JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Navlakha S, He X, Faloutsos C, Bar-Joseph Z. 2014 Topological properties of robust biological and computational networks. *J. R. Soc. Interface* **11**: 20140283.
http://dx.doi.org/10.1098/rsif.2014.0283

Received: 18 March 2014
Accepted: 7 April 2014

Subject Areas:
systems biology, computational biology, mathematical physics

Keywords:
biological networks, communication networks, robustness, security

Author for correspondence:
Ziv Bar-Joseph
e-mail: zivbj@cs.cmu.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2014.0283 or via http://rsif.royalsocietypublishing.org.

Royal Society Publishing

Topological properties of robust biological and computational networks

Saket Navlakha[1], Xin He[2], Christos Faloutsos[1] and Ziv Bar-Joseph[1,2]

[1]Machine Learning Department, and [2]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Network robustness is an important principle in biology and engineering. Previous studies of global networks have identified both redundancy and sparseness as topological properties used by robust networks. By focusing on molecular subnetworks, or modules, we show that module topology is tightly linked to the level of environmental variability (noise) the module expects to encounter. Modules internal to the cell that are less exposed to environmental noise are more connected and less robust than external modules. A similar design principle is used by several other biological networks. We propose a simple change to the evolutionary gene duplication model which gives rise to the rich range of module topologies observed within real networks. We apply these observations to evaluate and design communication networks that are specifically optimized for noisy or malicious environments. Combined, joint analysis of biological and computational networks leads to novel algorithms and insights benefiting both fields.

1. Introduction

Robustness to failures, environmental and signalling noise, and attacks is a key requirement in many biological, computational and engineered networks [1,2]. Protein interaction networks, in particular, are robust to most single and double node failures [3–5], expression and environmental noise [6] and viral and bacterial attacks [7].

There have been two primary approaches for explaining how connectivity affects the robustness of molecular interaction networks. The first highlights the importance of *structural redundancy*, often manifested as high connectivity, where the failure of one pathway can be compensated for by a redundant pathway. To support this view, researchers have highlighted several topological features observed in protein interaction networks, including backup mechanisms based on paralogues with similar interactions [3,8], parallel pathways [9,4] and dense subnetworks that may compensate for loss-of-function mutations [5]. On the other hand, high connectivity leads to functional coupling of different components making a network more susceptible to cascading local failures or attacks. Researchers have discussed several design properties that help biological networks overcome these issues, including bottlenecks [10], weakly linked modules [11] and network motifs [12]. While only implicit, such local structures imply *sparsely connected* networks. To observe this, consider the most redundant network possible: a clique (a complete network where every node has an edge to every other node). Clearly, such a network has several alternative pathways and backup nodes. However, it does not have any bottlenecks (defined as nodes or edges that bridge two otherwise disconnected sets of nodes), nor any modules (sets of nodes that are highly intra-connected *and* sparsely inter-connected), nor any statistically over-represented network motifs compared with a random network with the same number of edges (electronic supplementary material). Sparser networks, on the other hand, are much more likely to contain these three features. Thus, while these two approaches for explaining robustness have so far not been explicitly contrasted, they represent different expected topological structures.

Most prior work on the analysis of robustness in biological and computational networks has focused on *global* topology. For example, Jeong *et al*. [13] proposed the centrality-lethality principle, which stated that high-degree nodes in the global protein interaction network in yeast were not likely to be robust. In other words, these gene nodes tended to be fragile or essential (i.e. deletion of these genes, as determined by genetic knock-out experiments, results in cell death). These hubs contributed to the overall higher connectivity of the network, but they also served as bottlenecks through which many otherwise disconnected proteins interacted. This observation has been confirmed both in yeast [10] and in the protein interactomes of other organisms [14]. Further, theoretical models analysing global networks have shown that higher connectivity promotes robustness to perturbation under stabilizing selection [15]. However, adding new interactions may also introduce new degrees of freedom for perturbations, and when such costs are taken into account using a cost–benefit analysis, sparser networks have been shown to promote more robustness [16]. These computational models also assume that the likelihood of perturbation is equal for all nodes in the network, even though experimental work has shown that environmental noise affects some proteins more than others [6].

While studies of global molecular networks have identified important issues that affect robustness, processing in these networks typically occurs within local subnetworks or *modules* [17]. Anecdotal evidence suggests that module-level properties may be important in determining network robustness to mutations. For example, cellular localization was found to be a useful feature when predicting gene essentiality in yeast [18], and nuclear proteins in particular were shown to be enriched for such essential genes [19]. Other studies have also shown that involvement of genes in specific subnetworks is the most indicative of individual gene essentiality [20,21]. However, the exact relationship between the level of robustness necessary for a biological process to operate and the topological properties that give rise to this robustness has so far not been determined.

Here, we bridge the gap between the redundancy and sparsity hypotheses by studying how robustness is intricately linked to the function and topology of specific modules within global networks. Our results emphasize the role of the environmental exposure of a module in determining the appropriate network structure. For modules subject to large environmental influences (which we call *external modules*), lower connectivity (fewer edges) promotes robustness by limiting the functional coupling of components of the subnetwork. However, for modules that are relatively insulated from these external noise sources (called *internal modules*), robustness to environmental variability may not be a strong evolutionary driver, leading to higher and more efficient connectivity. Using these observations, we develop a computational method to predict network robustness from topology and show that it can distinguish between internal and external modules in several additional biological networks. We also extend a popular network generative model, based on gene duplication, to explain how modules with varying topologies and levels of robustness may have evolved. Finally, we use our robustness prediction method and generative model to improve the analysis and design of communication networks, demonstrating that insights from biological networks may be beneficial for designing robust networks in other domains.

2. Results

2.1. Robust genes and modules in *Saccharomyces cerevisiae*

The robustness of yeast to single gene deletion has been comprehensively analysed [3,5], and several global network features of a protein in the yeast interaction network (especially its degree) have been linked to its essentiality [13,20,21]. We collected a global interaction network for yeast by integrating protein–protein interactions and protein–DNA interactions (Material and methods). Of the 5796 proteins, 19.4% (1122) were determined to be essential in normal growth conditions (i.e. removal of any of these single gene nodes caused the entire network to fail in YPD).

First, to determine whether gene robustness is better predicted using global or module-level topology, we decomposed the global network into 50 modules corresponding to biological processes that are required for cell survival and growth (Material and methods). Each biological process is associated with a module subgraph consisting of all the nodes that belong to that process and their interactions with other genes in the process. Topological features were derived either from the global network or the local module topology. For example, a node's degree is the number of total neighbours it has in the entire network, or the number of neighbours it has among nodes in its module. The former definition is used for the global network-level analysis and the latter for the local module-level analysis.

We found that the correlation between gene essentiality and gene centrality increased by roughly 50% when using network features of a gene derived from its local module topology as opposed to its global network topology (table 1 and electronic supplementary material, table S4). This suggests that a gene's essentiality depends on both its module (i.e. its function) and its topological role within the module, and this analysis has also been shown to hold for various definitions of biological module [21]. Interestingly, the feature with the highest correlation using the global network was not the average degree (0.352) but the k-core number of the node (0.367). A k-core is a maximal subgraph that contains nodes of degree more than or equal to k, and the core number of a node u is defined as the largest value k of a k-core containing u. This feature is also implicitly a module-level attribute because it takes the density of the subgraph surrounding the gene into account.

