These algorithms focus on estimating how much more frequently a subgraph isomorphic to a motif graph (with n vertices and m edges) occurs relative to what would be expected by pure chance.

The number N_{mn} of subgraphs with n nodes and m interactions expected of a network of N nodes can be estimated from the two key topological parameters of a complex network – namely the power-law exponent β and the hierarchical exponent α as we discuss in equations (1 and 2) below. In general the subgraph motifs can be classified in two types: Type I motifs are those where $(m-n+1)\alpha-(n-\beta)<0$, and type II subgraph motifs are those that satisfy the reverse inequality. One can determine their numbers N^I and N^{II} approximately as a function of $(m-n+1)\alpha-(n-\beta)$ and n_{max}, the degree of the most connected node in the network. One can show that $N^I_{nm} >> N^{II}$. One can also show that the relative number of Type II subgraphs is vanishingly small compared to Type I.

2.2 Algebraic Invariants and Spectrum

The intuitive pictorial/combinatorial representation of graphs is an extremely useful aid to their understanding. However, computing the topological properties of graphs combinatorially is computationally challenging especially when the size of the graph becomes large. As noted earlier, indeed, most

combinatorial algorithms on biomolecular networks such as on PPI networks and GRNs are computationally complex problems (most of them fall in the NP-complete complexity class) Karp (2011). Therefore, in order to carry out any quantitative and computational analysis, graphs are better represented as algebraic objects. This representation allows us to use linear algebra and mathematical analysis techniques. The key to this representation is the adjacency matrix $A(G)$. It is defined as $\{0,1\}^{n\times n}$ matrix in which, $A_{ij}=1$ if the vertices i and j are connected ($\exists e\in E, o(e)=i, t(e)=j$) and 0 otherwise. The matrix is symmetric if the graph is undirected. For weighted graphs we can assign weights w_{ij} for existing edges.

Algebraic properties provide us with tools to deduce various properties of the biomolecular networks. In particular, the spectral representation of the graph is of importance for a number of applications such as graph classification, etc. We can think of the adjacency matrix A as operating on the space $V=C^n$ of complex n-tuples written as column vectors x,y as follows $Ax\rightarrow y$. It can be shown that there are directions left invariant in this space. That is to say, $A_ix_i=\lambda_i x_i$ where λ_i are the eigenvalues and corresponding x_i the eigenvectors (spanning invariant directions) of the adjacency matrix for $1\leq i\leq n$. The spectrum of the graph G is defined as the collection of eigenvalues of the adjacency matrix $\mathrm{Spec(G)}=\mathrm{Spec(A)}=\lambda_1,..,\lambda_\mathrm{n}$. Naturally, if A is a real symmetric matrix, then the eigenvalues of A are real.

In particular, one algebraic invariant of the graph is the *spectral gap* $\gamma(G)$. It can be shown that the spectral gap gives excellent bounds on a combinatorial invariant, the Cheeger constant $h(G)$ (see the Supplementary material).

3 NETWORK EVOLUTION

Starting with the seminal work of Erdös and Rényi Erdös and Rényi (1959), a number of mathematical frameworks have been developed to model the "evolution" of graphs, covering the family of biomolecular networks. These frameworks may prove useful in explaining why most biological networks have certain non-obvious properties: namely, (i) Small-world property; (ii) High clustering coefficients (varying with degree distribution); (iii) Emergence of "hubs." Such network models are ultimately expected to capture various observed properties of biomolecular networks, and the evolutionary trajectories leading up to them.

3.1 Random Network Models

Erdös and Rényi Model

The Erdös and Rényi model of random graphs (ER-graphs, denoted $G(n,p)$) is characterized by two parameters, the number of vertices in the network N and the fixed probability of choosing edges p Erdös and Rényi (1959). The graph G is generated by choosing N vertices and connecting each pair of vertices with probability p. The model yields a network with approximately $p\binom{N}{2}=O(pN^2)$ randomly distributed edges. The probability of choosing a specified graph G with N vertices and e edges is therefore $\binom{M}{e}p^e(1-p)^{M-e}$, where $M=\frac{N}{2}=$ the maximum number of possible edges connecting N vertices.

It can be shown that in such random graph the average vertex degree is $\langle k\rangle=p(N-1)=O(pN)$. The diameter of such graph is $d=\ln N/\ln\langle k\rangle\approx\ln N/(\ln N-ln(1/p))$ which is small compared to the graph size. Thus, random graphs exhibit "the small world property." The degree distribution for ER graphs is a binomial distribution $P[\deg(u)=k]=\binom{(N-1)}{k}p^k(1-p)^{N-k-1}$, which for large N (relative to $1/p$: where $N=\lambda/p$) converges to the Poisson distribution $P[\deg(u)=k]=e^{-\lambda}\frac{\lambda^k}{k!}$. Then the local clustering coefficient is $C_i=p$ is independent of the degree of the node and the average clustering coefficient

$C = p/N$ scales with the network size. Therefore, the standard ER random model seems not to capture either the properties of degree distribution or the clustering coefficient of biomolecular networks.

Typically, an ER random graph model is used as a "null model" for the evolutionary process. However, while deviations from randomness are frequently used as evidence for the direct action of natural selection, often non-randomness may reflect neutrally generated (non-adaptive) emergent phenomena Massey (2015). We emphasize here that many topological features of biomolecular networks are unlikely to be directly selected for, but instead are a side-product of network growth, and decay, captured by the dynamics of edge and node addition and removal.

Small World Model

The biomolecular networks have features that are not captured by the Erdös and Rényi random graph model. As we have seen, random graphs have low clustering coefficient and they do not account for formation of hubs. To rectify some of these shortcomings, the *small world model* or popularly known as the *six degree of separation model* was introduced as the next level of complexity for probabilistic model with features that are closer to the real world networks Watts and Strogatz (1998); Watts (1999). The evolution and dynamics of such networks have been discussed in detail Watts (2003), in particular in the diseases propagation literature Dodds and Watts (2005).

In this model the graph G of N nodes is constructed as a ring lattice, in which, (i) first, *wire*: that is, connect every node to $K/2$ neighbors on each side and (ii) second, *rewire*: that is, for every edge connecting a particular node, with probability p reconnect it to a randomly selected node.

The average number of such edges is $pNK/2$. The first step of the algorithm produces local clustering, while the second dramatically reduces the distance in the network. Unlike the random graph, the clustering coefficient of this network $C = 3(K-2)/4(K-1)$ is independent of the system size. Thus, the small world network model displays the small world property and the clustering of real networks, however, it does not capture the emergence of hubby nodes (e.g., P53 in biomolecular networks).

3.2 Scale-free Network Models

Most biomolecular networks are hypothesized to have a degree distribution, described as *scale-free*. In a scale free network the number of nodes n_k of degree k is proportional to a power of the degree, namely, the degree distribution of the nodes follows a *power-law*

$$n_k = k^{-\beta}, \tag{1}$$

where $\beta > 1$ is a coefficient characteristic of the network Barabási and Albert (1999). Unlike in random networks, where the degree of all nodes is centered around a single value – with the probability of finding nodes with much larger (or smaller) degree decaying exponentially, in scale-free networks there are nodes of large degree with relatively higher probability (*fat tail*). In other words, since the power low distribution decreases much more slowly than exponential, for large k (heavy or fat tails), scale-free networks support nodes with extremely high number of connections called "hubs." Power law distribution has been observed in many large networks, such as the Internet, the phone-call maps, the collaboration networks, etc. Képès (2007); Barabási (2009); Loscalzo and Barabási (2016). A caveat to these reports is that inappropriate statistical techniques are often been used to infer power law distributions, and alternative heavy tailed distributions may fit the data better Clauset et al. (2009a). However, the power law is a useful approximation

that allows mechanisms of network growth to be explored, such as Preferential Attachment, discussed next, while the examination of alternative heavy tailed distributions is set as an Open Problem.

