



# OPEN Input node placement restricting the longest control chain in controllability of complex networks

Samie Alizadeh<sup>1</sup>, Márton Pósfai<sup>2</sup> & Abdorasoul Ghasemi<sup>1</sup>✉

The minimum number of inputs needed to control a network is frequently used to quantify its controllability. Control of linear dynamics through a minimum set of inputs, however, often has prohibitively large energy requirements and there is an inherent trade-off between minimizing the number of inputs and control energy. To better understand this trade-off, we study the problem of identifying a minimum set of input nodes such that controllability is ensured while restricting the length of the longest control chain. The longest control chain is the maximum distance from input nodes to any network node, and recent work found that reducing its length significantly reduces control energy. We map the longest control chain-constraint minimum input problem to finding a joint maximum matching and minimum dominating set. We show that this graph combinatorial problem is NP-complete, and we introduce and validate a heuristic approximation. Applying this algorithm to a collection of real and model networks, we investigate how network structure affects the minimum number of inputs, revealing, for example, that for many real networks reducing the longest control chain requires only few or no additional inputs, only the rearrangement of the input nodes.

Network control aims to understand how network structure affects our ability to control the dynamics of complex systems<sup>1</sup>. Development of the field is motivated by a range of applications from biology to epidemiology or technological systems. Beyond potential applications, designing control strategies that leverage network structure also deepen our understanding of the dynamics of networked systems. Particularly useful for this are methods that clearly separate the role of network structure from the details of the dynamics on the network, examples of such methods include pinning control, feedback vertex control or structural control of linear systems<sup>2–6</sup>. These methods, however, necessarily make assumptions about the dynamics, limiting the scope of systems they apply to and the range of questions we can ask. For example, structural controllability of linear systems allows us to determine if a network with a set of input nodes is controllable or not relying on network structure only<sup>4</sup>. However, we cannot use it to design an input signal to actually control the network, nor can we rely on it to calculate the energy needed for control. To exactly answer these questions, we need a full description of the dynamics, which in the case of linear systems means knowing all link weights exactly. In addition, input node selection that minimizes control energy is an NP-hard problem<sup>7</sup>. Most work investigating this problem, therefore, model link weights and largely focus on approximate numerical methods<sup>8,9</sup>, for example, using properties of the controllability Gramian matrix that is often ill-conditioned or singular. To overcome these limitations, several authors developed heuristic methods that rely only on network structure only to investigate control energy related problems, and these heuristics perform reasonably well for realistic networks<sup>10</sup>. For example, references<sup>11,12</sup> found that input node placement that minimizes the longest control chain (LCC), i.e., the largest distance from input nodes to any node, can significantly reduce the energy necessary to control networks.

In this paper, we investigate a graph combinatorial optimization problem motivated by the LCC heuristic. Specifically, we search for a minimum set of input nodes in a directed network that ensures controllability such that the largest control chain is at most  $\ell$ . Relying on the framework of structural controllability, we show that the LCC-constrained minimum input problem is equivalent to solving a combined maximum matching and minimum dominating set problem. After showing that the problem is NP-complete, we develop a heuristic algorithm to approximate the optimal solution. We then investigate the effect of network structure on the optimal number of inputs and the cost of reducing the length of the LCC by applying the algorithm to a collection of

<sup>1</sup>Department of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran. <sup>2</sup>Department of Network and Data Science, Central European University, Vienna, Austria. ✉email: arghasemi@kntu.ac.ir

model and real complex networks. We demonstrate, for example, that the cost of reducing LCCs diminishes for very heterogeneous networks and very sparse or very dense networks. We also show that for many real complex networks reducing the LCC compared to the maximum matching based minimum input selection requires only few or no additional inputs, only the rearrangement of the input nodes.

The paper is organized as follows. Section 2 summarizes the theoretical background and introduces the problem setup. Next, in Sect. 3, we formally define the LCC-constrained minimum input problem, we prove that the problem belongs to the NP-complete class and formulate it as an integer linear programming to solve small size instances. In Sect. 4, we introduce and validate a heuristic algorithm to solve the LCC-constrained input problem for large networks. We use this algorithm and a collection of model and real networks in Sect. 5 to study how network structure affects minimum input selection. Finally, we conclude with a short discussion and outlook in Sect. 6.

## Problem setup and background

Linear time-invariant dynamics with open-loop control became a canonical model to study the control of complex networks, since the choice of linear dynamics often leads to tractable formulation of control problems and it allows focus on the role of network structure<sup>5</sup>. To formally define the model, consider a directed network  $\mathcal{G}(\mathcal{V}, \mathcal{E})$  where  $\mathcal{V} = \{v_i | i = 1, \dots, N\}$  is the set of nodes and  $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$  is the set of weighted directed links, i.e., each link  $(v_i \rightarrow v_j)$  has an associated weight  $a_{ji} \in \mathbb{R}$  representing the strength of the interaction. We assign a state  $x_i \in \mathbb{R}$  to each node  $v_i$  which evolves following the equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad (1)$$

where the first term on the right hand side represents the internal dynamics of the system and the second term expresses the external control imposed on the network. Specifically,  $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$  is the vector of node states,  $\mathbf{A} \in \mathbb{R}^{N \times N}$  is the weighted adjacency matrix,  $\mathbf{u}(t) = [u_1, u_2, \dots, u_M]^T \in \mathbb{R}^M$  is a vector of  $M$  time-dependent control signals, and matrix  $\mathbf{B} \in \mathbb{R}^{N \times M}$  defines how the control signals are coupled to the system.

A system  $(\mathbf{A}, \mathbf{B})$  is controllable if, with the proper choice of control signals, we can drive it from any initial state  $\mathbf{x}(0)$  to any final state  $\mathbf{x}(t_f)$  in finite time. Determining controllability, in general, is a numerically unstable problem that requires exact knowledge of the link weights, making it difficult to study for real large-scale networks directly. However, relying on some mild conditions on the link weights and the structural controllability framework, this inherently numerical problem can be mapped to a combinatorial graph problem. First, we define an extended network  $\mathcal{G}'(\mathcal{V} \cup \mathcal{U}, \mathcal{E} \cup \mathcal{E}')$  where we add additional nodes  $\mathcal{U} = \{u_i | i = 1, \dots, M\}$  representing the control signals and we add a link  $(u_i \rightarrow v_j)$  if the control signal  $u_i(t)$  is directly connected to network node  $v_j$ , i.e.,  $b_{ji} \neq 0$ , where  $b_{ji}$  is an element of  $\mathbf{B}$ . The network is controllable if in  $\mathcal{G}'$  (i) all nodes are accessible from the control signals; and (ii) there exists a matching such that only nodes representing control signals are unmatched<sup>5</sup>. A matching in a directed network is a subset of links  $\mathcal{E}_M \subset \mathcal{E}$  such that no two links in  $\mathcal{E}_M$  share start or endpoints, and a node is unmatched if no link in  $\mathcal{E}_M$  is pointing at it.

This mapping provides an efficient and robust method to determine the controllability of a network and a controller; however, we are often interested in designing a controller or characterizing control properties of a network without any predefined  $\mathbf{B}$ . For example, the minimum number of input signals  $N_i$  needed to control a network can serve as a quantity capturing the difficulty of controlling a network. If a control signal  $u_i(t)$  can couple to multiple nodes, finding  $N_i$  is equivalent to finding a maximum matching in  $\mathcal{G}$ : inserting an independent control signal to any unmatched node in  $\mathcal{G}$  ensures that only control signals are unmatched in  $\mathcal{G}'$  and the accessibility condition can be satisfied by adding additional connections between control signals and network nodes as needed. Finding a maximum matching in a directed network can be done in polynomial time; if, however, each control signal can only be connected to a single network node, i.e., each column of  $\mathbf{B}$  only contains one nonzero element, we cannot satisfy the accessibility condition easily, and the minimum input problem becomes NP-complete<sup>13</sup>.

Solving the minimum input problem provides a set of nodes to control the system; however, it does not provide any information about the signal  $\mathbf{u}(t)$  or the energy required to control the system. The control energy of a control signal  $\mathbf{u}(t)$  is defined as

$$\varepsilon = \int_0^{t_f} \mathbf{u}_t^T \mathbf{u}_t dt, \quad (2)$$

where  $t_f$  is the time at which a target state is reached<sup>14</sup>. We can express the optimal control signals minimizing the control energy explicitly as

$$\mathbf{u}(t) = \mathbf{B}^T e^{\mathbf{A}^T(t_f-t)} \mathbf{W}_B^{-1}(t_f) (\mathbf{x}(t_f) - e^{\mathbf{A}t_f} \mathbf{x}(0)), \quad (3)$$

where  $\mathbf{W}_B^{-1}(t_f) = \int_0^{t_f} e^{\mathbf{A}\tau} \mathbf{B}\mathbf{B}^T e^{\mathbf{A}^T\tau} d\tau$  is controllability Gramian matrix<sup>14</sup>. Several factors determine the optimal control energy, including the matrices  $\mathbf{A}$  and  $\mathbf{B}$ , the target state  $\mathbf{x}(t_f)$ , and control time  $t_f$ <sup>15-19</sup>. Also, several metrics were proposed to quantify the energy requirements of controlling a system  $(\mathbf{A}, \mathbf{B})$ , common choices include the average energy or the worst-case energy among all possible target states<sup>16</sup>. Several authors observed that independent of which metric is chosen, the energy required to control networks through minimum input sets is often prohibitively large<sup>11,15,16,20</sup> and adding additional input nodes can reduce the energy. This prompts the question: how can we select a small number of input nodes such that both controllability is ensured and control energy is reduced?

