

(?) Comparative Netomics - lessons from cross-disciplinary network comparison

TITLE

A signature of biology in the "omic" era is the shift of attention from a few individual components to comprehensive collections of constituents [1]. In the past structural biologists studied the binding of a few proteins, but now they are able to probe the interactions between thousands of proteins. Similarly, geneticists who would previously knockout a single gene for functional characterization can now employ high-throughput techniques in functional genomics to study the genetic relationships between all genes. In many cases, genome-scale information describing how components interact can be captured by a network representation [2]. While researchers have been astonished by the complexity of such networks found in genomics or systems biology, many are not able to gain any intuition from these hairballs [3].

What approaches might help decipher these hairballs? Throughout the history of science, many advances in biology were catalyzed by discoveries in other disciplines. For instance, the maturation of X-ray diffraction facilitated the discovery of the double helix, and later on, the characterization of structures of thousands of different proteins. One may wonder if ideas in other areas of science could help us to decipher the hairballs. In this essay, we argue that, while the influx of ideas in the age of reductionism mostly originated from specific subfields of physics or chemistry, to understand biology via a systems perspective, we need a new wave of catalysts coming from disciplines as diverse as engineering, behavioral science and sociology. These new ideas are centered on the concept of network. Toward this end, biologists should think about performing cross-disciplinary network comparison.

Drawing analogies is not new to biologists. For instance, to illustrate the principles of selection Dawkins came up with the idea of meme, which is a unit carrying cultural ideas analogous to a gene in biology [4]. This comparison has been further elaborated in the protofield of phylomemetics, which concerns itself with phylogenetic analysis of non genetic data [5]. Nevertheless, comparing a bio-molecular network with a complex network from a disparate field, say a social network, may sound like comparing apples to oranges. What kinds of comparison could truly deepen our understanding? We believe that it is useful to think of different descriptions of a cellular system as a spectrum.

**A spectrum of cellular descriptions**

Given the complexity of a cell, a certain level of simplification is necessary for useful discussion. The depth of description of cellular systems can be seen as a spectrum (Figure 1). On one extreme, there is a complete three or four-dimensional picture of how cellular molecules interact in space and time. On the other extreme, there is a simple parts list that simply enumerates each component without specifying any relationships. However, the parts list description is not fully informative. It is widely appreciated that the characteristics of a cellular system cannot be explained by the characteristics of individual components – the whole is greater than the sum of its parts. To describe the full picture, one would need the 3D structures of everything in the genome as well as representation of their dynamical movements. This level of detail is often too ambitious for the current state-of-the-art in data acquisition.

The network representation sits conveniently between these extremes. It captures some of the relationships between the components on the parts list in a flexible fashion, especially those where topology rather than exact location determines the consequence. There are two particularly useful ways to think about networks. One can view a network as an association network, which is essentially a process of abstraction in which entries are connected via purely mathematical association. Any mechanistic interaction could be abstracted as a mathematical association. However, the idea of association can be generalized to statistical relationships between two components. This approach is exemplified by disease networks [6] in which a gene (genotype) and a disease (phenotype) are connected via the statistical association between the existence of genomic variants and the occurrence of the disease. Networks derived from co-expression relationships provide another example. On the other hand, mechanistic networks represent a process of concretization. Unlike abstract association networks that move away from

CAN

ON THE OTHER EXTREME

CONNECTIVITY

SOCIOLOGY


SEE REWRITE

COMPONENTS &

KKY 10/19/2014 3:42 PM Deleted: abstract

KKY 10/19/2014 3:42 PM Deleted: Meanwhile


the complete 4D-picture, concrete mechanistic networks aim to more completely describe this picture. Mechanistic networks are intended to describe and integrate many of the physical processes happening inside a living system, for instance the processing of information, the chemistry of metabolites and the assembly of molecular machine and therefore focus on incorporating various details of interactions. Adding further mechanistic detail onto a simple nodes-and-edges skeleton can often be visualized by decorating edges with directionality, color, thickness etc. However, incorporating too much detail makes the system intractable. The network formalism generally breaks down if we try to load spatial or temporal details as well as higher-order interactions onto the diagram. At certain point, the actual four-dimensional picture is required.

On one hand, abstract association networks offer transferrable mathematical formalisms. Toward this end, by comparing similar network-based mathematical formalisms across disciplines, biologists will benefit in terms of algorithms or method development. On the other hand, mechanistic networks can serve as the skeletons for describing different complex systems in detail. In this case, because of systems-specific details, it is less likely that everything could be transferred from one discipline to another. Instead, it is important to focus on  conceptual resemblance instead of merely topological resemblance. Comparison of appropriately matched networks can enable biologists to gain intuition into the interactions between molecular components of cells by examining analogous interactions in cross-disciplinary complex systems.

### Comparison leverages mathematical formalism

The power of the network formalism lies in its simplicity. In the era of Big Data, the network is a very useful data structure with a wide variety of applications in both biology and other data intensive disciplines like computational social science. This is particularly true for abstract association networks.