Preferential Attachment

The original model of *preferential attachment* was proposed by Barabási–Albert Barabási and Albert (1999). The scheme consists of a local *growth rule* that leads to a global consequence, namely a power law distribution. The network grows through the addition of new nodes linking to nodes already present in the system. There is higher probability to preferentially link to a node with a large number of connections. Thus, this rule gives more preferences to those vertices that have larger degrees. For this reason it is often referred to as the "rich-get-richer" or "Matthew" effect.

With an initial graph G_0 and a fixed probability parameter p, the preferential attachment random graph model $G(p, G_0)$ can be described as follows: at each step the graph G_t is formed by modifying the earlier graph G_{t-1} in two steps – with probability p take a *vertex-step*; otherwise, take an *edge-step*:

(i) *Vertex step:* Add a new vertex v and an edge $\{u, v\}$ from v to u by randomly and independently choosing u proportional its degree;

(ii) *Edge step:* Add a new edge $\{r, s\}$ by independently choosing vertices r and s with probability proportional to their degrees.

That is, at each step, we add a vertex with probability p, while for sure, we add an additional edge. If we denote by n_t and e_t the number of vertices and edges respectively at step t, then $e_t = t + 1$ and $n_t = 1 + \sum_{i=1}^{t} z_i$, where z_i's are Bernoulli random variables with probability of success $= p$. Hence the expected value of nodes is $\langle n_t \rangle = 1 + pt$.

It can be shown that exponentially (as t asymptotically approaches infinity) this process leads to a scale-free network. The degree distribution of $G(p)$ satisfies a power law with the parameter for exponent being $\beta = 2 + \frac{p}{2-p}$. Scale-free networks also exhibit *hierarchicity*. The local clustering coefficient is proportional to a power of the node degree

$$C(k) \approx k^{-\alpha} \tag{2}$$

where α is called the *hierarchy coefficient*.

This distribution implies that the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through hubs. In other words, it means that the level of clustering is much larger than that in random networks.

Consequently, many of the network properties in a scale-free network are determined by the local structures – namely, by a relatively small number of highly connected nodes (hubs). A consequence of this structure of the scale-free network is its extreme robustness to failure, a property also displayed by biomolecular networks and their modular structures. Such networks are highly tolerant of random failures (perturbations); however, they remain extremely sensitive to targeted attacks.

Assortativity Network Model

Assortative mixing refers to the property exhibited by a preference of nodes to attach to similar (respectively, dissimilar) nodes; for example, high-degree vertices exhibit preference to attach to high-degree (resp. low-degree) vertices. Network models, discussed earlier and including the preferential attachment model, do not capture such important properties exhibited by real biomolecular networks

Girvan and Newman (2002). Assortativity can be measured by the Pearson correlation coefficient r of degrees of linked nodes Girvan and Newman (2002). Positive correlation means connections between nodes of similar degree (assortativity) and negative correlation means connections between nodes with different degree (disassortativity). Unlike technological networks and social networks (showing assortative mixing), biological networks appear to evolve in a disassortative manner.

Many genetic networks, especially the DNA networks, lead to directed graphs. Assortative mixing can be generalized to directed biological graphs Piraveenan et al. (2012). For directed networks two new measures, in-assortativity and the out-assortativity , can be defined measuring the correlation between the in-degree r_{in} and out-degree r_{out} of the nodes respectively. Biological networks, which have been previously classified as disassortative, have been shown to be assortative with respect to these new measures. Also it has been shown that in directed biological networks, out-degree mixing patterns contain the highest amount of Shannon information, suggesting that nodes with high local out-assortativity (regulators) dominate the connectivity of the network Piraveenan et al. (2012). The occurrence of assortativity in social networks has been attributed to a process of homophily (that is people tend to associate with others on the basis of ethnicity, religion, sports preferences etc McPherson et al. (2001); Newman (2003a)). The mechanisms that give rise to assortativity in biomolecular networks likely arises by a similar proximate mechanism of like nodes forming edges with like nodes, but the ultimate cause(s) remains unclear.

Duplication Model

Our earlier discussions suggest that biomolecular networks exhibit power-law degree distribution. However, unlike other complex networks, such as the Internet, the growth exponent of biomolecular networks typically falls into a lower range $1 < \beta < 2$, as opposed to $\beta \geq 2$. This discrepancy has been suggested to have resulted from evolution by gene duplication dominating evolutionary mechanism Chung et al. (2003). Various biomolecular networks have been studied using a partial duplication process, which proceeds in the following manner: Let the initial graph G_0 have N_0 vertices. In each step, G_t is constructed from its previous graph G_{t-1} as follows: A random vertex u is selected. Then a new vertex v is added in such a way that for each neighbor w of u, a new edge (u, w) is added with probability p. The process is then applied repeatedly. The full duplication model is simply the partial model with $p = 1$.

It has been shown that as the number N of vertices becomes infinitely large, the partial duplication model with selection probability p generates power-law graphs with the exponent satisfying the transcendental equation Chung et al. (2003)

$$p(\beta - 1) = 1 - p^{\beta - 1},$$

whose solution determines the scale-free exponent β as a function of p. In particular, if $1/2 < p < 1$ then $\beta < 2$.

For illustrative purposes, we describe below an abstract gene network growth model incorporating the processes of gene duplication and deletion, as described above (Mishra and Zhou (2004) and Zhou (2005)). Using a Markov chain model the following features were investigated: (i) the origination of the segmental duplication; (ii) the effect of the duplication on the genome structure; and (iii) the role of duplication and deletion process in the genomic evolutionary distance. Unlike standard models of stationary Markov chain models, most processes in evolutionary biology belong to the group of non-stationary Markov processes, in which the transition matrix changes over time, or depends upon the current state.

This model results in the neutral emergence of scale-free degree distributions. It shows that the genomes of different organisms exhibit different network properties, likely reflecting differences in the rates of gene

duplication and deletion Mishra and Zhou (2004). This analysis provides an example of how network topology can be used to provide insight into fundamental molecular evolutionary (neutral/Markov) processes in different species. Note that the model is relatively idealized, as it does not account for higher order interactions in a population involving: effective population size and allelic fixations; sex, diploidy and sex-chromosomes (e.g., X and Y in mammals or W and Z in birds, etc.); surveillance and repair in somatic cells; embryonic lethality; homologous recombination, etc. The mathematical model explored here is kept simple to motivate the machinery from graph theory developed later.

Hierarchical Network Models

Another interesting model, introduced by Ravasz and Barabasi and dubbed *hierarchical network model*, simulates the characteristics of many real life complex models and may be relevant. The resulting networks have modularity, high degree of clustering, and scale-free property. Modularity refers to the network phenomenon where many sparsely inter-connected dense subgraphs can be identified – "one can easily identify groups of nodes that are highly interconnected with each other, but have only a few or no links to nodes outside of the group to which they belong to." (from Ravasz and Barabási (2003)).

