www.sciencemag.org/content/343/6177/1373/suppl/DC1

Supplementary Materials for

Control Profiles of Complex Networks

Justin Ruths* and Derek Ruths

*Corresponding author. E-mail: justinruths@sutd.edu.sg

Published 21 March 2014, *Science* **343**, 1373 (2014)
DOI: 10.1126/science.1242063

This PDF file includes:

Materials and Methods

1 Background

We review relevant theoretical results from both the control of structured systems as well as the study of dynamical networks.

1.1 Dynamical Networks

We consider a linear dynamic model of the directed network graph $G(A)$, which is composed of N nodes and L directed edges between nodes in the network. Systematically analyzing linear models is a key step in generalizing to broader classes of models, such as nonlinear dynamics. In addition, there are a large number of network phenomena that exhibit a good fit to a linear time-invariant model of the form [1, 26, 27],

$$\frac{d}{dt}x(t) = Ax(t),$$

where the state $x(t) \in \mathbb{R}^N$ is the value of all of the nodes at time t and $A \in \mathbb{R}^{N\times N}$ is the transpose of the adjacency matrix of the network, such that the value $A_{i,j}$ is the weight of a directed edge from node x_j to node x_i and zero if there is no such edge. In what follows, we study the controlled network $G(A, B)$, corresponding to adding m control nodes yielding the form,

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \tag{1}$$

where the control $u(t) \in \mathbb{R}^m$ and $B \in \mathbb{R}^{N\times m}$ models the connection of the controls into the network.

1.2 Controllability of Dynamical Networks

As with any engineering system, we ultimately desire to interact with networks so as to modify them for new purposes or to optimize their function in some way (e.g., to achieve greater efficiency, robustness, or performance). Fundamental controllability properties of networked systems (equivalently structured systems) have existed in the literature for several decades [5, 28, 29, 30]. However, with the past focus on control now shifting to center on networks, the implications of such analysis has yielded new insights into the properties of dynamical networks.

The control of a networked system modeled by (1) is concerned with guiding the system state, the value of the nodes, from an initial configuration $x(t_0) = x_0 \in \mathbb{R}^N$ to a final configuration $x(t_1) = x_1 \in \mathbb{R}^N$. Due to their size and our inability to know all parameters, complex systems of interest are often modeled as interaction networks in which direction of influence may be known, but not the specific weighting of that influence. In those networks for which we do have weighting values, they often include significant uncertainty or variation. In response to these unknown or uncertain edge weights, the controllability of complex networks has focused on the generic properties, independent of specific parameter values.

1.3 Examples of Control of Dynamical Networks

The desire to control a complex network arises in wide ranging areas of application, including natural, social, and engineered systems. We discuss three examples below to frame the context of this work.

Neural Networks. The brain can be represented as a network on a variety of levels. For diagnosis and treatment of diseases, a macroscopic level of detail is most useful because clinical data is typically collected by EEG, which records the excitation of a macroscopic voxel of brain cortex. In this model each node in the network is a voxel of cortex and the state that is associated with this node is the level of excitation at a given point in time. Edges represent the overall connectivity of one brain region to another (typically this is found by looking at the correlation in the EEG signals, but can also be determined from anatomic studies). Electrodes can be used to stimulate brain voxels, usually in an attempt to generate a motion or sensation, or to eliminate pathological neuronal behavior, such as epileptic seizures. The question of control (electrode) placement is eminently important to accomplish this objective in a minimally invasive manner.

Social Networks. The interactions that occur in large-scale social systems are often modeled using networks. In such applications, people are nodes in the network and links represent the connections (friendship, association, trust) between people. Many behaviors can be studied with a social network; one avenue of investigation is to understand the propagation of sentiment within a population. In this case, the state of each person in the population is given by their sentiment towards a particular topic, which can be encoded as a real number, a positive value indicating favor and negative value indicating opposition. In order to maximize the effect of advertising, again the question of control placement - influential individuals to target the advertisements - is highly useful towards changing the minds of people in the population.

Transportation Networks. The flow of traffic on transportation infrastructures is very often modeled by networks. In an airport network, airports are nodes and flights between airports establish edges in the network. The state of the system then encodes the number of passengers traveling from one destination to another. Ideally an airport traffic management system would directly regulate as few points as possible, which could be identified as the control locations in the network.

1.4 Structural Controllability

Interest in structured systems originated from the need to analyze dynamical systems for which only the structure of zero and non-zero elements was known, but not the exact values of the nonzero elements [5]. We update our definition of (1) such that the matrices A and B are structured matrices, which have a given set of so-called fixed-zero elements and the remaining free elements which can take any real value, provided they remain uncorrelated with each other. We study the generic properties of this system, i.e., those that hold for *almost all* parameter values. It is understood that the analysis of a generic property rather than the standard definition of the property is equivalent for all parameter values except a proper algebraic variety in the parameter space, which is of Lebesgue measure zero [7].

In particular we are interested in the controllability of a structured system (A, B), which is determined by irreducibility and full generic rank conditions on $[A\ B]$ [30, 7]. These linear algebra conditions and methods have corresponding network analogs on the control-augmented graph $G(A, B)$ which provide efficient methods for dealing with large-scale networks and the sparsity inherent in the structural information. The control-augmented graph is formed by adding additional control "nodes", one for each control, to the original graph $G(A)$.

In his seminal paper, Lin formed a network representation of structural systems and subsequently defined key structures within the network that obey various rules for controllability. A *path* exists from $\mathcal{S}_0$ to $\mathcal{S}_1$ if there is an integer ℓ and nodes $x_0, x_1, \ldots, x_\ell$ such that the *bottom node* $x_0 \in \mathcal{S}_0$, the *top node* $x_\ell \in \mathcal{S}_1$, and $A_{i,i-1} \neq 0$ for $i = 1, 2, \ldots, \ell$. Similarly, a *cycle* exists if there is a path such that $x_0 = x_\ell$. We can refer to simple paths as those in which a node occurs no more than once and disjoint paths if they are composed of disjoint sets of nodes. Finally a U-rooted path has its bottom node in the set of control nodes.

With these elementary concepts we can build the higher level structures that influence controllability. A *bud* is a cycle with with an additional edge, the *distinguished edge*, that enters a node of the cycle from the so-called *initial node*. A *stem* is a U-rooted path. We call a collection of stems and buds a *cactus* in the following way. A stem is a cactus. In addition, a stem and collection of buds is a cactus if the initial node of each of the buds is not the top node of the stem and is the only node belonging to both the bud and the rest of the cactus [31].

A fundamental result from structural controllability states that the linear structured control system (A, B) is structurally controllable if and only if the control-augmented graph $G(A, B)$ is spanned by cacti. Finding these cacti is equivalent to confirming the irreducibility condition on $[A\ B]$, since the cacti is a minimal structure such that removing any edge will render the system uncontrollable.

Uncontrollable nodes arise from two scenarios, inaccessibility and dilations. A node is inaccessible if there is no directed path reaching the node from any of the input nodes. Inaccessible nodes are nodes that are simply not reachable from the input nodes, hence it is not possible to exert a controlling influence over them. A dilation exists in the graph $G(A, B)$ if a subset of nodes in $G(A)$, called S, can be found such that the number of nodes in the inbound neighborhood set of S, given by $|T(S)|$ is smaller than the number of nodes in S, given by $|S|$. The inbound neighborood set of S is the set of nodes with directed edges into S. Loosely speaking, dilations imply an expansion in the network whereby there is not a sufficient number of independent inputs to control all nodes in S. There are at most $|T(S)|$ independent controls leading into S and $|T(S)| < |S|$. The existence of cacti that span the graph $G(A, B)$ relies upon the control-augmented graph having no inaccessible nodes and no dilations.

In most large-scale networks, the structure of B is not declared a priori, hence a dominating interest in structural controllability of networks is to construct a B, in particular a minimal B ($B \in \mathbb{R}^{N\times m}$ with smallest possible m), such that there are no inaccessible nodes and there are no dilations. The maximum matching algorithm from graph theory finds the set of disjoint and simple paths and cycles that maximally cover the nodes of the network $G(A)$ [32]. The nodes without matched inbound edges identify the location of either inaccessible nodes or dilations in $G(A)$. By introducing control nodes which connect inputs to these unmatched nodes, we can create a spanning cacti for $G(A, B)$ without inaccessible nodes or dilations. Cycles which are not buds (no distingushed edge exists from a node in a stem to a node in the cycle) become buds with a new distinguished edge from a new control node to a node in the cycle. The matrix B encodes the new edges from the control nodes to the unmatched nodes and the new distinguished edges from the control nodes to the non-bud cycles. Fig. 1 in the manuscript shows a small network which exhibits one source (an inaccessible node, node A) and two dilations (occuring at the branching at nodes A and E). The maximum matching identifies the stems and cycles. A minimal control configuration B is chosen in Fig. 1 with three controls driving three stems (in blue) of length 2, 4, and 1 respectively. The matrix B would have two nonzero entries in the first column, which corresponds to the attachments of the first control, at the rows corresponding to nodes A and K. The second and third columns would each have one nonzero entry on the rows corresponding to nodes B and F, respectively. The first control (indicated by the number 1) is attached to the root of the stem (node A) to compensate for the inaccessbile node and also to the independent cycle (at node K, showing it is controlled as a bud). The second control is attached at node B to compensate for the dilation at node A (internal dilation). The third control is attached at node F to compensate for the dilation at node E (external dilation). The second control drives the bud (green) indirectly by way of the distinguished edge from D to H. Several control configurations exist for the same number

of minimum controls, only one is illustrated in Fig 1. Additional controls could be added, however, since the $G(A, B)$ is already controllable, it will not change the controllability of $G(A, B)$.