#### *Formalism focusing on network topology*

One of the key applications of the abstract network formalism is  compar<sup>ing</sup> the organizing principles of various complex systems. The earliest and probably most important observation is that networks organize themselves into scale free architectures in which a majority of the nodes contain very few connections (edges) while a few nodes (also called hubs) in the network are highly connected [7]. The behavior of scale-free networks is dominated by a relatively small number of nodes and this ensures that such networks are resistant to random accidental failures but are vulnerable to coordinated attacks at hub nodes [8]. In other words, just as the Internet functions without any major disruptions even though hundreds of routers malfunction at any given moment, different individuals belonging to the same biological species remain healthy in spite of considerable random variation in their genomic information. However, a cell is not likely to survive if a hub protein is knocked out. For example, highly connected proteins in the yeast protein-protein interaction network are three-times more likely to be essential than proteins with only a small number of links to other proteins [9]. A scale-free network is a kind of small-world network because hubs ensure that the distance between any two nodes in the network is small.[10][11]. For example, the presence of hubs in the airport network makes it possible to travel between any two cities in the world within a short interval of time. An example of a small-world network that is not scale-free is the mammalian cerebral cortex. The cortical neuronal network is subdivided into more than 100 distinct, highly modular, areas [12] that are dominated by connections internal to each area, with only ~20% of all connections being between neurons in different areas [13]. Each area is considered to have a primary feature, for example in processing sensory or cognitive signals, and is an excellent analogue of the modular characteristics of intra-cellular molecular networks in which proteins in tightly controlled functional groups coordinate as part of larger pathways to achieve well defined cellular functions. The cortical architecture has a high degree of clustering and small path-length and exhibits an exponential degree-distribution [14].

Though counting the number of neighbors is very useful in determining the centrality of a node, a more sophisticated way to define centrality is to take into account the importance of neighbors. The PageRank algorithm is a prominent example of this approach. Faced with a search query,

[Google must decide which set of results to rank higher and place on the first results page. Originally developed in social network analysis \[15\], PageRank utilizes an algorithm developed to rank relevant documents based on the rank of the websites that link to this document in a self-consistent manner - i.e. being linked to by higher ranking nodes has a larger impact on the document's ranking. This algorithm was then adopted in food webs to prioritize nodes that are in danger of extinction \[16\] and was also used to rank prognostic relevance for patients with cancers \[17\].](#) A second method of measuring a node's centrality is based on the number of paths passing through it. Similar in spirit to heavily used bridges, highways, or intersections in transportation networks, a few centrally connected nodes termed bottlenecks funnel most of the paths between different parts of the network and removal of these nodes could reduce the efficiency (increase the distance) of communication between nodes within these networks [18]. Indeed, it has been reported that changes to the sequences of bottlenecks in biological networks can be deleterious [19]. Apart from degrees and paths, one can easily observe that social networks tend to have communities within them due to the relatively larger number of interactions between people in the same neighborhood, school, or work place. People within the same social group tend to form strong ties in the form of cliques and form a single cohesive group. Analogous to closely-knit social groups, a large number of biological components can form a single functional macromolecular complex like the ribosome. More generally, a common feature of a large number of social, technological and biological networks is that they are composed of modules such that nodes within the same module have a larger number of connections with each other as compared to nodes belonging to different modules [20]. The quantity dubbed modularity tries to quantify this, comparing the number of intra and inter module links in a network.

#### *Formalisms focusing on the interplay between topologies and the properties of nodes*

Networks are useful in data science because they can be used as a reference for mapping additional properties or features of different nodes. Similar questions and solutions have arisen from dealing with biological data as well as data from disciplines like computational social science. An important example is the inference of missing data using the idea of "guilt by association", or the idea that nodes that have similar associations in the network tend to be similar in nature. For example, in a social context, if your friends in Facebook use Product Y, you are more likely to use product Y and the advertisements you view online are personalized based on these recommendation systems [21]. In a biological context, this assumption is based on observations that cellular components within the same module are more closely associated with the same set of cellular phenotypes than components belonging to different modules [22]. Furthermore, modules within gene coexpression networks also tend to contain genes that are in the same biological pathway or have similar functions [23]. As a result, one could infer the functions of a protein or a non-coding element based on the function of its neighbors in the underlying network. Networks play an important role in gene prioritization, an essential process for applications like disease gene discovery because of limited validation and characterization resources [24]. For example, network properties of individual genes have been used to distinguish functionally essential and loss-of-function tolerant genes [25]. One could prioritize the candidate genes based on how they are connected to the known genes. If a gene is one step away from a group of disease genes, it is very likely that the gene is associated with disease X. The influence of a node may not be restricted to its nearest neighbors; network flow algorithms are widely used to examine the long-range influence [26][27]. In a social science context, researchers use cascade structured models to capture the information propagation on web blog networks, and predict the blog's popularity [28].

High-throughput experiments can be quite noisy. The resultant networks may contain spurious links, and missing data is very common. Methods for link prediction and denoising are therefore very useful. Link prediction can be done using network information alone. For instance, in a protein-protein interaction network, defective cliques were used to find missing interactions and determine the parts required to form a functional macromolecular complex [29]. Whether two nodes are connected depends on their intrinsic properties. Consequently, one could employ machine-learning techniques to explore the relationships between connections and various features [30]. Recently, generative models of networks, such as stochastic block models [31],

KKY 10/19/2014 3:42 PM

Deleted: extremely

KKY 10/19/2014 3:42 PM

Deleted: when

have been popular in computational social science. Nevertheless, such models are not yet widely used in biological context, presumably because of the lack of gold standards for validation.