A generative process for hierarchical network model may be described as follows: For instance, consider an initial network H_0 of c fully interconnected nodes (e.g., $c = 5$). As a next step, create $(c-1)$ replicas of this cluster H_0 and connect the peripheral nodes of each replica to the central node of the original cluster to create H_1 with c^2 (e.g., $c^2 = 25$) nodes. This step can be repeated recursively and indefinitely, thereby for any k steps the number of nodes generating the graph H_k with c^{k+1} nodes. If the central nodes of H_0 is called a *hub* and other nodes *peripheral*, then each recursion replicates additional copies of hubs and peripheral nodes.

One can carry out carry out a recursive analysis and shows that one obtains a power-law (i.e. scale-free) network with exponent $\beta = 1 + \frac{\ln(c)}{\ln(c-1)}$. The local clustering coefficients (for the hub-nodes) follow $C(k) \approx \frac{2}{k}$. Also, one can show that this duplication feature of evolution leads to hierarchical behavior of the network. The networks are expected to be fundamentally modular, in other words, the network can be seamlessly partitioned into collection of modules where each module performs an identifiable task, separate from the function(s) of other modules. One can also show that the average clustering coefficient on N nodes at any given stage is about $C = .7419282..$ (for $c = 4$), $C = 0.741840$ (for $c = 5$), and a constant for a fixed c, independent of N (see Ravasz and Barabási (2003), and for exact computations Noh (2003)). While for the preferential attachment model of Barbasi-Albert has the average clustering coefficient C on N nodes decreases as $1/N$, in addition not exhibiting modularity.

4 OPEN PROBLEMS AND FUTURE CHALLENGES

The study of biomolecular networks is still a relatively young field and has thus far focused on a mechanistic perspective. As we begin to explore it from an evolutionary view point, we encounter a large array of promising areas of investigation – most of which focuses on how information asymmetries among the gene players ultimately sculpt the information flow, as necessary for an organism to navigate in a complex and fluctuating environment. In particular, at its core this program requires an explanation of how features of genome evolution and structure might be algorithmically inferred from a network science perspective.

The traditional approaches of phylogenetic study may be applied here, but examining specifically the family of species-specific biomolecular networks. Thus mathematically we would need the networks to be aligned, motifs to be mapped to each other and network-distances to be correlated to deep evolutionary

time. In order to account for the evolution by duplications, *orthologs* and *paralogs* of a gene (or gene families) are to be identified and connected to their roles in biochemical pathways. Ultimately, this analysis could be targeted at extracting the origin of various information-asymmetric signaling games and how they stabilized in their Nash equilibria.

Network analysis is used in disease studies, but there have been more focused studies with applications to disease processes in cancer. In Figure 4 we show part of an interactome network useful in deciphering aberrant interactions in diseases (Figure 2.3 from Loscalzo et al. (2017)).

4.1 Algorithmic Complexity Issues

A key problem central to this program would be in detecting isomorphism mappings among pairs of graphs or subgraphs, a problem of infeasible algorithmic complexity (assuming $P \neq NP$.) We start with a discussion of these issues and cite heuristics that can tame the problem, albeit computing the solutions approximately.

Intractability: NP-Completeness

Many combinatorial optimization problems seem impossible to solve except by brute-force searches evaluating all possible configurations in the search space. They belong to a complexity class called NP-complete and include such problems as whether a graph has a clique of size k. Since finding certain shared motifs in a class of networks shares many computational characteristics of the clique problem and since it could be central to discovering important evolutionary signatures (e.g., EBD), it seems unlikely that it would be possible to characterize the evolutionary trajectories precisely – especially when the number of genes involved are in the thousands. See the supplement for additional discussions on graph representations and to derive their algebraic invariants, that provide bounds on complexity of algorithms possibly leading to excellent approximate results in the study of sparse complex networks (see Chung (1997); Chung and Lu (2006).

Problem 4.A

Classify various computational problems involved in detecting evolutionary trajectories of biomolecular networks and characterize their algorithmic complexity.

Problem 4.B

Explore PTAS (Polynomial Time Approximation Schemes) for these problems – Especially when the graphs satisfy certain sparsity, modularity and/or hierarchy properties.

Algebraic Approximation

As described earlier, many interesting topological features of a graph can be computed efficiently (on both sequential and parallel computers) from their descriptions in terms of adjacency matrices. The resulting spectral methods have found recent applications in complex networks (e.g., communication, social, Internet) (see Spielman (2018), Spielman (1996), Spielman and Teng (2014), Spielman and Teng (2013), Spielman and Teng (2011a), Spielman and Teng (2004) Chung and Lu (2006), Chung (1997), Chung (2010), MacKay (2003)). These methods are efficient (linear time complexity) for sparse graphs, whose number of edges is roughly of the same order as the number of vertices. Thus, they are well suited to biomolecular networks (for example for clustering, community detection, hubs, robustness, assortative mixing, spreading and mixing, closeness, isomorphism, among others).

Thus, spectral graph theory may be expected to have many applications in the analysis of biomolecular networks, most prominently, in clustering, graph similarity and graph approximation, but also in smoothing analysis and sparsification. One can envisage that many, if not most, classical network algorithms in biomolecular networks can be made faster by spectral methods. Indeed, since most biomolecular networks are sparse – both in terms of sparse connections, and in precise algebraic sense (see the supplementary section), these algorithms likely lead to linear time algorithms. The smoothing analysis methods, as well as sparsification approximations are worth exploring in these contexts.

Another fruitful direction is in parallelizing these algorithms. As an illustration, in several studies of biomolecular networks it would be useful to identify when two networks X_1 and X_2 are "close." We may wish to say that two networks are close if $Spec(X_1)$ and $Spec(X_2)$ are close – a computational problem that is polynomially computable (and efficiently parallelizable) (see Spielman and Teng (2013)). We can now give a mathematical formulation of this closeness, which can also be incorporated into phylogenetic studies. These biomolecular networks may be annotated with weights that are linear or quadratic approximation of relations, as common in these studies. These analyses may identify subnetworks that have been influenced by EBD, in concert with selection.

Problem 4.C

Classify various algebraic problems involved in detecting evolutionary trajectories of biomolecular networks and characterize their ability to approximate. Explore their practical implementations on sequential and parallel computers.

4.2 Design Principles via Motif Analysis

The study of Systems Biology postulates that there are important design principles of biological circuits that provide a great deal of insight. The connections of gene and protein interaction networks are assumed to provide the necessary robustness and control to achieve cellular function in the face of chemical noise. However, it remains unclear how random variations alone provide such robustness. A possible explanation may come from a game-theoretic model that lead to stable equilibria and is expected to have precipitated from duplication of genes, interactions and motifs.

Machine Learning

The biomolecular networks of interest are derived from highly noisy data e.g., CHIP-Chip, CHIP-Seq (for GRN) or co-localization or two-hybrid (for PPI) and consequently, the inferred edges of the network may miss certain genuine interactions or include several spurious interactions. Various machine learning algorithms (with fdr, false discovery rates, control and regularization techniques) have been devised in order to improve the accuracy of such models. Biomolecular networks from related species (with orthologs and paralogs analysis) are often combined to improve the accuracies and cross-validate results. The accuracies may be further ascertained via various local properties.