We also tested another network feature to characterize gene robustness based on epidemiological theory. Previous theoretical work regarding virus propagation in networks has shown that the more connected a node is in the network, the more fragile or essential it is, and hence the more important it is to immunize to increase network robustness [22]. In these models, a single node becomes infected with a virus that then spreads through the network using a contagion model. In molecular networks, cells need to respond to noise (i.e. variable signals from the environment), which can similarly propagate through the network of interacting proteins. We found that a network feature based on a susceptible-infectious (SI) virus infection model (Material and methods) well-explains why some nodes are essential while others are not. Specifically, the larger the size of the affected subnetwork when propagating a virus from a node (called the *infect* size of the node), the more likely that this node is essential. For this feature, module-level infect sizes were also more predictive

Table 1. Predicting node and module robustness in biological networks. The yeast and *Escherichia coli* interaction networks were decomposed into 50 and 38 gene ontology modules, respectively. Individual gene essentiality was predicted using topological features of the gene computed within its local module or the global network, respectively. For each, the Kendall τ rank correlation coefficient against essentiality is shown. The top three features for each dataset are italicized. Several module-based features showed a large gain in predictive power versus global features. The third and fourth blocks show the Pearson correlation coefficients between module essentiality (percentage of essential genes in a module) and topological features of the module. For the *Caenorhabditis elegans* neuronal network, modules represented eight anatomically defined ganglia, and the biological feature was the percentage of internal neurons in the module. For the bacterial metabolic networks, we correlated niche breadth of the species with topological features of the network.

network	topological feature	module	global	gain (%)
yeast interaction versus gene essentiality	degree	*0.497*	0.352	41.1
	infect	*0.453*	0.302	50.0
	eigenvector centrality	*0.424*	0.270	57.0
	PageRank	0.404	0.363	11.3
	betweenness centrality	0.385	0.314	22.6
E. coli interaction versus gene essentiality	degree	0.216	0.233	−7.3
	infect	*0.274*	0.161	70.2
	eigenvector centrality	0.171	0.081	111.1
	PageRank	*0.304*	0.235	29.4
	betweenness centrality	*0.242*	0.187	29.4
yeast interaction versus module essentiality	density	*0.699*	—	—
	norm. eigenvalue	*0.647*	—	—
	avg. infect	*0.574*	—	—
	max. independent set	−0.567	—	—
	avg. degree	0.394	—	—
E. coli interaction versus module essentiality	density	*0.654*	—	—
	norm. eigenvalue	*0.557*	—	—
	avg. infect	*0.665*	—	—
	max. independent set	−0.170	—	—
	avg. degree	−0.061	—	—
C. elegans neural versus internal processing	norm. eigenvalue	*0.703*	—	—
	avg. infect	*0.672*	—	—
	density	*0.537*	—	—
	max. independent set	0.228	—	—
	avg. degree	−0.135	—	—
bacteria metabolic versus niche breadth	norm. eigenvalue	*0.431*	—	—
	avg. infect	*0.424*	—	—
	density	*0.421*	—	—
	max. independent set	−0.037	—	—
	avg. degree	−0.387	—	—

of essentiality than infect sizes computed using the global topology (table 1).

Having established that gene robustness is a module-level property, we next sought to differentiate between robust and fragile *modules* based on their topology and function. We computed, for each module, a *module essentiality* score equal to the proportion of essential genes in the module, and we sought topological features of these subnetworks that predicted their robustness. Here, degree, for example, is defined as the average degree of all nodes in the module (i.e. the average number of neighbours each node has with other nodes in the same module).

We found that highly essential (non-robust) modules were quickly swamped by noise in the SI infection model, and they were topologically denser with a higher eigenvalue compared with robust modules (table 1; electronic supplementary material, table S5; Material and methods). This is also in agreement with previous theoretical work in network epidemiology, which has proved that the epidemic threshold of a network (i.e. the likelihood that a virus infects every node) is proportional to the first eigenvalue of its adjacency matrix [23]: higher eigenvalue → higher likelihood of an outbreak. Interestingly, the average degree of the module (0.394) was much less correlated than its eigenvalue (0.647), density (0.699) and number of connected components (−0.506). The latter feature suggests that proteins belonging to robust modules are only weakly connected overall, which serves as another means to localize the spread of noise within these subnetworks.

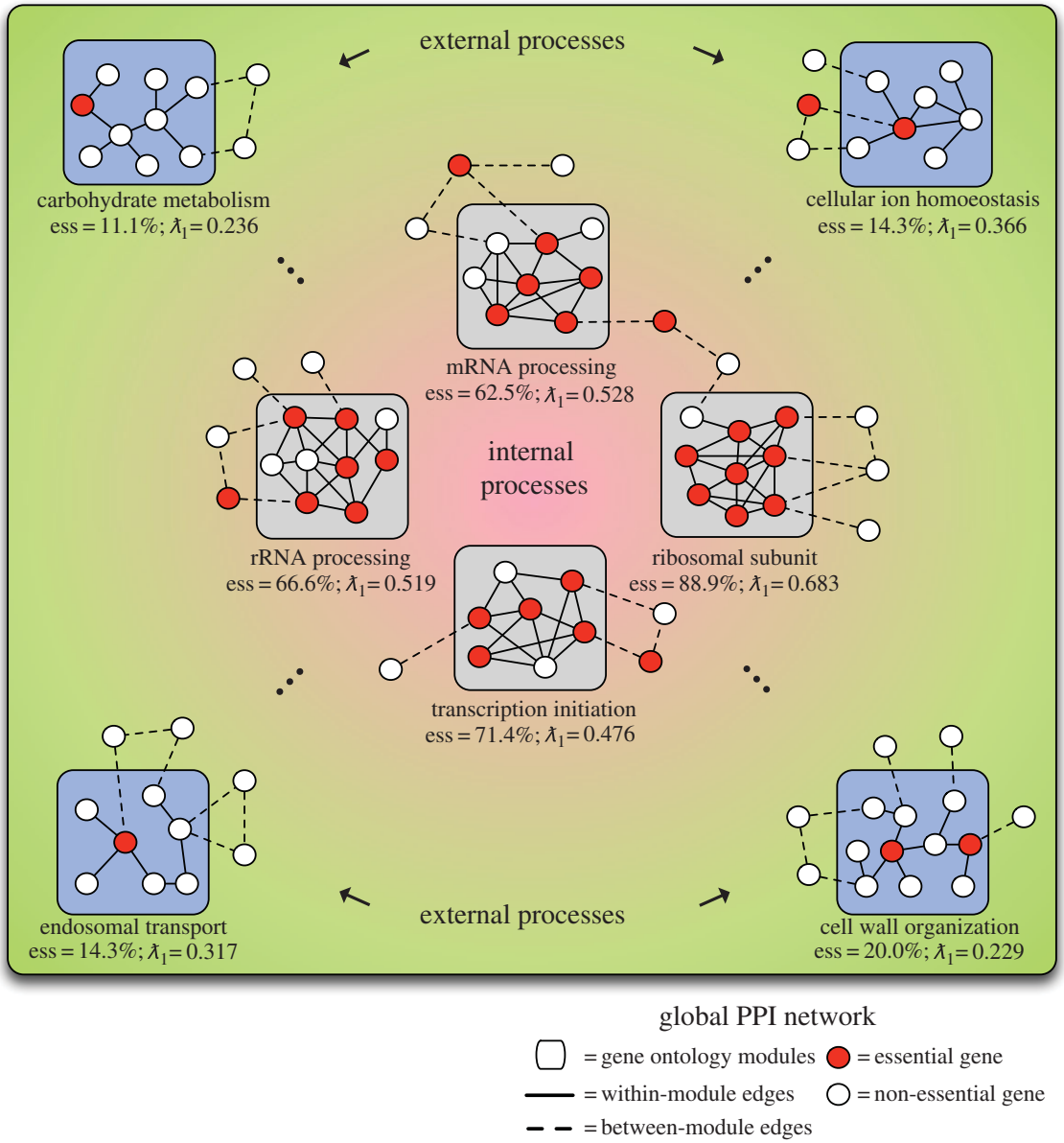

Figure 1. Topological differences between internal and external processes in the cell. Centre: examples of internal module topologies. Periphery: examples of external modules. For each module, we list its function, percentage of essential genes (ess, red) and eigenvalue of its adjacency matrix (λ). Internal modules are more dense with a higher percentage of essential genes.

