Advances in Complex Systems
Vol. 22, Nos. 7 & 8 (2019) 1950014 (15 pages)

DOI: 10.1142/S0219525919500140

CONTROL CONTRIBUTION IDENTIFIES TOP DRIVER NODES IN COMPLEX NETWORKS

YAN ZHANG*, ANTONIOS GARAS† and FRANK SCHWEITZER‡

Chair of Systems Design, ETH Zurich,
Weinbergstrasse 58, 8092 Zurich, Switzerland
**yanzhang0@ethz.ch*
†*antonios.garas@gmail.com*
‡*fschweitzer@ethz.ch*

Received 11 June 2019
Revised 11 September 2019
Accepted 24 October 2019
Published 27 December 2019

We propose a new measure to quantify the impact of a node i in controlling a directed network. This measure, called "control contribution" $\mathcal{C}_i$, combines the probability for node i to appear in a set of driver nodes and the probability for other nodes to be controlled by i. To calculate $\mathcal{C}_i$, we propose an optimization method based on random samples of minimum sets of drivers. Using real-world and synthetic networks, we find very broad distributions of C_i. Ranking nodes according to their C_i values allows us to identify the top driver nodes that can control most of the network. We show that this ranking is superior to rankings based on other control-based measures. We find that control contribution indeed contains new information that cannot be traced back to degree, control capacity or control range of a node.

Keywords: Complex network; structural controllability; driver nodes.

1. Introduction

One objective of system design is to obtain the ability to steer the dynamics of a system such that a desired system state is reached. One way of achieving this is rooted in control theory, which utilizes components of the system as drivers for the dynamics [1]. This relates to the fundamental question if and how a system can be controlled [2]. It requires to have a proper system representation. Here, we use a complex network approach, in which system elements are represented by nodes and their interactions by links. Liu *et al.* [3] introduced an analytical framework called *network controllability* that combines network theory and classical control theory. This framework has important applications for the study of empirical systems [4–10], if interactions between system elements are captured by means of a large-scale complex network topology.

Network controllability assumes that there is a dynamics *on the network*, specifically that the state of nodes can be described by a linear dynamics. Then, the

‡Corresponding author.

minimum set of driver nodes required to control the whole system can be efficiently identified using this framework. Given that not every node necessarily is a driver node, it is important to quantify to what extent a node contributes to controlling the whole network. To solve this problem is the aim of our paper. We want to rank driver nodes according to their contribution to controlling the network, in particular we want to identify the *top drivers*, i.e., driver nodes that together can control a large part of the network.

In the literature, various measures have been introduced to quantify the *topological importance* of nodes. Examples include degree centrality, PageRank, coreness [11, 12] and betweenness centrality [13]. These measures, however, do not help us in solving our problem because they do not consider a dynamics on the network. There are, indeed, other measures that take this dynamics into account, for instance *control range*, $\mathcal{R}$ [14], *control capacity*, $\mathcal{K}$ [15] and control centrality $\mathcal{H}$ [16]. $\mathcal{R}$ quantifies the size of the subnetwork, i.e., the number of nodes, controlled by one driver node together with other driver nodes, $\mathcal{K}$ quantifies the likelihood that one node is a driver node, and $\mathcal{H}$ quantifies the size of the subnetwork controlled by only a single node. In Sec. 2.1, we will give examples for that. These measures, however, separate individual aspects of drivers, therefore it is not clear which measure should be effectively used to identify top driver nodes.

In this work, we propose a new measure to identify top drivers, called *control contribution*, $\mathcal{C}$, which is a *node property*. Intuitively, the control contribution $\mathcal{C}_i$ of node i captures the probability for any node in a network to be controlled by node i joint with the probability that i becomes a driver. To calculate $\mathcal{C}_i$, we propose an optimization method based on random samples of minimum sets of drivers that will be explained in Sec. 2. Calculating the distribution $P(\mathcal{C})$ for different real-world and synthetic networks, we find that the distribution is always very broad and does not follow a specific pattern. Looking into relations to *topological* quantities such as degree distribution, we find that the degree distribution does not determine the distribution of control contributions. Further, no uniform pattern in the correlation exists between control capacity $\mathcal{K}$ and control range $\mathcal{R}$ that determine control contribution. Therefore, we argue that control contribution $\mathcal{C}$ indeed contains new information.

We demonstrate that driver nodes chosen according to their $\mathcal{C}_i$ values lead to a larger part of a network that can be controlled. In this respect, nodes with a higher $\mathcal{C}_i$ outperform nodes chosen according to their control capacity $\mathcal{K}_i$, control range $\mathcal{R}_i$ or control centrality $\mathcal{H}_i$ values. Therefore, our new quantity *control contribution*, $\mathcal{C}_i$, efficiently identifies the top driver nodes in a network.

2. Control Contribution

2.1. *Identifying sets of driver nodes*

To introduce the concept of control contribution, we consider a directed network of N nodes. In this network, each node i is captured by a state variable $x_i(t)$, and the

states of all nodes can be described by the state vector $\mathbf{X}(t) = [x_1(t), x_2(t), \ldots, x_N(t)]$ with $\mathbf{X} \in \mathbb{R}^N$. Some of these nodes are controlled directly by the control signals $u_k(t)$. We call these nodes *drivers* and N_c the number of driver nodes. We denote $\mathbf{U}(t) \in \mathbb{R}^{N_c}$ as the vector of control signals. The matrix $\mathbf{B} \in \mathbb{R}^{N \times N_c}$ maps these signals to driver nodes, $b_{ij} \neq 0$ when control signal j is attached to node i.