Although the criteria for structural controllability is well-defined, the local view imposed by checking for inaccessibility and dilations does not generate an understanding of how the function of various networks and their idiosyncratic structures yield different types of control architectures. So although the cause of controllability can be traced back to avoiding inaccessibility and dilations, we seek a framework that yields an understanding of the functional origin of controls in complex networks.

1.5 Comparisons to Existing Control Classification Systems

The central goal of our work is to classify each controlled node according to the property which justifies its designation as a control. To our knowledge, two control classification paradigms have been previously proposed within the context of complex systems [33, 3]. These are related to ours insofar as they categorize nodes/links in control structures. However, they do not approach the same question considered here. The work of Jia et al. classifies nodes according to whether they must, can, or are never included in a control configuration [3]. This perspective addresses the question of how many different nodes may participate in control, but not why (though they do point out that source nodes will always require controls). The work of Liu et al. proposes a framework for quantifying how many edges in the network are always, sometimes, or never part of the control structure [33]. This, again, gives insight into the amount of the network that participates in control, but not why. We suspect that, in future work, our control profile could be fruitfully coupled with these other classification schemes to offer explanatory insight into why nodes/links belong to the classes they do.

Another aspect that differentiates our work from previous classification schemes is the application of our classification to a wide array of real-world networks. In order to further underscore the relevance of our definition of control profile, in the paper we provide an extensive analysis of the control profiles of many real-world networks, showing that our framework provides a novel and meaningful way of comparing real-world networks. The analysis culminates in the finding that real-world network control profiles fall into three well-supported classes that have structural and functional significance. Among the frameworks mentioned above, ours is the only one so far shown to provide such a categorical way of approaching real-world systems.

2 Methods

2.1 Calculating the Number of Controls

2.1.1 Maximum Matching

In graph theory, the maximum matching problem seeks the largest set of nodes that can be uniquely paired amongst themselves using edges present in the network. This problem would be trivial were it not for the constraint that a node can be paired with at most one other node. The solution to the maximum matching problem is the subset of edges in the network that provide the pairing. [1]

A number of efficient algorithms are known for solving this problem on bipartite networks: networks (1) which have nodes of two "types" and (2) whose edges only connect nodes of different types. For sparse networks, the Hopcroft-Karp algorithm has the best worst-case performance of $O(L\sqrt{N})$, where N is the number of nodes in the network and L is the number of edges [32]. The algorithm works by identifying and extending *augmenting paths*, paths which are composed of alternating matching/non-matching edges. Such alternation is a requirement since two matching edges cannot be adjacent to one another; if they were, then the node they share in common would be matched twice, which invalidates the solution.

Of course, most natural networks of interest are not bipartite networks. While some maximum matching algorithms have been developed for other kinds of networks, the maximum matching problem can be solved on general directed and undirected networks by simply obtaining a bipartite representation of the network and then running a bipartite maximum matching algorithm on this bipartite network. This approach has been used in an array of other papers concerning controllability and networks (e.g., [7, 1]). This is the approach we use throughout this project as well.

The output of the maximum matching algorithm on a bipartite graph is the collection of matched nodes, in pairs, and a set of nodes for which no match could be found. When a directed network is represented as a bipartite graph and the maximum matching algorithm is run, the corresponding output is the set of edges that is contained in the maximum matching. These edges are necessarily non-overlapping and form paths and cycles through the network. From this, we can extract all nodes that do not have an incoming edge from the matching. These nodes are either not contained within any of the stems or cycles, or they lie at the beginning of a stem, since any other node in a stem or cycle will have an incoming edge from the set of matched edges. It is this group of nodes that must be directly controlled through the application of external controls, one for each unmatched node. Therefore, the matching gives the minimum number and location of the external controls. The stems and cycles of the matching can be seen to intuitively represent the propagation of control through the network, and creates the cacti that governs the control properties of the network.

[1] Note that the maximum matching problem is separate from the maximal matching problem, whose solution is a set of edges with no common nodes such that no additional edges may be added without having a node in common. Although a maximum matching is maximal, the converse is not true.

2.1.2 Special Case: $N_s = N_e = N_i = 0$

Networks with $N_s = N_e = N_i = 0$ (i.e., no source or sink nodes and no internal dilation points) arise due to being extremely well connected. In this extreme case, there are no source or sink nodes and the maximum matching is composed entirely of cycles (no stems). Although $N_s = N_e = N_i = 0$, the cycles must be driven by at least one common control, hence $N_c = 1$. Note that it is appropriate that $N_i = 0$ since there can be no dilations if there are no stems. Two real networks from our data sets meet this criteria: manuf-familiarity-rev (Intra-Organizational) and airports500 (Airports, Top 500). The airports500 network is a selectively chosen subset of all the airports, retaining the 500 busiest airports (in terms of passengers moved). It is intuitive that this core of the entire airport network will be well-connected. There are a number of Erdos-Renyi networks that also evidence this behavior, especially as average degree increases. Because the control profile of these networks is $(0, 0, 0)$, it does not satisfy $1 = \eta_s + \eta_e + \eta_i$ and does not lie within the orthogonal projection of this 2-simplex. From this reasoning, we omit these networks from the control profile plots.

This special case can also arise if we assume intrinsic nodal dynamics for each node (e.g., the change in a node's state is a function of its current state). The structural implication of such dynamics is that a self-loop is present at each node in the graph — resulting in a graph which consists of $|N|$ cycles of length 1. In this case, we also obtain $N_s = N_e = N_i = 0$ since all cycles can be driven by a single control [34]. There are at least two reasons why we may choose to analyze the network controllability in the absence of such self-loops. First, assigning a single control to hundreds or thousands of self-regulating nodes will generally not yield tractable, practical control schemes; in such situations, many more controls will be needed. Second, the timescale of such intrinsic nodal dynamics (e.g., molecule degradation and mortality) can be much longer than many of the interactions present in the network, again raising the question of whether these intrinsic mechanisms can serve as the basis for a practical control scheme. A more meaningful analysis would result by first culling edges with weights (timescale rate constants) below a certain threshold.

2.1.3 Control Allocation

It is important to note that the maximum matching is not unique - that is, although the number of controls stays the same the allocation of controls to nodes is different. Consider two simple cases: a symmetric "Y" shaped graph and a graph consisting of a single straight path and a disjoint cycle. These are both examples of "degenerate" multiplicity in the maximum matching. In the former case, it is necessary to place a control at the lowest node in the "Y", but it is not particularly important whether we consider the stem as continuing left and we allocate the second necessary control to control the right branch, or we consider the stem as continuing right and allocate the second control to control the left branch. Here there are two controls and two choices of control configurations. In the latter case, a single control drives two nodes - one at the base of the path (stem) and one in the cycle, however, it is not important which of the nodes in the cycle receives the control. Therefore, there is one control and $1+n_{\text{cyc}}$ number of potential control configurations (where n_{cyc} is the number of nodes in the cycle). More exotic configurations can certainly be imagined with more complex network structure. As a final example, consider a network with k stems and ℓ cycles. Because cycles self-regulate (from the perspective of structural control), any number of cycles can be driven by a single control. Therefore, we need k controls connected to $k + \ell$ nodes, however, we could arbitrarily choose to allocate the first control to drive the first stem and the first cycle, or drive the first stem and all the cycles. Considering all of the potential configurations in between these extreme alternatives yields a great many options.

2.2 Prediction by Source and Sink Nodes

The bases for control in some networks are well characterized through the input/output nature of their connectivity alone. In particular, we noted that many of the surveyed synthetic networks, ER, BA and LA, fall into this category, as well as certain classes of real networks. This is significant because it implies that some real world networks have a structure distinctly different from what is found through these synthetic construction mechanisms. It is furthermore significant that this feature arises in the context of control of networks, which is a new perspective to take in network analysis - especially with regard to functional properties.

One of the major contributions of Liu et al. 2011 was that they showed that the minimum number of controls required to drive a network was maintained under a degree-preserving shuffle. We are now able to shed more light onto this discovery by evaluating the control features due to the input/output structure of the networks. We find that we get a nearly identical trend if we only consider the in-degree=0 and out-degree=0 portions of the degree sequence, corresponding to the source and sink nodes of the network. Figure 2 illustrates that if we use only the source and sink nodes to predict the number of controls required ($n_c^{\text{se}} = n_s + n_e$) the prediction is already very close to number estimated by a degree-preserving shuffle of the original network (n_c^{deg}) even though we have ignored the majority of the degree sequence. Moreover, inspecting the more detailed Fig. S1, we note that if the prediction given by the source/sink calculation performs poorly on a given network, then generally the correspondence with the degree-shuffle is also poor. In fact, the non-zero terms of the degree sequence account for only 5.1% of the fraction of controls required in real networks and 2.7% (ER), 0.6% (BA), 0.7% (LA), and 7.3% (DD) synthetic networks. These percentages are calculated as the difference in average error of the prediction $|n_c - n_c^{\text{se}}|$ for source/sink and $|n_c - n_c^{\text{deg}}|$ for the full degree distribution, where both averages are taken across all the real or synthetic networks within each category. Note that because the degree shuffle preserves the number of sink and source nodes, that N_s and N_e are the same for the source/sink prediction and the

degree-shuffle correlation. Even though the degree-shuffle correlation performs marginally better than the source/sink prediction, it is important to point out that the source/sink calculation is both predictive and causal, and the correlation with degree sequences is neither. Degree-preserved shuffling cannot be used as a predictive tool because calculating the matching to yield the number of controls is a computationally equivalent task to calculating the matching on the original network.

It is important to note that, while the basis for control in many networks can be largely characterized by the input/output characteristics of their degree distributions, internal dilations, which are *not* accounted for by degree distributions alone, make substantial contributions to the basis for control in many networks. Thus, understanding the role of internal dilations is important within the present investigation and in the broader study of control in complex systems.

