Finding the right scale of a network: Efficient identification of causal emergence through spectral clustering

Ross Griebenow[1], Brennan Klein[2,3], and Erik Hoel[*4]

[1]*Drexel University, Philadelphia, PA, USA*
[2]*Network Science Institute, Northeastern University, Boston, MA, USA*
[3]*Laboratory for the Modeling of Biological and Socio-Technical Systems, Northeastern University, Boston, MA, USA*
[4]*Allen Discovery Center, Tufts University, Medford, MA, USA*

August 22, 2019

Abstract

All networks can be analyzed at multiple scales. A higher scale of a network is made up of macro-nodes: subgraphs that have been grouped into individual nodes. Recasting a network at higher scales can have useful effects, such as decreasing the uncertainty in the movement of random walkers across the network while also decreasing the size of the network. However, the task of finding such a macroscale representation is computationally difficult, as the set of all possible scales of a network grows exponentially with the number of nodes. Here we compare various methods for finding the most informative scale of a network, discovering that an approach based on spectral analysis outperforms greedy and gradient descent-based methods. We then use this procedure to show how several structural properties of preferential attachment networks vary across scales. We describe how meso- and macroscale representations of networks can have significant benefits over their underlying microscale, which include properties such as increase in determinism, a decrease in degeneracy, a lower entropy rate of random walkers on the network, an increase in global network efficiency, and higher values for a variety of centrality measures than the microscale.

1 Introduction

Networks can be used to represent a wide range of systems, and over the past decade their use has become more common throughout the sciences [1]. While network analysis is typically performed on the full, microscale representation of a network, recent research has shown that informative higher scales of networks can be identified and explicitly modeled [2]. Using these techniques, a network, G, can be recast into a new network, G_M, wherein subgraphs of the networks are grouped into individual *macro-nodes*. These macro-nodes summarize the behavior of the subgraph in a manner that recapitulates the dynamics of the original networks. Thus, higher scales act like accurate or approximate models of the original system.

There has long been the assumption in science that, whenever possible, coarse-grained models should be replaced by fine-grained models [3]. Due to the general success of this reductionist method, there has been little attention to the gains that accompany macroscale models. This is exacerbated by the lack of formal methods for dealing with systems across scales, as well as the computational cost of detecting an informative higher scale—the number of possible dimension reductions exponentially increases with the size of the system. So while explicitly modeling higher scales has been investigated in logic gates [4, 5] and Markov processes [6], these investigations have been limited by computational resources.

---

[*]erik.hoel@tufts.edu

Here we compare and evaluate different methods for finding informative higher scales of networks by searching the computationally challenging space of possible scales. We compare a greedy algorithm, an approach based on gradient descent, and an approach based on the spectral decomposition of a network. We use these algorithms to show that mesoscale structures are, in general, the most computationally difficult scales to identify in a network. However, this issue is mostly avoided by the adapted spectral approach introduced here, which has the strongest performance of the algorithms, and can find informative and complex higher scales even in large networks. This spectral analysis technique groups micro-nodes into macro-nodes based on a clustering and ordering approach commonly used in machine learning tasks [7]. We also explore how higher scales can possess more informative connectivity (a phenomenon called *causal emergence* [4]). Compared to their underlying microscale, the macroscales that are causally emergent are found to, among other properties, decrease the uncertainty of random walkers, increase the global efficiency of the network, and increase various centrality measures of nodes.

2 Methods

2.1 Causal emergence in networks

Every node, v_i, in a network is associated with an out-weight vector, W_i^{out}, which represents the possible outputs from v_i. W_i^{out} consists of weights, w_{ij}, between node v_i and its neighbors, v_j, such that if $w_{ij} = 0.0$ there is no edge from v_i to v_j. For these networks the W_i^{out} of each node sums to 1.0. Therefore, the edge weights w_{ij} are equivalent to the probability p_{ij} that a random walker on v_i will transition to v_j in the next time step.