We further assume that $\mathbf{X}(t)$ follows the linear dynamics

$$\ddot{\mathbf{X}}(t) = \mathbf{A}\mathbf{X}(t) + \mathbf{B}\mathbf{U}(t), \tag{1}$$

with time-invariant matrices $\mathbf{A}$. $\mathbf{B}$ and $\mathbf{U}$. $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the interaction matrix with elements a_{ij} $(i, j = 1, \ldots, N)$ describing the strength in which node j can influence node i. According to the Kalman rank condition [17], the linear system defined by Eq. (1) is controllable, if and only if the controllability matrix $\mathbf{C} = [\mathbf{B}, \mathbf{AB}, A^2B, \ldots, A^{N-1}B] \in \mathbb{R}^{N \times (N \cdot N_c)}$ has full rank, i.e., $\text{rank}(\mathbf{C}) = N$.

In many cases, we do not have the precise value of the nonzero elements in $\mathbf{A}$ and $\mathbf{B}$, therefore, it is not feasible to calculate $\text{rank}(\mathbf{C})$ to check for controllability. For these cases, we have the weaker requirement of structural controllability [18]. The key idea is to treat both the adjacency matrix $\mathbf{A}$ and the mapping matrix $\mathbf{B}$ as *structural matrices* whose nonzero elements are free parameters. Then the system is controllable iff the free parameters can be chosen such that $\text{rank}(\mathbf{C}) = N$.

In general, $\text{rank}(\mathbf{C})$ denotes the controllable subsystem size, N_b, i.e., the number of nodes in the network that can be controlled. N_b not always equals N, but is smaller, which means the system is partially controllable. We then define the fraction $n_b = N_b/N$ as the relative size of the controllable system.

Based on structural controllability, [3] combined tools from network theory and statistical physics to identify minimum sets of driver nodes that allow to control the whole network of N nodes. Their approach identifies drivers based on *maximum matching*, which denotes the largest set of directed links without common nodes. That means, this set contains only node-disjoint directed paths, and directed cycles. In a maximum matching, a node is unmatched if no link in the maximum matching points at it. These unmatched nodes form one minimum set of driver nodes. In the example of Fig. 1, which we explain further below, (b) and (d) correspond to two configurations of maximum matching. Additionally, if we attach a control signal to the top node in a path, we have a stem (see Figs. 1(b) and 1(d). If we add the minimum set of links that connects each cycle to only one of the stems, we have a *cactus structure* spanning the network (see Figs. 1(c) and 1(e). This cactus structure maintains the controllability of the whole network, with the unmatched nodes as one minimum set of drivers [18].

For an arbitrary network of size N, there are multiple sets of driver nodes. If N_c denotes the number of driver nodes, then MDS denotes the *minimum set* of driver nodes that is required to control the *full network*. N_d is the size of the set MDS, it can be larger or smaller than N_c. Again, there can be multiple minimum sets of driver nodes. We then define the fraction $n_d = N_d/N$ as the relative size of the

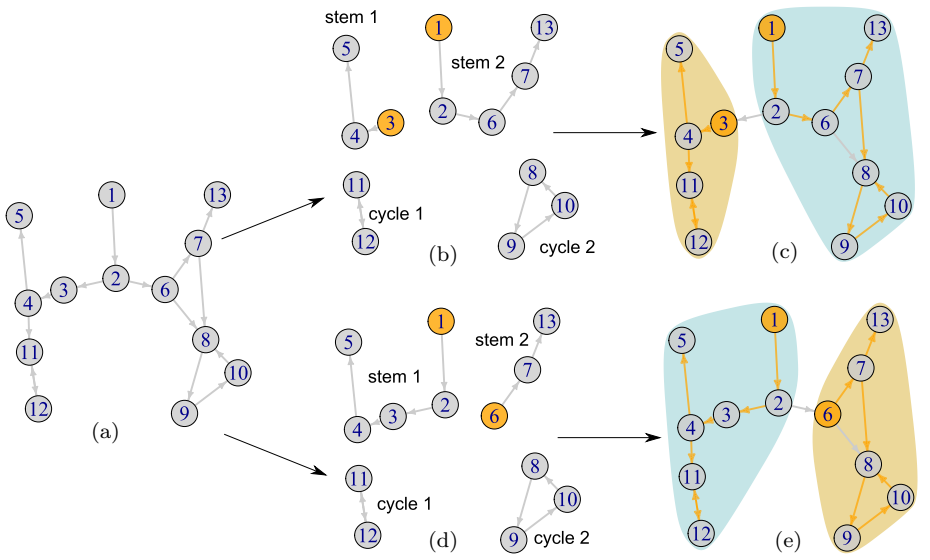

Fig. 1. (Color online) Controlling a network with minimum driver node sets. (a) A toy network with 13 nodes. (b) One maximum matching with unmatched nodes 1, 3. (c) The cactus structure corresponding to (b) with driver nodes 1, 3. (d) One maximum matching with unmatched nodes 1, 6. (e) The cactus structure corresponding to (d) with driver nodes 1, 6. In each cactus structure, orange nodes and gray nodes are driver nodes and non-driver nodes. The subnetwork each driver node controls is in shaded area.

minimum set of drivers. Each of these sets can guarantee controllability of the whole network but does not always contain the same nodes. Some nodes appear in every driver node set, while others are redundant and seldom become a driver. Denote the event node i appears in a minimum set of driver nodes as D_i, then the corresponding probability is $P(D_i)$. This probability is called *control capacity* $\mathcal{K}$ in the literature [15].