#### *Formalisms focusing on causal relationships and dynamics*

The construction of various association networks is an active area of research for both biology and computational social science. While correlational relationships could be easily calculated with the appropriate data, a fundamental question is the distinction between direct and indirect interactions. For example, if transcription factor X regulates gene Y and Z, one would expect pairs like X-Y, X-Z and Y-Z to be correlated, but the key is to identify the direct regulatory interactions X-Y and X-Z. Established mathematical machineries like Bayesian networks, Markov random fields and other information theoretical frameworks [32] have been used for this purpose.

The inference of causal relationships is greatly improved by time-series data. In social science, online retailers are interested in using purchase records to study how customers influence each other [33]. The same question is extremely common in biology, under the term "reverse engineering". For example, how can we infer the developmental gene regulatory network from temporal gene expression dynamics? Ideally, one could write differential equations to fit the temporal data. However, most genomics experiments do not contain enough time-points. To overcome this drawback, data mining techniques such as matrix factorization are employed. For instance, given the genome-wide expression profile at different time-points, one could project the high-dimensional gene expression data to low dimensional space and write differential equations to model the dynamics of the projections [34]. The inference of casual and direct relationships from statistical data points to the study of mechanistic networks.

KKY 10/19/2014 3:42 PM

Deleted: would be

In addition to the actual dynamic processes occurring in a network, one can explore the evolutionary dynamics of networks by comparing networks. In a biological context, pairs of orthologous genes (nodes) can be used to define conserved edges like interologs and regulogs. Furthermore, orthologous genes have been used to align networks from different species [35], and to detect conserved and specific functional modules across species [36]. More formally, a mathematical formalism has been developed to measure the evolutionary rewiring rate between networks across species using methods analogous to those quantifying sequence evolution. It was shown that metabolic networks rewire at a slower rate compared to various regulatory networks [37].

KKY 10/19/2014 3:42 PM

Deleted: could

#### **Comparison gains physical intuition**

Now we shift discussion to "mechanistic" networks. Here, the network framework serves as a skeleton for different complex systems. From the standpoint of a biologist, network comparison can bring intuition from other disciplines to bear on molecular biology.

#### *Looking for mechanistic insights*

The previous sections discussed universal frameworks and insights gained by applying the same formalisms to biological networks as well as to various social and technological networks. Such wide-ranging universal insights were possible only because the detailed identities of the nodes in the networks were neglected during the comparison. Only the abstracted "association" between the various nodes was considered. On the other hand, if details are added to this picture, insights about a system become more specific, and in a sense, more meaningful. However, it is in general harder to apply the same formalism to two networks. This transition from applying general formalisms to abstract networks to more mechanistic descriptions is well described when one tried to explain the scale-free degree distribution of various networks described above.

A number of different stochastic models and explanations can lead to the formation of scale-free graphs. First let's consider the hub-and-spoke system of the airline network, one of the paradigms of scale-free structure. How does this come about? Every time a new airport is created, the airlines have to create a balance between the resources and customer satisfaction, i.e., the cost of adding a new flight and customer comfort due to connectivity between the new airport and a larger number of airports. The most efficient use of these limited resources occurs if the new

airport connects to pre-existent hubs in the network as it reduces the travel time of the average customer. This model is called the preferential attachment as newly created nodes prefer to connect to pre-existent hubs in the network [7] and, in this case, it emphasizes the small-world property of scale-free networks [11]. In contrast, one explains the evolution and growth of the WWW, which is also scale free, in somewhat different way. Here, a random pre-existing node and its associated edges (for example, a webpage with all its pre-existing links) are duplicated [38]. After duplication, subtle changes to the connectivity pattern of both nodes may occur such that a large proportion of their edges are likely to be shared [39]. Such a duplication-divergence model leads to the formation of scale-free networks because the connectivity of a hub increases as one of its neighbors has a higher chance of getting duplicated. The same duplication-divergence mechanism can describe the patterns and occurrence of "memes" in online media [40]. As gene duplication is one of the major mechanisms for the evolution of protein families, the formation of scale-free behavior in the protein-protein interaction network was proposed to evolve via the duplication-divergence model [39][41]. However, for protein networks there are additional twists in this explanation because one can actually resolve each of the nodes in the network as molecules with specific 3D geometry. In particular, upon analyzing the structural interfaces involved in protein-protein interactions, there are great differences in hubs that interact with many proteins by reusing the same structural interface versus those that simultaneously use many different interaction interfaces. The duplication divergence model only applies to the former situation (with the duplicated protein reusing the same interface as its parent) [42].

Common scale free topology in biological and other networks lead to the emergence of universal patterns in biological and other complex systems. It has been reported that the frequency of appearance of individual enzymes across different bacterial genomes and the frequency of local installations of individual packages in multicomponent software projects follow a broad distribution [43]. Recently, it has been suggested that the observations can be explained by the scale free topology of the corresponding multi-levels dependency networks (enzyme A is connected to enzyme B if A is used to decompose the output metabolites of enzyme B; package A is connected to package B if the installation of package A depends on the installation of package B), because incorporation of an additional component requires the presence of the depending factors in the network [43].