Selecting input nodes that minimize the control energy is typically a hard computational problem, and most existing work focuses on numerical approximations that do not explicitly leverage the network structure of the

system. For example, reference<sup>7</sup> shows that identifying a matrix  $\mathbf{B}$  that minimize the average control energy while keeping  $\text{tr}(\mathbf{B}^T \mathbf{B}) = M$  fixed is an NP-hard problem. then propose an approximation algorithm based on the projected gradient method and test it using both synthetic and real networks. However, almost all elements of the matrix  $\mathbf{B}$  that their algorithm finds are non-zero, meaning that external signals are split up and connected to a significant fraction of network nodes, which is impractical for large networks. Several extensions were proposed that aim to constrain the number of nodes directly driven by external signals; these methods typically do not guarantee controllability and may suffer performance loss in real networks<sup>21,22</sup>.

The above methods take a generic matrix  $\mathbf{A}$  as input and rely on numerical optimization, not making explicit use of the network structure of the system. Acknowledging that  $\mathbf{A}$  represents an adjacency matrix of a complex network allows us to develop methods that use the typical structure of complex networks, potentially leading to useful approximations and providing insights into how these systems function. For example, for a given network and a set of input nodes, reference<sup>11</sup> defined the longest control chain (LCC) as the largest distance between input nodes and any node, i.e.,

$$l_{\text{LCC}} = \max_{w \in \mathcal{V}} \min_{v \in \mathcal{S}} d(v, w), \quad (4)$$

where  $\mathcal{S}$  is the set of input nodes and  $d(v, w)$  is the length of the shortest path connecting from  $v$  to  $w$ . The authors found that reducing the length of the LCC in a network reduces the energy requirements of control, later theoretical support for this observation was provided<sup>12,23</sup> and the role of LCC was explored for target control<sup>24</sup>. The relationship between  $l_{\text{LCC}}$  and  $\varepsilon$  offers a heuristic strategy to select input nodes that reduce the energy requirements of control: the algorithms proposed by references<sup>12,23</sup> take an already controllable system  $(\mathbf{A}, \mathbf{B})$  as input and modify  $\mathbf{B}$  by coupling signals to additional nodes or adding further input signals to reduce  $l_{\text{LCC}}$ . Yet, the question of identifying an optimal set of input nodes for a network  $\mathbf{A}$  without first providing a  $\mathbf{B}$  remains unexplored.

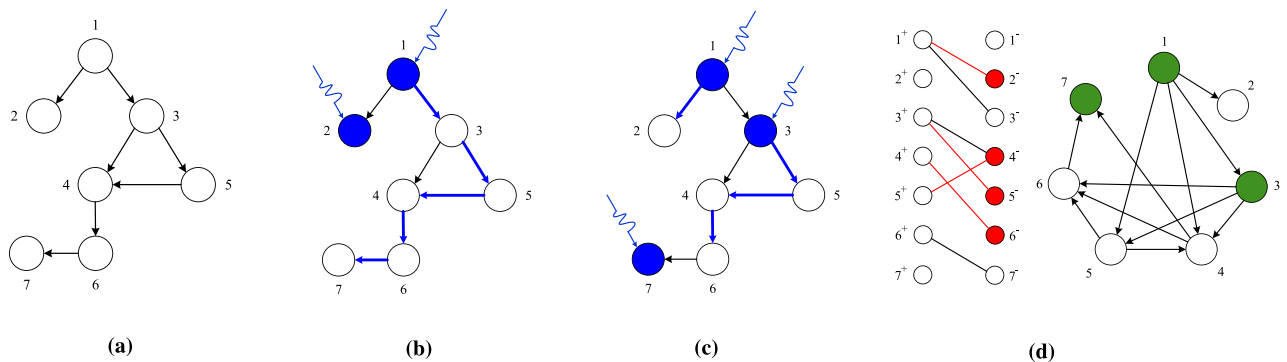
In this article, we investigate the LCC-constrained minimum input problem, more specifically, for a given directed network  $\mathbf{A}$  and a integer  $\ell > 0$ , we search for a minimum set of input nodes that simultaneously ensures structural controllability and  $l_{\text{LCC}} \leq \ell$  such that each input signal is coupled to exactly one node (Fig. 1). Both determining structural controllability of a linear system and calculating the length of the LCC only depend on network structure, hence the LCC-constrained minimum input problem is a graph combinatorial problem. A central quantity in our study is the minimum number input nodes needed to control a network while ensuring that  $l_{\text{LCC}} \leq \ell$ , which we denote by  $N_i(\ell)$ .

In the next section, we formally define the problem and we provide an exact algorithm to solve it using integer linear programming.

### LCC-constrained minimum input problem

We study the problem of identifying a minimum set of input nodes  $\mathcal{S}$  in a directed network  $\mathcal{G}$  such that the system is structurally controllable, the length of the LCC is at most  $\ell$ , and each input is connected to exactly one node in the network. To continue, we map the controllability condition to a bipartite matching problem and the LCC constraint to a dominating set problem:

- **Controllability constraint** The set of input nodes  $\mathcal{S}$  ensures structural controllability if (i) there exists a matching such that the set of unmatched nodes is  $\mathcal{S}$  and (ii) all nodes are accessible from  $\mathcal{S}$ . The condition imposed on the LCC implies accessibility; therefore if  $l_{\text{LCC}} \leq \ell$  then condition (ii) is also satisfied, and we only have to check condition (i). For this we create a bipartite representation  $\mathcal{B}$  of a directed network  $\mathcal{G}$ : We split each



**Figure 1.** LCC-constrained minimum input problem. (a) We aim to control a small directed network  $\mathcal{G}$ . (b) Finding a maximum matching in the network (blue links) and selecting the unmatched nodes 1 and 2 provides a minimum input node set. This input set, however, does not optimize the length of the LCC. In this case the LCC  $(1 \rightarrow 3 \rightarrow 4 \rightarrow 6 \rightarrow 7)$  has length four. (c) To ensure that  $l_{\text{LCC}} \leq 2$ , we need one additional input node (node 7) and we need to control node 3 instead of node 2. (d) The bipartite representation  $\mathcal{B}$  of digraph  $\mathcal{G}$  and the  $\ell$ -step accessibility graph  $\mathcal{G}_\ell$  for  $\ell = 2$  (defined in Sect. 3). Red links in  $\mathcal{B}$  represent a matching. The unmatched nodes in the  $\mathcal{V}^-$  side of  $\mathcal{B}$  form a dominating set in  $\mathcal{G}_\ell$  (green nodes), directly controlling these nodes simultaneously ensures controllability and satisfies the  $l_{\text{LCC}} \leq 2$  condition.

- node  $v \in \mathcal{V}$  into two copies  $v^+ \in \mathcal{V}^+$  and  $v^- \in \mathcal{V}^-$ . If there exists a directed link ( $v \rightarrow w$ ) in  $\mathcal{G}$ , we add an undirected link ( $v^+ - w^-$ ) to  $\mathcal{B}$ . A directed matching  $\mathcal{G}$  is equivalent to a bipartite matching in  $\mathcal{B}$  (Fig. 1d)<sup>5,25</sup>.
- **LCC-constraint** To satisfy the condition  $I_{LCC}$ , we must ensure that there exists a path connecting an input node to each node with length at most  $\ell$ . For this we define the  $\ell$ -step accessibility graph  $\mathcal{G}_\ell$  of a  $\mathcal{G}$ , as a directed graph with the same node set as  $\mathcal{G}$ , and where we connect node  $v$  and  $w$  in  $\mathcal{G}_\ell$  with a directed link ( $v \rightarrow w$ ) if there exists a path from  $v$  to  $w$  with at most  $\ell$  steps in the original network  $\mathcal{G}$ . A node set  $\mathcal{S}$  satisfies the LCC-constraint in  $\mathcal{G}$  if and only if  $\mathcal{S}$  is a dominating set in the accessibility graph  $\mathcal{G}_\ell$ , i.e., each node  $v$  in  $\mathcal{G}_\ell$  is either a member  $\mathcal{S}$  or there is a link pointing from a node in  $\mathcal{S}$  to  $v$  (Fig. 1d).

This mapping allow us to build on algorithms and analytical results originally developed for bipartite matchings and directed dominating sets. For example, in the following, we rely on the mapping to calculate a simple upper and lower bound for the minimum number of input nodes  $N_i(\ell) = |\mathcal{S}|$ , and to develop an exact algorithm to solve the minimum input problem.

To find the lower bound, first note that the input nodes  $\mathcal{S}$  are unmatched; therefore  $N_i(\ell)$  is at least equal to the number of unmatched nodes  $N_M$  in a maximum matching of  $\mathcal{G}$ . The input nodes  $\mathcal{S}$  are also a dominating set in  $\mathcal{G}_\ell$ ; therefore  $N_i(\ell)$  is at least equal to the size of a minimum dominating set  $N_{DS}$  of  $\mathcal{G}_\ell$ . These two observations lead to the lower bound

$$N_i(\ell) \geq \min\{N_M, N_{DS}\}. \tag{5}$$

To obtain the upper bound, we construct a non-optimal set of input nodes by finding a maximum matching and a minimum dominating set independently, and taking the union of unmatched nodes and the dominating set. In the worst case there is no overlap between the unmatched and dominating nodes, leading to the upper bound

$$N_i(\ell) \leq N_M + N_{DS}. \tag{6}$$

We can improve this bound by noting that source nodes, i.e., nodes with no incoming links, are both always unmatched and always dominating. The source nodes, therefore, are counted twice in Eq. (6), correcting for this provides

$$N_i(\ell) \leq N_M + N_{DS} - N_s, \tag{7}$$

where  $N_s$  is the number of source nodes.

To develop an algorithm to find the optimal set of input nodes, we first show that the LCC-constrained minimum input problem is NP-complete by reducing the minimum dominating set problem to it, for details see Sect. S1 in the Supplementary Information. This prompts us to reformulate the problem as an integer linear programming (ILP) problem, which allows us to use efficient branch-and-bound ILP solvers to find exact solutions to the minimum input problem.