In order to find the maximally informative scale of the network, as in [2], we make use of the *effective information* (EI), which is a network measure that quantifies the degree of certainty in the transitions of random walkers between nodes in a network, as well as how that certainty is distributed. Therefore, EI is based on two uncertainties. The first is the Shannon entropy of the average out-weight vector in the network, $\langle W_i^{out} \rangle$, which captures how distributed out-weights of the network are. The second is the average entropy of *each* node's W_i^{out}. Their difference is the EI of a network:

$$EI = H(\langle W_i^{out} \rangle) - \langle H(W_i^{out}) \rangle \tag{1}$$

When a network has a macroscale with greater EI than its underlying microscale, this is known as *causal emergence*. Causal emergence is when some recast network, G_M (the macroscale), is associated with a gain in EI relative to the original network, G (the microscale). The amount of causal emergence is the difference between the EI of the microscale and that of the macroscale. Ideally, one would find the macroscale mapping that maximizes the EI of the network, EI^{max}, and use the resulting macroscale network to model the system in question. This EI^{max} approximates the channel capacity of the system [6].

To measure causal emergence, one needs to define a higher scale (some dimension reduction of the original system). In networks, macroscales are networks that are comprised of *macro-nodes*. Each macro-node is a subgraph of micro-nodes that are grouped together and replaced with a single node, μ. A macro-node μ has some W_μ^{out} that replaces the corresponding microscale outputs of nodes in the subgraph.

Note that macroscales should generally be accurate, in that they will produce identical or approximately identical dynamics to those of the underlying microscale. In networks this can be assessed by comparing the dynamics of random walkers at each scale to see the extent to which the macroscale recapitulates the dynamics of the microscale. Different types of macro-nodes (constructed in different ways to summarize a subgraph's behavior) are sometimes needed to maintain accuracy, such as using higher-order properties [8, 2]. Here, all macro-nodes are of the $\mu|\pi$ type (based on the stationary distribution) since it has been shown that such nodes are in general highly accurate [2]. The $W_{\mu|\pi}^{out}$ is a weighted average of each node in the subgraph's W^{out}, weighted by the stationary distribution, π, of the micro-nodes in the subgraph that constitute the new macronode, μ.

Finding macro-nodes that produce a gain in EI can be thought of as an iterative procedure, wherein a subgraph is grouped in a macro-node and then the resulting change in EI is calculated; this is followed by

testing a new grouping of a subgraph, comparing its EI to the original, and so on. Below, we compare and detail algorithmic variants on this approach.

2.2 Algorithmic approaches to identifying causal emergence

2.2.1 Greedy algorithm

The greedy algorithm, which was first introduced in [2], is structured as follows: for each node, v_i, a list of neighboring nodes is constructed, $\{v_j\} \in B_i$, where B_i is the *Markov blanket* of v_i. The Markov blanket, B_i, of a node, v_i, is composed of the "parents", the "children", and the "parents of the children" of v_i. Therefore $\{v_j\} \in B_i$ includes only those nodes targeting v_i, and also nodes targeted by v_i, and then also the nodes that target those nodes targeted by v_i [9]. The algorithm assesses the change in EI after a node v_i, and another node, $v_j \in B_i$, are combined into a macro-node, v_μ. If this leads to a gain in EI, the algorithm stores this change. If necessary, it will change the queue of nodes, $\{v_j\}$, with any new neighboring nodes from v_j's Markov blanket that were not already in $\{v_j\}$, so as to expand the search. If a node, v_j, has already been combined into a macro-node via a grouping with a previous node, v_i, then it will not be included in new queues, $\{v_j\}$, of later nodes to check. Each pair of nodes is iteratively checked by the algorithm, starting with some node v_j, and pairing it with every node in it's Markov blanket, $\{v_j\} \in B_i$, and then starting on a new node, until every node is tested.

Given a network with n nodes, checking a single pair of nodes for causal emergence requires computing a macroscale network, which has $O(n^2)$ time complexity, and then computing the EI of the candidate macro network, G_M, which is also $O(n^2)$. In the worst case, $\binom{n}{2}$ pairs of nodes need to be checked, though in practice interesting networks typically require far fewer checks, so the overall runtime of this algorithm is $O(n^4)$. To use this greedy algorithm we made use of the publicly-available Python package at github.com/jkbren/einet.