Many networks that exhibit similar topologies are the result of significantly different underlying mechanisms. In the case of scale free networks, there exists a common mathematical formalism but somewhat different mechanistic explanations in many different domains (e.g. airline networks vs gene networks). Some of the domains share the same mechanistic explanation -- i.e. the scale-free structure in both protein-protein interaction and web-link networks can be explained by duplication and divergence. Moreover, this later commonality provides additional intuition about the protein interaction network through comparison to the web-link network, which is more commonplace and better connected with the average person's experience.

#### *Looking for common design principles*

The ability to gain intuition about the often-arcane world of molecular biology by comparison to commonplace systems is even stronger for comparisons with social networks, where people have very strong intuition for how a system can work. Transferring an intuitive understanding network hierarchy is a good example of this type of comparison (see Box 1). Many biological networks, such as transcription regulatory networks, have an intrinsic direction of information flow, forming a loose hierarchical organization. Likewise, many social structures are naturally organized into a hierarchical structure -- e.g. a military command chain or a corporate "org-chart" [44]. In the purest form of the military hierarchy each person on a lower-level reports to a single individual on a higher level and there are fewer and fewer individuals on the upper levels, eventually culminating in a single supreme commander at the top ruling over an entire army. This structure naturally leads to information flow bottlenecks as all the orders and information related to many low-rank privates must flow through a very limited number of mid-level corporeal and majors. In a biological hierarchy of TFs, one sees a somewhat similar pattern with a high betweenness of

Koon-Kiu Yan 10/19/2014 5:10 PM

**Deleted:** . The essence is, similar structures of these two distinct networks are due to incorporate the fact that

KKY 10/19/2014 3:42 PM

**Deleted:** ; one has to ensure

Koon-Kiu Yan 10/19/2014 5:11 PM

**Deleted:** Yet another mechanistic model giving rise to a power-law distribution is the components usage in both bacterial genomes and in software systems.

KKY 10/19/2014 3:42 PM

**Deleted:** For

KKY 10/19/2014 3:42 PM

**Deleted:** one

KKY 10/19/2014 3:42 PM

**Deleted:** This

bottlenecks in the middle. In many cases, these bottlenecks create vulnerabilities. Indeed, it has been shown that many of the bottlenecks are essential in gene knockout experiments [19]. Hierarchies can insulate themselves from this vulnerability by allowing middle managers to co-regulate those under them. This eases information flow bottlenecks in an obvious way (if one major gets knocked out, the privates under him can receive orders from a second major). Many commenters have mentioned that, in order to function smoothly, it is imperative for social hierarchies to have middle managers working together [45]. Strikingly, biological regulatory networks employ the same strategy by having two mid-level TFs co-regulate targets below them [46]. Thus, one can get an intuition for the reason behind a particular biological structure through analogies to commonplace social networks.

The goal of this comparison is the transfer of ideas on the relationship between network structure and "function" from a social context to a less intuitive biological context. This underscores a more general point: Lying at the heart of deciphering biological networks mediated by mechanistic interactions is this mapping between architecture and function. The mapping points to simple biological circuits that solve common functional problems – effectively a component toolbox for systems biology [47]. As it is often hard to define a "function", comparison with various technological or engineered components that possess basic and well-defined functions is particularly insightful. [As an example, consider analog electronic circuit architecture. The subthreshold behavior of transistors has been compared to the molecular dynamics occurring within biological systems. Researchers have used the parallels between reaction rates and current flows, the noise present in both systems, and the formalisms developed for electronic circuits to improve models of transcription factor binding networks \[48\]. Additionally, the intuition and models developed through this comparison has informed the construction of novel biological networks capable of implementing analog computations in vivo \[49\].](#) A decade ago, Uri Alon pointed out several common design principles in biological and engineering networks such as modular organization and robustness to perturbation [50]. Robustness is a preferred design objective because it makes a system tolerant of stochastic fluctuations, from either intrinsic or external sources. Modularity, on the other hand, makes a system more evolvable. For instance in software design, modular programming that separates functionality of a program into independent modules connected by an interface is widely practiced [51]. The same is for biological networks because modules can be readily reused to adapt new functions.

#### *Looking for the commonalities and differences between tinkerer and engineer*

The comparison of biological and technological networks is best performed in light of evolution. As Alon's phrase "the tinkerer as an engineer" [50] highlights, it is remarkable that "good engineering solutions" are found in biological systems that have evolved by random tinkering. Indeed, comparison between biological and technological networks should manifest the nature of these two very different approaches. Evolution is a tinkerer that neither consciously designs components nor systematically builds larger systems— it settles on systems that have historically conveyed a survival benefit (and adopts better methods if they come along). On the other hand, technological networks are essentially blueprints drawn by engineers who have a grand plan that makes sure everything works harmoniously. Biologists often tend to distinguish the two approaches cautiously so as to avoid the notion of intelligent design – the existence of an intelligent agent that constructs living organisms. However, the distinction is not clear-cut. Both biological networks and man-made technological ones like roadways and electronic circuits are complex adaptive systems; there are plenty of examples showing that many great innovations are results of trial and error, and all technological systems are subjected to selection such as users requirements. In a recent review, Wagner summarized nine commonalities between biological and technological innovation, including descent with modification, extinction and replacement, and horizontal transfer [52]. Thus, to a certain extent, an engineer is a tinkerer (see Box 2).