One important local property of networks are so-called network motifs, which are defined as recurrent and statistically significant sub-graphs or patterns. Thus, network motifs are sub-graphs that repeat themselves in a specific network or even among various networks. Each of these sub-graphs, defined by a particular pattern of interactions between vertices, may reflect a framework in which particular functions are achieved efficiently. Indeed, motifs are of notable importance largely because they may reflect functional properties. They have recently gathered much attention as a useful concept to uncover structural design principles of complex networks. Although network motifs may provide a deep insight into the network's functional abilities, their detection is computationally challenging. Thus an important challenge for both experimental

and computational scientists would be to study the evolutionary dynamics starting with the experimental data *ab initio*, as well as in improving the accuracy and efficiency of both the experimental and algorithmic techniques simultaneously.

Problem 4.D

Classify the species distributions of the different forms of heavy tailed distributions (e.g. power law, exponential, power law with exponential decay, lognormal), in different types of biomolecular network, and infer the mechanistic causes during network growth, and ultimate molecular evolutionary origins

Problem 4.E

Characterize the motifs in the biomolecular networks of closely related species starting with the noisy experimental data. Explain the structure of the motifs via their effect on the information flow. For instance, one may focus on DOR (Dense Overlapping Regulons) motifs and how they might have evolved from a simpler ancestral regulon Alon (2006).

Problem 4.F

Study Subgraph Isomorphism Algorithms (and heuristics) for sparse graphs and identify special cases most suitable for studying evolutionary trajectories, while relating them to biomolecular design principles.

Network Alignment

Critical to the evolutionary studies, described above, is the topic of network alignment and subsequent network tree building. Networks may be aligned in a pairwise fashion to calculate similarity, and from this a distance matrix calculated, and used for the construction of a network tree, showing the relationships between multiple networks. For example, in the case of meta-metabolic networks, such studies will reveal relationships between the meta-metabolic networks of different microhabitats. A plausible prediction is that the network tree should show convergent evolution in microbial communities from microhabitats with similar conditions (e.g., anaerobic habitats). Thus this approach could lead to a tool to study convergent evolution of microbial community structure in similar habitats Goldford et al. (2018).

From an algorithmic point of view, one may employ any of the three types of network alignment approaches:

1. where node identity is known;
2. where node similarity can be determined (based on sequence similarity for example); and
3. where node identity is unknown, here only network topology is used for alignment.

The first is a straightforward edge alignment. However, a refined expression is required that incorporates similarities in edge widths in addition to the basic edge alignment (presence / absence of common edges between networks). There do exist some first generation heuristics that utilize the second and third types of alignment approach (i.e., sequence similarity and topology, and only topology) Kuchaiev and Przulj (2011), but the underlying graph isomorphism problem is known to be #P-complete. But these heuristics, as would be expected, do not work well – a straightforward test for this problem is applying them to align the social networks of the Gospels of Luke and Matthew (Figure 3) - the Jesus node should always align, as it is rather obvious topologically; but often leads to failure.

Problem 4.G

Classify and characterize the graph alignment algorithms.

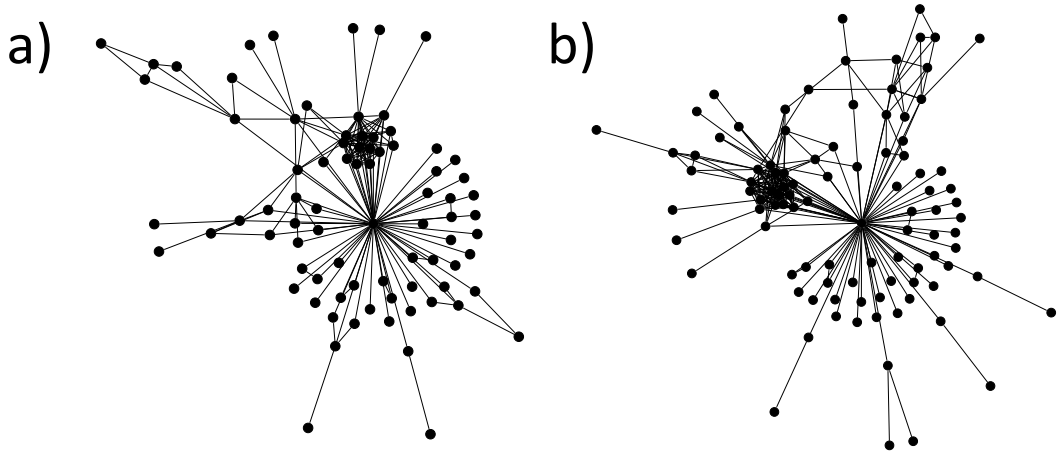

Figure 3. *Topological Alignment of Networks.* Similar Biomolecular networks could be topologically aligned and compared in order to express an evolutionary distance, which may then augment the traditional approaches of phylogenetic study. In order to account for the evolution by gene duplications, genes (or gene families) are to be identified and connected to their roles in biochemical pathways. Such an approach would lead to a program to understand the critical role of information asymmetries in driving evolution. Network alignment, a core problem in this program, is computationally intractable. To sharpen our intuition, we illustrate the problem using the social networks of the Gospels of Matthew and Luke. These networks represent social interactions between characters in the gospels of Matthew (a) and Luke (b). These were chosen as a basic test for topological alignment procedures, given that they share a similar number of nodes, and the highly connected node of Jesus. A straightforward test for the efficacy of a topological alignment algorithm therefore constitutes aligning both networks and verifying that the Jesus node from both networks is matched

4.3 Somatic Evolution and Cancer

Cancer is a complex disease, but governed by somatic genomic evolution, as propelled by mutation. Thus as a consequence GRNs may be used to better understand cancer susceptibility, map its progression, design better tailored therapies, and better understand the evolution of endogenous anti-cancer strategies. Cancer genes are often network hubs Karimzadeh et al. (2018), as they are often involved in critical developmental pathways. But a better network analysis will shed light on many natural questions: Why is it so? How does this come about from the process of network growth over evolutionary time? What clues do they provide to understand the somatic evolution in cancer and its progression?

During cancer progression, the disease reduces a cell's healthy genome into an aberrant mutant, where cancer eventually leads to metastasis, ultimately resulting in death of the patient. The healthy cells in the patient may be thought to possess a normal network, that is a gene network that engenders health and well-being. Cancer progression is reflected by a dynamic change of the normal network into an aberrant network. The aberrant network manifests itself by tumorigenesis, and finally metastasis. There is a substantial literature enumerating the identity of oncogenes and tumor suppressor genes, which aberrantly gain function (e.g., amplification of copy number) or lose function (e.g., deletion in copy number, hemi- or homozygously), respectively. They modify the cell biology of cancer progression, effected via the dynamics of GRN and PPI networks in cancer progression – all remain to be fully characterized.