Further, for one minimum set of driver nodes, each driver i controls a non-overlapping subnetwork of size N_i, which can be identified based on the corresponding cactus structure. Dependent on which cactus structure is obtained, N_i can vary and its distribution is $f(N_i)$. This allows us to define the average $\langle N_i \rangle$ over all minimum sets of drivers in which node i is a driver.

We eventually denote the event that one node is part of the controllable subnetwork of driver node i as S_i, and the corresponding probability as $P(S_i)$. To illustrate this, we look at the toy network presented in Fig. 1(a). For this network, there are two minimum sets of driver nodes of size $N_d = 2$, i.e., there are only two drivers shown in Figs. 1(b) and 1(c). The subnetwork each driver node controls is highlighted with the shaded area. Obviously, in each cactus configuration, every node appears only in one of the subnetworks controlled by a driver node. In this toy example, we observe that node 1 appears in both sets of driver nodes, and the size of the subnetwork it controls is larger than that of the other two driver nodes.

In the following, we combine the two information about $P(D_i)$ and $P(S_i)$ to define our measure *control contribution*, $\mathcal{C}_i$. Let MDS_i denote the set of all driver node sets which include node i. The probability that node i becomes a driver node in MDS is $P(D_i) = |\mathrm{MDS}_i|/|\mathrm{MDS}|$. The conditional probability that one node appears in the subnetwork controlled by i given that i is a driver is expressed by $P(S_i|D_i) = \langle N_i \rangle / N$. Based on Bayes' rule, we finally calculate the joint probability

$$P(D_i \cap S_i) = P(S_i|D_i) * P(D_i). \tag{2}$$

To calculate $P(D_i \cap S_i)$, it is required to know all possible cactus structures of the network, which is computationally prohibitive for large networks because it can increase with $N!$ Therefore, instead of calculating the precise value of $P(S_i|D_i)$, we restrict ourselves to identifying its *upper bound*,

$$P(S_i|D_i) \leq \arg \max_{m \in \mathrm{MDS}_i} \left(\frac{N_i(m)}{N} \right) = \mathcal{R}_i. \tag{3}$$

This upper bound is the *control range*, $\mathcal{R}$, normalized by the system size. Control range gives the maximum size of the subnetwork controlled by node i when i is a driver in one minimum set of drivers. Based on this, we now define the most central measure of our study as

$$\mathcal{C}_i = \mathcal{R}_i \mathcal{K}_i \geq P(S_i|D_i) * P(D_i). \tag{4}$$

We denote this as the *control contribution* $\mathcal{C}_i$ of node i.

2.2. *Calculating control contributions of driver nodes*

To calculate control contribution $\mathcal{C}$, let us first look at the toy network presented in Fig. 1(a). There are only two unique cactus structures with three driver nodes in total: node 1, 3 and 6. We remind that control range $\mathcal{R}$ is the number of nodes controlled by a given node, normalized to the system size and control capacity $\mathcal{K}$ is the probability for a node to become a driver node.

With reference to the two cactus structures, we have $\mathcal{R}_1 = 1$, $\mathcal{R}_3 = 0.5$, $\mathcal{R}_6 = 0.5$, and $\mathcal{K}_1 = 8/13$, $\mathcal{K}_2 = 5/13$, $\mathcal{K}_3 = 6/13$. The control contribution for these three driver nodes are: $\mathcal{C}_1 = 8/13$, $\mathcal{C}_2 = 5/26$, $\mathcal{C}_3 = 3/13$. $\mathcal{C}_1$ is the highest among all three driver nodes.

For an arbitrary network, we can calculate the control contribution using a random sampling approach. This takes into account the fact that, in different samples of cactus structures, the controllable subnetwork of a driver node can be the same. After accumulating a sufficient number of samples, new cactus structures can hardly produce new controllable subnetworks of larger size for any potential driver node, as demonstrated in [14]. This property allows us to efficiently approximate control contribution $\mathcal{C}$ using random samples of cactus structures.

Starting from an initial configuration for the minimum driver set, we generate random samples of cactus structures based on random samples of minimum sets of

drivers. Note that for a given minimum set of drivers, different cactus structures can be sampled in this process, as indicated in [14, 15]. By definition, for a minimum set of drivers, each driver node i is the top node in a stem. The corresponding cactus structure controlled by node i can be constructed by adding cycles that are attached to the stem via a directed link. In some cases, if a cycle is not accessible from any driver, following [3], we connect an arbitrary driver to the cycle with an additional link.

For example, Fig. 1(b) shows one maximum matching for the toy network in Fig. 1(a). This maximum matching contains two stems and two cycles. For the driver node 3, the corresponding stem is stem 1. It contains the nodes 3, 4, 5, in which node 3 is the top node colored in orange. To construct the cactus structure, we now need to connect the cycles to the stems. In the toy network Fig. 1(a) we see that cycle 1 is pointed to by node 4. Therefore, we add cycle 1 to the subnetwork controlled by node 3. On the other hand, because there is no link from node 4 and node 5 to cycle 2, we cannot add this cycle to the subnetwork controlled by node 3. This results in the cactus structure shown in Fig. 1(c).

Note that in an arbitrary network, there can be more than one minimum set of drivers that lead to different cactus structures. For example, a second configuration of the cactus structure can be seen in Figs. 1(d) and 1(e). Because we limit ourselves to the upper bound of N_i, we focus on the cactus structure that identifies the largest controllable subnetwork for each driver node. In our example, for driver node 1 we only consider the subnetwork shown in Fig. 1(c). Repeating this construction process for different minimum sets of drivers, we have random samples of cactus structures.