Under such a united framework, we could picture that both the engineer and tinkerer are working on an optimization problem with similar underlying design objectives. Like all optimization problems, there is no method that optimizes all objectives and thus tradeoffs are unavoidable in both biological and technological systems. This is essentially the conventional wisdom – there's

no free lunch [53][54]. Despite their similarities, tinkerers and engineers take different views when balancing constraints and tradeoffs. Their optimal choices are exhibited in the topology of their corresponding networks. In biological networks, for example, more connected components (as measured by their hubbiness or betweenness) tend to be under stronger constraint than less connected ones. This is seen in numerous studies that have analyzed the evolutionary rate of genes in many networks (e.g. protein interaction networks and transcription regulatory networks) in many organisms (e.g. humans, worms, yeast, *E. coli*) using many different metrics of selection (e.g. variation within population or dN/dS for fixed differences) [55][56][57][58]. To some degree constraint is connected to connectivity in biological systems. One's intuition here is obvious: the more connected components are more vulnerable to changes, particularly since mutation and change occurs randomly in biology. But is this finding true in general? Comparison can provide intuition.

Consider software systems, software engineers tend to reuse certain code, leading to modularity. Intuitively one would expect that the robustness of software would be reduced if a piece of code is highly called by many disparate processes – i.e. if it's highly connected. Analysis of the evolution of a canonical software system, the Linux kernel, revealed that the rate of evolution of its functions (routines) is distributed in a bimodal fashion and thus a significant fraction of functions are updated often [59]. Therefore, unlike biological systems in which the majority of components are rather conserved and thus prefer a more independent organization to maintain robustness, software engineers pay the price of reusability and robustness by constantly tweaking the system. Indeed, further analysis of the underlying network of Linux kernel, the so-called call graph, shows that more central components in the call graph actually require more fine-tuning. These patterns seem to hold for other software systems. [For instance, in packages dependency network of the statistical computing language R, packages that are called by many others are updated more often](#) (Figure 2). In other words, unlike biological networks whose hubs tend to evolve slowly, hubs in the software system evolve rapidly. This seems to run counter to the intuition that an engineer should not meddle too much with highly connected components. However, there is another factor to consider: rational designers may believe that they can modify a hub without disrupting it. This is in contrast to a situation in which random changes dominate. Moreover, the central points in a system are often those in the greatest use and hence are in the most need of the designer's attention. This situation is analogous to road networks: one sees comparatively more construction on highly used bottlenecks (e.g. the George Washington Bridge) compared to out of the way thoroughfares (see Box 2).

### Conclusion

Biology is a subject with a strong tradition of utilizing comparative methods. One hundred years ago, biologists compared the phenotypes of different species. Since the discovery of DNA, biologists have been comparing the sequences of different genes, and then all sorts of 'omes' across species. Perhaps, it is a time to extend this tradition even further to compare networks in biology to those in other disciplines. Here, we have tried to describe how these comparisons are beginning to take place. First, we have described how association networks that just show simple connections between entities are abstract enough to allow the application of mathematical formalisms across disciplines. Then, we show how mechanistic details can be placed onto these simple networks and enable them to better explain a real process such as transcriptional regulation or software code development. In this case, the networks are often too detailed to allow for direct transfer of formalisms but often one can gain meaningful intuition about a biological system through comparing it to a more commonplace network such as a social system via the same mechanistic description.

What's next? We envision that these cross-disciplinary network comparisons will become increasingly common. Networks are one of the key structures used for the analysis of large datasets in the emerging field of data science. These datasets are becoming increasingly common in many fields. We anticipate that this will enable further fruitful comparisons with biology. One area that is especially ripe for comparison is multiplex networks. Over the past few years, efforts have been spent on concatenating networks together to form a multiplex structure

Koon-Kiu Yan 10/19/2014 5:22 PM

~~Deleted:~~ like the organization of packages in the

Koon-Kiu Yan 10/19/2014 5:24 PM

~~Deleted:~~

[60][61]. This framework is commonly used in social science in which an individual may participate in multiple social circles: family, friends, colleagues, or in online setting: Facebook, LinkedIn and Twitter. However, it has not been very well explored in biology. Nevertheless, this direction is of particular interest to biology because of rapid advancements in data acquisition. The structure of biological data now extends beyond a single network to a multiplex structure: the multiple layers could be formed by different categories of relationships (co-expression, genetic interactions, etc.). Furthermore, biological regulation occurs at multiple levels: transcriptional, post-transcriptional, and post-translational regulation in a manner in analogous to a city with electrical networks, water pipes, and cell phone lines. We are looking forward to some of the methods developed in other contexts to be applied in biology.

So far we have focused on leveraging the ideas and methods developed in multiple disciplines through comparison. We can even imagine that these comparisons will lead to new real connections between biological networks and those in other disciplines. For instance, there is an increasing amount of attention among biologists and sociologists on the connection between genomics information and sociological information such as whether phenotypes or genotypes are correlated in friendship networks [62]. Indeed, various scientific disciplines form a network in the intellectual universe where knowledge emerges when things connect.