To define the ILP problem, we assign a binary variable  $e_{i \rightarrow j} \in \{0, 1\}$  to each link ( $i \rightarrow j$ ), such that  $e_{i \rightarrow j} = 1$  if link ( $i \rightarrow j$ ) is in the matching, otherwise  $e_{i \rightarrow j} = 0$ . Using this binary variable we reformulate the LCC-constrained minimum input problem as

$$\min_{e_{i \rightarrow j} \in \mathcal{E}} \sum_{v \in \mathcal{V}} (1 - \sum_{i \in \mathcal{V}_v^-} e_{i \rightarrow v}) \tag{8a}$$

subject to

$$\sum_{j \in \mathcal{V}_v^+} e_{v \rightarrow j} \leq 1 \quad \forall v \in \mathcal{V} \tag{8b}$$

$$\sum_{i \in \mathcal{V}_v^-} e_{i \rightarrow v} \leq 1 \quad \forall v \in \mathcal{V} \tag{8c}$$

$$\sum_{k \in \mathcal{V}_v^\ell} (1 - \sum_{i \in \mathcal{V}_k^-} e_{i \rightarrow k}) \geq 1 \quad \forall v \in \mathcal{V}, \tag{8d}$$

where  $\mathcal{V}_i^+$  and  $\mathcal{V}_i^-$  are the set of neighbors of node  $v_i^+$  and  $v_i^-$  in  $\mathcal{B}$ , and  $\mathcal{V}_i^\ell$  is the set of predecessors or in-neighbors of node  $v_i$  in the accessibility graph  $\mathcal{G}_\ell$ . The cost function (8a) counts the number of input nodes, i.e., unmatched nodes in  $\mathcal{V}^-$ . Constraints (8b) and (8c) enforce the matching criteria, i.e., that each node in  $\mathcal{B}$  is adjacent to at most one link in the matching. While constraint (8d) enforces the dominating set criteria, i.e., that each node in  $\mathcal{G}_\ell$  is an input node or has an input node as a predecessor. Using Eq. (8) we can use off-the-shelf ILP solvers to find exact solutions to the LCC-constrained minimum input problem. Note, however, that many ILP representations of the same problem may exist and some of the representations might be more suitable for generic ILP solvers. Indeed, in Sect. S2 of the Supplementary Information we provide an alternative representation that is more complicated to formulate than Eq. (8), yet leads to superior performance.

Although we rely on efficient ILP solvers, due to hardness of the problem we are able to find minimum input sets for networks with upto a few hundred nodes and links. We overcome this limitation in the next section by developing a heuristic algorithm and we use the exact algorithm to test the performance of the approximation on small networks.

## Approximate algorithm

Since the LCC-constrained minimum input problem is NP-complete, exact solution is limited to networks with a few hundred links, prompting us to search for approximate solutions. In this section, we develop a greedy algorithm based on a set of locally optimal rules. To find an input set, we iteratively apply these rules to a network: if all nodes are processed we find an optimal solution, if the process gets stuck, we proceed following a heuristic and the final solution is no longer guaranteed to be optimal.

The LCC-constrained minimum input problem is a combination of the maximum matching and the minimum dominating set problems. Next, we review existing greedy methods for these two optimization problems, then we describe how to combine the two algorithms to address the LCC-constrained minimum input problem.

**Greedy maximum matching.** A greedy construction of a maximal matching is possible using the Karp-Sipser leaf-removal algorithm<sup>26,27</sup>, also known as core percolation in the network science literature<sup>28,29</sup>. For consistency within the paper, we refer to the algorithm as matching leaf-removal (MLR). Although polynomial time exact algorithms exist for maximum matching, it is still useful to consider this greedy construction, as it serves as the basis for approximating the NP-complete LCC-constrained minimum input problem.

To apply the MLR to a directed network  $\mathcal{G}$ , we first convert  $\mathcal{G}$  to its undirected bipartite network  $\mathcal{B}$ . The MLR algorithm then constructs a maximum matching by iteratively applying the following rule to  $\mathcal{B}$ :

- *Rule-M* Select a leaf node  $v$ , i.e., a node with a single neighbor  $w$ , and place the link  $(v - w)$  into the matching. Remove the link  $(v - w)$  from the network together with any other link connected to  $w$ .

If there are no more leaf nodes in the network the MLR halts, and we call the set of remaining links the matching-core (M-core) of the network. If the M-core of the network is empty, we are done and the obtained matching is optimal. If the M-core is non-empty, we have to continue relying on a heuristic and the constructed matching is no longer guaranteed to be optimal. The most simple way to proceed is to add a random link to the matching, remove it together with any adjacent links, and apply the MLR algorithm again to the remaining network. We repeat this until all links are removed.

Note that the order of leaf-removal is not specified by the algorithm and while the size of the M-core, and, if there is no M-core, the cardinality of the obtained matching does not depend on the order of leaf-removal, the specific links that are added to the matching do. Therefore, different leaf-removal order leads to different matchings even if the M-core is empty and the cardinality of the matching is optimal. Choosing different leaf-nodes to remove lead to equivalent outcomes when searching for a maximum matching, but this is no longer true when we are simultaneously considering the dominating set problem.

**Greedy minimum dominating set.** Finding the minimum dominating set of a general network is itself NP-complete and therefore a harder problem than maximum matching. A large number of approximate algorithms exist<sup>30–35</sup>, here we focus on a greedy generalized leaf-removal algorithm introduced by Habibulla et al., which in spirit is similar to the Karp-Sipser leaf-removal<sup>36</sup>.

The generalized leaf-removal scheme, which we refer to as the dominating set leaf-removal (DSLRL), iteratively deletes nodes and links from a directed graph  $\mathcal{G}(\mathcal{V}, \mathcal{E})$  to find a minimal set of dominating nodes. During the DSLRL each node  $v$  can have three labels: (i)  $v$  is dominating if it has been placed in the dominating set, (ii)  $v$  is observed if  $v$  or at least one of its predecessors is a dominating node, or (iii) otherwise  $v$  is unobserved. Initially all nodes are unobserved and we apply the following rules iteratively:

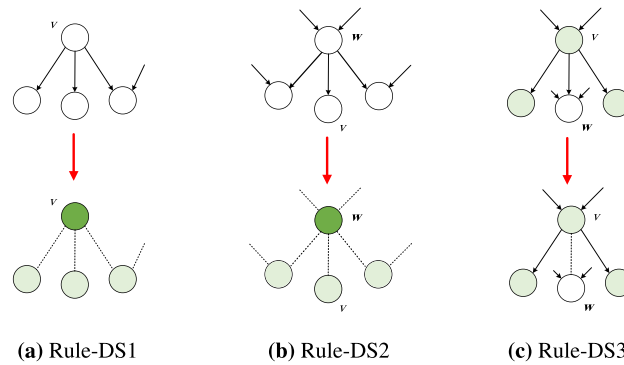
- *Rule-DS1* If an unobserved node  $v$  has no predecessor (i.e. has in-degree zero), add it to the dominating set. All successors of  $v$  become observed (Fig. 2a).
- *Rule-DS2* If an unobserved node  $v$  has only a single predecessor  $w$  and no unobserved successor, add node  $w$  to the dominating set. All the successors of  $w$  then become observed (Fig. 2b).
- *Rule-DS3* If a node  $v$  is observed and has only a single unobserved successor  $w$ , selecting  $v$  as a dominating node is not better than selecting  $w$ , therefore the link  $(v, w)$  is deleted from the network. The node  $w$  remains unobserved after the link deletion (Fig. 2c).

When applying the rules, links that point to dominating nodes and their successors are removed, i.e., incoming links to observed nodes. If no more links can be removed from the network, we call the remaining set of links the dominating set-core (DS-core) of the network. Reference<sup>36</sup> showed that if the DS-core of the network is empty, the obtained dominating set (DS) is optimal. If the DS-core is non-empty, we must proceed relying on a heuristic and the solution is no longer guaranteed to be optimal. Specifically, we continue by adding the node with the highest degree to the dominating set, setting all of its successors to be observed, then applying the DSLRL rules to the updated network. We repeat this until all links are removed.

Similarly to maximum matching, the dominating set produced by the DSLRL algorithm is not unique, even if the network has no DS-core. The choice of the order that the rules are applied to the network does not affect the outcome when minimizing the dominating set, but again, this is no longer true if we are simultaneously aiming to satisfy the matching problem.

**LCC-constrained minimum input problem.** To identify an LCC-constrained minimum input set, we search for a set of nodes  $\mathcal{S}$  such that both (i)  $\mathcal{S}$  is a dominating set in the accessibility graph  $\mathcal{G}_\ell$  and (ii) there exists a matching in  $\mathcal{B}$  such that  $\mathcal{S}$  is the set of unmatched nodes in  $\mathcal{V}^-$ . Our strategy is to use the MLR to identify





**Figure 2.** Local rules to construct a minimal dominating set. White nodes are unobserved, light green nodes are observed, and dark green indicates dominating nodes. Removed links are drawn with a dashed line.

unmatched nodes in  $\mathcal{B}$  and update  $\mathcal{G}_\ell$  by adding these nodes to the dominating set. If we get stuck, we proceed by applying the DSLR algorithm to  $\mathcal{G}_\ell$  and update  $\mathcal{B}$ . We iterate these steps until both networks are processed, if both MLR and DSLR get stuck, we rely on heuristics to proceed.

The construction of the matching and the dominating set are coupled together, i.e., every time, we modify  $\mathcal{B}$  we need to update  $\mathcal{G}_\ell$  and vice versa:

- $\mathcal{B} \rightarrow \mathcal{G}_\ell$ : If a node  $v \in \mathcal{V}^-$  becomes isolated in  $\mathcal{B}$ , it becomes unmatched and hence a control node. Therefore the corresponding  $v$  node in  $\mathcal{G}_\ell$  is added to the dominating set, any successor  $w$  of  $v$  becomes observed, and all links connected to  $v$  and any incoming links connected to  $w$  are deleted.
- $\mathcal{G}_\ell \rightarrow \mathcal{B}$ : If a node  $v$  is a dominating node in  $\mathcal{G}_\ell$ , it has to be a control node. Therefore the corresponding node  $v^-$  in  $\mathcal{B}$  does not have to be matched, hence we remove any links connected to  $v^-$ .