2.2.2 Spectral analysis

Historically, spectral methods have been successful in obtaining partitions of graphs with desirable properties and good theoretical guarantees [10, 11, 12]. We use a novel variation of classical spectral algorithms to identify causal emergence accurately and efficiently. Given the transition probability matrix W_{out} of a network, our spectral algorithm calculates the eigendecomposition $\Lambda = \{\lambda_i\}$, $E = \{e_i\}$ of W_{out}, where λ_i is the ith eigenvalue of W_{out} and e_i is the corresponding eigenvector. We obtain a basis E' for the span of W_{out} by removing the kernel and weighting the vectors by their associated eigenvalues: $E' = \{\lambda_i(e_i) \mid \lambda_i \neq 0\}$. Intuitively, disregarding the kernel in our analysis makes sense because it corresponds to degeneracy in the corresponding network. Therefore, considering the span gives us a description of the network topology without the components that generate degeneracy. Additionally, the nonzero eigenvalues and corresponding eigenvectors of W_{out} contain rich information about the topological structure of the network.

We use E' to associate each node v_j in a network with a vector composed of the entry in each eigenvector corresponding to v_j. We calculate a distance metric for all pairs of nodes in the network by taking the cosine similarity of these vectors. If a pair of nodes are not in each other's Markov blankets, then grouping them together cannot increase EI, so we define the distance between them to be ∞. We then apply the OPTICS clustering algorithm [7] to this distance matrix to obtain a clustering over the nodes of the network, the output of which we interpret as a mapping from the microscale to the new macroscale. The quality of this coarse-graining (i.e., the amount of causal emergence it discovers) depends on the distance threshold ϵ used in the clustering, and the optimal value for ϵ depends on the topology of the network and is difficult to select *a priori*. Therefore we check the EI gain over a range of ϵ values to find the best clustering, which can be done efficiently by the OPTICS algorithm.

Regarding algorithmic complexity, given a network with n nodes, eigendecomposition can be performed in $O(n^3)$, computing the OPTICS reachability graph is $O(n \log n)$, and computing the clustering for a given ϵ is $O(n)$. Only a constant number of different ϵ's are considered, so the overall complexity is $O(n^3)$, though

in practice the runtime is dominated by computing the OPTICS reachability graph for networks at least up to 5000 nodes. Additionally, both the eigendecomposition and clustering operations are parrallelizable.

We speculate that there is a deep connection between the kernel of the adjacency matrix and the scale of the corresponding network, and this is why analyzing the adjacency matrix works for finding causal emergence. For example, Erdős–Rényi random graphs are almost always optimal at the microscale and exhibit no causal emergence, and almost always have kernel dimension zero [13]. Likewise, star graphs are optimal at the macroscale and exhibit maximal causal emergence, and a star graph of size n has kernel dimension $n - 2$ [14]. Comparatively, both star graphs and typical Erdős–Rényi graphs have Laplacian kernel dimension zero.

The presence of macroscales in a given network is determined by the amount of indeterminism and degeneracy present in the network. Using tools from linear algebra, here we derive a connection between the basic properties of adjacency matrices and the EI of networks that motivates the use of spectral analysis for determining scale.

Given of a graph G of size n, the *adjacency matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$ of G is the matrix where the entries a_{ij} are equal to the weight of the edge from node i to node j in G if this edge exists, and zero otherwise (for review of the following relevant topics in linear algebra, see [15]). The *normalized adjacency matrix* is given by dividing the columns of A by the degrees of the corresponding nodes, and can be thought of as the transition probability matrix of a random walk on G. The *kernel* of a matrix $\mathbf{M} \in \mathbb{C}^{n \times n}$ is the set of vectors $\{\mathbf{v} \in \mathbb{C}^n : \mathbf{M}\mathbf{v} = \mathbf{0}\}$, where $\mathbf{0}$ is the vector of all zeros. The kernel of $\mathbf{M}$ forms a linear subspace of $\mathbb{C}^n$.