2.3 Control Profiles

Control profiles offer a way to capture the origin of control in complex networks. We motivated this perspective partially due to the fact that two networks can have similar statistics in terms of the number of controls, but the reasons (structures) giving rise to these controls are substantially different.

2.3.1 Control Profiles of Synthetic Networks

We observed that all four models of network formation considered generated networks whose control profiles were dominated by source controls. The absence of external dilations can be explained in terms of two features of the formation mechanisms employed in the models.

New node attachment. As mentioned in the text, both the Barabasi-Albert (BA) and local attachment (LA) models employ a new node attachment model: at each time step, a new node is added to the network and an outbound neighborhood is formed. Thus, at every time step, a source is created. Particularly in models with preferential attachment characteristics, attachment to sources is unlikely, giving rise to a large population of source nodes, even at steady state.

Source/sink creation with equal probability. In the Erdos-Renyi (ER) formation model, each edge exists with equal probability. Thus a node has equal probability to be a source and a sink. Thus, we observe, on average, approximately an equal number of sources and sinks being created. Recall that $N_e = \max(0, N_t - N_s)$. So, a network with an appreciable number of external dilations requires that many *more* sinks be created than sources. This simply does not happen often, explaining why ER network control profiles are characterized by high source controls. As an aside, it is also relatively common for ER networks to be spanned by a set of cycles - in which case only one control is required which drives all the cycles. Because these networks have zero control profiles, they are omitted from the present analysis.

Duplication-divergence also lacks any degree of external dilations on account of generating sources and sinks with equal probability — consider that nodes are imperfectly duplicated, meaning that a random subset of the original node's edges are copied over to the new node. Thus sources are created with equal probability to sinks.

Internal dilations. What is distinctly different about duplication-divergence (DD) networks, however, is the presence of substantial numbers of internal dilations. Structural controllability establishes dilations as having a very specific structure in which there is a larger number of nodes receiving input from a smaller number of nodes: effectively a point of expansion in the network. Models with preferential attachment characteristics favor attaching nodes to a small population of high-degree nodes: this implements the exact opposite of spreading. Thus, it is not surprising that BA and LA networks have diminishingly few internal dilations. Furthermore, while ER networks do not focus linking activities on a subset of nodes, it does not favor creating expansion points either. As with creating sources and sinks, it neither creates narrowing or expansion structures, which explains why it does not generate a notable number of internal dilations.

Duplication-divergence alone yields appreciable numbers of internal dilations largely due to the fact that duplicating a node multiple times (or duplicating its duplicates) can generate precisely the kind of expansion structures which give rise to dilations: consider a node with an inbound edge. If that node is copied (with no other connectivity being added), a dilation is created. Thus, the duplication-divergence model employs a formation mechanism that favors the generation of internal dilations.

Note that, while DD does reliably generate internal dilations, there is still no way of tuning the number (or fraction) of internal dilations present in networks generated. This same issue is present in the other models considered as well as all other existing network formation models. Control structures are not yet parameterized in formation models, making this a direction for interesting and important future work.

2.3.2 Clustering of Control Profiles

The control profiles of all real networks were found to form three clusters through both visual and computational analysis. Figure S2 shows that the vast majority of real networks fall in the corners of the control profile plot. To validate these clusters, we applied k-means clustering to the control profiles, using $k = 3$ and $k = 4$ [35]. For $k = 3$, k-means returned centroids placed in

each of the three corners (source, external-, and internal-dilation extremes). For $k = 4$, the method returned centroids in the same extreme positions as well as a fourth near the middle of the space — this fourth cluster, however, has a very poorly defined boundary with the internal-dilation-dominated cluster. Further, when the numeric values of the control profiles are considered, the membership of the three clusters identified by k-means are always consistent with the names we applied to them: e.g., profiles in the source-dominated cluster have source components that are larger than the other two components (this holds for the other two clusters as well).

External-dilation-dominated networks. We observed that external-dilation-dominated networks exhibited strong aspects of top-down control. Corporate ownership, transcriptional networks, and P2P systems are known to have very pyramidal shapes [15, 17, 36]. Co-purchase networks, obtained from Amazon, presumably have this structure due to the fact that the purchase of many products (e.g., accessories) is driven by the acquisition of core products which need to be accessorized, outfitted, or complemented. Systems in which structured or institutionalized social influence occurs may also have a top-down structure, as is the case with physicians, the Wikipedia community, and classrooms [16]. We consider the messaging networks to be outliers from the other social networks due to the fact that the messaging platform from which these networks were derived was unique to a specific university and, therefore, may not have been fully adopted. This would explain the high number of sinks (people who received messages, but did not respond) [37].

As noted, top-down architecture generally implies that controls at the sources cannot drive the system to any desired state. The presence of numerous external dilations in the networks indicates that sinks outnumber sources. As a result, controls applied to sources will yield correlated behavior within the network — in particular among nodes that are downstream of common sources. This correlated behavior is, in many instances, by design: e.g., in classrooms a teacher's influence synchronizes student behavior, companies expect subsidiaries to act in concert, and accessories of products are expected to sell in proportion to that of the core product line. Thus, if we seek to *fully control* such a system, it is reasonable to expect that we will have to add controls beyond those provided as sources.

Source-dominated networks. In the manuscript, we observe that the key distinction between networks in this cluster and those in the external-dilation-dominated cluster is that the ratio of sinks to sources is less than 1 — there are fewer sinks than sources. Crucially, zero external dilations does *not* imply that there are no sinks; rather, it indicates that there simply is at least one distinct source which reaches each sink through a directed path. In such situations, the pyramidal, top-down structure described for the external-dilation networks above does not exist in networks in this cluster. A direct implication of this is that various components of such systems may have more freedom to explore state spaces than those with a top-down structure, giving rise to greater distributed processing capacities. The networks that belong to this cluster are consistent with this interpretation: both social and neuronal networks characteristically exhibit highly decentralized processing and behavioral patterns at a system-level [20, 21].

Social networks[2] exhibit extensive community structure, in which individuals will belong to a small subset of all the communities present in the system [21]. While behaviors may be correlated within such communities, behaviors can appear quite independent (or at least different) across communities [38]. Thus, a particular stimulus (e.g., the introduction of some piece of information) to the network can result in a broad range of responses in different communities.

At a macroscale, neuronal systems segment into modules that perform different, relatively independent functions [20]. While, certainly, there is communication among these modules, evidence suggests that modules enjoy the freedom to explore a large range of states, irrespective of the state of other modules. This larger state space enables individual modules (and, as a result, the overall system) to perform more complex computations than might be found in a more typical top-down topology.

Internal-dilation-dominated networks. Networks with controls that originate primarily through internal dilations lack both sources, which indicate clear inputs, and sinks, which indicate clear system outputs. One compelling interpretation of such a system is that it does not require inputs and was not designed to produce outputs: it is a closed system. Admittedly, this is not the only interpretation. For example, it is possible that proper inputs into or outputs from the system exist, but simply are involved in feedback loops from the system (giving them non-zero in-bound and out-bound neighborhoods, respectively). Despite this plausible alternative, the set of networks which belong to this cluster support the former interpretation. Each system can be characterized as a (mostly) closed system, which we operationalize by considering whether they implement laws (or weaker principles) of conservation.

Food webs are, perhaps, the most obvious example of a closed system. Designed to track the flow of biomass through an ecosystem, these typically satisfy a conservation of biomass. Electronic circuits, also, satisfy well-known conservation laws — best known is Kirchhoffs law for the conservation of current. That said, electronic circuits certainly are not *completely* closed systems since they do receive some limited amount of input. Nonetheless, it might be considered "mostly closed" on the grounds that the amount of input a substantial circuit receives is small compared to the complexity of the system itself and the complex logic it implements.

Airports, internet routing networks, intranets, and complete maps of the political blogosphere satisfy less stringent notions

[2]Social networks are different from social influence networks, which we define to be networks designed or sampled to highlight the propagation of some quantity through a population.

of conservation. Consider that airport networks are designed to satisfy a weak condition of reciprocal travel: people who make a trip in one direction will generally make the trip back (the notion of a round-trip). Internet routing (autonomous) systems satisfy a similar criterion for digital packets: the TCP/IP protocol effectively requires that traffic going from machine X to machine Y eventually be reciprocated by traffic moving from Y back to X. Intranet web portals can be considered closed in at least two regards. First, membership is restricted to only certain pages which satisfy the physical and content criteria of the portal. Second, and more importantly, member pages serve the primary purpose of allowing users to circulate among pages within the intranet — not to link out of the intranet itself [39]. A complete map of the political blogosphere (all political blogs that link to one another) may be considered closed in a way analogous to web portals: readers will circulate primarily among the blogs [40, 41]. Moreoever, the structure of the blogosphere also maps the paths of information propagation as information is digested, interpreted, and re-interpreted by cascades of bloggers that, eventually, feed back into the system for further processing [42].

One-control networks. To ensure our classification is complete, we add an additional "trivial" classification for networks with $N_s = N_e = N_i = 0$ (i.e., no source or sink nodes and no internal dilation points). Because the control profile of these networks is (0,0,0), it does not fall into one of the above categories. Although these networks demonstrate the interesting property that they can be controlled with a single control, they do not reveal meaningful information about the flow of information in the network and they also seldom appear in real-world networks. The two real-world networks that fall into this category (*airports500* and *manuf-familiarity-rev*) are a result of biased sampling.

In each of the networks above, it is worth noting that many internal-dilations would give way to sources and external-dilations were the connectivity subsampled. Crucially, subsampling a closed system renders the portion sampled no longer closed, a fact that would be immediately evidenced by the fall in dominance of internal-dilations.