#### **Potential exhibits:**

##### **Figure 1 Caption**

##### **Figure 2 Caption**

?A table showing examples of the two types networks.  
(Give more examples of association networks, like genetic interaction networks.)

?A table highlighting problems studied in the framework of association networks, and the corresponding problems arise in computational social science.

Table/Figure summarizing all comparisons/references.

Box 0 Network science 101?

##### **Box 1 Hierarchical organization of networks**

Many biological networks possess an intrinsic direction of information flow, forming a hierarchical network organization. The hierarchical organization in biological networks resemble the chain of command in human society, like in military context and corporate hierarchy [44]. For instance, in a transcriptional regulatory network more influential transcription factors (regulators whose expression are more highly correlated with the expression of target genes) tend to be better connected (have more interacting partners) and higher in the hierarchy. Moreover, in order to avoid information bottlenecks, the transcription factors in the middle layer tend to be more cooperative [46], resulting at many cross-links between pathways. Such a situation has been well studied in management science, where in certain corporate settings middle managers interact the most with peers to manage subordinates below them [45]. These observations reflect a democratic hierarchy as opposite to a conventional autocratic organization [63].

Of particular interest for hierarchical organization is the so-called bow-tie structure, meaning the intermediate layers have fewer components than the input and output layers. For example, in a signaling network, a large number of receptors corresponding to diverse stimuli and many transcription factors form the input and output layers, whereas the intermediate layer refers to a few key molecules like calcium and cAMP that mediate the inputs and outputs [64]. Similarly, in the networking architecture of the Internet, various protocols in the input/link layer (ARP, RARP, NDP etc) and various application protocols in the application/output layer (HTTP, FTP, DHCP etc) are essentially connected by only IPv4, the primary protocols in the internet layer. The reason for



the emergence of such a common pattern is still widely open, a recent paper suggested bow-tie is a result of information compression [65].

#### Box 2 Tinkerer versus engineer

Despite the apparent differences, the similarity between biological systems and technological systems draws a parallel between tinkerer and engineer, and the parallel points to a common framework to unite them. Wagner further proposed an analogy between the genotype space for a biological system and the design space for a technological system. These spaces contain all the possible networks in the corresponding systems. In biology, many attempts have been made to search for solutions of common functional problems such as adaptation, oscillation and cell polarization [47]. Similar studies were performed in the context of circuit design, where a set of logic gates was evolved via rewiring in order to perform a predefined computational task [66][67]. These studies suggested that in both kinds of systems, the solution networks are close together in the genotype/design space. As each solution in genotype/design has multiple neighbors, robustness of a solution to mutation facilitates the evolvability of these systems [68][69]. Indeed, it has been demonstrated that electronic circuits can be evolved to fulfill a fluctuating evolutionary goal [66]. Similarly, metabolic networks of bacteria living in multiple habitats are evolved to decompose multiple food sources [70][71]. Both of these networks show a level of modular organization.

Very often we picture engineers design things from scratch. In reality, as a technological system evolves, engineers are subjected to various constraints like tinkerer. In the example of internet architecture, while there are frequent innovations at the input layer that interact with a variety of networking hardware and output layers that connect with many different software applications, the internet layer with very few protocols is the bottleneck under heavy constraints and such protocols can hardly be replaced [72]. The observed rapid innovation at the top and bottom layers but constraint at the middle is very common in biological system. Consider the metabolic networks of different bacteria, the anabolic and catabolic components are much more diverse whereas there are less variations between central pathways [73].

- [1] M. Baker, "Big biology: The 'omes puzzle," *Nature*, vol. 494, no. 7438, pp. 416–419, Feb. 2013.
- [2] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat. Rev. Genet.*, vol. 5, no. 2, pp. 101–113, Feb. 2004.
- [3] A. D. Lander, "The edges of understanding," *BMC Biol.*, vol. 8, no. 1, p. 40, Apr. 2010.
- [4] R. Dawkins, *The selfish gene*, New ed. Oxford ; New York: Oxford University Press, 1989.
- [5] C. J. Howe and H. F. Windram, "Phylomemetics—Evolutionary Analysis beyond the Gene," *PLoS Biol*, vol. 9, no. 5, p. e1001069, May 2011.
- [6] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proc. Natl. Acad. Sci.*, vol. 104, no. 21, pp. 8685–8690, May 2007.
- [7] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
- [8] R. Albert, H. Jeong, and A.L. Barabasi, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, Jul. 2000.
- [9] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, May 2001.
- [10] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