We also found that highly essential modules represent biological processes that occur internal to the cell, whereas less essential (robust) modules represent processes that interact with the environment or other cells (figure 1). For example, several of the most essential processes involve transcription and translation including: DNA replication (50% of genes are essential), rRNA processing (66%) and ribosomal subunit export (81%). By contrast, the least essential modules include: cellular ion homoeostasis (9%), endosomal transport (9%) and cell wall organization or biogenesis (10%) (electronic supplementary material, table S5). We scored proteins as internal (1) or external (0) using protein localization data (Material and methods) and indeed found that robust external modules tended to contain proteins on the cell wall, membrane or other extracellular regions (correlation of 0.435 between module essentiality and average localization score of the module). Further support for the relationship between robustness and expected environmental noise is found in the correlation between module essentiality and the average coefficient of variation in expression for genes in the module (−0.380, i.e. more essential modules exhibit less variation in expression). Stochasticity in gene expression is a well-known mechanism to overcome noise [24], and higher variability has also been previously linked to 'environmental exposure' [6].

Thus, internal modules are less susceptible to noise and promote higher connectivity, whereas external modules are more exposed to environmental variability and promote sparser connectivity to localize the spread of noise.

2.2. Relationship between module essentiality and topology in other conditions and species

While gene deletion has been extensively studied in normal growth conditions (YPD), there are also systematic studies of gene deletions in other conditions. Depending on the condition, different modules may become more important to the cell than others (e.g. heat shock response modules in heat shock experiments). To see whether our findings regarding topological features related to robustness in YPD carried over to other conditions in yeast, we evaluated four additional

conditions in which genome-wide deletion studies have been performed [25]. For each, we selected the relevant modules (electronic supplementary material, table S6) and as before correlated module essentiality with topological features of the module. Results for three of the four conditions we tested were similar to our YPD results: essential modules were more highly connected than the robust modules (electronic supplementary material, table S3).

To test whether a similar relation between module essentiality and topology existed in molecular interaction networks in other species, we collected protein–protein interactions and protein–DNA interactions for the bacteria *Escherichia coli* (Material and methods). Of the 2915 proteins, 21.1% (616) were essential. We decomposed this network into the 38 processes (of the 50 used in yeast) that are relevant for prokaryotes. As in yeast, we found that predicting individual gene essentiality was often better using the module-level topology as opposed to global topology, and there was a similar distinction between robust (external) and fragile (internal) processes based on their function and topology (table 1; electronic supplementary material, tables S7 and S8).

2.3. Additional biological networks support the focus on modules

To determine the generality of the topological differences we observed between internal and external subnetworks, we investigated two additional types of biological networks.

We first analysed the *Caenorhabditis elegans* neuronal network, where nodes are neurons and directed edges represent chemical synapses. The network is naturally decomposed into eight modules (ganglia) representing the physical location of cell bodies. Neurons can be labelled as external neurons involved in input or output processing (sensory and muscle neurons), or internal neurons (interneurons) that regulate locomotion (electronic supplementary material). Similar to the yeast network, we found that modules with higher percentages of internal neurons have a significantly higher eigenvalue, density and infect size than external modules (table 1 and electronic supplementary material, table S9). Neurons in external modules directly interact with the environment via input stimuli or output motor response and are therefore more sensitive to environmental perturbations, whereas internal neurons handle complex processing that integrates sensory information and coordinates downstream processes. Thus, we see a similar efficiency–robustness trade-off within *C. elegans*, which suggests that noise buffering within local modular structures may also be a key component of neural circuit design.

Next, we looked at a collection of metabolic networks for 75 bacterial species. Nodes in these networks correspond to metabolites and edges imply an enzymatic reaction transforming one metabolite to another. Each species is assigned a niche breadth score that relates to the environment in which it operates (Material and methods). As above, we found that bacterial networks that thrive in stable and narrow environments have a significantly higher eigenvalue, density and infect size than those that inhabit dynamic and variable environments (table 1 and electronic supplementary material, table S10). Interestingly, and in contrast to results obtained from the analysis of global yeast networks, this correlation is not driven by an increase in node degree (which actually has the opposite correlation with niche breadth [26]).

2.4. A computational model for generating modules with varying topologies

One canonical model for the evolution of molecular interaction networks comes from the duplication–divergence principle [27]. In this model, gene duplication results in two equivalent proteins, which initially share interaction partners. This equivalence diverges over time leading to specialized subtasks and interactions. While this and other models have been shown to reproduce coarse features of global networks [28,29], they cannot explain the wide differences observed in the local topology across modules. Thus, a revised evolutionary model may be needed to better explain the diverse range of topological properties we observe among modules in molecular interaction networks. As robustness plays a major role in engineering, such a model can also be used to design communication networks based on the level of environmental noise and attacks they are expected to face.

We found that a small change to a standard duplication–divergence model can give rise to the complex range of topological features observed across yeast modules. In the model of Vazquez *et al*. [27], in each iteration, an existing random node u is duplicated into a topological equivalent u' that is initially connected to all neighbours of u. Then, for each common neighbour x of u and u', with probability q_{mod}, we remove either (u, x) or (u', x). The two duplicates are connected themselves with probability q_{con}. To determine how to adjust the duplication model to account for varying module topologies, we analysed how common-neighbour retainment (characterized by q_{mod}) following duplication varies across paralogues in different modules. For each module, we computed the average Jaccard coefficient between each gene in the module and all of its paralogues (including both whole-genome duplicates and small-scale duplicates [30]) in the network. We found that the higher the average Jaccard coefficient, the more likely the module was essential (correlation of 0.330).

Based on these observations, we modified the duplication model by varying q_{mod}, the divergence probability after duplication, between 0 and 1 (electronic supplementary material, table S1). This modification still adheres to the original duplication principle but can produce different network topologies by only varying one parameter. Low values of q_{mod} generate dense, clique-like networks characteristic of internal modules, whereas higher values of q_{mod} will generate sparse, chain-like networks characteristic of robust external modules. Indeed, we found that this model reproduces the rich range of topologies observed in the real yeast modules (figure 2*a*), whereas the standard duplication model, which uses a single value of q_{mod} for the entire network, does not (electronic supplementary material). We also observed a transition in the degree distribution of real modules starting from sparse and power-law-like (low essentiality) to pockets of cliques (high essentiality), and this trend was also visible in the networks produced by our generative model (figure 2*b*).