Degeneracy is an indication of attractor dynamics in a system—it measures the number of states which converge onto the same state in the future, a phenomena which we claim can also be quantified using the algebraic properties of adjacency matrices. Given a network G with n nodes, let $\mathbf{x_t}$ be a vector representing a distribution of random walkers on a graph G: $\mathbf{x_t}[i] = \Pr(\textit{walk is on node i at time t})$; note that $\mathbf{x_t}$ could represent a deterministic state if all components except for one are zero. Let $\mathbf{A}$ be the degree-normalized adjacency matrix of G, then the distribution at the next time step is given by $\mathbf{x_{t+1}} = \mathbf{A}\mathbf{x_t}$. In order to quantify degenerate behavior, we can use the kernel of $\mathbf{A}$ to construct distributions $\mathbf{w}$ that differ from $\mathbf{x_t}$ and also transition to $\mathbf{x_{t+1}}$, but care is required to ensure that $\mathbf{w}$ is a valid probability distribution. As such, let $\mathbf{v}$ be any real vector in the kernel of $\mathbf{A}$ where $\mathbf{v}[i]$ is non-negative if $\mathbf{x_t}[i]$ is zero, and let $\beta = \max_{i \leq n} \frac{\mathbf{x_t}[i]}{\mathbf{v}[i] - \mathbf{x_t}[i]}$. Then for any $b \in [0, \beta]$ we can construct $\mathbf{w} = \mathbf{x_t} + b\mathbf{v}$ satisfying $\mathbf{A}\mathbf{w} = \mathbf{A}(\mathbf{x_t} + b\mathbf{v}) = \mathbf{A}\mathbf{x_t} + b\mathbf{A}\mathbf{v} = \mathbf{A}\mathbf{x_{t+1}} + \mathbf{0} = \mathbf{A}\mathbf{x_{t+1}}$. Scaling $\mathbf{v}$ by b ensures that the entries of $\mathbf{w}$ are non-negative. By definition, $\mathbf{x_t}$ is a valid probability distribution, so its components sum to 1. Multiplication of vectors by Markov matrices preserves the sum of their components, and since $\mathbf{A}\mathbf{v} = \mathbf{0}$, the components of $\mathbf{v}$ must also add up to 0. Thus we have $\sum_{i=1}^{n} \mathbf{w}[i] = \sum_{i=1}^{n} \mathbf{x_t}[i] + b\sum_{i=1}^{n} \mathbf{v}[i] = 1 + 0$, and we can conclude that $\mathbf{w}$ is a valid probability distribution. If we additionally assume that all components of $\mathbf{x_t}$ are nonzero, the set of such degenerate distributions $\mathbf{w}$ forms a convex region of a linear subspace of the same dimension as the kernel of $\mathbf{A}$. Conversely, if we have two distributions $\mathbf{x_t^1}$ and $\mathbf{x_t^2}$ that both transition to $\mathbf{x_{t+1}}$, then the vector $\mathbf{x_t^1} - \mathbf{x_t^2}$ is in the kernel of $\mathbf{A}$: $\mathbf{A}(\mathbf{x_t^1} - \mathbf{x_t^2}) = \mathbf{A}\mathbf{x_t^1} - \mathbf{A}\mathbf{x_t^2} = \mathbf{x_{t+1}} - \mathbf{x_{t+1}} = \mathbf{0}$.

This tells us that degenerate dynamics in a network correspond directly with the kernel dimension of the adjacency matrix. For any arrangement of degenerate nodes in G, there is a corresponding element in the kernel of $\mathbf{A}$, and for every element in the kernel of $\mathbf{A}$, there is a set of degenerate nodes. Therefore, we conclude that there is a strong relationship between the connectivity and algebraic structures of graphs, and the effectiveness of spectral analysis for determining scale is a result of this connection.

2.2.3   Gradient descent

Gradient descent is a powerful approach for solving a wide range of optimization problems, and it is a ubiquitous approach in machine learning [16, 17]. However, it is not immediately applicable to the problem of finding good macroscale networks, because we need to optimize EI as a function of a set partition (coarse-graining), but this function is not differentiable, which is a requirement for performing gradient descent. Given a network with n nodes, we relax the problem by replacing the set partition with a matrix $M \in \mathbb{R}^{n \times n}$ with entries $m_{i\mu} = \Pr(v_i \in v_\mu)$ for micro-node, v_i, and macro-node, v_μ. For the purposes

of optimization, this matrix is represented as unconstrained real numbers, to which the softmax function (which can be interpreted as a differentiable approximation of the argmax function) is applied to obtain normalized probability distributions. Using this relaxation, the EI of the "probabilistic" coarse graining becomes a differentiable function of M and the network adjacency matrix, allowing M to be optimized using gradient descent with momentum. M is initialized randomly, and updated until convergence, or up until a certain number of iterations. When it converges, the result is a "deterministic" coarse graining which can be interpreted in the same way as the outputs of the previous two algorithms. A disadvantage of this approach is that the convergence of M depends on the random initialization, so multiple runs on the same network may produce different results. Performance also depends on the learning rate used in gradient descent, and the maximum number of iterations allowed.