2.3.3 Validating the Control Profile Classification

Characterizing the quality and accuracy of a clustering method is widely recognized to be a difficult problem when ground truth can't be established. This is certainly the case with the clusters proposed here because we do not have unambiguous indicators of which networks belong together in the same cluster. One proxy we may use is the notion that the three-cluster system groups together networks of the same type. While this is not an intended guarantee of the proposed framework, we would expect this to hold very often based on the intuition that networks of the same narrowly-defined type have similar properties. With this in mind, we can leverage this proxy to characterize the performance (and statistical quality) of the proposed control profile clusters.

The 68 real-world networks we considered were organized into conceptual types: e.g., neural, corporate ownership, peer-to-peer, and web-links. These types and their membership were largely adopted from earlier work in the network science literature (e.g., [1]) - crucially, they were not created to fit the observed clusters. Thus, one way of assessing the goodness of our three-cluster system is to look at how consistent it is with the various network types: if it often splits up network types, this might be a cause for concern. Of the 13 types, 11 contained more than one network. Of these, only one category (Social Influence) was split: the Epinions network was placed in the Internal- rather than External-dilation category. As a point of contrast, when the networks are assigned randomly to the types (maintaining the type sizes and the individual network-to-cluster assignments), there are, on average, 9.4 types split across categories ($\sigma = 1.05$). This difference is statistically significant (P-score $< 10^{-}6$).

Another, related, way of scoring the quality of the three-cluster system would assume that each network type "belongs" to the cluster into which the majority of its networks fall. We could then ask what percent of network-to-cluster assignments conflict with the network-type-to-cluster assignments. As mentioned above, only one network in one type was placed into a category different from its type (Epinions from the Social Influence type). Thus, the overall categorization error is $1/28 = 0.036$ (the raw categorization error is $1/68 = 0.014$, but this must be adjusted since we cannot include any network belonging to types with less than three members (6) as well as the number of networks required to form a majority for each type (32)). Beyond being a small error (3.6%), consider that it originates entirely from only one type (Social Influence). As shown in Fig. S2, the vast majority of real-world networks (and network types) fall far from the boundaries separating the different control profile classes.

2.3.4 Control Profiles of Randomized Real Networks

The control profile enables us to investigate the similarities and differences between the the origin of controls in real world networks and the randomized versions of these networks. Figure S3 presents the profiles of the network classifications defined above along with the profiles of their degree-, IO-, and random-shuffled counterparts. A degree preserving shuffle randomizes the network subject to the criteria that the overall degree sequence does not change. An input/output (IO) shuffle randomizes the network while maintaining the same number of inputs (source nodes) and outputs (sink nodes). This new shuffling paradigm was inspired by the observation made in this paper that the number of controls required by a network was largely explained through the number of source and sink nodes. A random shuffle randomizes the network without any constraints. The data in Fig. S3 is derived from an average of 10 shuffles of each type (this number seems to be sufficient, as the standard deviation of the number of controls is already quite low, see Table S2).

Although a full investigation into the control profiles of randomized networks is part of our future work, a few observations can be made quite readily. In particular, the control profile is qualitatively conserved best by the degree shuffle, then by the IO shuffle, then the random shuffle, as would be expected based on the literature and our findings. Note that both the degree and IO shuffle methods necessarily conserve the numbers of source and sink nodes. In any of the cases it is possible for the shuffling process to generate either more or fewer internal dilations. For example, the degree shuffle increases the number of internal dilations in the social networks (slashdot networks, see Table S2) and decreases the number of internal dilations in many of the food web networks (see Table S2). This implies that these real world networks have somewhat unique or idiosyncratic architectures which yield different, or simply a different abundance of, dilation structures than what is predicted by the degree distribution. The random shuffle seems to introduce a surprising number of internal dilations, but this is due to the fact that this type of shuffle produces relatively few source and sink nodes so the majority of controls will arise from internal dilations.

2.3.5 Control Profile Figure Construction

Figure 3 in the main text presents the control profiles of various collections of networks. The figure compares the control profiles of synthetic networks with real networks and also provides a comparision between categories of real networks. The equation $1 = \eta_s + \eta_e + \eta_i$ describes a plane in $\mathbb{R}^3$. When restricted to $\eta_s, \eta_e, \eta_i \geq 0$ (first octant), the plane forms a triangle with normal vector $(1, 1, 1)$. The figures of control profiles are projections of this triangle where the "η_s corner" corresponds to $(1, 0, 0)$, the "η_e corner" corresponds to $(0, 1, 0)$, and the "η_i corner" corresponds to $(0, 0, 1)$ (see Fig. S4). Heat maps were chosen to represent the data because they make the contrast between networks more visually apparent and because a heatmap is able to capture overlapping data points, unlike a typical scatter plot where two overlapping dots just look like one data point. The challenge we faced was that we had unequal representations of types of networks. For example, our cumulative data set includes 3 electronic circuit networks and 22 food web networks. It would be misleading to make a simple heatmap from these 25 points - specifically giving food web networks a dominant influence in the figure. To overcome this, we created heatmaps corresponding to each distinct type of network and recorded the underlying histogram that gave rise to the heatmap. When we combined types of networks, the combined histogram had bin values that were the average of the same bin across the types of networks. In the example above, the bins of the circuit networks would be averaged with the bins of the food web networks, creating a representation that values the 3 circuit networks equally to the 22 food web networks. For the real networks control profile this combination happened three times: once to combine networks of the same type, then to combine these types together within each cluster, and finally to combine the clusters into the aggregate plot. The "types" of networks that we chose aligned with the function and domain of the network; the network types are listed in the data table in Table S2. It should be noted that any heatmap will have a bias based on the way it was normalized. Our effort to normalize across types and across clusters was so that overrepresentation was avoided and so that the clusters can be identified clearly.

As mentioned in section 2.1.2, the networks satisfying $N_s = N_e = N_i = 0$ are not shown on these plots.

2.4 Software

This work was carried out using a Python-based network modeling and analysis Zero-Effort Network (Zen) library. This open-source library is maintained by one of the authors, Derek Ruths. Zen is the fastest Python network library available, with benchmarks against other popular libraries (NetworkX and igraph) available online. The synthetic network generation methods (ER, BA, LA, DD) are available in Zen, as are all the tools for both standard network analysis and control-related network analysis used in this study. To the authors' knowledge, this is the only network library which has control analysis features as part of the core functionality. The specialized routines to compute minimum control sets and cacti structures have been deployed with the publication of this article. Zen and its documentation is available online at: `http://zen.networkdynamics.org/`.

3 Data

The data we used for this study is a collection of freely and publicly available data sources. We briefly describe these collections below. All of these networks are considered directed.

3.1 Synthetic Networks

The synthetic networks were generated using the Erdos-Renyi, Barabasi-Albert, local attachement, and duplication-divergence algorithms for network generation. In all of these cases, we sampled networks from an extremely wide parameter set. It should be noted that many of the networks that do not comply with the input/output prediction tend to lie at the extreme parameter values. Hence, we consider our findings to be conservative when applied to thinking about networks (real or synthetic) that have more standard parameters.

Erdos-Renyi (ER) Following the random connection model presented in [10], we created networks with the following characteristics. Networks with 100, 250, 500, 1000, 2500, 5000 total nodes (n) were synthesized, each with an average

degree (k) of 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 34, 38, 42, 46, 50, 54, 58. In addition networks with 10000, 25000, 50000 total nodes were synthesized, each with an average degree of 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50. In both cases candidate random networks were generated until the number of edges (L) was within an acceptable tolerance: $|L - kn| < 0.001kn$. For each (n,k) pair, 10 networks were generated.

Barabasi-Albert (BA) Following the preferential connection model presented in [11], we created networks of the follow characteristics. Networks with 100, 250, 500, 1000, 2500, 5000 total nodes were synthesized, each with an average degree of 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 34, 38, 42, 46, 50, 54, 58. In addition networks with 10000, 25000, 50000 total nodes were synthesized, each with an average degree of 2, 3, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50. From n and k we can back calculate the number of nodes needed to add (and attach) at each time point (q), which yields an average degree as close as possible to the desired k value: $q = (n - \sqrt{n^2 - 4L})/2$. For the discriminant to be positive, $n > 4k$, therefore, some combinations of n and k are not possible. For each feasible (n,k) pair, 10 networks were generated.

Local Attachment (LA) Networks are created using a local attachment model in which the n nodes are added incrementally with m edges each [12]. Out of the m edges, r of them are connected randomly to existing nodes and the remaining $m - r$ edges are connected to the neighbors of the randomly chosen nodes. By increasing the latter fraction, clustering can be promoted in the network. We created networks with the following characteristics. Networks with 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 50000 total nodes were synthesized each with an average degree of 2, 3, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50. Within this networks with clustering values 0, 0.25, 0.5, 0.75. The clustering value, in this case, means that at $c = 1$, only 1 edge is randomly added, the rest are added to neighbors of the one randomly chosen node. This leads to a high amount of clustering. Because m is different for different k values, we calculate the r edges by $r = (1 - c)m$.

Duplication Divergence (DD) Networks are created using a duplication-divergence model in which a node is copied and its edges are kept with a probability s [13]. The only adjustment in order to accommodate a directed graph is to start the network with an edge from both 0 to 1 and 1 to 0 (the undirected case only has an edge from 0 to 1). We created networks with the following characteristics. Networks with 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 50000 total nodes were synthesized each with a probability to copy edges (s) of 0.1, 0.3, 0.5, 0.7, 0.9. The only exception was $n = 50000$ and $s = 0.9$; due to computational constraints, this pairing was not included.

Edge Orientation Sensitivity

Many of the synthetic network models above were originally proposed for undirected graphs. Since then they have been generalized for use in generating directed graphs. Although the orientation conventions are consistent with the definitions provided in the literature, it is worth briefly observing the sensitivity to the choice in edge orientation in the various models. For ER and DD networks, due to the symmetry in the construction of the graphs, reversing the edge orientation has no effect on the presented results. In BA and LA networks, reversing the edge orientation convention would lead to an exchange in the numbers of source and sink nodes. More importantly, however, this reversal does not change our larger observations that null models yield networks with few internal dilation points and, more importantly, lack the ability to tune the control profile of produced networks.