Now, we approximate the control contribution $\mathcal{C}$ based on these samples. Concretely, we denote Q as the sum of the control contributions of all nodes, $Q(t) = \sum_i \mathcal{C}_i(t)$. t refers to the number of iterations because the cactus structures are

Algorithm 1. Random sampling procedure to calculate control contribution.

```
Initialization;
Q(0) = 0
t = 1
marker = 0
while marker < δ do
    produce a random cactus structure
    calculate Q(t) and Δ(t)
    if Δ(t) == 0 then
        marker = marker + 1
    else
        marker = 0
    end
    t = t + 1
end
```

identified in an optimization process. For each iteration, we identify one cactus structure. Then, the quality function which measures the relative increase in Q at each step t can be defined as $\Delta(t) = (Q(t) - Q(t-1))/Q(t-1)$. $\Delta(t) = 0$ indicates that the cactus structure generated at iteration t does not contain larger controllable subnetworks for each driver and therefore will not change with further iterations. As shown in Algorithm 1, if the condition $Q(t) = 0$ continues to hold for enough iteration steps δ, this implies we have already generated sufficient samples of cactus structures to approximate the control contribution. The algorithm works as follows.

3. Results

3.1. *Distribution of control contribution*

Data sets. To explore the properties of our measure $\mathcal{C}_i$, we consider both synthetic networks and empirical networks as summarized in Table 1. The synthetic networks represent two important classes of network topologies, namely random networks with (a) a (narrow) Poissonian degree distribution and (b) a (broad) power law degree distribution with degree exponent 2.1. The latter is generated using the algorithm developed by Chung and Lu [19]. The empirical networks include the `SmaGri citation` network, obtained from the `Pajek` dataset. It shows how papers cite each other according to the `WebofScience`. The second empirical network is an ownership network constructed from the `ORBIS` 2007 dataset [20, 21]. It contains information about millions of firms and their ownership relations. Therefore, in our analysis we restrict ourselves only to the strongly connected component, which contains 1318 firms. The ownership network is a dense network, while the citation network is sparse. In accordance with the requirement of structural controllability, both the empirical and the synthetic networks are directed.

Distribution of $\mathcal{C}$. Figure 2 shows the distributions of our measure $\mathcal{C}$ for the four networks studied. We find that, for all networks, the maximum value of $\mathcal{C}$ is less than 0.1. This indicates that driver nodes can at most control a small part of the network.

Looking at the form of the distributions, we find that for the Erdős–Rényi network, the scale-free network and the citation network, the distribution of $\mathcal{C}$ is skewed to the left with its peak near zero. This differs from the distribution for the ownership network, where $\mathcal{C}$ has a peak at the center. This can be partly attributed to the fact

Table 1. Statistics of networks analyzed in this paper: network size N, number of links E, average degree k, degree correlation r, and clustering coefficient c.

Name	N	L	k	r	c
Erdős–Rényi	1000	1500	3.0	0.021	0.004
Scale-free	1000	4000	8.0	−0.030	0.017
SmaGri citation	1024	4918	9.6	−0.18	0.094
Ownership	1318	12184	18.4	0.13	0.032

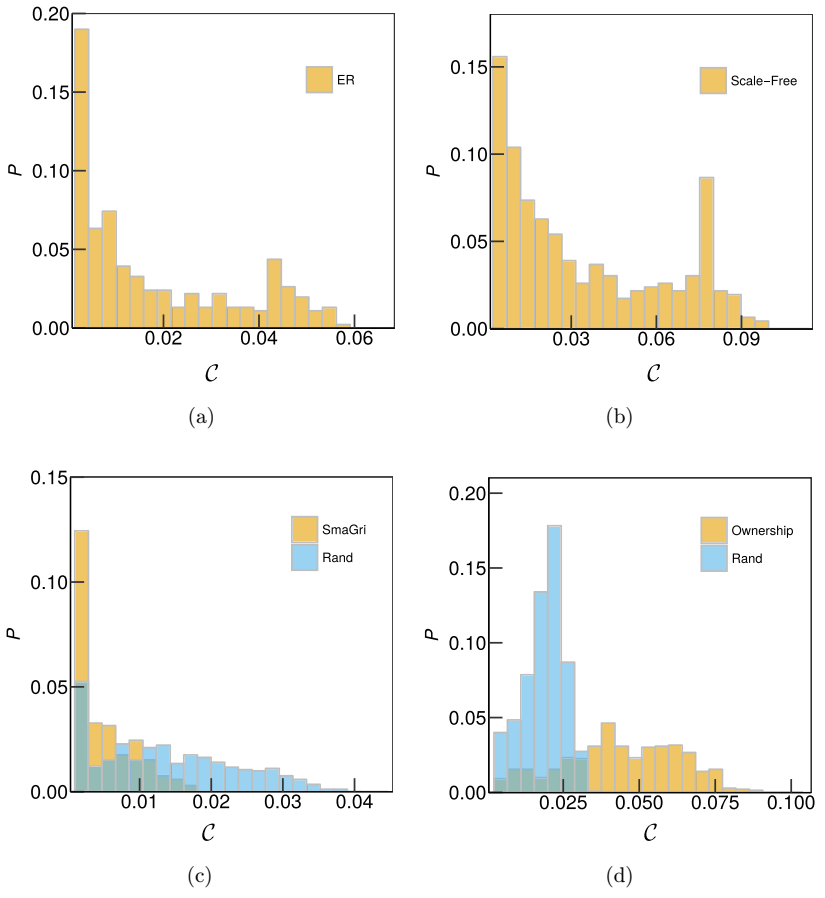

Fig. 2. Distribution of control contribution $\mathcal{C}$ for driver nodes in real and synthetic networks. In ((c)–(d)), we also show the distribution for the randomized version of the network preserving the degree sequence.

that the ownership network is very dense, as it only considers the strongly connected component.