The time complexity of a single iteration of the gradient descent algorithm is dominated by a constant number of matrix multiplications, where the matrix sizes correspond to the size of the network being coarse-grained, so a single iteration can be done naively in $O(n^3)$. Since at most only a constant number of iterations are performed (while in practice this constant is large), the overall time complexity of this approach is also $O(n^3)$. While the performance of this algorithm is asymptotically equivalent to that of the spectral approach, in practice the spectral algorithm is much faster on all instances of reasonable size.

3 Results

3.1 Comparing methods of finding macroscales

What is the best way to find the scale at which EI is high, along with the associated changes to network properties? A brute-force search is impossible due to the number of subgraphs (the same as the number of partitions). However, the challenge faced is no more difficult than the challenge of finding communities of nodes, which is common in network science, even though it too is made difficult by the number of partitions. Here we compare three methods for finding EI^{max}: a greedy algorithm (described in Section 2.2.1), a method based on gradient descent (Section 2.2.3), and a method based on spectral analysis (Section 2.2.2).

First, we analyzed the computational runtime of these different algorithms for networks of 150 nodes at varying degrees of preferential attachment (Fig. 1A). What we noticed is that extreme microscales and extreme macroscales do not require much computational resources to discover. These can be viewed as cases where reduction or emergence are very clear based on the system architecture. However, networks with significant mesoscales ($1.0 < \alpha < 2.0$) require significant runtime. Notably, this is less so for the spectral analysis, and also the runtime is several orders of magnitude lower in all conditions. This is true even when the node number of the network is significantly increased (Fig. 1B).

Next we examined whether the algorithms successfully captured causal emergence and found informative higher scales (Fig. 1C). While all algorithms could identify cases of causal emergence, both the spectral and the greedy seemed to perform better and find equivalent cases of causal emergence. This pattern continued even when the number of nodes was increased (Fig. 1D).

3.2 Network properties of macroscales

To explore how to find informative higher scales of networks, we repeatedly simulate networks grown under preferential attachment rules [18]. In a preferential attachment growth model, the network grows each time step by adding a new node with its m new edges to the network. Each new edge connects to nodes already in the network, v_j, with a probability proportional to k_j^α. k_j is the degree of node v_j and α tunes the amount of preferential attachment ($\alpha = 1.0$ corresponds to linear preferential attachment and produces the network is "scale free" [18]).

Using these networks, we review some of the changes that occur to commonly-studied network properties. To identify cases of causal emergence, we made use of the methods in Section 2.1 for calculating EI and creating macro-nodes. As originally shown in [2], in a significant domain of preferential attachment causal emergence cannot happen to any real degree, for example, if $\alpha < 1.0$. This corresponds to the region
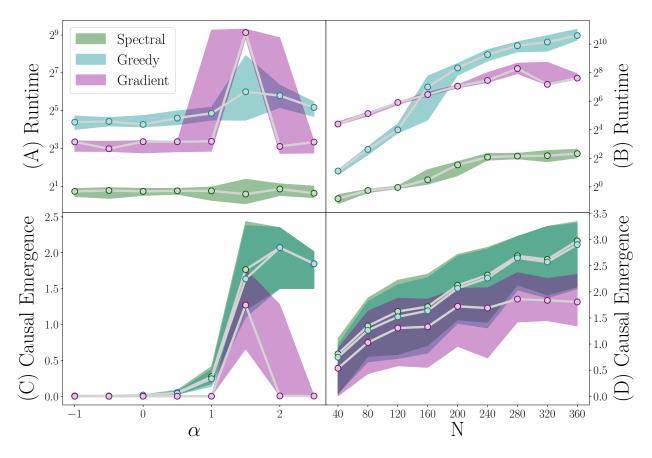
Figure 1: Comparison of Methods for Computing Causal Emergence (A) Comparing algorithm runtimes (in seconds) for different values of α, calculated by adding the system and user CPU time used by the process during calls to each function. Both the greedy and gradient descent algorithms struggle with detecting mesoscales around $\alpha = 1.5$ while the spectral algorithm remains relatively constant. **(B)** Runtime vs. network size. While the greedy algorithm is more efficient than the gradient descent algorithm on small instances, gradient descent scales better to larger networks. The spectral algorithm outperforms both. **(C)** Causal emergence found by different algorithms for preferential attachment networks with different values of α. The greedy and spectral algorithms behave similarly, while gradient descent struggles in the more subtle $\alpha < 1$ cases. **(D)** Detected causal emergence compared to network size, n. The spectral algorithm slightly outperforms the greedy algorithm, and gradient descent lags behind both.