3.2 Real Networks

The table preceding this section lists the real networks used in this study. They were acquired from a variety of sources online and all are freely available. The table is segmented (by horizontal lines) into the fundamental types of networks we used to classify the different networks. These correspond to the types used to construct the control profile figures in section 2.3.5.

Incomplete Data

Records of complex system are incomplete due to the very means with which we record and survey them, which means that the network representations of these systems are also partially incomplete. For example, in social surveys, participants are often asked to list their top 3 or top 5 connections, which clearly misses potential links. In other cases, our lack of understanding of the system, e.g., a biological system, means that we miss components (e.g., transcription factors) of the system. No matter the mechanism, the incompleteness of data necessarily effects our analysis of these network representations. We discuss two implications particularly relevant to this work. First, incompleteness of links (or nodes) has the potential to lead to dangling nodes - nodes that appear to be either sinks or sources but are actually neither if we had complete information. This has the potential to skew our results by increasing the expected number of controls needed to drive the system. Although we can do little in our analysis to compensate for this information it is important to be aware of this possibility. Second, while it is possible that incompleteness may seemingly create "extra" controls, we posit that the input/output structure of the system is likely to be the most well-known aspects of the system. As observers of these systems, we are capable of reliably measuring the inputs and outputs of a system because they are readily apparent, however, uncovering the internal structure of the black box that lies between the inputs and outputs is a significantly more challenging process during which we are likely to miss information. Moreover, agents in the system that appear to be inputs

or outputs (sources or sinks) based on observation would be obvious flaws in our model since they had not already been identified as external signals to or from the system. Therefore, we suggest that using source and sink nodes as an estimate for the number of controls needed to drive the network, as presented in this work, is a reliable method because it depends on the aspects of complex systems which we know with the greatest level of certainty.

Edge Orientation Convention

The typical construction of some networks is defined such that if a trusts/eats/knows b, there is an edge in the graph pointing from a to b. While this relationship encodes the graph, it does not necessarily capture the network that is most meaningful for control. In particular, in many trust networks an edge pointing from a to b actually implies that b can influence a, but not necessarily that a can influence b. For this reason, in several of the networks below (where indicated) we have reversed the direction of edges in the original graph so that the network represents the path of influence. In some cases, there are alternate interpretations of what controlling the system means. For example, the food web networks below indicate that biomass flows in the direction of the arrows, i.e., prey points to predators. An alternate food web network could encode instead predators pointing to prey in a "what-eats-what" type network. Control in each case means something different - the former case would be controlling the quantities of biomass in each organism of the network (modulating the flow of biomass from fish to sharks for example) and the latter case would be controlling the populations of each organism (changes in shark population will effect fish population). Similarly, for web networks, we can either consider controlling web traffic (a points to b if a has a hyperlink to b since a passes web traffic onwards to b) or controlling website content/ideas (b points to a if a has a hyperlink to b since a gets inspiration from b). In cases where there are reasonable dual interpretations, the network was not reversed. This was only done to networks in which the reverse interpretation made significantly more sense.

Data Used in this Study

Airport Networks [43] This collection provides two graphs of airport networks. In both, vertices represent airports, and edges exist wherever there were flight(s) between the airports. The first graph represents the complete US airport network in 2010. The second is a graph of international airports. These are available online: `http://toreopsahl.com/datasets/`.

Airport Networks (Top 500)[44] This collection contains the flight paths between the 500 busiest airports worldwide, measured by number of passengers. This network is available online: `http://www.biological-networks.org/`.

Amazon Copurchase [45] This collection represents the relationship between copurchasing patterns on Amazon.com, namely indicating that the source of an edge is often copurchased with the target. This network is available online: `http://snap.stanford.edu/data/`.

Autonomous System [46] This collection contains 122 CAIDA AS graphs, from January 2004 to November 2007. Due to their similarity, we have selected the final snapshot (November 5, 2007) to be included in our analysis. All of the networks are available online: `http://snap.stanford.edu/data/`.

C. Elegans [47] This collection contains a graph of Caenorhabtitis elegans (C. elegans) worm's neural network. Neurons are vertices and edges indicate existance of at least one synapse or gap junction between neurons. This network is available online: `http://toreopsahl.com/datasets`.

Corporate Ownership [48] This collection consists of ownership relations among companies, where a directed link indicates that the source is an owner of the target. This network is available online: `http://vlado.fmf.uni-lj.si/pub/networks/data/econ/Eva/Eva.htm`.

E-coli [49] This collection contains a transcriptional regulation network for E. coli encoding 577 interactions between 116 transcription factors and 419 operons. It represents significant augmentation on top of the existing RegulonDB database. This network is available online: `http://www.weizmann.ac.il/mcb/UriAlon/`.

Electronic Circuits [50] This collection is composed of three network representations of electronic circuits parsed from the ISCAS89 benchmark collection (S8). They are available online: `http://www.weizmann.ac.il/mcb/UriAlon/`.

Email-EU [51] This collection includes the network generated from email data from a large European research institution over a period from October 2003 to May 2005. This network is available online: `http://snap.stanford.edu/data/`.

Epinions [52, 53] This collection contains a who-trust-who online social network of a general consumer review site Epinions.com. The edges in this network have been reversed in order to capture the true direction of influence. This network is available online: `http://snap.stanford.edu/data/`.

Food Web [54, 55, 56, 57, 58] This collection is a collection of various food web networks mostly from marine ecosystems. The orientation of this network is such that directed edges point towards the flow of biomass, e.g., edges point from prey to predators. They are part of the Pejak data collections and are available online: http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm.

Gnutella Networks [51, 59] This collection includes a sequence of snapshots of the Gnutella peer-to-peer file sharing network from August 2002. There are total of 9 snapshots of Gnutella network collected in August 2002. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts. These are available online: http://snap.stanford.edu/data/.

Intra-Organizational [60] This collection consists of four networks describing the relationships between employees of a consulting company and a research company relating how relevant the work/knowledge of the target is to the source. The edges in these networks have been reversed in order to capture the true direction of influence. This network is available online: http://toreopsahl.com/datasets/.

Macaque Neural [61, 62, 63] This collection contains three networks from separate sources, all representing structural (axon projections) cortical connectivity in Macaque monkeys. These networks are available online: CoCoMac: http://cocomac.g-node.org/, Mac-95: https://sites.google.com/site/bctnet/datasets, Macaque-71: http://www.biological-networks.org/.

Physician [64] This collection includes three networks relating to the interactions of physicians in four towns in Illinois, namely who they sought advice from, who they shared cases with, and who are their friends. The edges in these networks have been reversed in order to capture the true direction of influence. This network is available online: http://moreno.ss.uci.edu/data.html.

Pokec [65] This collection includes a network of the entire Pokec online social network, representing the relationships between members. The edges in this network have been reversed in order to capture the true direction of influence. This network is available online: http://snap.stanford.edu/data/.

Political Blog [40] This datset provides a directed graph of links between blogs on US politics in 2005. The edges in this network have been reversed in order to capture the true direction of influence. This network is available online: http://www-personal.umich.edu/~mejn/netdata/.

Slashdot [66, 52] This collection consists of networks containing friend/foe links between the users of Slashdot. We consider here the two unsigned networks available from this source. The edges in these networks have been reversed in order to capture the true direction of influence. This network is available online: http://snap.stanford.edu/data/.

Teacher-Student [67] This collection is a network that captures the teacher-student relationships among the founding members of the International Network for Social Network Analysis. This network is available online: http://moreno.ss.uci.edu/data.html.

UC Irvine [37] This collection represents the messaging patterns between users of a Facebook-like social network at University of California, Irvine. This network is available online: http://toreopsahl.com/datasets/.

Web [66, 68] This collection contains four web networks: berkely.edu, stanford.edu, nd.edu, and google.com. We did not include the nd.edu network because it demonstrated characteristics which made it seem that it was not as well formed as the others. This network is available online: http://snap.stanford.edu/data/.

Wikipedia [69, 52] This collection includes two networks representing the interactions of editors and administrators supporting Wikipedia. Wiki-Talk captures the communications via users' talk page discussing updates to various articles on Wikipedia (who-writes-on-who). Wiki-Vote captures the vote history data for administrator elections (who-votes-for-who). The edges in the Wiki-Vote network have been reversed in order to capture the true direction of influence. This network is available online: http://snap.stanford.edu/data.

Yeast [70] This collection describes the directed interactions in the yeast transcription network. This network is available online: http://www.weizmann.ac.il/mcb/UriAlon/.

As a side note, we omitted citation networks from this analysis on the grounds that the role of control in such networks is unclear. The networks, while always growing, do not have a well-defined dynamical element: i.e., what resources move along them? This said, while they are not reported here, we did test them in our study of real networks and found that they were split between the source-dominated (1 network) and external-dilation-dominated (2 citation networks) clusters.