To reflect the impact of the particular topology, we compare the two empirical networks with random networks generated by reshuffling connections. This randomization process not only keeps the degree distribution, but also the correlation between in-degree and out-degree. As we observe in the bottom panel of Fig. 2, even though the degree distributions are the same, the distributions of $\mathcal{C}$ change significantly, both with respect to the maximum value of $\mathcal{C}$ and the position of the peak. These differences indicate that the degree distribution alone is not sufficient to determine the distribution of control contribution. It is important to note that this observation is different from previous results [15, 16] in which the distributions of controllability-based measures, such as control range and control capacity, are mainly determined by the degree distribution. As a brief summary, here we provide

Table 2. Correlations between control contribution and degree.

	Erdős–Rényi	Scale-free	SmaGri citation	Ownership
In-degree	−0.43	−0.38	−0.20	−0.21
Out-degree	0.17	0.07	0.48	−0.24
Degree	−0.18	−0.22	0.36	−0.24

Table 2 to show that correlation between control contribution and in-degree, out-degree, and degree. From this table, it is clear that control contribution is negatively correlated with in-degree.

3.2. *Identifying top driver nodes*

We further compare control contribution $\mathcal{C}$ to other measures that are used to identify top drivers. Let us recap here that top driver nodes are nodes that together can control a larger part of the network. Precisely, given a small set of N_c drivers, the size N_b of the controlled subnetwork becomes as large as possible because these are the top driver nodes.

In our comparison, we consider seven measures in total: four control-based measures, control contribution $\mathcal{C}$, control range $\mathcal{R}$, control capacity $\mathcal{K}$ and control centrality $\mathcal{H}$, two degree-based measures, in-degree and out-degree, and one measure based on a completely random sampling of driver nodes, as a reference case. Each of these measures, except for the random case, provides us with a ranking of the nodes. From this ranking, we deterministically choose the top N_c nodes. Here, we have to consider that there are differences between degree-based measures and control-based measures. Driver nodes are more likely to be low-degree nodes [3, 15]. Therefore, if we rank nodes for degree-based measures, we have to rank them in *increasing* order of in(out)-degree, instead of decreasing order, which is done for all control-based measures.

With this set of top driver nodes for the seven different measures, we calculate the size of the subnetwork N_b that can be controlled. Based on this approach, the best measure to identify top driver nodes is the one that leads to the largest controllable system size N_b among all the measures.

Before we come to the results, we emphasize that calculating control range $\mathcal{R}$ and control capacity $\mathcal{K}$ is computationally expensive because it is necessary to generate random samples of cactus structures. Degree-based measures instead can be easily calculated.

To facilitate the comparison, we focus on the relative size of the controllable subnetwork, i.e., we calculate $n_b = N_b/N$ for each ranking scheme given a fraction of driver nodes $n_c = N_c/N$. For a given set of drivers, the relative size of the controllable subnetwork n_b can be efficiently determined via linear programming, as shown in [16, 22, 23]. Note that when there is only one driver node, n_b is defined as control centrality in [16].

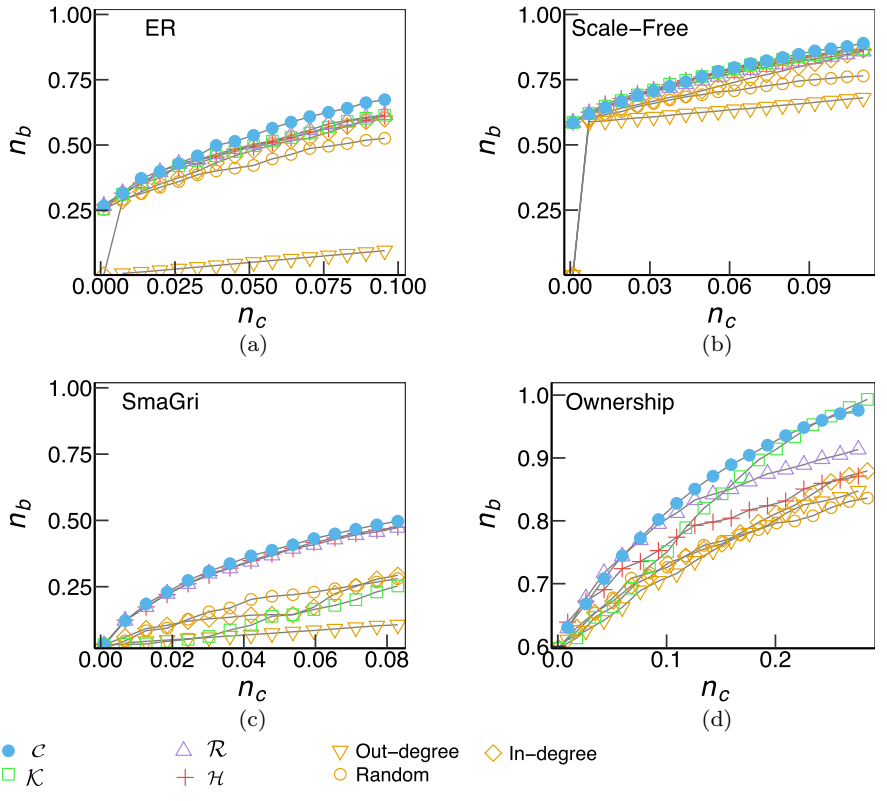

Fig. 3. n_b as a function of n_c for different ranking schemes.