Due to the coupling between the matching and dominating set problems, we cannot simply apply the rules described in Sects. **Greedy maximum matching** and **Greedy minimum dominating set** to  $\mathcal{B}$  and  $\mathcal{G}_\ell$  independently, since, for example, the order that rule-M is applied to  $\mathcal{B}$  may produce different outcomes in  $\mathcal{G}_\ell$ . Instead, when applying the local rules, we must take into account their effect on both networks, and we only proceed if there is no alternative step or it does not affect the other network. Specifically, we rely on the following rules:

- **Rule-M** Control nodes are unmatched nodes in  $\mathcal{V}^-$ , hence placing a node  $v \in \mathcal{V}^+$  into the matching does not affect the set of control nodes. Therefore we apply rule-M to any leaf  $v \in \mathcal{V}^+$  without constraint. On the other hand, removing a leaf  $v \in \mathcal{V}^-$  does affect the set of control nodes. For example, consider the set of leaves  $\mathcal{L} = \{v_i^- \mid i = 1, 2, \dots, k\}$ , where all leaves connect to the same node  $w^+$ . Matching and removing  $(v_j^- - w^+)$  causes leaves  $\mathcal{L} \setminus v_j^-$  to become isolated and hence unmatched nodes. The unmatched nodes in  $\mathcal{B}$  become dominating nodes in  $\mathcal{G}_\ell$ , affecting the dominating set. To make sure that the rule-M is optimal for both the matching and dominating set problems, we only match a leaf  $v^-$  if we already know that it does not have to be a dominating node, i.e.,  $v$  is observed and has zero out-degree in  $\mathcal{G}_\ell$ .
- **Rule-DS1** Any node  $v$  that becomes a source in  $\mathcal{G}_\ell$  deterministically has to become a dominating node; therefore, Rule-DS1 is applied without constraints.
- **Rules-DS2,3** When we apply the second or third rule, we choose a node  $v$  not to be a dominating node (Fig. 2b and c). This is a useful step if only the dominating set is considered, because it is guaranteed that there is no better choice and it decimates  $\mathcal{G}_\ell$ . In the combined matching and dominating set problem, however, the step is only optimal, if  $v^-$  is matched in  $\mathcal{B}$ , i.e., it does not become a control node due to the matching condition.

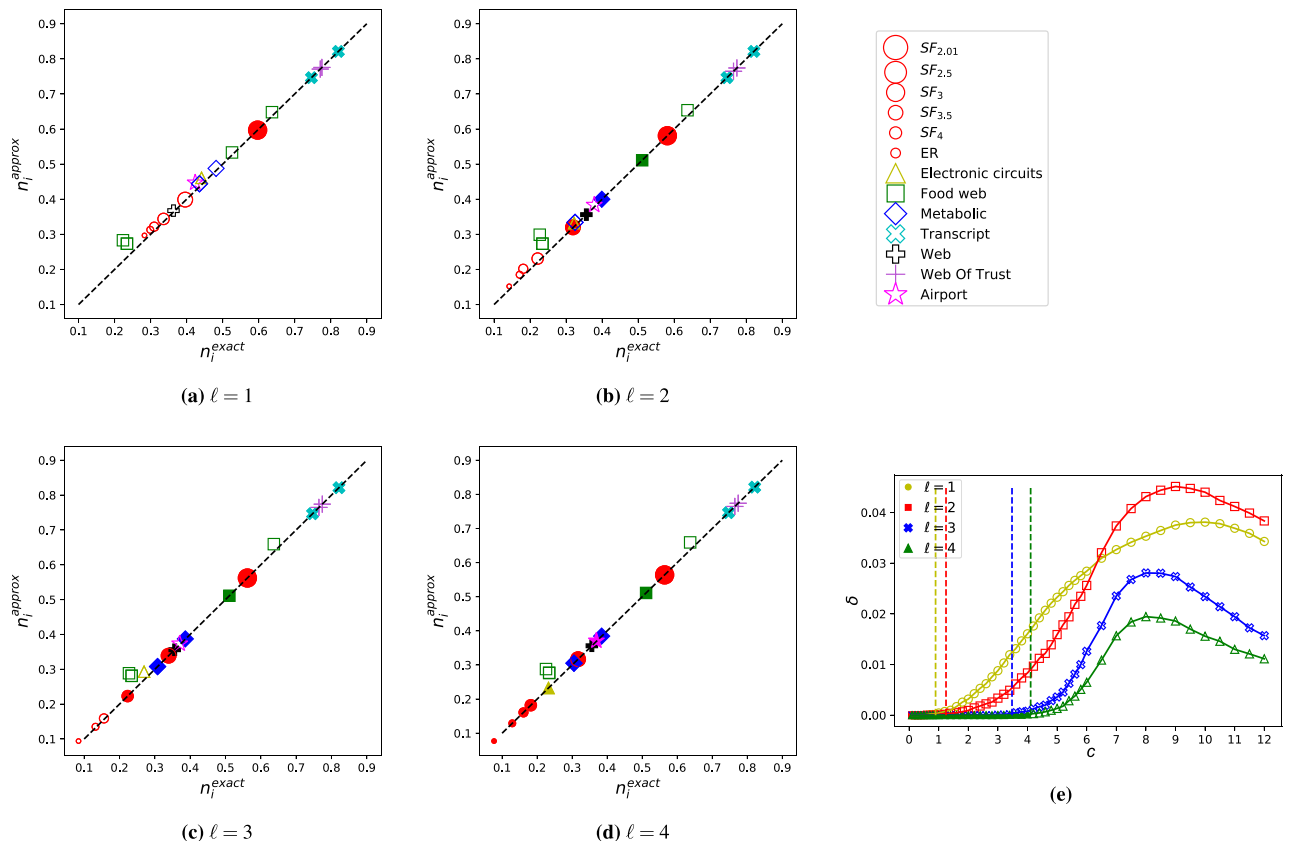
We apply the above rules to  $\mathcal{B}$  and  $\mathcal{G}_\ell$  iteratively. If both networks are processed entirely, the obtained solution is optimal. If we get stuck before resolving all nodes, we call the remainder of links the M-core and the DS-core of  $\mathcal{B}$  and  $\mathcal{G}_\ell$ , respectively. To proceed, we rely on heuristics to continue, and the solution is no longer optimal. Specifically, we match the node in  $\mathcal{B}$  with the least chance to be a dominating node, i.e. matching the node  $v^-$  that  $v$  has the least degree in  $\mathcal{G}_\ell$ . If  $\mathcal{B}$  is empty, we select the node with the highest degree as dominating node in  $\mathcal{G}_\ell$ . Then, we apply the rules to the remaining networks. We repeat this until all links are removed from both  $\mathcal{B}$  and  $\mathcal{G}_\ell$ . An implementation of the algorithm in Python is available online<sup>37</sup>.

**Performance of the algorithm.** We assess the performance of our algorithm using Erdős-Rényi (ER) and scale-free (SF) random networks and a collection of real networks that are small enough for the exact algorithm introduced in Sect. **LCC-constrained minimum input problem**. Figure 3a–e compares the fraction of input nodes  $n_i(\ell) = N_i(\ell)/N$ , where  $N$  is the total number of nodes in the network, obtained using the exact and the approximate algorithm for varying  $\ell$ . We find that (i) for networks without a core the solution returned by the greedy algorithm is indeed optimal, as expected, and that (ii) even for networks with a core the difference  $\delta = n_i^{\text{approx}}(\ell) - n_i^{\text{exact}}(\ell)$  is with less than 0.02 in real networks, except for food webs (the maximum difference 0.07 is measured for *magwet*), and the difference is less than 0.05 for synthetic networks for the parameters that

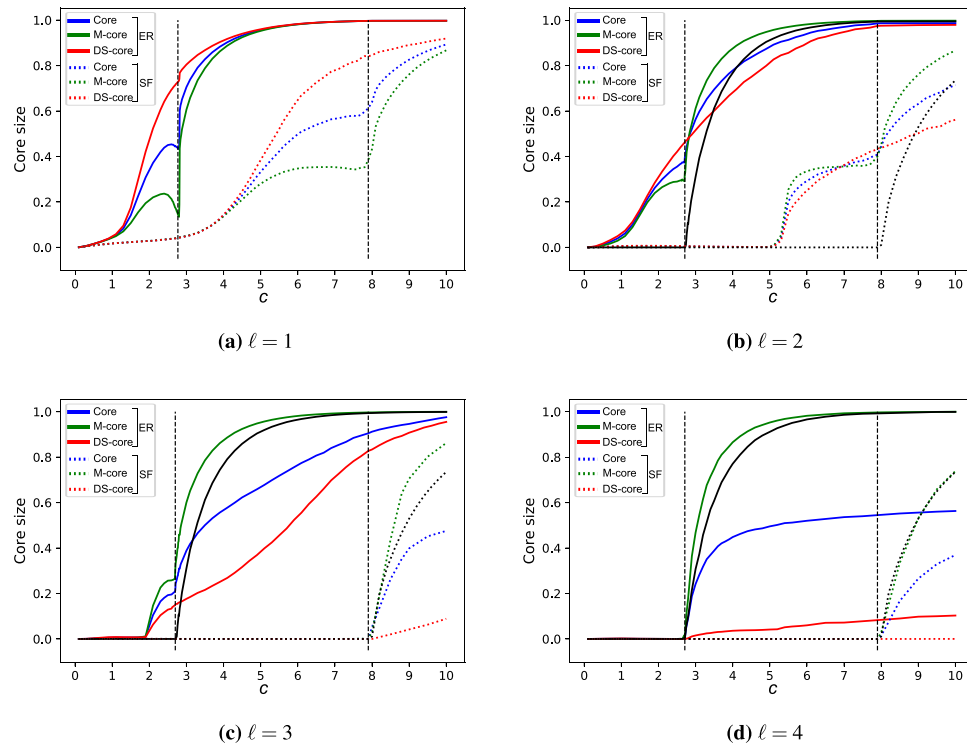
we checked. Figure 3f shows  $\delta$  as a function of the average degree  $c$  and the length of the LCC  $\ell$  for SF networks with a given degree exponent  $\gamma$ . We find that  $\delta$  is the highest for intermediate  $c$  values:  $\delta$  becomes non-zero at the point when the core forms and initially increases as the size of the core increases. However,  $\delta$  starts to drop for large  $c$ , since adding links to a network increases the maximum matching and reduces the minimum dominating set, making the LCC-constraint minimum input problem easier to approximate. Increasing the length of the LCC  $\ell$  increases the density of  $\mathcal{G}_\ell$ ; we, therefore, see a similar pattern: the error  $\delta$  is the highest for  $\ell = 2$ , and further increasing  $\ell$  makes the problem easier to approximate.