- [11] L. a. N. Amaral, A. Scala, M. Barthélemy, and H. E. Stanley, "Classes of small-world networks," *Proc. Natl. Acad. Sci.*, vol. 97, no. 21, pp. 11149–11152, Oct. 2000.
- [12] D. C. V. Essen, M. F. Glasser, D. L. Dierker, and J. Harwell, "Cortical Parcellations of the Macaque Monkey Analyzed on Surface-Based Atlases," *Cereb. Cortex*, vol. 22, no. 10, pp. 2227–2240, Oct. 2012.
- [13] N. T. Markov, M. Ercsey-Ravasz, D. C. V. Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy, "Cortical High-Density Counterstream Architectures," *Science*, vol. 342, no. 6158, p. 1238406, Nov. 2013.
- [14] D. S. Modha and R. Singh, "Network architecture of the long-distance pathways in the macaque brain," *Proc. Natl. Acad. Sci.*, vol. 107, no. 30, pp. 13485–13490, Jul. 2010.
- [15] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, Mar. 1953.
- [16] S. Allesina and M. Pascual, "Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions?," *PLoS Comput Biol*, vol. 5, no. 9, p. e1000494, Sep. 2009.
- [17] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert, M. Niedergethmann, W. Weichert, M. Bahra, H. J. Schlitt, U. Settmacher, H. Friess, M. Büchler, H.-D. Saeger, M. Schroeder, C. Pilarsky, and R. Grützmann, "Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes," *PLoS Comput Biol*, vol. 8, no. 5, p. e1002511, May 2012.
- [18] M. E. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 64, no. 1 Pt 2, p. 016132, Jul. 2001.
- [19] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics," *PLoS Comput. Biol.*, vol. 3, no. 4, p. e59, Apr. 2007.
- [20] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
- [21] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithm for Collaborative Filtering," in *Proceedings of the 14 th Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 43–52.
- [22] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011.
- [23] J. M. Stuart, "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules," *Science*, vol. 302, no. 5643, pp. 249–255, Oct. 2003.
- [24] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nat. Rev. Genet.*, vol. 13, no. 8, pp. 523–536, Jul. 2012.
- [25] E. Khurana, Y. Fu, J. Chen, and M. Gerstein, "Interpretation of genomic variants using a unified biological network approach," *PLoS Comput. Biol.*, vol. 9, no. 3, p. e1002886, 2013.

- [26] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, Apr. 2010.
- [27] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating Genes and Protein Complexes with Disease via Network Propagation," *PLoS Comput Biol*, vol. 6, no. 1, p. e1000641, Jan. 2010.
- [28] E. Adar and L. A. Adamic, "Tracking Information Epidemics in Blogspace," 2005, pp. 207–214.
- [29] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein, "Predicting interactions in protein networks by completing defective cliques," *Bioinforma. Oxf. Engl.*, vol. 22, no. 7, pp. 823–829, Apr. 2006.
- [30] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008.
- [31] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed Membership Stochastic Blockmodels," *J Mach Learn Res*, vol. 9, pp. 1981–2014, Jun. 2008.
- [32] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, 2006.
- [33] P. Domingos and M. Richardson, "Mining the Network Value of Customers," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2001, pp. 57–66.
- [34] D. Wang, A. Arapostathis, C. O. Wilke, and M. K. Markey, "Principal-Oscillation-Pattern Analysis of Gene Expression," *PLoS ONE*, vol. 7, no. 1, p. e28805, Jan. 2012.
- [35] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proc. Natl. Acad. Sci.*, vol. 105, no. 35, pp. 12763–12768, 2008.
- [36] K.-K. Yan, D. Wang, J. Rozowsky, H. Zheng, C. Cheng, and M. Gerstein, "OrthoClust: an orthology-based network framework for clustering data across multiple species," *Genome Biol.*, vol. 15, no. 8, p. R100, Aug. 2014.
- [37] C. Shou, N. Bhardwaj, H. Y. K. Lam, K.-K. Yan, P. M. Kim, M. Snyder, and M. B. Gerstein, "Measuring the Evolutionary Rewiring of Biological Networks," *PLoS Comput Biol*, vol. 7, no. 1, p. e1001050, Jan. 2011.
- [38] K. Evlampiev and H. Isambert, "Conservation and topology of protein interaction networks under duplication-divergence evolution," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 29, pp. 9863–9868, Jul. 2008.
- [39] R. Pastor-Satorras, E. Smith, and R. V. Solé, "Evolving protein interaction networks through gene duplication," *J. Theor. Biol.*, vol. 222, no. 2, pp. 199–210, May 2003.
- [40] M. P. Simmons, L. A. Adamic, and E. Adar, "Memes online: Extracted, subtracted, injected, and recollected," in *In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.
- [41] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, "Modeling of Protein Interaction Networks," *Complexus*, vol. 1, no. 1, pp. 38–44, 2003.