$n_b = 1$ implies that we can control the whole network, and $n_c = 1$ indicates that we choose all nodes as drivers directly. Since there is always one *minimum set* of driver nodes of size N_d that we can use to control the whole network, we set $n_d = N_d/N$ as the upper bound for n_c.

Figure 3 compares n_b for the seven ranking schemes as a function of n_c. Here, we highlight two observations: (i) Among the seven measures that we use to rank driver nodes, control contribution $\mathcal{C}$ leads to the largest n_b, followed by control capacity $\mathcal{K}$. In comparison, out-degree leads to the smallest n_b. We further discuss this in Sec. 3.3. (ii) Whichever measures we use to rank driver nodes, n_b never reaches 1. This is expected because of the condition $n_c < n_d$.

In conclusion, control contribution $\mathcal{C}$ outperforms the other tested measures when identifying top driver nodes, because nodes chosen according to their $\mathcal{C}_i$ values always control a larger part of the network. We are aware that the above results are obtained for the particular set of networks used in our study. However, we have also checked the robustness of our findings with an ensemble approach presented in the supporting information.

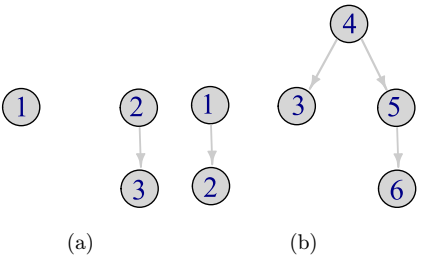

Fig. 4. Toy examples to illustrate why control capacity and control range should be considered together in choosing driver nodes.

3.3. *Control contribution and other structural properties*

So far, we have observed that driver nodes chosen according to their $\mathcal{C}_i$ values lead to the largest n_b. Now, we explore why this is the case. We remind that $\mathcal{C}$ is composed of control range $\mathcal{R}$ and control capacity $\mathcal{K}$, according to Eq. (4). Therefore, we limit our exploration to these two measures.

We start from two toy examples, shown in Fig. 4. Example (a) contains 3 nodes in which node 1 is isolated. Based on structural controllability theory, we know that this node is always a driver node, because it has to be included into the set of driver nodes if one wants to control the *whole* network. Such an isolated node always has the highest value of control capacity $\mathcal{K}$. However, with respect to our performance measure, N_b, we do not gain a lot from this isolated node, because it only controls itself, which increases N_b by one. In comparison, node 2 can control both itself and node 3 and it is always a driver node, therefore it should be the top driver node in this network. This makes clear why control capacity $\mathcal{K}$ alone is a bad measure for top driver nodes.

In example (b), there are two minimum sets of drivers. One set contains node 1, 4 and 3, the other set contains node 1, 4 and 5. Node 5 controls both itself and node 6. Therefore, it has a control range $\mathcal{R}$ of 2. However, because it appears in only one of two minimum sets of drivers, it has a low control capacity value $\mathcal{K}$ of 0.5, and the subnetwork controlled by it can be fully controlled by node 4 as well. This is different from node 1 which has the same control range but a higher control capacity of 1. Therefore, node 1 should be the top driver in the network. This example indicates that $\mathcal{R}$ alone is also a bad measure for top driver nodes.

The above two examples demonstrate that, in order to identify top driver nodes, we should consider both control capacity $\mathcal{K}$ and control range $\mathcal{R}$. However, one could still argue that for an arbitrary network, because both $\mathcal{K}$ and $\mathcal{R}$ are calculated based on random samples of cactus structure, a strong positive correlation could be observed. If this was the case, there would be no need for a new measure. To investigate this conjecture, we explore the correlation of control capacity $\mathcal{K}$ and control range $\mathcal{R}$ with the scatter plot of driver nodes shown in Fig. 5.