The algorithm finds the exact solution if there is no core in the network; therefore, another way to understand the performance of the approximation is to investigate how network properties affect the formation of the core. For this we numerically measure the size of the core for ER and SF random networks as a function of the average degree. For the maximum matching problem by itself the M-core emerges through a second order phase transition at a critical average degree  $c_M^*$ , for ER random networks  $c_{M,ER}^* = e$ , and in SF random networks the formation of the core is typically delayed, i.e.,  $c_{M,SF}^* > c_{M,ER}^*$ . Figure 4 shows that for  $\ell = 1$  the core is already present for any  $c > 0$ , and we only observe the phase transition for  $\ell \geq 3$  for ER, and  $\ell \geq 2$  for SF networks, further increasing  $\ell$  delays the emergence of the core, supporting our previous observation that for  $\ell > 2$  increasing  $\ell$  improves the performance of the approximation. We also observe that the emergence of the core happens at larger average degree  $c$  for SF networks than for ER network, i.e.,  $c_{SF}^*(\ell) \leq c_{ER}^*(\ell)$ , providing evidence that heterogeneous networks are easier to approximate. For details about the core formation in real networks, see Table SII in the Supplementary Information.

Overall, we find that the approximation algorithm estimates  $n_i(\ell)$  within 5% error for the networks that we tested. The performance of the algorithm is enhanced by degree heterogeneity, an ubiquitous feature of complex networks, and high  $\ell$ , and it is inhibited by the emergence of the core. We also find that our algorithm



**Figure 3.** The performance of the approximation. (a–d) We compare the approximate fraction of input nodes  $n_i^{\text{approx}}(\ell) = N_i^{\text{approx}}(\ell)/N$  to the exact solution  $n_i^{\text{exact}}(\ell) = N_i^{\text{exact}}(\ell)/N$  for model and real networks with a few hundred nodes. For networks without a core (solid markers)  $n_i^{\text{approx}}(\ell) = n_i^{\text{exact}}(\ell)$ , as expected. We observe a strong correlation between the approximate and exact solutions for networks with a non-zero core (hollow markers), indicating that the approximation works well for complex networks. The synthetic networks are SF and ER model networks with  $N = 1000$  and average degree  $c = 4$ . Each data point is an average of 100 independent networks. For details of the real networks see Table SI in the Supplementary Information. (e) The difference  $\delta = n_i^{\text{approx}}(\ell) - n_i^{\text{exact}}(\ell)$  as a function of the average degree shows that below the core percolation threshold the algorithm finds the exact solution, after the threshold  $\delta$  increases, and the approximation becomes better for very dense networks. The data points represent the average of 1000 independently generated SF networks with  $N = 300$  and  $\gamma = 3$ .



**Figure 4.** Core percolation in ER and SF networks. We measure the relative size of core in the network (blue), which is the average of the M-core (green) and the DS-core (red), for ER and SF random networks as a function of the average degree. We find that for  $\ell = 1$  the core is present for any  $c > 0$ , and the core appears at a finite critical  $c^*$  for  $\ell \geq 3$  for ER and  $\ell \geq 2$  for SF networks. For the maximum matching problem by itself the M-core emerges at average degree  $c_M^*$  (vertical dashed line), this critical point always effects the core, and numerical evidence suggests that  $c^* = c_M^*$  for  $\ell \geq 3$  for ER and  $\ell \geq 2$  for SF networks. The core size is calculated for networks with  $N = 10^6$  and the SF networks have degree exponent  $\gamma = 2.5$ . The data points are an average of 100 independent networks, and the error bars are smaller than the line width.

significantly outperforms the naïve approach of finding a maximum matching and a minimum dominating set independently and taking the union of the unmatched and the dominating nodes, the naïve algorithm finding upto three times the optimal required nodes (Supplementary Information Sect. S4). While the exact algorithm is limited to networks with a few hundred links, in the next section we apply the leaf-removal approximation to networks with up to five million links.

## Results

In the previous section we developed an algorithm to approximately solve the LCC-constrained control problem for large complex networks. Relying on this algorithm we now systematically explore how network structure affects the minimum fraction of inputs  $n_i(\ell) = N_i(\ell)/N$  and the cost of enforcing  $l_{LCC} \leq \ell$ .

**Number of input nodes.** First, we measure  $n_i(\ell)$  for Erdős-Rényi (ER) and scale-free (SF) model networks to explore the role of the degree distribution. The SF model we use is a variant of the hidden parameter model, which allows us to control both the average degree  $c$  and the degree exponent  $\gamma$  characterizing the tail of the degree distribution<sup>38</sup>. Figure 5 shows  $n_i(\ell)$  as a function of  $c$ ,  $\gamma$ , and  $\ell$ , to help interpret the results we also include the lower bound 5 and upper bound 7). We find that dense networks are easier to control than sparse networks for any  $\ell$  Fig. 5a–c, similarly to previous results not taking into account the LCC<sup>5</sup>. We also observe that the upper and lower bound converge towards each other as the average degree  $c$  increases. To explain this, note that the maximum matching rapidly grows with increasing  $c$ , hence the number of unmatched nodes rapidly decreases. For example, in ER networks, the number of unmatched nodes drops exponentially as a function of  $c$ <sup>5</sup>. The minimum dominating set also decreases with increasing  $c$ , but the decay is significantly slower. For example, for ER networks, the minimum dominating set decreases slower than  $c^{-136}$ . It means that for large  $c$ , both the lower and the upper bound approach  $n_{DS}$  (the fraction of nodes in a minimum dominating set in  $\mathcal{G}_\ell$  without considering the matching), and matching has a diminishing effect on  $n_i(\ell)$ .

The role of degree heterogeneity is also consistent with previous results: Figs. 5d–f show that for any  $\ell$  homogeneous networks require fewer inputs  $n_i(\ell)$  than heterogeneous networks with the same average degree  $c$ . We find that the upper and lower bounds converge towards each other as the degree exponent  $\gamma$  decreases, which corresponds to increasing degree heterogeneity. Both the number of hubs and low-degree nodes increase with



increasing degree heterogeneity, which in turn decreases the size of the maximum matching and increases  $n_M$  (the fraction of unmatched nodes in  $\mathcal{B}$  without considering the dominating set). The effect of  $\gamma$  on the minimum dominating set is more complex: the presence of hubs generally reduces  $n_{DS}$ , while the presence of sources  $n_s$ , i.e., nodes with zero in-degree, increases  $n_{DS}$ . The net effect is that  $n_{DS}$  converges to  $n_s$  as  $\gamma$  decreases; source nodes, however, are also always unmatched nodes, hence in the limit of very heterogeneous networks, the LCC problem is determined by the maximum matching of  $\mathcal{B}$ .

Figure 5g shows the affect of the LCC  $\ell$ : increasing  $\ell$  initially lowers  $n_i(\ell)$ , however once  $\ell$  exceeds the diameter  $D$  of the network,  $n_i(\ell)$  becomes constant. For large  $\ell$ , we again observe that the lower and the upper bounds approach each other since with increasing  $\ell$ , the accessibility graph  $\mathcal{G}_\ell$  becomes more densely connected and the minimum input problem becomes largely determined by the maximum matching problem.

So far we studied the role of the degree distribution using network models where we fix the expected degree of each node, but are otherwise random. To reveal the role of higher-order structural properties, such as degree correlations or community structure<sup>39–42</sup>, we first measure  $n_i(\ell)$  for a collection of real networks, for details see Table SI in the Supplementary Information. We then randomize the networks while preserving their degree distribution but otherwise randomly rewiring their links, thus removing any higher-order structure. Finally we measure  $n_i^{\text{rand}}(\ell)$ , the number of input signals needed to control these randomized networks. Figure 6 compares  $n_i(\ell)$  to its randomized counterpart  $n_i^{\text{rand}}(\ell)$ : If the degree sequence of a network would completely determine the number of control signals, the two values would be equal, any difference is due to the structural features removed by randomization. We observe a strong correlation between  $n_i(\ell)$  and  $n_i^{\text{rand}}(\ell)$ , indicating that degree distribution indeed plays an important role in determining the controllability of real complex networks. Some markers, however, deviate from the diagonal, food webs being the strongest outliers. These food webs are relatively small and strongly heterogeneous, and their control characteristics are often observed to deviate from what is expected based on their degree distribution alone<sup>42,43</sup>.