2.5. Biological insights for the analysis of secure communication networks

Similar to biological networks, communication networks also constantly face failures and attacks [31]. For example, the Internet is regularly targeted with worms that infect and compromise machines [32], and transport networks such as

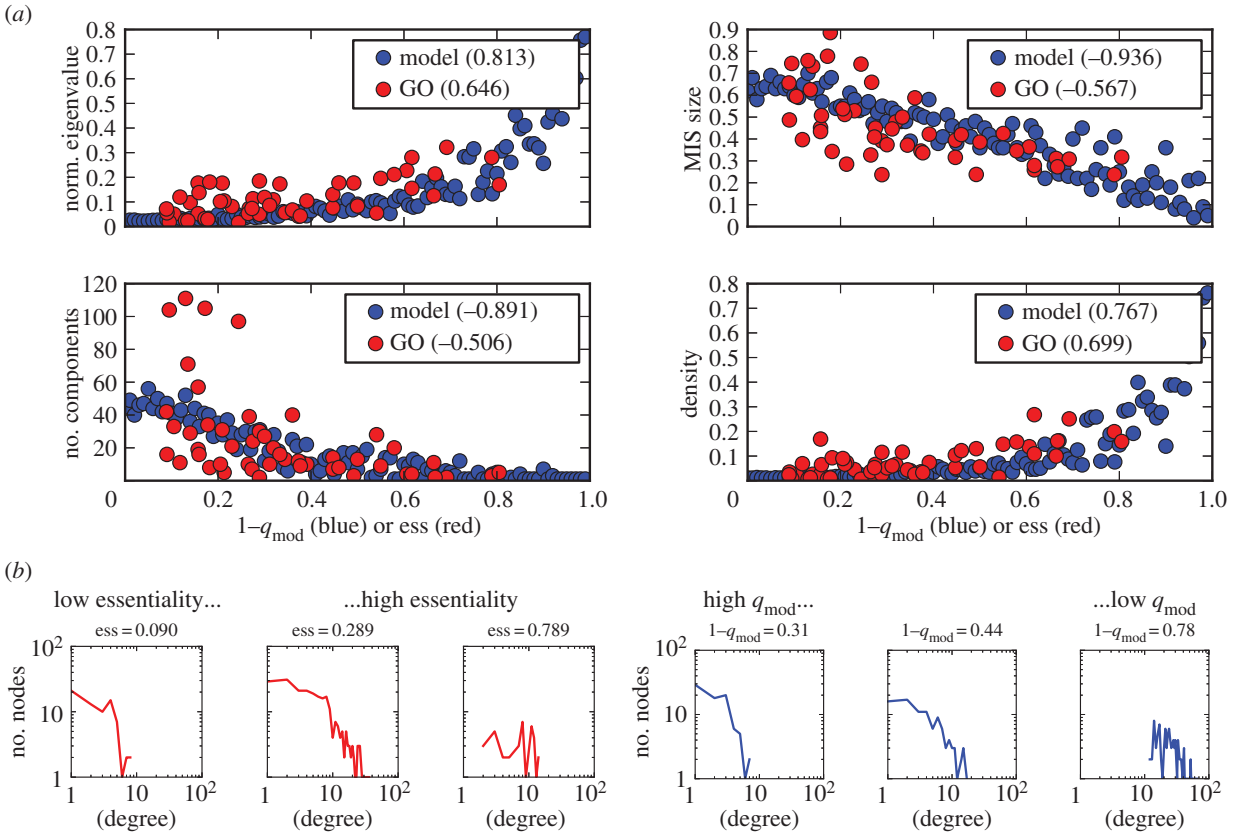

Figure 2. Duplication–divergence generative model. (*a*) Overlap of network features between real and model-generated modules: for the yeast gene ontology (GO) modules (red), the *x*-axis corresponds to essentiality, and for the generative model (blue), the *x*-axis is $1 - q_{mod}$. In both cases, highly robust modules with low essentiality or high q_{mod} have significantly lower density and eigenvalue and significantly more connected components. Correlations shown in the legend. (*b*) Degree distribution of real modules (red) versus the generative model for different values of q_{mod} (blue). Both degree distributions transition from sparse and power-law-like to more uniform.

the power grid have also experienced similar failures that have led to sudden and widespread blackouts [33].

We developed a computational model to benchmark the ability of a network to maintain connectivity following an attack that spreads from a single infected node. Typically, infected machines are detected following an attack and then isolated from the network for maintenance (e.g. applying anti-virus software) [32,34,35]. A key question is how does such removal affect the ability of all nodes to communicate? This requires a delicate balance: densely connected topologies will have low *residual connectivity* in the graph after infection because the virus will have spread to many nodes (Material and methods); however, if the initial topology is too sparse, many pairs of nodes will become disconnected from each other and residual connectivity will be equally low. Unlike previous models that consider an attack as only the removal of a single node [9], in our model hubs may not be safe from infection even if not directly targeted due to the spreading of the virus along edges in the graph.

We first applied the security benchmark to the yeast modules to test whether genetically robust modules are also determined to be robust under our benchmark. We found that highly essential modules had the lowest (worst) residual connectivity after infection (−0.430 correlation), which is consistent with our finding that internal modules have topologies that promote efficiency more than robustness.

We next applied our benchmark to two time-evolving communication networks. The first is a sequence of five snapshots of the Gnutella peer-to-peer (P2P) file-sharing network from several consecutive days in August 2002 [36]. The second is a sequence of five snapshots (each separated by six months) of the graph of routers (autonomous systems) representing the Internet [36]. Using our evaluation criteria, we found that the P2P networks had much higher residual connectivity, and hence robustness, than the Internet (figure 3). Given that P2P networks provide a robust storage mechanism and constantly deal with the addition and removal of nodes (users) and edges, this robustness is in agreement with prior work analysing P2P networks [37]. By contrast, the Internet is characterized by a few large service providers whose failure can significantly damage network connectivity, as has been observed recently by the large Amazon failure [38] and other catastrophic failures [9]. Identifying vulnerabilities in a network is also a key security challenge from the perspectives of defence and immunization [39], information dissemination [40] and network control [41].

For the P2P network, it is notable that a robust topology emerged, even though no centralized controller is in charge of this process. We next apply our biologically inspired generative model to construct such robust distributed networks.

2.6. Designing robust communication networks for varying environmental conditions

Like in biological networks, in many real-world scenarios, noise and susceptibility to attacks may not affect all nodes equally. For example, some subnetworks may be highly controlled or insulated from the outside and therefore require less protection [42], whereas others may exist in highly variable or unknown environments, requiring more

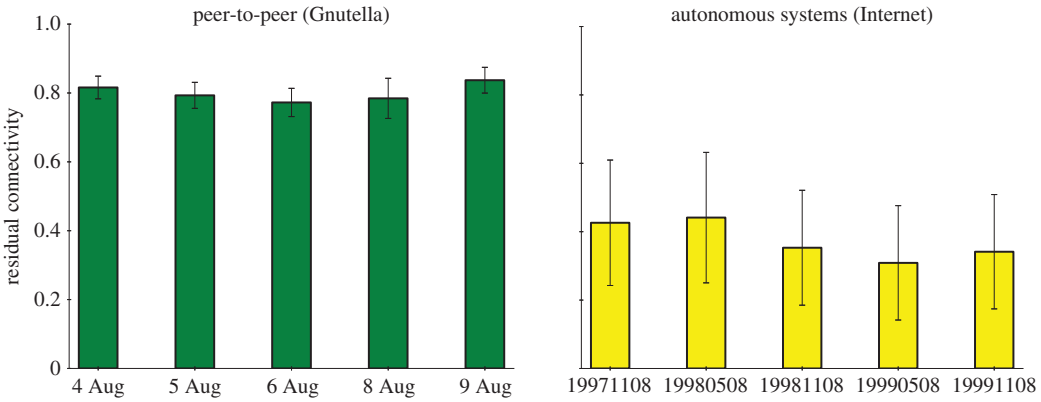

Figure 3. Evaluating the robustness of real-world communication networks. Residual connectivity (i.e. the percentage of nodes that remain connected after a cascading failure) for two time-evolving networks. The Gnutella P2P network, which still supports connectivity between roughly 80% of users, is much more resilient than the Internet. Results shown here use $\beta = 0.3$ in the SI virus propagation model. (Online version in colour.)

protection [43]. A key question is: given an expected infection rate γ for a subnetwork (i.e. an expected level of attack on the nodes), how should the network be designed to minimize the *residual distance* (figure 4*a*; Material and methods) between nodes following the attack?