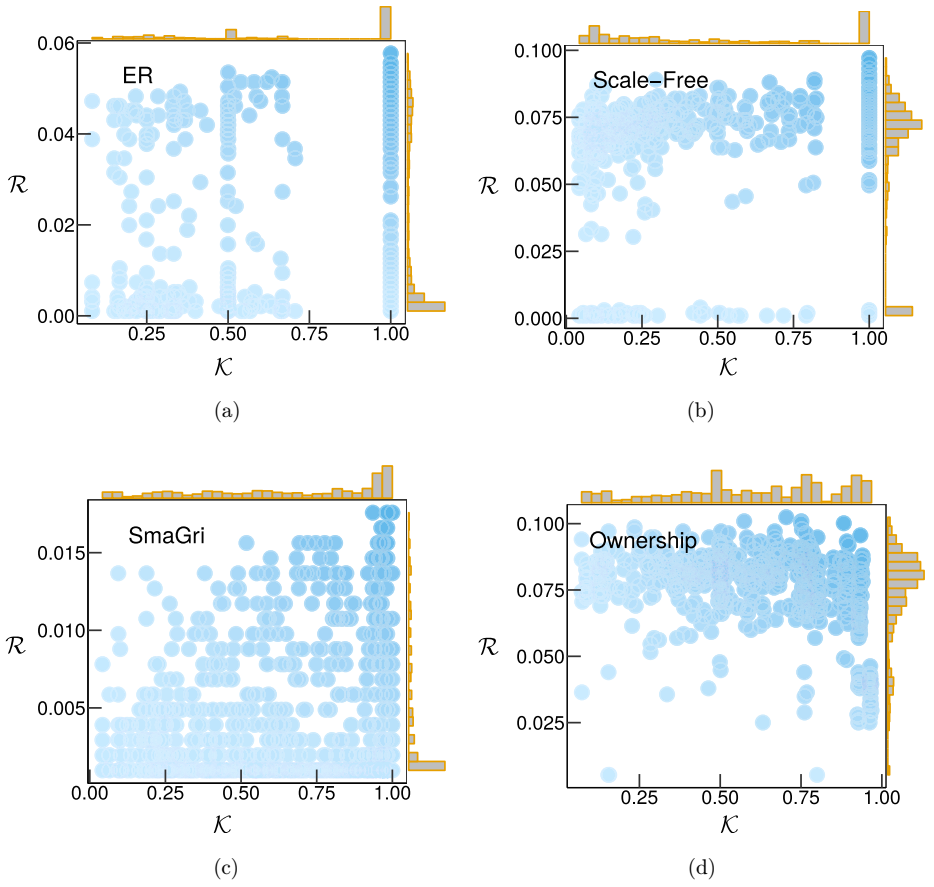

Fig. 5. (Color online) Scatter plot to illustrate the absence of strong positive correlation between control capacity and control range. Here, we color each point according to the control contribution of the corresponding driver node.

We observe that, for all four networks, the values of $\mathcal{R}(\mathcal{K})$ are very broadly scattered. Importantly, there are no uniform patterns or strong correlation between control range $\mathcal{R}$ and control capacity $\mathcal{K}$. This means that we can hardly predict one measure based on the other measure, and both measures capture individual aspects of drivers. This also justifies the motivation for our proposed measure, which combines the two different aspects and provides new information that cannot be covered by either of them, alone.

4. Discussion

Network controllability helps us to identify the minimum set of driver nodes, MDS, needed to control the whole network. Under practical circumstances, however, we may not have access to all of these driver nodes or do not want to control the whole

network. Then, the question arises how to choose a smaller set of driver nodes such that, given this number, the largest possible subset of the network can be controlled. If we have to restrict to this smaller set, we should have a ranking of driver nodes that allows us to pick those that have the largest impact on controlling the network.

Existing measures for such a ranking, e.g., control capacity, $\mathcal{K}$, and control range, $\mathcal{R}$, are not best suited because they only focus on one aspect of driver nodes, either their probability to become a driver or the size of the subnetwork they control. As one contribution of this paper, we provide a new measure, *control contribution* $\mathcal{C}$, that combines these two aspects. We demonstrate that driver nodes chosen according to their $\mathcal{C}$ values always perform better in controlling the network.

As a second contribution, we verify that $\mathcal{C}$ indeed contains information that cannot be traced back to the degree, control capacity or control range of a node. This was shown both by studying the correlations between these measures and by means of arguments related to the network topology (see Sec. 3.3).

In conclusion, using control contribution $\mathcal{C}$ allows us to identify the driver nodes with the most impact in controlling the network, without the pain of a "brute force" approach to cope with combinatorial explosion.

Acknowledgments

All authors designed and performed the research and wrote the paper. Y. Z. performed all the simulations.

Appendix A. Significance in the Difference of Controllable Subspace N_b

As Fig. 3 indicates, some of the curves are very close. Therefore, we test whether the difference in n_b is significant. For this, we use an ensemble approach based on 100 synthetic networks with the same network configuration parameters. Concretely, we calculate the areas $S_{\mathcal{C}}$, $S_{\mathcal{K}}$, $S_{\mathcal{R}}$ under the respective curves of n_b. To facilitate the comparison, we define two measures, $RS_0 = S_{\mathcal{C}}/S_{\mathcal{K}} - 1$ and $RS_1 = S_{\mathcal{C}}/S_{\mathcal{R}} - 1$ to capture the difference in the area sizes. Obviously, if the measure $\mathcal{C}$ for choosing driver nodes outperforms $\mathcal{K}$ and $\mathcal{R}$, then RS_0 and RS_1 should be larger than 0.

Figure A.1 displays the distribution of RS_0 and RS_1 obtained from the 100 synthetic networks, both for Erdős–Rényi and scale-free networks. RS_1 is always positive, i.e., choosing driver nodes according to control contribution always leads to a larger controllable subnetwork, in comparison to control range. For RS_0, the major part of the distribution is above zero, with an associated P value of 0.0001 at the 95% confidence interval. Therefore, on average, choosing driver nodes with respect to control contribution leads to a larger controllable subnetwork, in comparison to control capacity. This means that based on our simulations, $\mathcal{C}$ is the best measure to rank driver nodes compared with existing control-based measures.