**Cost of the LCC constraint.** To quantify the cost of ensuring that the LCC is at most  $\ell$ , we calculate

$$C(\ell) = n_i(\ell) - n_i(\infty), \quad (9)$$

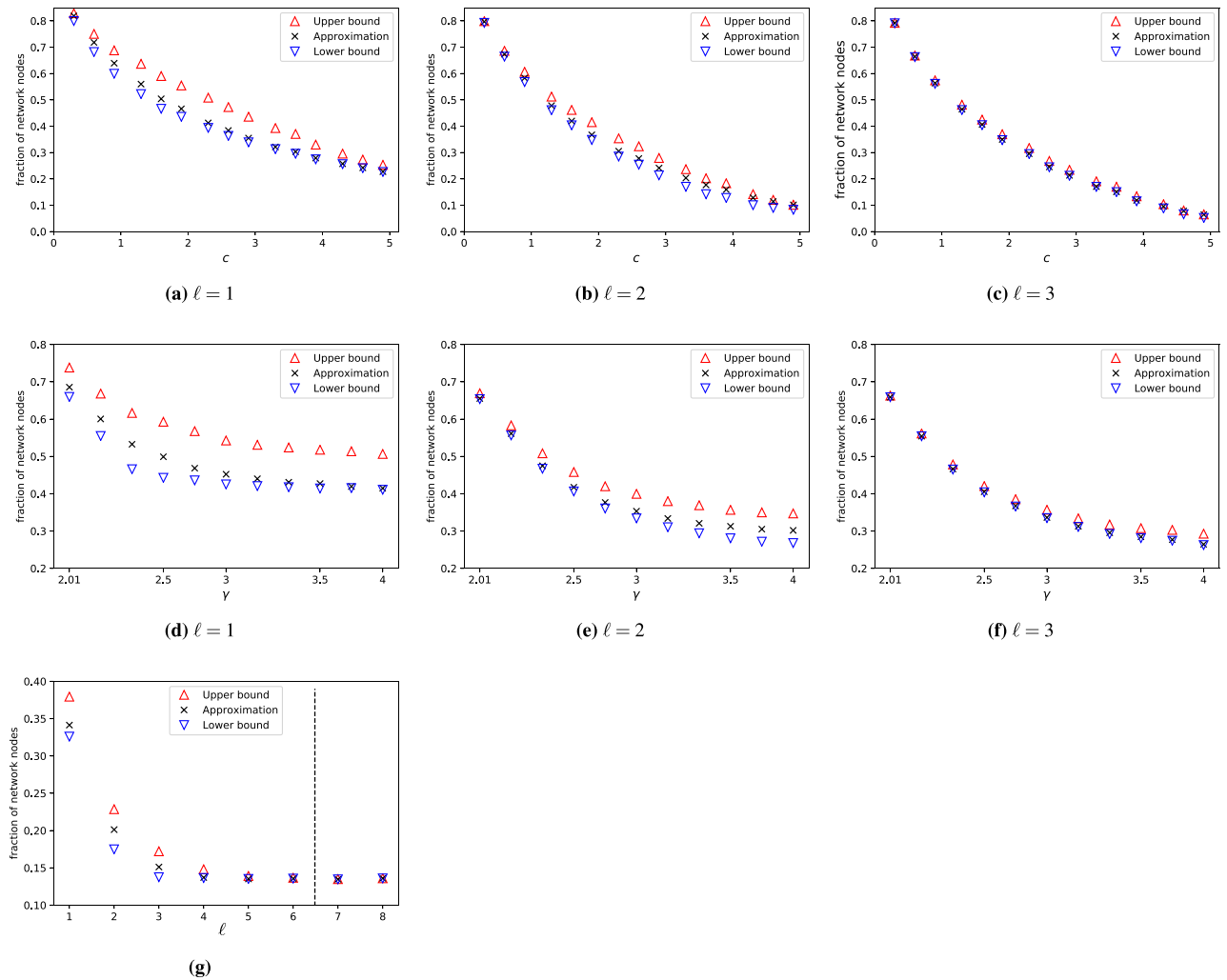
where  $n_i(\infty)$  is the fraction of inputs needed to control the network without any restrictions on the length of LCC. Note that  $n_i(\infty)$  is related but distinct from the number of driver nodes introduced by Liu et al., since in case of driver nodes the input signals are allowed to connect to multiple network nodes, while here each input is connected to exactly one node.

Figure 7 shows the cost  $C(\ell)$  for model networks as a function of degree heterogeneity and average degree. A first general observation is that if  $n_i(\infty)$  is large, i.e., the network requires a large fraction of inputs to control without restrictions on the LCC, the cost  $C(\ell)$  is low. This is natural since the LCC is the maximum distance between the set of inputs nodes and any other nodes, the more inputs there are, the shorter the expected LCC is even without explicitly tuning it. For example, increasing degree heterogeneity increases the number inputs<sup>5</sup>, correspondingly Fig. 7a the cost of ensuring  $l_{LCC} \leq 1$  for SF networks with  $\gamma = 4$  is 0.15, while for  $\gamma = 2.1$  the cost is only 0.03. The role of the average degree  $c$  is more complex: on one hand sparse networks require high  $n_i(\infty)$ , lowering the cost, on the other hand the diameter of dense networks is small making it easier to reduce the LCC. The net effect is that  $C(\ell)$  tends to zero for both  $c \rightarrow 0$  and  $c \rightarrow \infty$ , and  $C(\ell)$  has a maximum at finite  $c$ . Finally, we note that increasing  $\ell$  dramatically reduces  $C(\ell)$ . To understand this note that complex networks are small-world, meaning that the typical number of nodes at distance  $k$  from any node grows exponentially as a function of  $k$ . Hence the density of the accessibility graph  $\mathcal{G}_\ell$  also grows exponentially with  $\ell$ , greatly reducing the cost.

We also calculate  $C(\ell)$  for the collection of real networks (See Table SII in the Supplementary Information), finding that for a significant number of them we can ensure the constraint on the LCC at zero cost, particularly for  $\ell = 3$  the majority of networks do not require additional inputs. This means that reducing the LCC can be achieved solely by selecting the right input node set out of the possible minimum input node sets that ensure controllability and we do not need to increase their number. Figure 8 compares the  $C(\ell)$  of real networks to the cost  $C^{\text{rand}}(\ell)$  of their degree-preserved randomized counterparts. We find a positive correlation between  $C(\ell)$  and  $C^{\text{rand}}(\ell)$ , although not as strong as the correlation observed for the total number of inputs  $n_i(\ell)$ , indicating that degree distribution is an important factor determining  $C(\ell)$ . The strongest outliers are again the food web networks, which require close to zero cost, yet their randomized counterparts may require us to directly control more than 0.1 fraction their nodes.

## Conclusion

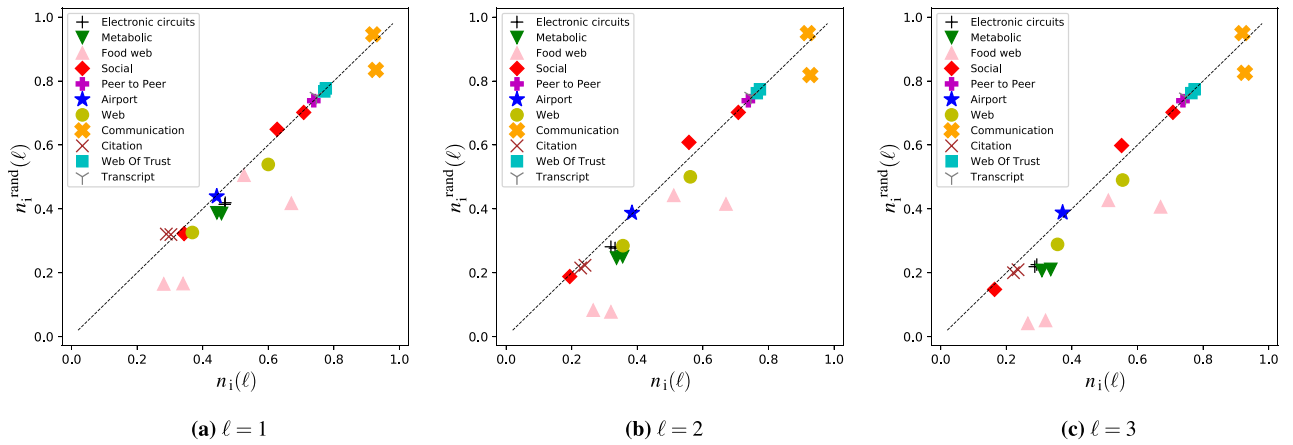
The minimum number of input nodes required to ensure controllability became a central metric to quantify the difficulty of controlling networks<sup>5,43,45</sup>. Controlling networks through a minimum set of nodes, however, often leads to unattainable control energy requirements, highlighting the need for methods of input node selection that reduce control energy. It was recently shown that reducing the length of the longest control chain provides a useful heuristic for this which only depends on the network structure. Motivated by this connection, we developed an algorithm to identify a minimum set of input nodes such that the network is controllable and the longest control chain is at most  $\ell$ . Figure 9 shows for a directed chain network and a small real network that selecting the input nodes using this algorithm indeed reduces the control energy by orders of magnitude compared to random placement of the same number of input nodes. Relying on the algorithm, we systematically investigated how features of complex networks affect the cost of control, showing, for example, that the cost of reducing the longest control chain decreases with increasing degree heterogeneity and is low for both very sparse and very dense networks. Future work may explore alternative approximations either to improve accuracy of the solution