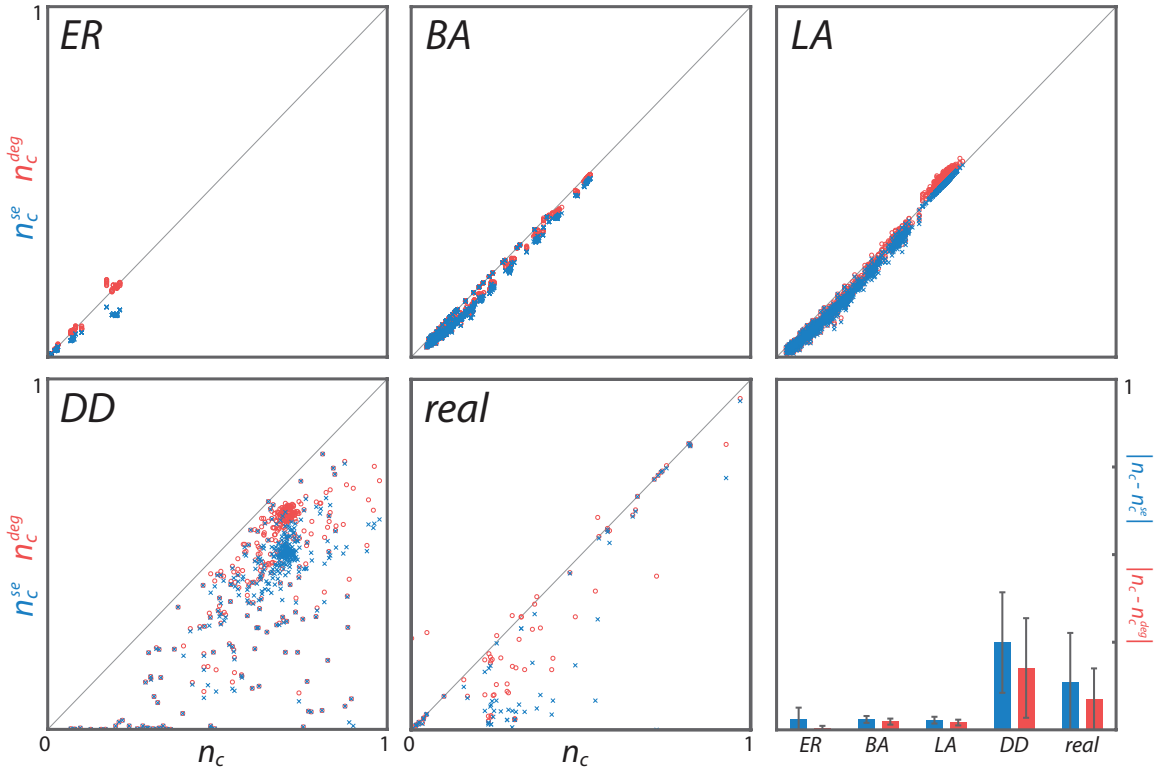

Figure S1: The fraction of controls required (n_c) for synthetic networks, Erdos-Renyi (ER), Barabasi-Albert (BA), Local Attachment (LA), Duplication-Divergence (DD), and real networks is predicted well simply by the source and sink nodes of the network, $n_c^{\text{se}} = n_s + n_e$ [9]. The improvement gained by incorporating non-zero terms from the degree distribution ($n_c^{\text{deg}} = N_c^{\text{deg}}/N$ is the fraction of controls required after a degree-preserving shuffle) is shown by the difference in bar heights in the final subplot. This contribution is small, accounting for only 5.1% of the fraction of controls required in real networks and 2.7% (ER), 0.6% (BA), 0.7% (LA), 7.3% (DD) required in synthetic networks. Synthetic networks were generated using a comprehensive set of parameters with network sizes, N, ranging from 100 to 50000; average degree, L/N, ranging from 2 to 58; for DD networks a probability to retain copied edges from 0.1 to 0.9.

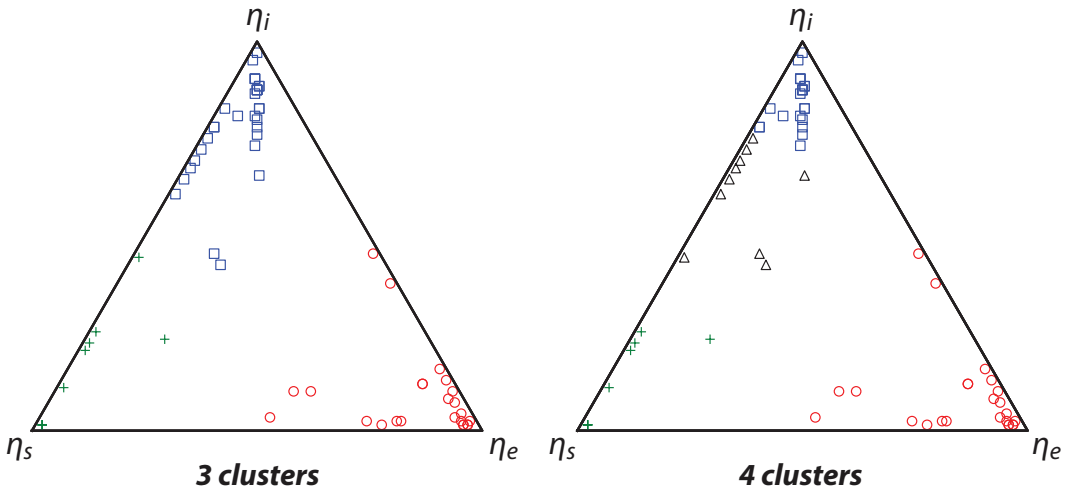

Figure S2: The distribution of control profiles of all the real networks analyzed in this study. The content of both panels are identical, showing the grouping of the profiles into three clusters and four clusters. As can be seen, the three clusters reveal the strong association of control profiles to the individual components of the control profile. The four clusters induce this same overall division, but create a rather arbitrary separation among the internal-dilations-dominated networks.

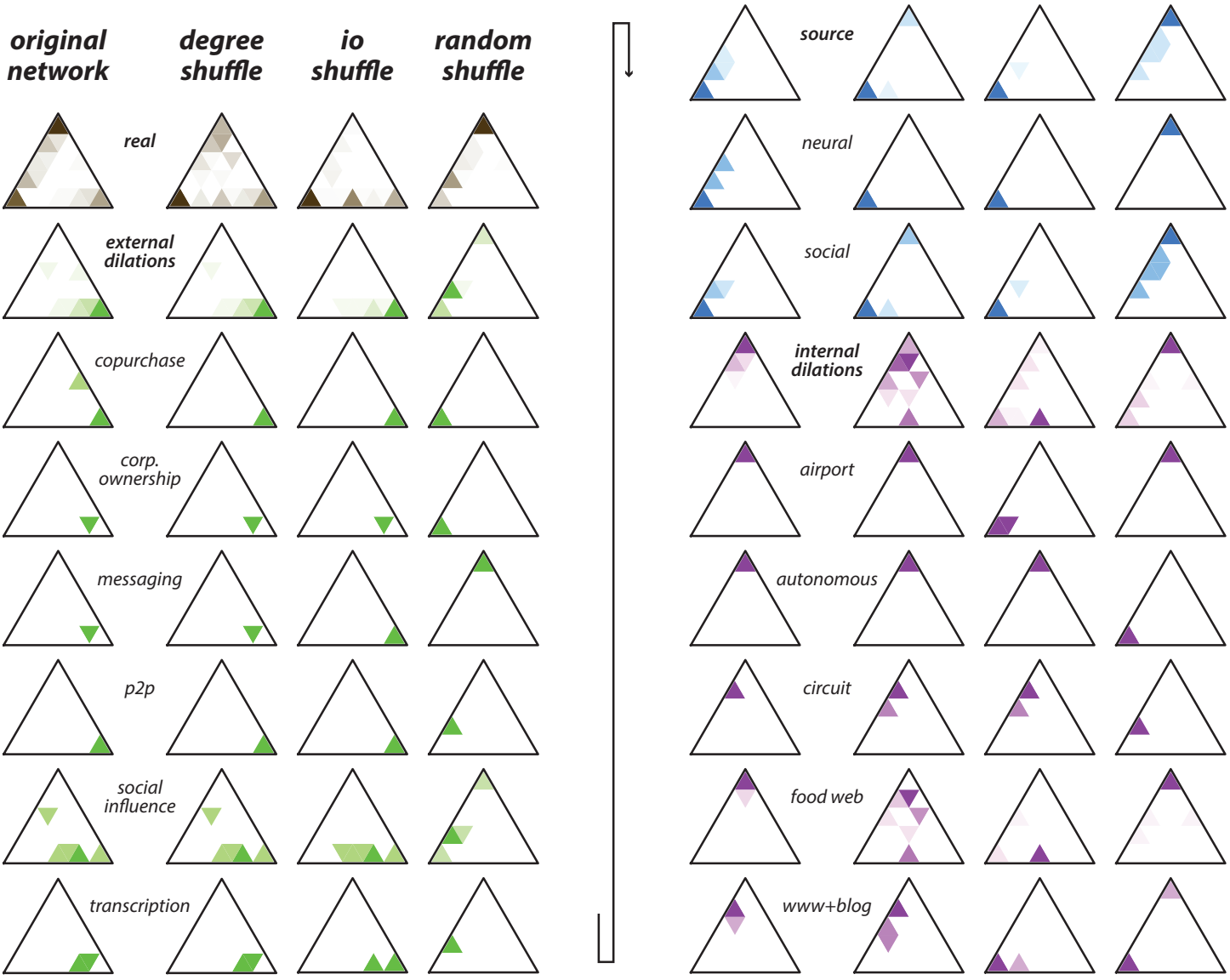

Figure S3: Real world networks not only differ compared to synthetic network profiles, they also exhibit differences between various randomized versions of their structure. In particular, we compare the control profiles of the original network with the control profiles of the network after a degree preserving shuffle, an IO shuffle, and a random shuffle. A degree preserving shuffle maintains the entire degree distribution of the network; an IO shuffle preserves only the number of inputs (source nodes) and outputs (sink nodes) of the network; and a random shuffle randomizes the network without any conservation requirements.