- [42] P. M. Kim, L. J. Lu, Y. Xia, and M. B. Gerstein, "Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights," *Science*, vol. 314, no. 5807, pp. 1938–1941, Dec. 2006.
- [43] T. Y. Pang and S. Maslov, "Universal distribution of component frequencies in biological and technological systems," *Proc. Natl. Acad. Sci.*, vol. 110, no. 15, pp. 6235–6239, Mar. 2013.
- [44] H. Yu and M. Gerstein, "Genomic analysis of the hierarchical structure of regulatory networks," *Proc. Natl. Acad. Sci.*, vol. 103, no. 40, pp. 14724–14731, Oct. 2006.
- [45] S. W. Floyd and B. Wooldridge, "Middle management involvement in strategy and its association with strategic type: A research note," *Strateg. Manag. J.*, vol. 13, no. S1, pp. 153–167, Jun. 1992.
- [46] N. Bhardwaj, K.-K. Yan, and M. B. Gerstein, "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels," *Proc. Natl. Acad. Sci.*, vol. 107, no. 15, pp. 6841–6846, Mar. 2010.
- [47] W. A. Lim, C. M. Lee, and C. Tang, "Design Principles of Regulatory Networks: Searching for the Molecular Algorithms of the Cell," *Mol. Cell*, vol. 49, no. 2, pp. 202–212, Jan. 2013.
- [48] R. Sarpeshkar, "Analog synthetic biology," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 372, no. 2012, p. 20130110, Mar. 2014.
- [49] R. Daniel, J. R. Rubens, R. Sarpeshkar, and T. K. Lu, "Synthetic analog computation in living cells," *Nature*, vol. 497, no. 7451, pp. 619–623, May 2013.
- [50] U. Alon, "Biological Networks: The Tinkerer as an Engineer," *Science*, vol. 301, no. 5641, pp. 1866–1867, Sep. 2003.
- [51] M. A. Fortuna, J. A. Bonachela, and S. A. Levin, "Evolution of a modular software network," *Proc. Natl. Acad. Sci.*, vol. 108, no. 50, pp. 19985–19989, Dec. 2011.
- [52] A. Wagner and W. Rosen, "Spaces of the possible: universal Darwinism and the wall between technological and biological innovation," *J. R. Soc. Interface*, vol. 11, no. 97, p. 20131190, Aug. 2014.
- [53] A. D. Lander, "Pattern, growth, and control," *Cell*, vol. 144, no. 6, pp. 955–969, Mar. 2011.
- [54] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, "Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space," *Science*, vol. 336, no. 6085, pp. 1157–1160, Jun. 2012.
- [55] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, "Evolutionary Rate in the Protein Interaction Network," *Science*, vol. 296, no. 5568, pp. 750–752, Apr. 2002.
- [56] G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili, "Interaction network containing conserved and essential protein complexes in *Escherichia coli*," *Nature*, vol. 433, no. 7025, pp. 531–537, Feb. 2005.

- [57] H. B. Fraser, D. P. Wall, and A. E. Hirsh, "A simple dependence between protein evolution rate and the number of protein-protein interactions," *BMC Evol. Biol.*, vol. 3, p. 11, May 2003.
- [58] M. W. Hahn and A. D. Kern, "Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks," *Mol. Biol. Evol.*, vol. 22, no. 4, pp. 803–806, Apr. 2005.
- [59] K.-K. Yan, G. Fang, N. Bhardwaj, R. P. Alexander, and M. Gerstein, "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks," *Proc. Natl. Acad. Sci.*, vol. 107, no. 20, pp. 9186–9191, May 2010.
- [60] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community Structure in Time-Dependent, Multiscale, and Multiplex Networks," *Science*, vol. 328, no. 5980, pp. 876–878, May 2010.
- [61] P. Holme and J. Saramäki, "Temporal networks," *Phys. Rep.*, vol. 519, no. 3, pp. 97–125, Oct. 2012.
- [62] J. H. Fowler, J. E. Settle, and N. A. Christakis, "Correlated genotypes in friendship networks," *Proc. Natl. Acad. Sci.*, p. 201011687, Jan. 2011.
- [63] Y. Bar-Yam, D. Harmon, and B. de Bivort, "Attractors and Democratic Dynamics," *Science*, vol. 323, no. 5917, pp. 1016–1017, Feb. 2009.
- [64] N. Polouliakh, R. Nock, F. Nielsen, and H. Kitano, "G-Protein Coupled Receptor Signaling Architecture of Mammalian Immune Cells," *PLoS ONE*, vol. 4, no. 1, p. e4189, Jan. 2009.
- [65] T. Friedlander, A. E. Mayo, T. Tlusty, and U. Alon, "Evolution of bow-tie architectures in biology," *ArXiv14047715 Q-Bio*, Apr. 2014.
- [66] N. Kashtan and U. Alon, "Spontaneous evolution of modularity and network motifs," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 39, pp. 13773–13778, Sep. 2005.
- [67] K. Raman and A. Wagner, "The evolvability of programmable hardware," *J. R. Soc. Interface*, vol. 8, no. 55, pp. 269–281, Feb. 2011.
- [68] A. Wagner, "Neutralism and selectionism: a network-based reconciliation," *Nat. Rev. Genet.*, vol. 9, no. 12, pp. 965–974, Dec. 2008.
- [69] J. Masel and M. V. Trotter, "Robustness and Evolvability," *Trends Genet.*, vol. 26, no. 9, pp. 406–414, Sep. 2010.
- [70] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin, "The evolution of modularity in bacterial metabolic networks," *Proc. Natl. Acad. Sci.*, vol. 105, no. 19, pp. 6976–6981, May 2008.
- [71] S. Maslov, S. Krishna, T. Y. Pang, and K. Sneppen, "Toolbox model of evolution of prokaryotic metabolic networks and their regulation," *Proc. Natl. Acad. Sci.*, vol. 106, no. 24, pp. 9743–9748, Jun. 2009.
- [72] S. Akhshabi and C. Dovrolis, "The Evolution of Layered Protocol Stacks Leads to an Hourglass-shaped Architecture," in *Proceedings of the ACM SIGCOMM 2011 Conference*, New York, NY, USA, 2011, pp. 206–217.
- [73] M. Csete and J. Doyle, "Bow ties, metabolism and disease," *Trends Biotechnol.*, vol. 22, no. 9, pp. 446–450, Sep. 2004.