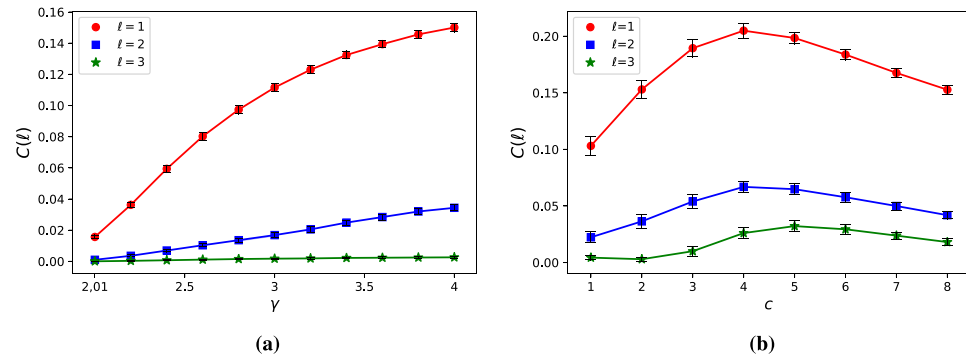
**Figure 5.** Minimum number of inputs  $n_i(\ell)$  – model networks. **(a–c)** Increasing average degree  $c$  decreases  $n_i(\ell)$  for any  $\ell > 0$ , similarly to previous results obtained without considering the LCC. The lower bound (5) and upper bound (7) converge towards each other in the dense network limit, since  $n_M$  decreases rapidly and both bounds are determined by  $n_{DS}$ . Plots show results obtained for SF networks with  $\gamma = 3$  and  $N = 10^4$ ; we observe similar behavior for other  $\gamma$  values and ER networks. **(d–f)** Measuring  $n_i(\ell)$  for SF networks as a function of the degree exponent  $\gamma$  shows that homogeneous networks are easier to control for any  $\ell > 0$ , consistent with previous results. Plots show SF networks with  $c = 2$  and  $N = 10^4$ ; we observe similar behavior for other  $c$  values. **(g)** Increasing  $\ell$ , the maximum allowed length of the LCC, decreases  $n_i(\ell)$ , since the accessibility graph  $\mathcal{G}_\ell$  becomes denser for increasing  $\ell$ ; and therefore the accessibility condition becomes easier to satisfy. This also means that the lower and upper bounds are both determined by  $n_M$ . The horizontal dashed line indicates the average shortest path in the network, above  $n_i(\ell)$  becomes independent of  $\ell$ . The plot shows ER networks with  $N = 10^3$  and  $c = 2.5$ , we observe similar behavior for other  $c$  values and SF networks. Data points in the figure indicate the average of 100 independent networks, error bars are smaller than the markers.

or to simplify the algorithm. Particularly, simplified algorithms may yield to analytical treatment, allowing us to better understand the core percolation transition observed numerically in Fig. 4.

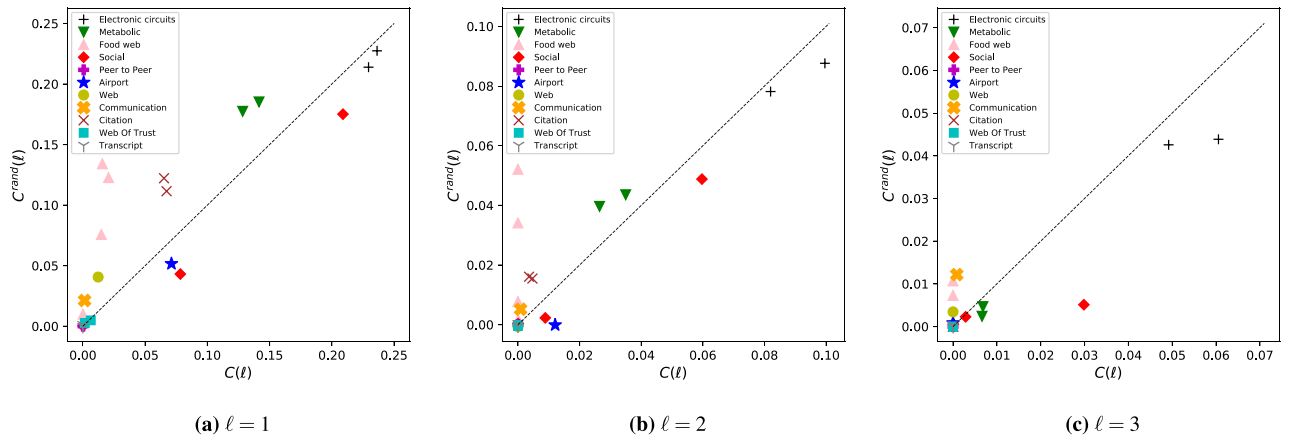
Input node selection based on structural controllability and the longest control chain necessarily has its limitations. For example, here we focused on fully controlling all degrees of freedom of simple directed networks with linear dynamics by coupling each input signal to a single node. It would be interesting to relax some of these assumptions, for example, to consider undirected networks or to couple inputs to multiple nodes. Furthermore, some complex systems are better described by generalized networks, such as hyper-graphs or multi-layer networks<sup>46–51</sup>; and often the goal of control is to influence only a subset of nodes<sup>52–54</sup> or to stabilize a non-linear system in a steady state<sup>2,3,55,56</sup>. Future work may, therefore, aim to extend our method to generalized networks, or to understand how to leverage structural network features to reduce control energy beyond the assumptions of structural controllability.



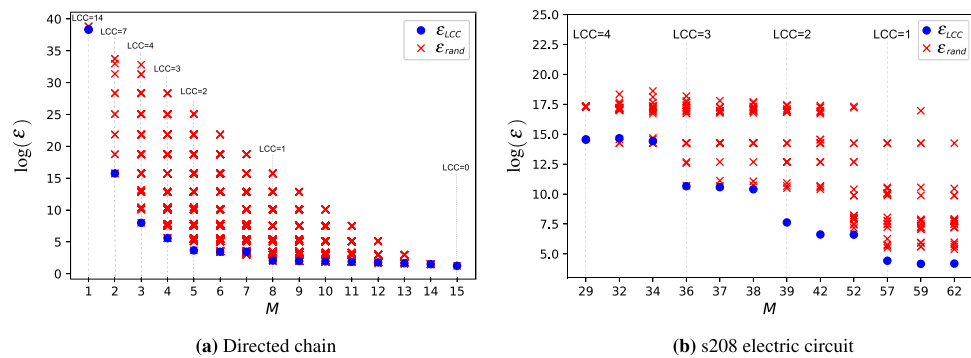
**Figure 6.** Minimum number of inputs  $n_i(\ell)$  – real networks. We measure  $n_i(\ell) = N_i(\ell)/N$  for a collection of real networks, randomize the networks preserving only their degree distribution and again measure the number of inputs needed for control  $n_i^{\text{rand}}(\ell)$ . We find a strong correlation between  $n_i(\ell)$  and  $n_i^{\text{rand}}(\ell)$ , indicating that degree distribution is an important property determining  $n_i(\ell)$ , and higher-order network properties, such as degree correlations and community structure, play a limited role. We also observe that the correlation between  $n_i(\ell)$  and  $n_i^{\text{rand}}(\ell)$  becomes slightly weaker for increasing  $\ell$ . Data markers indicate the average of twenty randomized networks, and for each randomization  $(|\mathcal{E}|/2) \times \ln(1/\epsilon)$  (where  $\epsilon$  is between  $10^{-6}$  and  $10^{-7}$ ) number of link rewiring trials were performed<sup>44</sup>. Error bars are smaller than the markers.



**Figure 7.** Cost of reducing the LCC – model networks. **(a)** High degree heterogeneity, corresponding to low  $\gamma$ , reduces the diameter and increases  $n_i(\infty)$ , both reducing the cost. The result shows that by decreasing the heterogeneity, i.e., increasing  $\gamma$  in SF networks, the cost of ensuring the LCC constraint increases. **(b)** For average degree  $c \rightarrow 0$  the network becomes a set of isolated nodes, and as consequence  $n_i(\infty) \rightarrow 1$  and  $C(\ell) \rightarrow 0$ . On the other hand, for  $c \rightarrow \infty$  the network becomes fully connected and  $l_{\text{LCC}} \rightarrow 1$  even if there is only a single input, hence  $C(\ell) \rightarrow 0$ . The figure shows SF networks with  $N = 10^4$  with  $c = 2$  in panel **(a)** and  $\gamma = 3$  in panel **(b)**. We observe similar behavior for other parameter values. Data points are an average of one hundred independent realizations and errorbars indicate the standard error of the mean.



**Figure 8.** Cost of reducing the LCC – real networks..We find a positive correlation between  $C(\ell)$  and  $C^{\text{rand}}(\ell)$ , indicating that the degree distribution of real networks plays an important role in determining the cost of reducing the LCC. We also find that for increasing  $\ell$ , an increasing number of real networks require zero cost, meaning that we can ensure  $l_{\text{LCC}} \leq \ell$  by selecting the right input node set, but without increasing the number of inputs. Data markers indicate the average of twenty randomized networks, and for each randomization  $(|\mathcal{E}|/2) \times \ln(1/\epsilon)$  (where  $\epsilon$  is between  $10^{-6}$  and  $10^{-7}$ ) number of link rewiring trials were performed<sup>44</sup>. Error bars are smaller than the markers.



**Figure 9.** Reduction of control energy. We compare the energy requirement of control using an input node selection that constrains the longest control chain (LCC) and a random input selection strategy with the same number of inputs. Specifically, for a given number of inputs  $M$ , we first use the algorithm introduced in Sect. 4 to place  $N_i(\ell)$  inputs, where  $\ell$  is selected such that  $N_i(\ell) \leq M < N_i(\ell + 1)$ , and we select the remaining  $M - N_i(\ell)$  inputs randomly. For random input selection, we select a minimum input node set that ensures controllability based on maximum matching and randomly select the remaining inputs. The energy requirement of control is quantified by the mean energy  $\langle \epsilon \rangle = \text{tr}(\mathbf{W}_B^{-1})$  with control time  $t_f = 1$ . We calculated the control energy for (a) a directed chain with 15 nodes and (b) a real network s208 representing an electric circuit. Vertical dashed lines indicate  $M$  values where  $\ell$  changes. The energy requirement of LCC constrained input selection (blue markers) can reduce the control energy by orders of magnitude compared to random selection (red markers).

### Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 8 December 2022; Accepted: 1 March 2023

Published online: 07 March 2023

### References

1. Liu, Y.-Y. & Barabási, A.-L. Control principles of complex systems. *Rev. Mod. Phys.* **88**, 035006 (2016).
2. Wang, X. F. & Chen, G. Pinning control of scale-free dynamical networks. *Physica A* **310**, 521–531 (2002).
3. Informative and determining nodes in regulatory networks. Fiedler, B., Mochizuki, A., Kurosawa, G. & Saito, D. Dynamics and control at feedback vertex sets. *J. Dyn. Differ. Equations* **25**, 563–604 (2013).
4. Lin, C.-T. Structural controllability. *IEEE Trans. Autom. Control* **19**, 201–208 (1974).
5. Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. *Nature* **473**, 167–173 (2011).