To address this question using the methods derived from the biological networks, we extended our security benchmark to allow for variable infection rates: for each node in a module, with probability γ, it becomes infected and spreads the virus as usual, but with probability $1 - \gamma$, it does not get infected and does not spread the virus to any neighbour. When γ is low, few nodes become infected and as γ increases, more nodes become infected.

To design networks in accordance with γ, we varied the value of q_{mod} in the generative model (figure 4*b*). We found that when γ is low, clique-like networks (generated using low values of q_{mod}) confer the lowest (best) residual routeing distance after infection (figure 4*c*); on the other hand, as γ increases, sparser networks are preferred. However, even in the most noisy of environments ($\gamma = 1$), the best value of q_{mod} is 0.5, which implies that higher values of q_{mod} result in networks that are initially too sparse to withstand the attack (figure 4*c*). Thus, our generative model can be used to distributedly design on-the-fly networks that balance robustness and efficiency based on the expected security risk (e.g. in cases where wireless devices enter areas in which security requirements change [44]).

3. Discussion

There have been two primary approaches for explaining the robustness of molecular interaction networks, the first highlighting the importance of high connectivity (e.g. redundancy and parallel pathways) and the second highlighting sparsity (e.g. bottlenecks and weakly linked modules). These two directions partly stem from the fact that most previous works studied robustness with respect to global properties of the network as opposed to properties of local subnetworks or modules.

We bridged the gap between the redundancy and sparsity hypotheses by showing how robustness is intricately linked to the function and topology of modules within the global network. We showed that robustness to single gene deletion is better predicted using module topology specific to the gene's function as opposed to global network topology, and we observed a similar gain in accuracy for predicting robustness to double node deletion, i.e. genetic interactions (electronic supplementary material, table S11). For external modules subject to large environmental perturbations, robustness promotes sparse topologies to localize the spread of failures and noise [11]. For internal modules that are relatively insulated from the environment, efficiency is promoted via more connected and redundant structures. While we believe that the most likely explanation for the topological differences we observed in yeast between internal and external modules is due to different robustness requirements, we cannot completely rule out alternative explanations that may be based on the compartments the proteins are active in, or on other issues. Still, the fact that we observed similar efficiency–robustness trade-off between internal and external subnetworks at the cellular and ecological [45] levels suggests that these principles may be deeply ingrained into overall biological network design.

An important question is the evolutionary origin of robustness of biological systems [46]. It is unlikely that a biological system evolves to become robust to genetic knockouts, as these are rare in nature. An interesting hypothesis is that biological systems are constantly subject to environmental insults; thus being robust to non-genetic variations could provide a key evolutionary advantage [47,48]. The same design may also help the organism withstand genetic variations, including experimental genetic manipulations. Our generative model may help explain how different module topologies emerge using a simple evolutionary process (gene duplication). One characterization of evolutionary rate uses the dN/dS ratio, which measures the selection pressure acting on a gene by computing the ratio of non-synonymous substitutions to synonymous substitutions per site; values more than 1 indicate positive (Darwinian) selection and values less than 1 indicate purifying (stabilizing) selection. We observed that the dN/dS ratio of genes was lower in internal modules (-0.218 correlation with module localization score and 0.207 correlation with module gene expression variation). This indicates that the value of the q_{mod} parameter, which we varied in the duplication model, may be linked to the level of environmental exposure of genes in a subnetwork (high $q_{\mathrm{mod}} \rightarrow$ external module) and may provide a mechanistic explanation to the observed topological differences between internal and external modules. Further, our observation that environmental noise is a key determinant of both module topology and robustness also provides support to the hypothesis that genetic robustness is