Of particular interest is the question whether there is an identifiable phase transition in network topology associated with metastasis. Figure 2 shows a simple model for how the evolution of p53 and its paralogs may affect GRN topology; such molecular evolutionary signaling games approaches may help to better understand the motifs associated with oncogenes in GRNs. An additional important factor in cancer is the

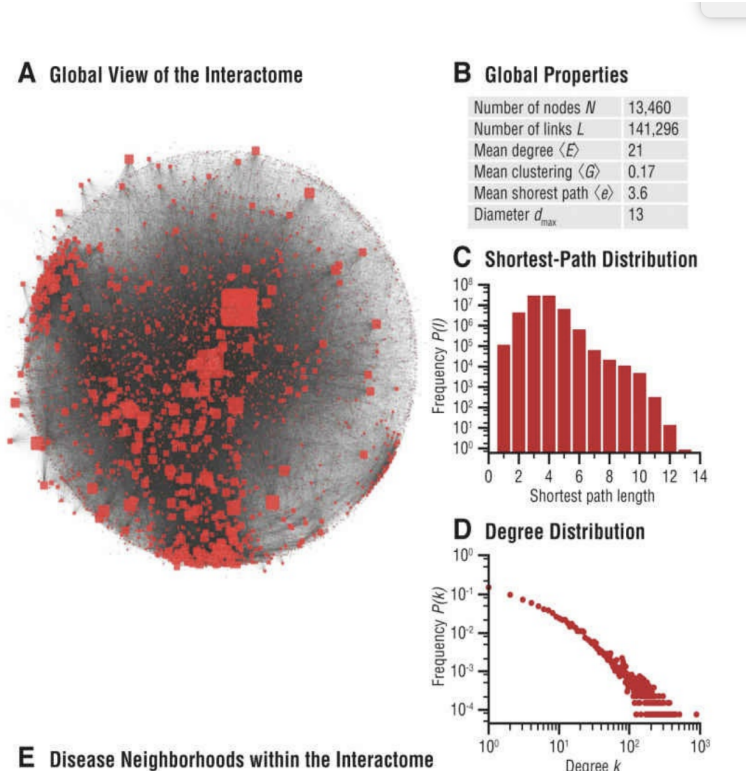

Figure 4. *Interactome Networks Used in the study of Diseases.* Undesirable interactions within a biomolecular network result in various disease states. Disease neighborhoods within the interactome can then be mapped to understand the progression of the disease. Progression of cancer have been studied using analysis of functionalization of oncogenes and dysfunctionalization of tumor suppression genes via copy number fluctuations, but much more can be learned from the topological features of these genes in their interaction neighborhood. (A) Global map of the interactome, illustrating its heterogeneity. Node sizes are proportional to their degree, that is, the number of links each node has to other nodes. (B) Basic characteristics of the interactome. (C) Distribution of the shortest paths within the interactome. The average shortest path is $\langle d \rangle = 3.6$. (D) The degree distribution of the interactome is approximately scale-free." (from Figure 2.3 in Loscalzo et al. (2017))

pervasive occurrence of molecular deception Bhatia and Kumar (2013), which from a signaling games perspective is consistent with cancer's conflict of interest with somatic cells. The identity of deceptive macromolecular signals may be incorporated into the network, potentially shedding a novel light on the mechanism of carcinogenesis. The genesis of deceptive signals therefore is expected to impact and drive carcinogenesis.

An additional factor to understanding this biology are copy number variants (CNVs) – types of gene mutations where a number of large sections of genomic DNA may be duplicated (or deleted), resulting in dosage effects of the resident gene sequences, which are exactly duplicated (or deleted). The numbers of CNVs can commonly vary substantially within a population, and have been shown to have significant roles in the propensity to develop cancer Krepischi et al. (2012). An increase in the number of CNVs would have the effect of enhancing the weight of an edge, which represents the interaction of the CNV gene product with its macromolecular binding partner. Such a network variant represents an increased disposition to develop cancer, and can be understood as occupying a position in 'network space' (the space of all possible network topologies) in greater proximity to an aberrant network, than a normal network.

Problem 4.H

Study Cancer progression models in terms of GRN's and identify the role of driver *and* passenger *genes in the somatically evolving networks.*

4.4 Gene Regulation and 3D Networks

In the genome of the ancestral life form, once a number of genes with separate function had evolved, it then would have become beneficial to evolve gene regulation. Therefore, genes with the dedicated function of regulating other genes in the genome would have arisen (transcription factors). The combination of regulatory and functional genes would have comprised the first gene regulation network. Increases in organismal complexity have been facilitated by an increase in the complexity of the gene regulation network Burton (2014).

Recent work has outlined the importance of three-dimensional proximity of genes to genes on other chromosomes, in addition to their immediate neighborhood on their own chromosome Li et al. (2018). This effect implies that gene proximity and spatial relationships within the nucleus can be meaningfully represented as a network. Such a network would be comprised of two types of edge: 1) linear distance on the same chromosome (centimorgans), 2) physical distance with genes on other chromosomes (nanometers). Such networks may be termed 3D gene orientation networks.

Gene regulation and co-regulation may be better understood by the construction and analysis of 3D gene orientation networks. This is because the proximity of regulatory modules to a gene has an influence of gene expression. Most genes have a regulatory region 5' of the transcription start site, the promoter. In addition, regulatory enhancers and other regulatory elements may be located distant from the gene, generally on the same chromosome Gondor and Ohlsson (2018). It is thought that the bending and juxtaposition of chromosomes within the nucleus may bring such elements into physical proximity to the gene Gondor and Ohlsson (2018). Clearly, the physical distance, and frequency with which the element is brought into contact with the gene will influence the nature of its regulatory input. Using 3D gene orientation networks, additional information may be incorporated into edges, such as whether physical proximity is static, or has movement. If there is movement, this may be coordinated (or not) with other regulatory elements affecting the same gene. Likewise, interactions with regulatory elements may show some coordination between genes.

Problem 4.I

Describe the Gene Duplication process and their utilities in terms of the genome's 3D structure.

5 CONCLUSION

Here, we have outlined graph theoretical approaches that may reveal some novel aspects of the molecular evolutionary process, which become manifest at the level of the phenome. Further work is required to link the diverse features of network topology with network evolution and growth. While the evolutionary aspects shaping individual gene-gene interactions has been addressed by geneticists and molecular evolutionists, we believe that a multi-disciplinary effort combining game theory, graph theory, and algebraic/statistical analysis will provide a more informative omnigenic model of gene interactions, in contrast to the traditional homogenic view. Given our view that biomolecular networks may be modeled using evolutionary game theory, and game theoretical approaches in the study of social networks, we expect that some surprising similarities and convergences between the topologies of the two might be observed. Finally, we note that

the field of statistics gained impetus from the consideration of biological problems, from workers such as Fisher, Haldane, Rao, Wright, Kimura, Crow and others, and so we suggest that consideration of the open problems listed here might also lead to a similar development of new mathematics.