6. Pang, S.-P., Wang, W.-X., Hao, F. & Lai, Y.-C. Universal framework for edge controllability of complex networks. *Sci. Rep.* **7**, 1–12 (2017).
7. Tzoumas, V., Rahimian, M. A., Pappas, G. J. & Jadbabaie, A. Minimal actuator placement with bounds on control effort. *IEEE Trans. Control Netw. Syst.* **3**, 67–78 (2015).
8. Li, G. *et al.* Minimum-cost control of complex networks. *New J. Phys.* **18**, 013012 (2015).
9. Deng, L., Li, G., Pei, J. & Huang, J. L0 norm constraint based external control source allocation for the minimum cost control of directed networks. *ISA Trans.* **76**, 88–96 (2018).
10. Klickstein, I. & Sorrentino, F. Selecting energy efficient inputs using graph structure. *Int. J. Control* <https://doi.org/10.1080/00207179.2021.2022218> (2022).
11. Chen, Y.-Z., Wang, L.-Z., Wang, W.-X. & Lai, Y.-C. Energy scaling and reduction in controlling complex networks. *R. Soc. Open Sci.* **3**, 160064 (2016).
12. Klickstein, I. S. & Sorrentino, F. Control distance and energy scaling of complex networks. *IEEE Trans. Netw. Sci. Eng.* **7**, 726–736 (2018).
13. Olshevsky, A. Minimal controllability problems. *IEEE Trans. Control Netw. Syst.* **1**, 249–258 (2014).
14. Rugh, W. J. & Rugh, W. J. *Linear system theory*, vol. 2 (prentice hall Upper Saddle River, NJ, 1996).
15. Yan, G. *et al.* Spectrum of controlling and observing complex networks. *Nat. Phys.* **11**, 779–786 (2015).
16. Pasqualetti, F., Zampieri, S. & Bullo, F. Controllability metrics, limitations and algorithms for complex networks. *IEEE Trans. Control Netw. Syst.* **1**, 40–52 (2014).
17. Yan, G., Ren, J., Lai, Y.-C., Lai, C.-H. & Li, B. Controlling complex networks: How much energy is needed?. *Phys. Rev. Lett.* **108**, 218703 (2012).
18. Kim, J. Z. *et al.* Role of graph architecture in controlling dynamical networks with applications to neural systems. *Nat. Phys.* **14**, 91–98 (2018).
19. Lindmark, G. & Altafini, C. Minimum energy control for complex networks. *Sci. Rep.* **8**, 1–14 (2018).
20. Sun, J. & Motter, A. E. Controllability transition and nonlocality in network control. *Phys. Rev. Lett.* **110**, 208701 (2013).
21. Deng, L., Li, G., Pei, J. & Huang, J. L0 norm constraint based external control source allocation for the minimum cost control of directed networks. *ISA Trans.* **76**, 88–96 (2018).
22. Ding, J., Wen, C. & Li, G. Key node selection in minimum-cost control of complex networks. *Physica A* **486**, 251–261 (2017).
23. Chen, Y.-Z., Wang, L., Wang, W. & Lai, Y.-C. The paradox of controlling complex networks: control inputs versus energy requirement. *arXiv preprint arXiv:1509.03196* (2015).
24. Klickstein, I., Kafle, I., Bartaula, S. & Sorrentino, F. Energy scaling with control distance in complex networks. In *2018 IEEE International symposium on circuits and systems (ISCAS)*, 1–5 (IEEE, 2018).
25. Hopcroft, J. E. & Karp, R. M. An  $n^2/2$  algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput.* **2**, 225–231 (1973).
26. Karp, R. M. & Sipser, M. Maximum matching in sparse random graphs. In *22nd Annual symposium on foundations of computer science (sfcs 1981)*, 364–375 (IEEE, 1981).
27. Zhao, J.-H. & Zhou, H.-J. Controllability and maximum matchings of complex networks. *Phys. Rev. E* **99**, 012317 (2019).
28. Liu, Y.-Y., Csóka, E., Zhou, H. & Pósfai, M. Core percolation on complex networks. *Phys. Rev. Lett.* **109**, 205703 (2012).
29. Bauer, M. & Golinelli, O. Core percolation in random graphs: A critical phenomena analysis. *Eur. Phys. J. B-Condens. Matter Complex Syst.* **24**, 339–352 (2001).
30. Pang, C., Zhang, R., Zhang, Q. & Wang, J. Dominating sets in directed graphs. *Inf. Sci.* **180**, 3647–3652 (2010).
31. Molnár, F., Sreenivasan, S., Szymanski, B. K. & Korniss, G. Minimum dominating sets in scale-free network ensembles. *Sci. Rep.* **3**, 1–10 (2013).
32. Zhao, J.-H., Habibulla, Y. & Zhou, H.-J. Statistical mechanics of the minimum dominating set problem. *J. Stat. Phys.* **159**, 1154–1174 (2015).
33. Simonetti, L., Salles da Cunha, A. & Lucena, A. The minimum connected dominating set problem: Formulation, valid inequalities and a branch-and-cut algorithm. In *International conference on network optimization*, 162–169 (Springer, 2011).
34. Habibulla, Y. Minimal dominating set problem studied by simulated annealing and cavity method: Analytics and population dynamics. *J. Stat. Mech: Theory Exp.* **2017**, 103402 (2017).
35. Alipour, S., Futuhi, E. & Karimi, S. On distributed algorithms for minimum dominating set problem, from theory to application. *arXiv preprint arXiv:2012.04883* (2020).
36. Habibulla, Y., Zhao, J.-H. & Zhou, H.-J. The directed dominating set problem: Generalized leaf removal and belief propagation. In *International workshop on frontiers in algorithmics*, 78–88 (Springer, 2015).
37. LCC-constrained minimum input problem, Python implementation. <https://github.com/Samie-Alizadeh/LCC-constrained-minimum-input-problem> (2022).
38. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **80**, 1275 (2008).
39. Xiang, J., Hu, K., Zhang, Y., Hu, T. & Li, J.-M. Analysis and perturbation of degree correlation in complex networks. *EPL (Europhys. Lett.)* **111**, 48003 (2015).
40. Mayo, M., Abdelzaher, A. & Ghosh, P. Long-range degree correlations in complex networks. *Comput. Soc. Netw.* **2**, 1–13 (2015).
41. Newman, M. E. Communities, modules and large-scale structure in networks. *Nat. Phys.* **8**, 25–31 (2012).
42. Pósfai, M., Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Effect of correlations on network controllability. *Sci. Rep.* **3**, 1–7 (2013).
43. Ghasemi, A., Pásfai, M. & D'Souza, R. Diversity of structural controllability of complex networks with given degree sequence. *IEEE Trans. Netw. Sci. Eng.* **7**, 2667–2679 (2020).
44. Ray, J., Pinar, A. & Seshadhri, C. Are we there yet? when to stop a Markov chain while generating random graphs. In *International workshop on algorithms and models for the web-graph*, 153–164 (Springer, 2012).
45. Ramos, G., Pequito, S., Kar, S., Aguiar, A. P. & Ramos, J. On the np-completeness of the minimal controllability problem. *arXiv preprint arXiv:1401.4209* (2014).
46. Pósfai, M., Gao, J., Cornelius, S. P., Barabási, A.-L. & D'Souza, R. M. Controllability of multiplex, multi-time-scale networks. *Phys. Rev. E* **94**, 032316 (2016).
47. Kivela, M. *et al.* Multilayer networks. *J. Complex Netw.* **2**, 203–271 (2014).
48. Boccaletti, S. *et al.* The structure and dynamics of multilayer networks. *Phys. Rep.* **544**, 1–122 (2014).
49. Jiang, L., Tang, L. & Lü, J. Controllability of multilayer networks. *Asian J. Control* **24**, 1517–1527 (2022).
50. Menichetti, G., Dall'Asta, L. & Bianconi, G. *Control of multilayer networks*. *Sci. Rep.* **6**, 1–8 (2016).
51. Wang, D. & Zou, X. Control energy and controllability of multilayer networks. *Adv. Complex Syst.* **20**, 1750008 (2017).
52. Gao, J., Liu, Y.-Y., D'Souza, R. M. & Barabási, A.-L. Target control of complex networks. *Nat. Commun.* **5**, 1–8 (2014).
53. Klickstein, I., Shirin, A. & Sorrentino, F. Energy scaling of targeted optimal control of complex networks. *Nat. Commun.* **8**, 1–10 (2017).
54. Guo, W.-F. *et al.* Constrained target controllability of complex networks. *J. Stat. Mech: Theory Exp.* **2017**, 063402 (2017).
55. Wang, B., Liu, H., Xu, J. & Liu, J. Pining control algorithm for complex networks. In *2019 Chinese control conference (CCC)*, 964–969 (IEEE, 2019).
56. Lü, J., Chen, G., Ogorzalek, M. J. & Trajković, L. Theory and applications of complex networks: Advances and challenges. In *2013 IEEE International symposium on circuits and systems (ISCAS)*, 2291–2294 (IEEE, 2013).



### Acknowledgements

M.P. acknowledges funding from ERC grant No. 810115-DYNASET. A.G. gratefully acknowledges funding from the Alexander von Humboldt Foundation (Ref. 3.4 - IRN -1214645 - GF-E) for his research fellowship at the University of Passau, Germany.

### Author contributions

A.G., S.A. conceived the project. S.A., M.P., A.G. designed the research. S.A. performed simulations and analyzed the empirical data. S.A., M.P., A.G. wrote the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-30810-w>.

**Correspondence** and requests for materials should be addressed to A.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023