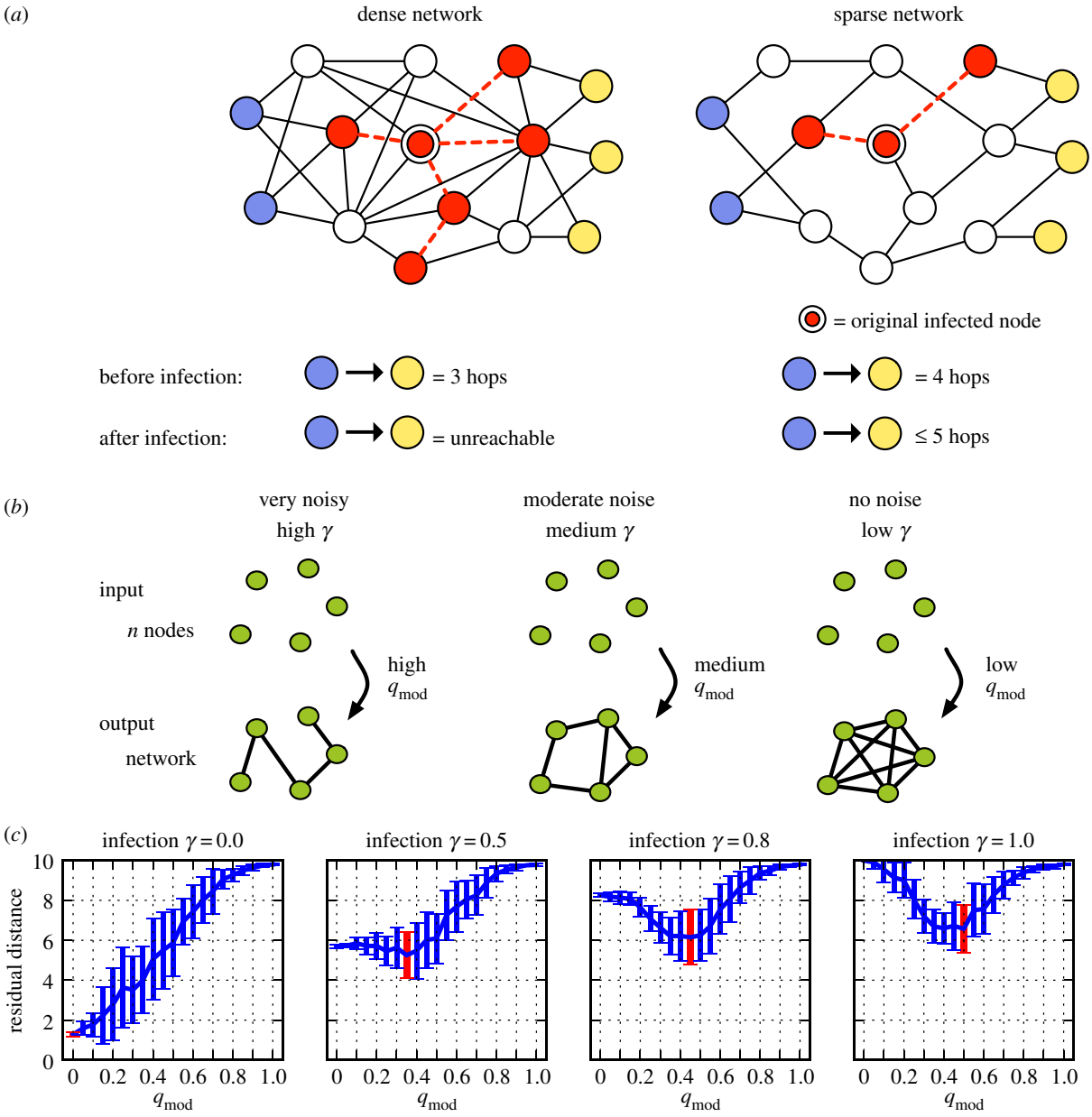

Figure 4. Security benchmark and robust network design. (*a*) Computing the residual distance of a graph. Dense network topologies are more sensitive to spreading infections than sparse topologies. Prior to infection, paths in the dense network are shorter, but after removal of infected nodes, the residual graph contains many disconnected pairs of nodes (e.g. purple–yellow nodes), leading to a large residual distance. (*b*) Examples of module topologies that can arise using three different values of q_{mod}, depending on the environmental variability, γ. (*c*) The effect of noise variability (γ) on optimal network design. For varying values of infection γ, we grow networks using our duplication-based generative model and plot the residual distance between pairs of nodes after infection. For $\gamma = 0$, cliques emerge as optimal. As γ increases, sparser designs are preferred.

evolutionarily selected as a by-product of selection of tolerance to environmental (non-genetic) variability.

Our modified duplication–divergence model also opens the door to a hierarchical network organization by changing the q_{mod} parameter for different potential sets of neighbours. Further analysis and discussion of this possibility is beyond the scope of this paper and is left for future work. Moreover, while the modified generative model fits the observed biological modules well and can be used to explain some evolutionary aspects of these networks, it is not intended to model the evolution of online social networks. On the other hand, our model is designed to generate networks with varying levels of robustness depending on the amount of noise the network is expected to encounter.

Recent work has also looked at how coupling between multiple networks may affect network performance and robustness. For example, power stations on the grid depend on communication nodes on the Internet for control, and communication nodes depend on power stations for their electricity supply [49]. Buldyrev *et al*. [49] assume random failure of nodes, which as we show, is not likely in biology: some proteins are more likely to fail than others and adjusting topology to account for this fact is important. Further, theoretical analyses of dependencies [50] have sought to increase robustness by maximizing algebraic connectivity, which is related to node and edge connectivity, but such measures can be optimized using all-to-all topologies (cliques), which as we discussed, is highly vulnerable to cascading failures in biological systems. Nonetheless, dependencies between entities across biological networks (e.g. proteins and metabolites) are likely to be an important area of future investigation. Previous works have also measured robustness (of single

networks) by determining the size of the largest connected component after some number of nodes has been removed from the graph [51]. As mentioned above, given enough resources, this measure can also be optimized using a clique. Further, this measure does not penalize a network for being inefficient (in terms of path lengths between nodes) despite being connected, which our residual distance captures.

While previous studies have documented the topological similarities of biological and computational networks (e.g. power-law degree distributions [52] and network motifs [12]), few have attempted to use insights derived from the structure of biological networks to inform the design of robust computing networks. Our security attack benchmark can be used to identify vulnerabilities in a network, and our generative model can be used to design communication networks that are specifically optimized to operate in environments with variable noise. Our results also connect with recent findings in computational security, which has shown that networks with limited connectivity are the most robust for overcoming sybil attacks (where a computational system is overtaken by users who forge fake identities) [53]. By contrast, when communication can be trusted (e.g. when using encrypted messages), redundancy provides several advantages including efficiency and reduced information leakage [54]. Thus, our findings add to recent work that highlights the advantages of joint biological–computational studies which benefit both fields [55,56].

4. Material and methods

4.1. Networks, modules, annotations and correlation coefficients

We constructed a yeast interaction network from several databases and recent publications [57,58]. Essentiality information was taken from the Saccharomyces Genome Deletion Project [59] and modules for yeast were based on gene ontology categories that are known to be involved in key processes in YPD. Both gene essentiality and gene expression are measured independently from the network.

Metabolic networks for 75 bacterial species were from Ma & Zeng [60]. Each species was annotated with a niche breadth score from the GenomeMine database [61]. This score ranges from 1 to 5 and indicates whether the species inhabits a narrow and stable environment (5) or a highly complex and dynamic environment (1) [26].

The *C. elegans* neuronal network was collected from Varshney *et al*. [62]. Neurons were divided into two categories—internal (interneurons) and external (sensory or muscle neurons)—using annotations from the Worm Atlas [63].

The correlation coefficients of gene features versus gene essentiality are computed using the Kendall tau correlation. We used this correlation measure because the target variable (gene essentiality) is binary 0 or 1. For correlations of module features versus module essentiality, we used the Pearson correlation because module essentiality can be any real value between 0 and 1.

The five time-evolving snapshots of the Gnutella file-sharing P2P networks from August 2002 were obtained from Leskovec *et al*. [36], the largest of which contained 10 876 nodes and 39 994 edges (nodes correspond to Gnutella hosts and undirected edges represent connections between hosts). The five autonomous system networks were obtained from Leskovec *et al*. [36], the largest of which contained 6127 nodes and 13 257 edges (nodes represent routers on the Internet). The six snapshots are equally spaced six months apart from 8 November 1997 to 8 November 1999.

See the electronic supplementary material for complete details of all data collected and analysed.

4.2. Infection model

We adopt a SI model for virus propagation. We are given a graph $G = (V, E)$ with n nodes and an initial infected node $u \in V$. Our model proceeds in discrete time steps and in each step, all infected nodes attempt to pass the virus to each uninfected neighbour independently with probability β. The state of the network at time t depends only on the state of the network at time $t-1$, and hence this model represents a Markov chain. Let $P_{i,t}$ be the probability that node i is infected at time t. Let $\zeta_{i,t}$ be the probability that node i does not receive the virus from any of its neighbours in the next iteration. This occurs if each infected neighbour fails to pass on the virus, or if the neighbour is not infected. Assuming the infection probabilities are independent: $\zeta_{i,t} = \prod_{j \in N(i)} P_{j,t-1}(1-\beta) + (1 - P_{j,t-1})$, where $N(i)$ represents the neighbours of i in G.

A node at time t will be uninfected if it was uninfected at time $t-1$ *and* it did not receive the infection from any of its neighbours, i.e. $P_{i,t} = 1 - (1 - P_{i,t-1})\zeta_{i,t}$. Initially, $P_{u,0} = 1$ and for all $v \neq u$, $P_{v,0} = 0$. Let η_t be the expected number of nodes infected at time t. The size of an outbreak caused by the infection of u can now be measured as: E [no. of infected nodes after t iterations] $= \sum_{i=1}^{n} P_{i,t}$.