6 BIBLIOGRAPHIC NOTES

We recommend the following articles for further reading: (Liu et al. (2013), Song et al. (2010), Davis et al. (2010), Vazquez et al. (2008), Candia et al. (2008), Goh and Barabási (2008), Barabási et al. (2004), Barabási et al. (2003), Barabási (2003), Farkas et al. (2002), Barabási et al. (2002), Schwartz et al. (2002), Albert and Barabási (2002)), Chung and Lu (2004), Chung and Lu (2006) and Janwa and Rangachari (2015). For other important sources (especially with respect to directed graphs), we refer to Zhang et al. (2017),Zhang et al. (2016), (Karrer and Newman (2010), Newman (2010), Clauset et al. (2009b), Moore et al. (2006), Newman (2006), Meyers et al. (2006),Newman (2004),Newman (2003c),Newman (2003d),Newman (2003b), Girvan and Newman (2002),Newman (2001),Newman and Watts (1999)), Newman et al. (2011). For network alignments and evolution of networks see for example Sharan et al. (2005); Pinter et al. (2005); Kalaev et al. (2008); Mazurie et al. (2010) . For bipartite networks (Hø holdt and Janwa (2012) and Janwa and Lal (2003)). For Spectral methods (Cvetković et al. (1980), Chung (1997),Spielman and Teng (2011a), Spielman and Teng (2011b),Chung and Lu (2006),Lubotzky (1994), Janwa and Rangachari (2015),Lubotzky et al. (1988),Sarnak (2004), Davidoff et al. (2003), and Lubotzky (2012)).

AUTHOR CONTRIBUTIONS

B.M. conceived of and structured the presented ideas at a high level. S.M. and B.M. developed the biological theories and H.J. & J.V. developed the computational, quantitative and mathematical theories. All authors discussed the open problems and contributed to the final manuscript.

FUNDING

This work was supported by National Science Foundation Grants CCF-0836649 and CCF-0926166, and a National Cancer Institute Physical Sciences-Oncology Center Grant U54 CA193313-01 (to B.M.).

ACKNOWLEDGMENTS

We acknowledge our colleagues in UPR and NYU, who have generously provided many constructive criticisms.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

DATA AVAILABILITY STATEMENT

The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY] [LINK].

REFERENCES

693 Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* 74,
694 47–97. doi:10.1103/RevModPhys.74.47
695 Alexander, J., Skyrms, B., and Zabell, S. (2012). Inventing new signals. *Dynamic Games and Applications*
696 2, 129–145
697 Alon, U. (2006). *An introduction to systems biology: design principles of biological circuits* (Chapman and
698 Hall/CRC)
699 Barabási, A.-L. (2003). Emergence of scaling in complex networks. In *Handbook of graphs and networks*
700 (Wiley-VCH, Weinheim). 69–84
701 Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *Science* 325, 412–413. doi:10.1126/
702 science.1173299
703 Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509–512.
704 doi:10.1126/science.286.5439.509
705 Barabási, A.-L., Dezső, Z., Ravasz, E., Yook, S.-H., and Oltvai, Z. (2003). Scale-free and hierarchical
706 structures in complex networks. In *Modeling of complex systems* (Amer. Inst. Phys., Melville, NY), vol.
707 661 of *AIP Conf. Proc.* 1–16. doi:10.1063/1.1571285
708 Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the
709 social network of scientific collaborations. *Phys. A* 311, 590–614. doi:10.1016/S0378-4371(02)00736-7
710 Barabási, A.-L., Oltvai, Z. N., and Wuchty, S. (2004). Characteristics of biological networks. In
711 *Complex networks* (Springer, Berlin), vol. 650 of *Lecture Notes in Phys.* 443–457. doi:10.1007/
712 978-3-540-44485-5_20
713 Belyi, V., Ak, P., Markert, E., Wang, H., Hu, A., W. Puzio-Kuter, and Levine, A. (2010). The origins and
714 evolution of the p53 family of genes. *Cold Spring Harbor Perspect Biol* 2, a001198
715 Bhatia, A. and Kumar, Y. (2013). Cellular and molecular mechanisms in cancer immune escape: A
716 comprehensive review. *Expert Rev Clin Immunol* 10, 758–762
717 Biggs, N. (1993). *Algebraic graph theory*. Cambridge Mathematical Library (Cambridge: Cambridge
718 University Press), second edn.
719 Burt, A. and Trivers, R. (2006). *Genes in Conflict: The Biology of Selfish Genetic Elements* (Cambridge,
720 Massachusetts: Harvard University Press)
721 Burton, Z. (2014). The old and new testaments of gene regulation. *Transcription* 5, e28674
722 Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering
723 individual and collective human dynamics from mobile phone records. *J. Phys. A* 41, 224015, 11.
724 doi:10.1088/1751-8113/41/22/224015
725 Chang, H., Pannunzio, N., Adachi, N., and Lieber, M. (2017). Non-homologous dna end joining and
726 alternative pathways to double-strand break repair. *Nature Reviews Molecular Cellular Biology* 18,
727 495–506
728 Chung, F. (2010). Graph theory in the information age. *Notices Amer. Math. Soc.* 57, 726–732
729 Chung, F. and Lu, L. (2004). The small world phenomenon in hybrid power law graphs. In *Complex*
730 *networks* (Springer, Berlin), vol. 650 of *Lecture Notes in Phys.* 89–104. doi:10.1007/978-3-540-44485-5_
731 4
732 Chung, F. and Lu, L. (2006). *Complex graphs and networks*, vol. 107 of *CBMS Regional Conference Series*
733 *in Mathematics* (Published for the Conference Board of the Mathematical Sciences, Washington, DC; by
734 the American Mathematical Society, Providence, RI). doi:10.1090/cbms/107
735 Chung, F., Lu, L., Dewey, T. G., and Galas, D. J. (2003). Duplication models for biological networks.
736 *Journal of computational biology : a journal of computational molecular cell biology* 10, 677–87

737 Chung, F. R. K. (1997). *Spectral graph theory*, vol. 92 of *CBMS Regional Conference Series in Mathematics*
738 (Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American
739 Mathematical Society, Providence, RI)
740 Clauset, A., Shalizi, C., and Newman, M. (2009a). Power-law distributions in empirical data. *SIAM Rev*
741 51, 661–703
742 Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009b). Power-law distributions in empirical data.
743 *SIAM Rev.* 51, 661–703. doi:10.1137/070710111
744 Cotterell, R., Vylomova, E., Khayrallah, H., Kirov, C., and Yarowsky, D. (2017). Paradigm completion
745 for derivational morphology. *Proceedings of the 2017 Conference on Empirical Methods in Natural*
746 *Language Processing* , 714–720
747 Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica* 50, 1431–1451
748 Cvetković, D. M., Doob, M., and Sachs, H. (1980). *Spectra of graphs*, vol. 87 of *Pure and Applied*
749 *Mathematics* (New York: Academic Press Inc. [Harcourt Brace Jovanovich Publishers]). Theory and
750 application
751 Davidoff, G., Sarnak, P., and Valette, A. (2003). *Elementary number theory, group theory, and Ramanujan*
752 *graphs*, vol. 55 of *London Mathematical Society Student Texts* (Cambridge University Press, Cambridge).
753 doi:10.1017/CBO9780511615825
754 Davis, D. A., Chawla, N. V., Christakis, N. A., and Barabási, A.-L. (2010). Time to CARE: a
755 collaborative engine for practical disease prediction. *Data Min. Knowl. Discov.* 20, 388–415.
756 doi:10.1007/s10618-009-0156-z
757 Demuth, J., De Bie, T., Stajich, J., Cristianini, N., and Hahn, M. (2018). The evolution of mammalian gene
758 families. *PLoS One* 1
759 Dodds, P. S. and Watts, D. J. (2005). A generalized model of social and biological contagion. *J. Theoret.*
760 *Biol.* 232, 587–604. doi:10.1016/j.jtbi.2004.09.006
761 Dokholyan, N., Shakhnovich, B., and Shakhnovich, E. (2002). Expanding protein universe and its origin in
762 from the biological big bang. *Proc Natl Acad Sci USA* 99, 14132–14136
763 Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae* 6, 290–297
764 Farkas, I., Derényi, I., Jeong, H., Néda, Z., Oltvai, Z. N., Ravasz, E., et al. (2002). Networks in life: scaling
765 properties and eigenvalue spectra. *Phys. A* 314, 25–34. doi:10.1016/S0378-4371(02)01181-0. Horizons
766 in complex systems (Messina, 2001)
767 Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc.*
768 *Natl. Acad. Sci. USA* 99, 7821–7826. doi:10.1073/pnas.122653799
769 Goh, K.-I. and Barabási, A.-L. (2008). Burstiness and memory in complex systems. *Europhys. Lett. EPL*
770 81, Art. 48002, 5. doi:10.1209/0295-5075/81/48002
771 Goldford, J., Lu, N., Bajic, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A., et al. (2018). Emergent
772 simplicity in microbial community assembly. *Science* 361, 1390–1396
773 Gondor, A. and Ohlsson, R. (2018). Enhancer functions in three dimensions: beyond the flat world
774 perspective. *F1000Research* 7, 681
775 Govindarajan, S. and Goldstein, R. (1997). Evolution of model proteins on a foldability landscape. *Proteins*
776 29, 461–466
777 Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell
778 biology. *Nature* 402, C47–52
779 Hawking, S. and Hertog, T. (2018). A smooth exit from eternal inflation? *arXiv:1707.07702*
780 Hø holdt, T. and Janwa, H. (2012). Eigenvalues and expansion of bipartite graphs. *Des. Codes Cryptogr.*
781 65, 259–273. doi:10.1007/s10623-011-9598-6