of sublinear preferential attachment, before a network develops higher scales. However, once preferential attachment is no longer "scale-free" but instead superlinear, causal emergence becomes significant (Fig. 2A). That is, networks must not be "scale free" to have informative higher scales. As α increases, the number of nodes grouped into the macro-nodes increases until it is $n - 1$ at high levels of α [2].

What other changes accompany this increase in EI? As defined in [2], $EI = $ determinism $-$ degeneracy, where the determinism is just $\log 2(n) - indeterminism$. The indeterminism is $\langle H(W_i^{out}) \rangle$, while the degeneracy is $log_2(n) - H(\langle W_i^{out} \rangle)$. The increase in determinism and decrease in degeneracy is the cause of the increase in EI at the macroscale. As can be seen in Fig. 2B, the indeterminism of causally-emergent macroscale networks decreases, and Fig. 2C shows the decrease in degeneracy at the macroscale.

In Fig. 2D, we show how the entropy rate of random walkers is much lower at a causally-emergent macroscale compared to its microscale. As defined in [8], the entropy of random walkers on the network is $H\big(X_{t+1}|X_t = \sum_{i,j} \pi(i)p(i \rightarrow j) \log_2 p(i \rightarrow j)\big)$, where $\pi(i)$ is the stationary probability, and $p(i \rightarrow j)$ is the
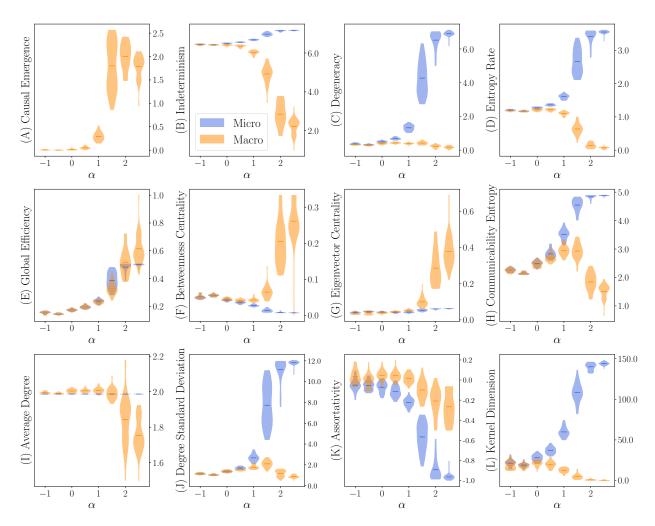
Figure 2: Preferential attachment network properties at different scales. (A) The amount of causal emergence found in 150 node preferential attachment networks for different values of α, calculated using a sample of 40 graphs per value of α. For small α the microscale gives a near optimal description. After $\alpha = 1.0$ mesoscales emerge, and above $\alpha = 2.0$, the macroscale dominates. **(B)** Determinism of node outputs increases at higher scales. **(C)** Similarly, networks are less degenerate (sharing targets) at the macroscale. **(D)** The entropy rate of random walkers over the stationary distribution of the network at both micro- and macro-scales. **(E)** Networks with a distinct mesoscale around $\alpha = 1$ are slightly more efficient at the microscale, but once the macroscale becomes dominant for $\alpha > 1.5$ the macro network is much more efficient. **(F)** Betweenness centrality is a measure of how important individual nodes are to the overall topology of the network. **(G)** Eigenvector centrality of the resulting macronodes increases as α gets larger. **(H)** Entropy of the communicability sequence of nodes in the network, with differences between the microscale and macroscale emerging at $\alpha \approx 1.0$ **(I)** Average degree of the network decreases as more macronodes emerge. **(J)** The variance of the node degree increases as α increases. **(K)** In general, the degree assortativity decreases as α increases, which corresponds to the emergence of more disassortative. However, the macroscale assortativity decreases less than the nodes in the microscale. **(L)** The kernel dimension relates to the degeneracy of a network. From a probabilistic perspective, it quantifies the space of distributions of random walkers that converge to the same distribution from one timestep to the next.