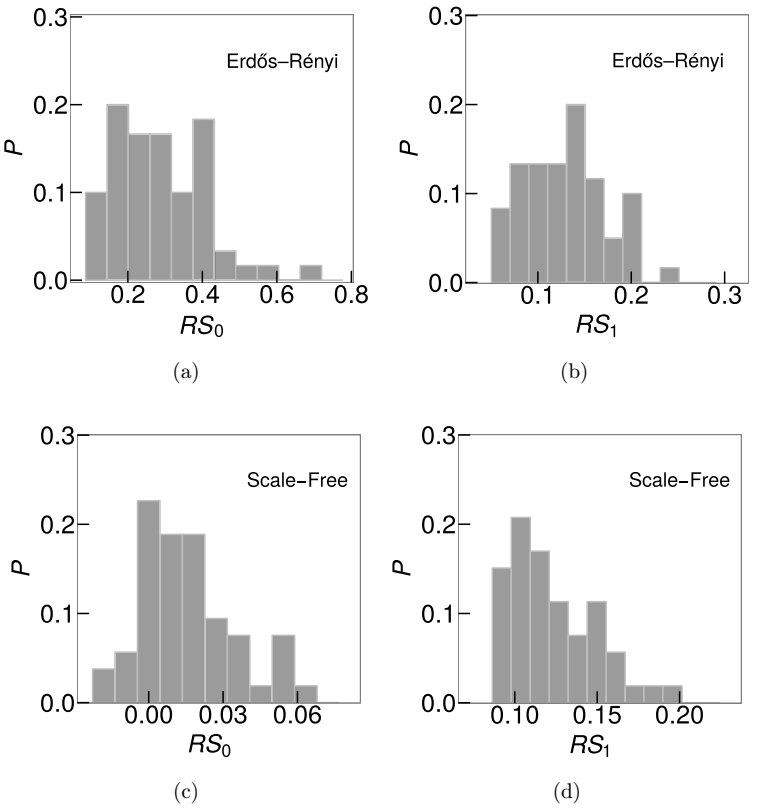

Fig. A.1. Distributions of RS_0 and RS_1 for ((a), (b)) Erdős–Rényi networks, ((c), (d)) Scale-free networks. Note the changes of the x scale.

References

[1] Luenberger, DG, *Introduction to Dynamic Systems Theory, Models, and Applications* (Wiley, 1979).
[2] Liu, Y.-Y. and B. Albert-László, Control principles of complex systems, *Rev. Mod. Phys.* **88**(3) (2016) 35006.
[3] Liu, Y.-Y., Slotine, J.-J. and Barabasi, A.-L., Controllability of complex networks, *Nature* **473**(7346) (2011) 167–173.
[4] Nacher, J. C. and Akutsu, T., Structural controllability of unidirectional bipartite networks, *Sci. Rep.* **3** (2013) 1647.
[5] Yuan, Z., Zhao, C., Wang, W. X., Di, Z. and Lai, Y. C., Exact controllability of multiplex networks, *New J. Phys.* **16** (2014) 103036.
[6] Srihari, S., Raman, V., Leong, H. W. and Ragan, M., Evolution and controllability of cancer networks: A Boolean perspective, *IEEE/ACM Trans. Comput. Biol. Bioinform.* **6** (2013) 83–94.
[7] Liu, X. and Pan, L., Detection of driver metabolites in the human liver metabolic network using structural controllability analysis, *BMC Syst. Biol.* **8** (2014) 51.
[8] Xiang, L., Chen, F., Ren, W. and Chen, G., Advances in network controllability, *IEEE Circuits Syst. Mag.* **19**(2) (2019) 8–32.

[9] Wuchty, S., Controllability in protein interaction networks, *Proc. Natl. Acad. Sci. USA* **111** (2014) 7156–1760.
[10] Vinayagam, A., Gibson, T. E., Lee, H.-J., Yilmazel, B., Roesel, C., Hu, Y., Kwon, Y., Sharma, A., Liu, Y.-Y., Perrimon, N. and Barabási, A.-L., Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets, *Proc. Natl. Acad. Sci. USA* **113** (2016) 1603992113.
[11] Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E. and Makse, H. A., Identification of influential spreaders in complex networks, *Nat. Phys.* **6** (2010) 888.
[12] Garas, A., Schweitzer, F. and Havlin, S., A κ-shell decomposition method for weighted networks, *New J. Phys.* **14** (2012) 083030.
[13] Freeman, L. C., A set of measures of centrality based on betweenness, *Sociometry* **40** (1977) 35–41.
[14] Wang, B., Gao, L. and Gao, Y., Control range: A controllability-based index for node significance in directed networks, *J. Statist. Mech.* **2012** (2012) P04011.
[15] Jia, T. and Barabási, A.-L., Control capacity and a random sampling method in exploring controllability of complex networks, *Sci. Rep.* **3** (2013) 2354.
[16] Liu, Y.-Y., Slotine, J.-J. and Barabási, A.-L., Control centrality and hierarchical structure in complex networks, *PLoS One* **7** (2012) e44459.
[17] Kalman, R. E., Mathematical description of linear dynamical systems, *J.S.I.A.M. Control* **1**(2) (1963) 152–192.
[18] Lin, C. T., Structural controllability, *IEEE Trans. Automat. Control* **19**(3) (1974) 201–208.
[19] Chung, F. and Lu, L., Connected components in random graphs with given expected degree sequences, *Ann. Combin.* **6** (2002) 125–145.
[20] Vitali, S., Glattfelder, J. and Battiston, S., The network of global corporate control, *PLoS One* **6**(10) (2011) e25995.
[21] Zhang, Y. and Schweitzer, F., The interdependence of corporate reputation and ownership: A network approach to quantify reputation, *R. Soc. Open Sci.* **6** (2019) 190570.
[22] Poljak, S., On the generic dimension of controllable subspaces, *IEEE Trans. Automat. Control* **35** (1990) 367–369.
[23] Zhang, Y., Garas, A. and Schweitzer, F., Value of peripheral nodes in controlling multilayer scale-free networks, *Phys. Rev. E* **93**(1) (2016) 1–6.