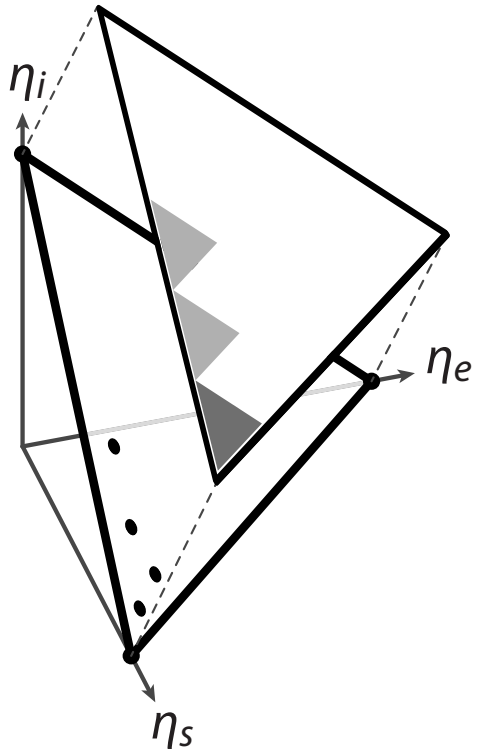

Figure S4: A single control profile plot is a triangular plot in which the corners correspond to the points (1,0,0), (0,1,0), and (0,0,1) in the three-dimensional (η_s, η_e, η_i) space. The equation $\eta_s + \eta_e + \eta_i = 1$ constrains all control profiles to lie on the plane going through these three points. The triangular plot is the part of this plane such that η_s, η_e, and η_i are all nonnegative. Here we use a heatmap to accurately represent the distribution of profiles when control profile points overlap. Deeper shades of the heatmap indicate a greater density of networks with control profiles located in that region.

Table S1: The complete assignment of networks to the three clusters identified. Two important attributes of these clusters are that (1) entire categories of networks are typically grouped into the same cluster and (2) multiple categories are grouped together that suggest shared traits. To emphasize this, cluster members are identified by their categories. All the datasets used and the categories they are grouped into are given in the table in Section 3. When all members of a category belong to the same cluster, just the category name is given. If membership is split, the category name is given along with the names of those particular networks from the category that belong to the cluster. Parentheses indicate the number of cluster members contributed by a given category.

Source	**External-dilation**	**Internal-dilation**
Neural (4) Social (8)	Co-purchase (3) Corporate ownership (1) Gnutella P2P (9) Messaging (2) Social Influence (5): Wiki, Physician, Teacher Transcription (2)	Airports (2) Autonomous System (1) Electronic Circuits (3) Food webs (23) Political blogs (1) Social Influence (1): Epinions Web (3)

Table S2: This table summarizes the real networks used in this study. We have grouped these networks into 13 "types" which we use to normalize the figures to eliminate bias from over-/under-representation by any one type.

Data Set	Network	Nodes	Edges	n_c	n_c^{deg}	η_s	η_e	η_i
Airports	us	1574	28236	0.37	0.38 ± 0.00	0.12	0.04	0.83
	intl	2939	30501	0.30	0.33 ± 0.01	0.02	0.00	0.98
Airports, Top 500	airports500*	500	24009	0.00	0.00 ± 0.00	0.00	0.00	0.00
Autonomous	as-caida20071105	26475	106762	0.72	0.44 ± 0.00	0.00	0.00	1.00
Electronic Circuit	circuit-s208	122	189	0.24	0.20 ± 0.01	0.34	0.00	0.66
	circuit-s420	252	399	0.23	0.20 ± 0.01	0.31	0.00	0.69
	circuit-s838	512	819	0.23	0.19 ± 0.01	0.29	0.00	0.71
Food Web	Chesapeake	39	177	0.31	0.18 ± 0.04	0.08	0.08	0.83
	ChesLower	37	178	0.30	0.18 ± 0.02	0.09	0.09	0.82
	ChesMiddle	37	209	0.22	0.10 ± 0.03	0.12	0.12	0.75
	ChesUpper	37	215	0.24	0.10 ± 0.02	0.11	0.11	0.78
	CrystalC	24	125	0.42	0.20 ± 0.02	0.20	0.00	0.80
	CrystalD	24	100	0.54	0.40 ± 0.03	0.15	0.00	0.85
	Everglades	69	916	0.30	0.03 ± 0.01	0.05	0.05	0.90
	Florida	128	2106	0.23	0.04 ± 0.01	0.03	0.03	0.93
	Maspalomas	24	82	0.25	0.13 ± 0.03	0.17	0.17	0.67
	Michigan	39	221	0.33	0.17 ± 0.03	0.08	0.08	0.85
	Mondego	46	400	0.41	0.17 ± 0.03	0.05	0.05	0.89
	Narragan	35	220	0.29	0.11 ± 0.03	0.10	0.10	0.80
	Rhode	19	53	0.26	0.21 ± 0.04	0.20	0.00	0.80
	StMarks	54	356	0.24	0.10 ± 0.02	0.08	0.08	0.85
	baydry	128	2137	0.23	0.04 ± 0.01	0.03	0.03	0.93
	baywet	128	2106	0.23	0.04 ± 0.01	0.03	0.03	0.93
	cypdry	71	640	0.28	0.11 ± 0.02	0.05	0.05	0.90
	cypwet	71	631	0.28	0.13 ± 0.02	0.05	0.05	0.90
	gramdry	69	915	0.30	0.03 ± 0.00	0.05	0.05	0.90
	gramwet	69	916	0.30	0.03 ± 0.00	0.05	0.05	0.90
	mangdry	97	1491	0.23	0.03 ± 0.01	0.05	0.05	0.91
	mangwet	97	1492	0.23	0.02 ± 0.01	0.05	0.05	0.91
Web	web-BerkStan	685230	7600595	0.38	0.29 ± 0.00	0.26	0.00	0.74
	web-Google	875713	5105039	0.48	0.32 ± 0.00	0.38	0.00	0.62
	web-Stanford	281903	2312497	0.32	0.26 ± 0.00	0.23	0.00	0.77
Political **Blog**	polblogs-rev	1224	19025	0.36	0.29 ± 0.01	0.36	0.17	0.46
Neural								
C. elegans	celegans	297	2345	0.16	0.10 ± 0.00	0.55	0.00	0.45
Macaque	cocomac	193	12051	0.02	0.02 ± 0.00	0.75	0.00	0.25
	mac95	94	2390	0.10	0.10 ± 0.00	1.00	0.00	0.00
	macaque71	71	746	0.01	0.01 ± 0.00	1.00	0.00	0.0
Social								
Email-EU	email-EuAll	265214	420045	0.93	0.81 ± 0.00	0.78	0.00	0.22
Intra-Organizational	cons-frequency-rev	46	879	0.04	0.04 ± 0.00	1.00	0.00	0.00
	cons-quality-rev	46	858	0.02	0.02 ± 0.00	1.00	0.00	0.00
	manuf-famliarity-rev*	77	2326	0.01	0.01 ± 0.00	0.00	0.00	0.00
	manuf-frequency-rev	77	2228	0.01	0.01 ± 0.00	1.00	0.00	0.00
Physician	physician-friend-rev	228	506	0.23	0.22 ± 0.01	0.60	0.17	0.23
Slashdot	soc-Slashdot0811-rev	77360	905468	0.00	0.26 ± 0.00	0.90	0.00	0.10
	soc-Slashdot0902-rev	82168	948464	0.05	0.28 ± 0.00	1.00	0.00	0.00
Pokec	pokec-rev	1632803	30622564	0.15	0.14 ± 0.00	0.80	0.00	0.20
Amazon **Copurchase**	amazon0302	262111	1234877	0.03	0.02 ± 0.00	0.00	0.54	0.46
	amazon0312	400727	3200440	0.04	0.03 ± 0.00	0.00	0.88	0.12
	amazon0505	410236	3356824	0.04	0.03 ± 0.00	0.00	0.91	0.09
	amazon0601	403394	3387388	0.02	0.01 ± 0.00	0.01	0.11	0.88
Corporate Ownership	corp-own	7253	6726	0.82	0.81 ± 0.00	0.16	0.83	0.01
Transcription								
E. coli	ecoli	418	519	0.75	0.75 ± 0.00	0.24	0.75	0.01
Yeast	yeast	688	1079	0.82	0.81 ± 0.00	0.17	0.82	0.01
Gnutella **P2P**	p2p-Gnutella04	10876	39994	0.55	0.55 ± 0.00	0.00	0.99	0.01
	p2p-Gnutella05	8846	31839	0.58	0.57 ± 0.00	0.02	0.95	0.02
	p2p-Gnutella06	8717	31525	0.58	0.57 ± 0.00	0.02	0.97	0.01
	p2p-Gnutella08	6301	20777	0.65	0.61 ± 0.00	0.02	0.91	0.07
	p2p-Gnutella09	8114	26013	0.66	0.63 ± 0.00	0.01	0.93	0.06
	p2p-Gnutella24	26518	65369	0.72	0.72 ± 0.00	0.02	0.98	0.00
	p2p-Gnutella25	22687	54705	0.73	0.73 ± 0.00	0.02	0.98	0.00
	p2p-Gnutella30	36682	88328	0.74	0.74 ± 0.00	0.01	0.99	0.00
	p2p-Gnutella31	62586	147892	0.74	0.74 ± 0.00	0.01	0.99	0.00
Social Influence								
Physician	physician-advice-rev	215	480	0.36	0.36 ± 0.01	0.33	0.58	0.09
	physician-discuss-rev	231	565	0.25	0.27 ± 0.01	0.47	0.52	0.02
Epinions	soc-Epinions1-rev	75879	508837	0.55	0.61 ± 0.00	0.37	0.20	0.43
Teacher-Student	teacher-student	60	94	0.58	0.57 ± 0.01	0.37	0.54	0.09
Wikipedia	wiki-Talk	2394385	5021410	0.97	0.95 ± 0.00	0.01	0.96	0.03
	wiki-Vote-rev	7115	103689	0.67	0.67 ± 0.00	0.21	0.79	0.0
UC Irvine **Messaging**	one-mode-char	1899	20296	0.32	0.32 ± 0.00	0.06	0.83	0.11
	one-mode-message	1899	20296	0.32	0.32 ± 0.00	0.06	0.83	0.11

References

[1] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Controllability of complex networks," *Nature*, vol. 473, no. 7346, pp. 167–173, 2011.