Typically, the steady state of this equation is solved numerically in the limit as $t \to \infty$ [23] by defining the following dynamical system: $\boldsymbol{P}_t = \boldsymbol{P}_{t-1} \cdot \beta\mathbf{A}$, where $\mathbf{A}$ is the adjacency matrix of G. Chakrabarti *et al*. [23] proved that for any arbitrary network and for various virus propagation models, in addition to SIS [39], the epidemic threshold is exactly related to $1/\lambda_{1,A}$, i.e. the higher the first (maximum) eigenvalue of the adjacency matrix, the more likely it is that a single node infection will lead to a global infection.

4.3. A limited-duration infection model

We prove that if we do not allow recovery of individual nodes (either because there is no 'individual' treatment or because treatments are done by taking nodes offline) then for connected networks, any infection of a single node will always converge to a full epidemic state (electronic supplementary material, theorem 1). In practice, however, infection can be detected after some time. Our model assumes an infection is detected and removed after t steps (i.e. we expect transient noise in biological systems which after some short time leads to correct behaviour of nodes even without any intervention; in cases of attacks on computer networks, nodes are taken offline once the attack is detected). We thus simulate the infection process for t rounds: in each round, all infected nodes spread the virus to each of its uninfected neighbours independently with probability β.

4.4. Topological features for nodes and modules

Using the infection model, we define the *infect* size of a node as η_t/n, i.e. the percentage of nodes that become infected at the end of round t (averaged over 10 restarts). We also define the infect size of a module or network G as the average of the infect sizes of each node in the module or network. In this paper, we used $t = 5$ and $\beta = 0.1$: we selected $t = 5$ based on analysis of typical pathway lengths in protein interaction databases [64], and we set $\beta = 0.1$ to centre the average score [23] ($\beta \times \overline{\lambda_{1,A}}$) of a virus on a PPI module subgraph around 1.0. This model is similar to a SIR model with $\beta = 0.5$ and $\delta = 1$ (where δ is the probability that a node recovers in the next time step after being infected).

We also computed several additional metrics of a given node in a network, including its degree, centrality (eigenvector and betweenness) and PageRank.

For the network G, we also computed its density, defined as the number of edges divided by the number of possible edges; its normalized eigenvalue, defined as the first (maximum) eigenvalue of its adjacency matrix divided by the number of nodes in G; its normalized maximal independent set size (i.e. the size of the largest set $S \subseteq G$ such that no two nodes in S are adjacent, divided by the number of nodes in G); and its average degree over all nodes in G. Normalization is performed to account for variability in the size of the modules. All features are either computed within the global network or the subnetwork induced by a module. See the electronic supplementary material, tables S4 and S5, for full list of features compared.

4.5. Security attack benchmark

To model the tension between a propagating virus and the connectivity of the residual graph, we performed the following steps. First, an initial node is infected and the virus is spread using. The SI model above (which is commonly used to simulate attacks on computer networks [32,23]) and all resulting infected nodes are isolated. Then, we compute the *residual connectivity* defined as the probability that there exists a path between any random pair of nodes (including both infected and uninfected nodes) in the residual graph, and we average this probability over 100 randomly chosen initial nodes. Similarly, the *residual distance* is the shortest path routeing distance between all pairs computed within the residual graph after infection. All residual measures are computed starting from the maximal component of the initial (uninfected) graph. If a pair of nodes is disconnected after infection, their distance incurs a fixed penalty of 10. See the electronic supplementary material for full details.

Acknowledgements. The authors thank Itamar Simon and Stuti Agrawal for help selecting the biological processes most relevant to the different conditions in yeast.

Data accessibility. Data tables and algorithm implementations are available by request.

Funding statement. This work was supported in part by the National Institutes of Health award no. F32-MH099784 to S.N.; and by the James S. McDonnell Foundation programme on Studying Complex Systems and the US National Science Foundation award no. DBI-0965316 to Z.B.-J.

References

1. Kitano H. 2004 Biological robustness. *Nat. Rev. Genet.* **5**, 826–837. (doi:10.1038/nrg1471)
2. Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J. 2004 Robustness of cellular functions. *Cell* **118**, 675–685. (doi:10.1016/j.cell.2004.09.008)
3. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003 Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66. (doi:10.1038/nature01198)
4. Kelley R, Ideker T. 2005 Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566. (doi:10.1038/nbt1096)
5. Li J, Yuan Z, Zhang Z. 2010 The cellular robustness by genetic redundancy in budding yeast. *PLoS Genet.* **6**, e1001187. (doi:10.1371/journal.pgen.1001187)
6. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS. 2006 Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–846. (doi:10.1038/nature04785)
7. Kitano H, Oda K. 2006 Robustness trade-offs and host–microbial symbiosis in the immune system. *Mol. Syst. Biol.* **2**, 2006.0022. (doi:10.1038/msb4100039)
8. Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, Bar-Joseph Z. 2009 Backup in gene regulatory networks explains differences between binding and knockout results. *Mol. Syst. Biol.* **5**, 276. (doi:10.1038/msb.2009.33)
9. Albert R, Jeong H, Barabasi AL. 2000 Error and attack tolerance of complex networks. *Nature* **406**, 378–382. (doi:10.1038/35019019)
10. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. 2007 The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* **3**, e59. (doi:10.1371/journal.pcbi.0030059)
11. Kirschner M, Gerhart J. 1998 Evolvability. *Proc. Natl Acad. Sci. USA* **95**, 8420–8427. (doi:10.1073/pnas.95.15.8420)
12. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002 Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827. (doi:10.1126/science.298.5594.824)
13. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001 Lethality and centrality in protein networks. *Nature* **411**, 41–42. (doi:10.1038/35075138)
14. Hahn MW, Kern AD. 2005 Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**, 803–806. (doi:10.1093/molbev/msi072)
15. Siegal ML, Bergman A. 2002 Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl Acad. Sci. USA* **99**, 10 528–10 532. (doi:10.1073/pnas.102303999)
16. Leclerc RD. 2008 Survival of the sparsest: robust gene networks are parsimonious. *Mol. Syst. Biol.* **4**, 213. (doi:10.1038/msb.2008.52)
17. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999 From molecular to modular cell biology. *Nature* **402**(Suppl. 6761), 47–52. (doi:10.1038/35011540)
18. Acencio ML, Lemke N. 2009 Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinform.* **10**, 290. (doi:10.1186/1471-2105-10-290)
19. Seringhaus M, Paccanaro A, Borneman A, Snyder M, Gerstein M. 2006 Predicting essential genes in fungal genomes. *Genome Res.* **16**, 1126–1135. (doi:10.1101/gr.5144106)
20. Zotenko E, Mestre J, O'Leary DP, Przytycka TM. 2008 Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* **4**, e1000140. (doi:10.1371/journal.pcbi.1000140)
21. Song J, Singh M. 2013 From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLoS Comput. Biol.* **9**, e1002910. (doi:10.1371/journal.pcbi.1002910)
22. Tong H, Prakash BA, Tsourakakis C, Eliassi-Rad T, Faloutsos C, Chau DH. 2010 On the vulnerability of large graphs. In *Proc. 10th Int. Conf. on Data Mining*, pp. 1091–1096. Washington, DC: IEEE Computer Society. (doi:10.1109/ICDM.2010.54)
23. Chakrabarti D, Wang Y, Wang C, Leskovec J, Faloutsos C. 2008 Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.* **10**, 1:1–1:26.
24. Kaern M, Elston TC, Blake WJ, Collins JJ. 2005 Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464. (doi:10.1038/nrg1615)
25. McGary KL, Lee I, Marcotte EM. 2007 Broad network-based predictability of *Saccharomyces cerevisiae* gene loss-of-function phenotypes. *Genome Biol.* **8**, R258. (doi:10.1186/gb-2007-8-12-r258)
26. Morine MJ, Gu H, Myers RA, Bielawski JP. 2009 Trade-offs between efficiency and robustness in bacterial metabolic networks are associated with niche breadth. *J. Mol. Evol.* **68**, 506–515. (doi:10.1007/s00239-009-9226-5)
27. Vazquez A, Flammini A, Maritan A, Vespignani A. 2003 Modeling of protein interaction networks. *Complexus* **1**, 38–44. (doi:10.1159/000067642)
28. Middendorf M, Ziv E, Wiggins CH. 2005 Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. *Proc. Natl Acad. Sci. USA* **102**, 3192–3197. (doi:10.1073/pnas.0409515102)