7

probability of v_j given v_i. Note that, the "reverse" entropy rate (the entropy derived if the random walkers are walking backwards across the network) shows a similar reduction at the macroscale (not shown due to its similarity with the entropy rate).

Global network efficiency is a measure of network connectedness based on path length, often used to quantify small-world network dynamics [19]. The global efficiency of a network, G, is the inverse of the average shortest path between all pairs of nodes in a network. The global network efficiency can also be greater at macroscales in cases of causal emergence (Fig. 2E).

The betweenness centrality of a node identifies how many shortest paths of the network traverse that node [20, 1]. Nodes that receive many edges are likely to have a high centrality and thus exert control over network dynamics. Comparing the centrality of G to G_m we see that that the average centrality of the network's nodes increases significantly after causal emergence (Fig. 2F). Since betweenness centrality can be interpreted as quantifying how much control nodes in the network have over information transmission, this indicates that macroscales have a higher degree of control over dynamics than their underlying microscales. Another common notion of centrality is the *eigenvector centrality* of nodes in a network, which not only corresponds to the degree of a given node but also considers nodes whose neighboring nodes have a high degree [21]. In Fig. 2G, we again see that the average eigenvector centrality scores for the macro-nodes begins to increase as α increases.

The communicability of a network is a generalization of the shortest path between any two nodes in a network [22], and the entropy of the communicability sequence has recently been used to characterize and compare networks [23]. We show in Fig. 2H the behavior of the communicability sequence entropy as α increases, showing that the communicability sequence of the macro-nodes begins to decrease after $\alpha = 1.0$, suggesting that the macro-nodes' communicability with other nodes in the network becomes less uniform (i.e., they have higher determinism).

Creating macro-nodes has a slight, though increasingly negative, effect on the node degree as α increases (Fig. 2I), but we really see the effect of macro-nodes when we observe the variance of the degree distribution (Fig. 2J). The detection of macro-nodes in a preferential attachment network dramatically increases the variance of the degree of the remaining micro-nodes. Similarly, we see the network becoming more degree disassortative as α increases, which is more pronounced in the micro-nodes than in the macro-nodes.

Finally, as discussed in Section 2.2.2, the kernel dimension of the network changes as α increases, with the kernel dimension of the micro-nodes increasing rapidly, while the macro-nodes decrease (Fig. 2L). This behavior offers insights into the possible mechanisms behind why the spectral approach to discovering causal emergence in networks is as effective as it appears to be. There is an inherent separation between nodes in the microscale and macroscale, which the spectral algorithm uses to inform which nodes to partition into macro-nodes.

4 Discussion

Networks can possess macroscales that often have different network properties than their underlying microscales. Finding informative higher scales is a search procedure very similar to community detection, as it is sorting through the number of possible partitions of the network. While in this sense finding macro-nodes is similar to identifying communities of nodes, generally community detection is focused on subgraphs that have more in-group connectivity than out-group [24], whereas macro-nodes represent subgraphs that possess a viable summary statistic in terms of their behavior in the network, and therefore macro-nodes can be over a range of connectivity. Additionally, after finding appropriate subgraphs, macro-nodes are a recasting of the network itself.