[2] M. Pósfai, Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Effect of correlations on network controllability," *Scientific reports*, vol. 3, 2013.

[3] T. Jia, Y.-Y. Liu, E. Csoka, M. posfai, J.-J. Slotine, and A.-L. Barabási, "Emergence of bimodality in controlling complex networks," *Nature Communications*, vol. 4, pp. 1–6, June 2013.

[4] In this work we discuss the minimum number of independent controls (denoted N_c). In some previous literature this number was called the minimum number of driver nodes (denoted N_D) [1].

[5] C.-T. Lin, "Structural Controllability," *IEEE Transactions on Automatic Control*, vol. AC-19, no. 3, pp. 201–208, 1974.

[6] The structure capturing the propagation of control, composed of stems, cycles, and buds, is called *cacti*. Each *cactus* within the cacti is contains only one stem and any number of cycles and buds (see supplementary online text).

[7] C. Commault, J.-M. Dion, and J. W. Van der Woude, "Characterization of generic properties of linear structured systems for efficient computations," *Kybernetika*, vol. 38, no. 5, pp. 503–520, 2002.

[8] An immediate extension can be made to accommodate isolated nodes as stems of unit length, acting as both a source and sink.

[9] In the exceptional case where $N_s = N_e = N_i = 0$, at least one control must be present, so $N_c = 1$. Formally we could revise Eq. 1 to: $N_c = \max(1, N_s + N_e + N_i)$. See supplementary online text for more discussion.

[10] P. Erdös and A. Renyi, "On Random Graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[11] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, pp. 509–512, Oct. 1999.

[12] M. O. Jackson and B. W. Rogers, "Meeting strangers and friends of friends: How random are social networks?," *The American economic review*, pp. 890–915, 2007.

[13] I. Ispolatov, P. Krapivsky, and A. Yuryev, "Duplication-divergence model of protein interaction network," *Physical Review E*, vol. 71, p. 061911, June 2005.

[14] Edge orientation is a matter of convention; see supplementary online text for discussion on the sensitivity of our results to edge reversals.

[15] S. Vitali, J. B. Glattfelder, and S. Battiston, "The Network of Global Corporate Control," *PLoS ONE*, vol. 6, p. e25995, Oct. 2011.

[16] N. E. Friedkin, *A Structural Theory of Social Influence*. Cambridge University Press, 2006.

[17] 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," *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 6841–6846, 2010.

[18] H. Yu and M. Gerstein, "Genomic analysis of the hierarchical structure of regulatory networks," *Proceedings of the National Academy of . . .*, 2006.

[19] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, Jan. 2002.

[20] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, pp. 186–198, 2009.

[21] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[22] F.-J. Müller and A. Schuppert, "Few inputs can reprogram biological networks," *Nature*, vol. 478, pp. E1–E1, Oct. 2011.

[23] W. K. Kroeze, "G-protein-coupled receptors at a glance," *Journal of Cell Science*, vol. 116, pp. 4867–4869, Dec. 2003.

[24] J. S. Gutkind, "Regulation of Mitogen-Activated Protein Kinase Signaling Networks by G Protein-Coupled Receptors," *Science Signaling*, vol. 2000, p. re1, July 2000.

[25] M. J. Marinissen and J. S. Gutkind, "G-protein-coupled receptors and signaling networks: emerging paradigms," *Trends in Pharmacological Sciences*, vol. 22, pp. 368–376, July 2001.

[26] H. G. Tanner, "On the controllability of nearest neighbor interconnections," *CDC 43rd IEEE Conference on*, vol. 3, pp. 2467–2472, 2004.

[27] A. Rahmani, M. Ji, M. Mesbahi, and M. Egerstedt, "Controllability of Multi-Agent Systems from a Graph-Theoretic Perspective," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 162–186, 2009.

[28] R. W. Shields and J. B. Pearson, "Structural Controllability of Multi-Input Linear Systems," *Automatic Control, IEEE Transactions on*, vol. AC-21, no. 2, pp. 203–212, 1976.

[29] K. Glober and L. M. Silverman, "Characterization of Structural Controllability," *IEEE Transactions on Automatic Control*, vol. AC-21, no. 4, pp. 534–537, 1976.

[30] S. Hosoe and K. Matsumoto, "On the Irreducibility Condition in the Structural Controllability Theorem," *IEEE Transactions on Automatic Control*, vol. AC-24, no. 6, pp. 963–966, 1979.

[31] J.-M. Dion, C. Commault, and J. van der Woude, "Generic properties and control of linear structured systems: a survey," *Automatica*, vol. 39, no. 7, pp. 1125–1144, 2003.

[32] J. E. Hopcraft and R. M. Karp, "An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs," *SIAM Journal on Computing*, vol. 2, no. 4, pp. 225–231, 1973.

[33] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Control centrality and hierarchical structure in complex networks," *PLoS ONE*, vol. 7, no. 9, p. e44459, 2012.

[34] N. Cowan, "Nodal dynamics determine the controllability of complex networks," *arXiv*, pp. 1–5, July 2011.

[35] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 ed., 2010.

[36] F. Wang, Y. Moreno, and Y. Sun, "Structure of peer-to-peer social networks," *Physical Review E*, vol. 73, p. 036123, Mar. 2006.

[37] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Networks*, 2009.

[38] L. Huang, K. Park, and Y.-C. Lai, "Information propagation on modular networks," *Physical Review E*, vol. 73, p. 035103, Mar. 2006.

[39] A. V 'azquez, "Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations," *Physical Review E*, vol. 67, p. 056104, May 2003.

[40] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog," in *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.

[41] M. Gotz, J. Leskovec, M. McGlohon, and C. Faloutsos, "Modeling Blog Dynamics," in *Proceedings of the International Conference on Weblogs and Social Media*, 2009.

[42] T. Furukawa, Y. Matsuo, I. Ohmukai, K. Uchiyama, and M. Ishizuka, "Social Networks and Reading Behavior in the Blogosphere," in *Proceedings of the International Conference on Weblogs and Social Media*, 2007.

[43] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, 2010.

[44] J. Marcelino and M. Kaiser, "Critical paths in a metapopulation model of H1N1: Efficiently delaying influenza spreading through flight cancellation," *PLoS Currents*, 2012.

[45] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (ACM TWEB)*, 2007.

[46] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.

[47] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *Nature*, 1998.

[48] K. Norlen, G. Lucas, and M. Gebbie, "EVA: Extraction, visualization and analysis of the telecommunications and media ownership network," *Proceedings of International Telecommunications Society 14th Biennial Conference*, 2002.

[49] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," *Nature genetics*, 2002.

[50] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and S. Shen-Orr, "Superfamilies of evolved and designed networks," *Science*, 2004.

[51] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 2007.

[52] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *CHI*, 2010.

[53] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the semantic web," *The Semantic Web-ISWC 2003*, 2003.

[54] R. R. Christian and J. J. Luczkovich, "Organizing and understanding a winter's seagrass foodweb network through effective trophic levels," *Ecological Modelling*, 1999.

[55] M. E. Monaco and R. E. Ulanowicz, "Comparative ecosystem trophic structure of three US mid-Atlantic estuaries," *Marine Ecology Progress Series*, 1997.

[56] J. Almunia, G. Basterretxea, and J. Arstegui, "Benthic-pelagic switching in a coastal subtropical lagoon," *Estuarine*, 1999.

[57] D. Baird, J. Luczkovich, and R. R. Christian, "Assessment of spatial and temporal variability in ecosystem attributes of the St Marks National Wildlife Refuge, Apalachee Bay, Florida," *Estuarine*, 1998.

[58] D. Baird and R. E. Ulanowicz, "The seasonal dynamics of the Chesapeake Bay ecosystem," *Ecological Monographs*, 1989.

[59] M. Ripeanu and I. Foster, "Mapping the gnutella network: Macroscopic properties of large-scale peer-to-peer systems," *Peer-to-Peer Systems*, 2002.

[60] R. Cross and A. Parker, *The Hidden Power of Social Networks*. Understanding how Work Really Gets Done in Organizations, Harvard Business Press, June 2004.

[61] R. Bakker, T. Wachtler, and M. Diesmann, "CoCoMac 2.0 and the future of tract-tracing databases," *Frontiers in neuroinformatics*, 2012.

[62] M. Kaiser and C. C. Hilgetag, "Nonoptimal Component Placement, But Short Processing Paths, Due to Long-distance Projections in Neural Systems," 2006.

[63] M. P. Young, "The organization of neural systems in the primate cerebral cortex," in *Proceedings of the Royal Society of London. Series B: Biological Sciences*, pp. 13–18, 1993.

[64] R. S. Burt, "Social contagion and innovation: Cohesion versus structural equivalence," *American journal of Sociology*, 1987.

[65] L. Takac and M. Zabovsky, "Data Analysis in Public Social Networks," in *International Scientific Conference & International Workshop Present Day Trends of Innovations*, (Lomza, Poland), May 2012.

[66] J. Leskovec, K. J. Lang, and A. Dasgupta, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, 2009.

[67] K. Reitz and D. R. White, "Rethinking the role concept: Homomorphisms on social networks," in *Research Methods in Social Network Analysis* (L. C. Freeman, D. R. White, and A. K. Romney, eds.), pp. 429–488, New Brunswick, NJ: George Mason Press, 1989.

[68] R. Albert, H. Jeong, and A. L. Barabási, "Internet: Diameter of the world-wide web," *Nature*, 1999.

[69] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *WWW*, 2010.

[70] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.