Huang, L., Liao, L., and Wu, C. (2017). Evolutionary analysis and interaction prediction for protein-protein interaction network in geometric space. *PLoS One* 12, e0183495

Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* 11, 97–108

Janwa, H. and Lal, A. K. (2003). On Tanner codes: minimum distance and decoding. *Appl. Algebra Engrg. Comm. Comput.* 13, 335–347. doi:10.1007/s00200-003-0098-4

Janwa, H. and Rangachari, S. (2015). Ramanujan graphs and their applications

Joerger, A. and Fersht, A. (2006). The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches. *Annual Review of Biochemistry* 85, 375–404

Kalaev, M., Smoot, M., Ideker, T., and Sharan, R. (2008). Networkblast: comparative analysis of protein networks 24, 594–596. doi:10.1093/bioinformatics/btm630. Exported from https://app.dimensions.ai on 2018/11/21

Karimzadeh, M., Jandaghi, P., Papadakis, A., Trainor, S., Gonzalez-Porta, M., Scelo, G., et al. (2018). Aberration hubs in protein interaction networks highlight actionable targets in cancer. *Oncotarget* 9, 25166–25180

Karp, R. M. (2011). Heuristic algorithms in computational molecular biology. *J. Comput. System Sci.* 77, 122–128. doi:10.1016/j.jcss.2010.06.009

Karrer, B. and Newman, M. E. J. (2010). Message passing approach for general epidemic models. *Phys. Rev. E (3)* 82, 016101, 9. doi:10.1103/PhysRevE.82.016101

Képès, F. e. (2007). *Biological Networks*. Complex Systems and Interdisciplinary Science (World Scientific)

Krepischi, A., Pearson, P., and Rosenberg, C. (2012). Germline copy number variations and cancer predisposition. *Future Oncology* 8, 681

Kuchaiev, O. and Przulj, N. (2011). Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics* 27, 1390–1396

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research* 12, 1048–1059

Lewis, D. (1969). *Convention: a philosophical study* (Cambridge: Harvard University Press)

Li, Y., Hu, M., and Shen, Y. (2018). Gene regulation in the 3D genome. *Human Mol Genet* 27, R228–233

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2013). Observability of complex systems. *Proc. Natl. Acad. Sci. USA* 110, 2460–2465. doi:10.1073/pnas.1215508110

Loscalzo, J. and Barabási, A.-L. (2016). *Network Science* (Cambridge University Press, Cambridge), 1st edn.

Loscalzo, J., Barabási, A.-L., and Silverman, E. K. e. (2017). *Network Medicine: Complex Systems in Human Disease and Therapeutic* (Harvard University Press), 1st edn.

Lu, W.-J., Amatruda, J., and Abrams, J. (2009). p53 ancestry: gazing through an evolutionary lens. *Nature Reviews Cancer* 9, 758–762

Lubotzky, A. (1994). *Discrete groups, expanding graphs and invariant measures*, vol. 125 of *Progress in Mathematics* (Basel: Birkhäuser Verlag). With an appendix by Jonathan D. Rogawski

Lubotzky, A. (2012). Expander graphs in pure and applied mathematics. *Bull. Amer. Math. Soc. (N.S.)* 49, 113–162. doi:10.1090/S0273-0979-2011-01359-3

Lubotzky, A., Phillips, R., and Sarnak, P. (1988). Ramanujan graphs. *Combinatorica* 8, 261–277. doi:10.1007/BF02126799

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms* (Cambridge University Press, New York)

827 Massey, S. (2015). Genetic code evolution reveals the neutral emergence of mutational robustness, and
828 information as an evolutionary constraint. *Life* 5, 1301–1332
829 Massey, S. and Mishra, B. (2018). Origin of biomolecular games: deception and molecular evolution. *J*
830 *Royal Soc Interface* 15, 20180329
831 Mazurie, A., Bonchev, D., Schwikowski, B., and Buck, G. A. (2010). Evolution of metabolic network
832 organization. *BMC Systems Biology* 4, 59. doi:10.1186/1752-0509-4-59
833 McCloskey, D., Palsson, B., and Feist, A. (2013). Basic and applied uses of genome-scale metabolic
834 network reconstructions of escherichia coli. *Mol Syst Biol* 9, 661
835 McPherson, M., Smith-Lovin, L., and Cook, J. (2001). Birds of a feather: homophily in social networks.
836 *Ann Rev Sociol* 27, 415–444
837 Meyers, L. A., Newman, M. E. J., and Pourbohloul, B. (2006). Predicting epidemics on directed contact
838 networks. *J. Theoret. Biol.* 240, 400–418. doi:10.1016/j.jtbi.2005.10.004
839 Mishra, B. and Zhou, Y. (2004). *Models of genome evolution* (Springer Verlag). Natural Computing Series,
840 Lecture Notes in Computer Science. 287–304
841 Moore, C., Ghoshal, G., and Newman, M. E. J. (2006). Exact solutions for models of evolving networks
842 with addition and deletion of nodes. *Phys. Rev. E (3)* 74, 036121, 8. doi:10.1103/PhysRevE.74.036121
843 Newman, M. (2003a). Assortative mixing in networks. *Phys Rev Lett* 89, 758–762
844 Newman, M., Barabasi, A.-L., and Watts, D. J. (2011). *The structure and dynamics of networks* (Princeton
845 University Press)
846 Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*
847 98, 404–409. doi:10.1073/pnas.021544898
848 Newman, M. E. J. (2003b). Mixing patterns in networks. *Phys. Rev. E (3)* 67, 026126, 13. doi:10.1103/
849 PhysRevE.67.026126
850 Newman, M. E. J. (2003c). Random graphs as models of networks. In *Handbook of graphs and networks*
851 (Wiley-VCH, Weinheim). 35–68
852 Newman, M. E. J. (2003d). The structure and function of complex networks. *SIAM Rev.* 45, 167–256.
853 doi:10.1137/S003614450342480
854 Newman, M. E. J. (2004). Who is the best connected scientist? A study of scientific coauthorship
855 networks. In *Complex networks* (Springer, Berlin), vol. 650 of *Lecture Notes in Phys.* 337–370.
856 doi:10.1007/978-3-540-44485-5_16
857 Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices.
858 *Phys. Rev. E (3)* 74, 036104, 19. doi:10.1103/PhysRevE.74.036104
859 Newman, M. E. J. (2010). *Networks* (Oxford University Press, Oxford). doi:10.1093/acprof:oso/
860 9780199206650.001.0001. An introduction
861 Newman, M. E. J. and Watts, D. J. (1999). Renormalization group analysis of the small-world network
862 model. *Phys. Lett. A* 263, 341–346. doi:10.1016/S0375-9601(99)00757-4
863 Nogales, E., Louder, R., and He, Y. (2017). Structural insights into the eukaryotic transcription initiation
864 machinery. *Ann Rev Biophys* 46, 59–83
865 Noh, J. D. (2003). Exact scaling properties of a hierarchical network model. *Phys. Rev. E* 67, 045103.
866 doi:10.1103/PhysRevE.67.045103
867 Ohno, S. (1970). *Evolution by gene duplication* (Berlin: Springer-Verlag)
868 Pellegrini, M. (2019). Community detection in biological networks. In *Encyclopedia of Bioinformatics*
869 *and Computational Biology*, eds. S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford:
870 Academic Press). 978 – 987. doi:https://doi.org/10.1016/B978-0-12-809633-8.20428-7

Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M. (2005). Alignment of metabolic pathways. *Bioinformatics* 21, 3401–3408. doi:10.1093/bioinformatics/bti554

Piraveenan, M., Prokopenko, M., and Zomaya, A. Y. (2012). On congruity of nodes and assortative information content in complex networks. *Netw. Heterog. Media* 7, 441–461. doi:10.3934/nhm.2012.7.441

Poulos, R., Sloane, M., Hesson, L., and Wong, J. (2015). The search for *cis*-regulatory driver mutations in cancer genomes. *Oncotarget* 6, 32509–32525

Ravasz, E. and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E* 67, 026112

Rodgers, K. and McVey, M. (2016). Error-prone repair of dna double-strand breaks. *J Cell Physiol* 231, 15–24

Sarnak, P. (2004). What is. . . an expander? *Notices Amer. Math. Soc.* 51, 762–763

Schuster, P., Fontana, W., Stadler, P., and Hofacker, I. (1994). From sequences to shapes and back: A case-study in rna secondary structures. *Proc R Soc Lond B* 255, 279–284

Schwartz, N., Cohen, R., ben Avraham, D., Barabási, A.-L., and Havlin, S. (2002). Percolation in directed scale-free networks. *Phys. Rev. E (3)* 66, 015104, 4. doi:10.1103/PhysRevE.66.015104

Serre, J.-P. (1980). *Trees* (Berlin: Springer-Verlag). Translated from the French by John Stillwell

Shapley, L. (1969). A value for n person games. In *The Shapley Value* (Cambridge: Cambridge University Press)

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., et al. (2005). Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences* 102, 1974–1979. doi:10.1073/pnas.0409522102

Smeenk, L., van Heeringen, S., Koeppel, M., van Driel, M., Bartels, S., Akkers, R., et al. (2008). Characterization of genome-wide p53-binding sites upon stress binding. *Nuc Acids Res* 36, 3639–3654

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science* 327, 1018–1021. doi:10.1126/science.1177170

Spielman, D. (2018). *Spectral Graph Theory and Its Applications* (http://www.cs.yale.edu/homes/spielman)

Spielman, D. A. (1996). Linear-time encodable and decodable error-correcting codes. *IEEE Trans. Inform. Theory* 42, 1723–1731. doi:10.1109/18.556668. Codes and complexity

Spielman, D. A. and Teng, S.-H. (2004). Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. *J. ACM* 51, 385–463. doi:10.1145/990308.990310

Spielman, D. A. and Teng, S.-H. (2011a). Spectral sparsification of graphs. *SIAM J. Comput.* 40, 981–1025. doi:10.1137/08074489X

Spielman, D. A. and Teng, S.-H. (2011b). Spectral sparsification of graphs. *SIAM J. Comput.* 40, 981–1025. doi:10.1137/08074489X

Spielman, D. A. and Teng, S.-H. (2013). A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM J. Comput.* 42, 1–26. doi:10.1137/080744888

Spielman, D. A. and Teng, S.-H. (2014). Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.* 35, 835–885. doi:10.1137/090771430

Taylor, P. and Jonker, L. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–156

Thompson, D., Regev, A., and Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annual Review of Cell and Developmental Biology* 31, 399–428

916 Thulasiraman, K., Arumugam, S., Brandstadt, A., and Nishizeki, T. (2015). *Handbook of Graph Theory,*
917 *Combinatorial Optimization, and Algorithms*. Chapman & Hall/CRC Computer and Information Science
918 Series (Taylor & Francis)
919 Vazquez, A., de Menezes, M. A., Barabási, A.-L., and Oltvai, Z. N. (2008). Impact of limited solvent
920 capacity on metabolic rate, enzyme activities, and metabolite concentrations of *s. cerevisiae* glycolysis.
921 *PLoS Comput. Biol.* 4, e1000195, 6. doi:10.1371/journal.pcbi.1000195
922 Wagner, A. (1994). Evolution of gene networks by gene duplications: a mathematical model and its
923 implications on genome organization. *Proc Natl Acad Sci USA* 91, 4387–4391
924 Watts, D. J. (1999). *Small worlds*. Princeton Studies in Complexity (Princeton University Press, Princeton,
925 NJ). The dynamics of networks between order and randomness
926 Watts, D. J. (2003). *Six degrees* (W. W. Norton & Co. Inc., New York). The science of a connected age
927 Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393,
928 440–442
929 Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011). ipath2.0: interactive pathway
930 explorer. *Nuc Acids Res* 39, W412–W415
931 Zhang, P., Moore, C., and Newman, M. E. J. (2016). Community detection in networks with unequal
932 groups. *Phys. Rev. E* 93, 012303, 12. doi:10.1103/physreve.93.012303
933 Zhang, X., Moore, C., and Newman, M. E. J. (2017). Random graph models for dynamic networks. *Eur.*
934 *Phys. J. B* 90, Paper No. 200, 14. doi:10.1140/epjb/e2017-80122-8
935 Zhang, Z., Luo, Z., Kishino, H., and Kearsey, M. (2005). Divergence pattern of duplicate genes in
936 protein-protein interactions follows the power law. *Mol Biol Evol* 22, 501–505
937 Zhou, Y. J. (2005). *Statistical Analyses and Markov Modeling of Duplication in Genome Evolution*. Thesis
938 (Ph.D.)–New York University (NYU)