29. Navlakha S, Kingsford C. 2011 Network archaeology: uncovering ancient networks from present-day interactions. *PLoS Comput. Biol.* **7**, e1001119. (doi:10.1371/journal.pcbi.1001119)
30. Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW. 2013 The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. *PLoS Genet.* **9**, e1003176. (doi:10.1371/journal.pgen.1003176)
31. Helbing D. 2013 Globally networked risks and how to respond. *Nature* **497**, 51–59. (doi:10.1038/nature12047)
32. Moore D, Shannon C, Voelker GM, Savage S. 2003 Internet quarantine: requirements for containing self-propagating code. In *Proc. 22nd Annual Joint Conf. of IEEE Computer and Communications Societies*, vol. 3, pp. 1901–1910. (doi:10.1109/INFCOM.2003.1209212)
33. Albert R, Albert I, Nakarado GL. 2004 Structural vulnerability of the North American power grid. *Phys. Rev. E* **69**, 025103. (doi:10.1103/PhysRevE.69.025103)
34. Chan H, Perrig A. 2003 Security and privacy in sensor networks. *Computer* **36**, 103–105. (doi:10.1109/MC.2003.1236475)
35. Brumley D, Liu LH, Poosankam P, Song D. 2006 Design space and analysis of worm defense strategies. In *Proc. ACM Symp. on Information, Computer and Communications Security*, pp. 125–137. New York, NY: ACM. (doi:10.1145/1128817.1128837)
36. Leskovec J, Kleinberg J, Faloutsos C. 2007 Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* **1**, 2. (doi:10.1145/1217299.1217301)
37. Ripeanu M, Foster I, Iamnitchi A. 2002 Mapping the Gnutella network: properties of large-scale peer-to-peer systems and implications for system design, *IEEE Internet Comput. J.* **6**, 2002.
38. Aron J. 2011 Amazon server failure knocks out social web. See http://www.newscientist.com/blogs/onepercent/2011/04/amazon-server-failure-knocks-o.html (accessed 17 March 2014).
39. Prakash BA, Chakrabarti D, Faloutsos M, Valler N, Faloutsos C. 2011 Threshold conditions for arbitrary cascade models on arbitrary networks. In *Proc. 11th IEEE Int. Conf. on Data Mining*, pp. 537–546. Washington, DC: IEEE Computer Society. (doi:10.1109/ICDM.2011.145)
40. Kempe D, Kleinberg J, Tardos E. 2003 Maximizing the spread of influence through a social network. In *Proc. 9th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 137–146. New York, NY: ACM. (doi:10.1145/956750.956769)
41. Liu YY, Slotine JJ, Barabasi AL. 2011 Controllability of complex networks. *Nature* **473**, 167–173. (doi:10.1038/nature10011)
42. Chapman DB, Zwicky ED. 1995 *Building internet firewalls*, 1st edn. Sebastopol, CA: O'Reilly and Associates, Inc.
43. Carman DW, Kruus PS, Matt BJ. 2000 *Constraints and approaches for distributed sensor network security*. Glenwood, MD: NAI Labs, The Security Research Division Network Associates, Inc.
44. Yang H, Luo H, Ye F, Lu S, Zhang L. 2004 Security in mobile *ad hoc* networks: challenges and solutions. *Wireless Commun.* **11**, 38–47. (doi:10.1109/MWC.2004.1269716)
45. Haldane AG, May RM. 2011 Systemic risk in banking ecosystems. *Nature* **469**, 351–355. (doi:10.1038/nature09659)
46. Hartman JL, Garvik B, Hartwell L. 2001 Principles for the buffering of genetic variation. *Science* **291**, 1001–1004. (doi:10.1126/science.291.5506.1001)
47. Wagner A. 2007 *Robustness and evolvability in living systems*. Princeton, NJ: Princeton University Press.
48. Costanzo M *et al*. 2010 The genetic landscape of a cell. *Science* **327**, 425–431. (doi:10.1126/science.1180823)
49. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S. 2010 Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028. (doi:10.1038/nature08932)
50. Filippo Radicchi AA. 2013 Abrupt transition in the structural formation of interconnected networks. *Nat. Phys.* **9**, 717–720. (doi:10.1038/nphys2761)
51. Schneider CM, Moreira AA, Andrade JS, Havlin S, Herrmann HJ. 2011 Mitigation of malicious attacks on networks. *Proc. Natl Acad. Sci. USA* **108**, 3838–3841. (doi:10.1073/pnas.1009440108)
52. Barabasi AL, Albert R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)
53. Danezis G, Mittal P. 2009 SybilInfer: detecting sybil nodes using social networks. In *Proc. Network and Distributed System Security Symp. San Diego, CA, USA, 8–11 February 2009*. Reston, VA: Internet Society.
54. Mittal P, Borisov N. 2008 Information leaks in structured peer-to-peer anonymous communication systems. In *Proc. 15th Conf. on Computer and Communications Security*, pp. 267–278. New York, NY: ACM. (doi:10.1145/1455770.1455805)
55. Tero A, Takagi S, Saigusa T, Ito K, Bebber DP, Fricker MD, Yumiki K, Kobayashi R, Nakagaki T. 2010 Rules for biologically inspired adaptive network design. *Science* **327**, 439–442. (doi:10.1126/science.1177894)
56. Navlakha S, Bar-Joseph Z. 2011 Algorithms in nature: the convergence of systems biology and computational thinking. *Mol. Syst. Biol.* **7**, 546. (doi:10.1038/msb.2011.78)
57. Chatr-Aryamontri A *et al*. 2013 The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* **41**, D816–D823. (doi:10.1093/nar/gks1158)
58. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. 2006 An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinform.* **7**, 113. (doi:10.1186/1471-2105-7-113)
59. Giaever G *et al*. 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391. (doi:10.1038/nature00935)
60. Ma H, Zeng AP. 2003 Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277. (doi:10.1093/bioinformatics/19.2.270)
61. Gray T, Wilson G, Tett A, Thurston M, Day J. 2013 GenomeMine. See http://www.genomics.ceh.ac.uk/cgi-bin/gmine/gminemenu.cgi.
62. Varshney LR, Chen BL, Paniagua E, Hall DH, Chklovskii DB. 2011 Structural properties of the *Caenorhabditis elegans* neuronal network. *PLoS Comput. Biol.* **7**, e1001066. (doi:10.1371/journal.pcbi.1001066)
63. Altun ZF, Herndon LA, Crocker C, Lints R, Hall DH. 2002–2012 Worm Atlas database. See http://www.wormatlas.org/.
64. Gitter A, Klein-Seetharaman J, Gupta A, Bar-Joseph Z. 2011 Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res.* **39**, e22. (doi:10.1093/nar/gkq1207)