In order to find informative higher scales we used metrics and methods previously used to identify higher scales of networks via identifying cases of causal emergence [2]. Specifically, causal emergence is the degree to which effective information, (the amount of information in the connectivity of a network), increases at higher scales. Identifying causal emergence in networks has numerous benefits in terms of reducing the dimension of networks, but also improving various network properties, from the structural to the informational. Macroscales that are causally emergent can show, in relation to their original microscale:

a decrease in indeterminism, a decrease in degeneracy, a decrease in the entropy rate of random walkers at the stationary distribution, an increase in global network efficiency, an increase in average betweenness centrality, and a decrease in the kernel dimension. Therefore, it is important that network scientists, if they wish to understand the function and structure of a network, explicitly model higher scales. This requires finding them.

We therefore compared three algorithms: one based on gradient descent, another on a greedy approach, and the other on spectral analysis. An evolutionary algorithm was also attempted but its performance was significantly worse (both in finding causal emergence and in computational time) that we did not include it in the results. Upon comparing the search methods, we can pinpoint the space of connectivity that is most difficult, but also rewarding, to find higher scales at. This is over the mesoscale of the network wherein the higher scale is composed of a complex array of macro-nodes of different sizes. Notably, this indicates that the systems that are most difficult to find an informative scale for possess mesoscales, which may explain the difficulties in understanding the functional architecture of complex systems like the brain [25, 26].

Of the three investigated algorithms we found that spectral analysis performed orders of magnitude better in terms of runtime and got equal to or better results in terms of identifying cases of causal emergence. Therefore, we recommend those that want to find informative higher scales use a clustering algorithm; specifically, a modified form of spectral clustering introduced in 2.2.2 shows promise for larger networks of thousands of nodes. For this purpose, this spectral analysis function was added to the publicly-available Python package for calculating causal emergence at github.com/jkbren/einet.

Overall, our results indicate that network macroscales can be more informative in terms of a higher EI, but also in terms of measures like the entropy rate. As the topology of a network changes across scales, its network properties change as well, often in ways that suggest that networks that possess informative higher scales should be understood as operating at that scale itself, since macroscales can entail a peak of efficiency, centrality, or information transmission.

5   Acknowledgments

References

[1]   Albert-László Barabási. *Network Science*. Cambridge University Press, 2016. ISBN: 9781107076266.

[2]   Brennan Klein and Erik Hoel. "Uncertainty and causal emergence in complex networks". In: *arXiv preprint arXiv:1907.03902* (2019). URL: https://arxiv.org/abs/1907.03902.

[3]   Richard Gallagher and Tim Appenzeller. "Beyond reductionism". In: *Science* 284.5411 (1999), pp. 79–80. DOI: 10.1126/science.284.5411.79.

[4]   Erik Hoel, Larissa Albantakis, and Giulio Tononi. "Quantifying causal emergence shows that macro can beat micro". In: *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19790–19795. DOI: 10.1073/pnas.1314922110.

[5]   Erik Hoel et al. "Can the macro beat the micro? Integrated information across spatiotemporal scales". In: *Neuroscience of Consciousness* 2016.1 (2016). DOI: 10.1093/nc/niw012.

[6]   Erik Hoel. "When the map is better than the territory". In: *Entropy* 19.5 (2017), p. 188. DOI: 10.3390/e19050188.

[7] Mihael Ankerst et al. "OPTICS: Ordering points to identify the clustering structure". In: *Proc. ACM SIGMOD'99 Int. Conf. on Management of Data*. ACM Press, 1999, pp. 49–60. DOI: 10.1145/304182. 304187.

[8] Jian Xu, Thanuka L Wickramarathne, and Nitesh V Chawla. "Representing higher-order dependencies in networks". In: *Science Advances* 2.5 (2016), e1600028. DOI: 10.1126/sciadv.1600028.

[9] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, 2014. ISBN: 9781558604797.

[10] Miroslav Fiedler. "Laplacian of graphs and algebraic connectivity". eng. In: *Banach Center Publications* 25.1 (1989), pp. 57–70. URL: http://eudml.org/doc/267812.

[11] Stephen Guattery and Gary L. Miller. "On the performance of spectral graph partitioning methods". In: *SODA*. 1995.

[12] Daniel A. Spielman and Shang-Hua Teng. "Spectral partitioning works: Planar graphs and finite element meshes". In: *Linear Algebra and its Applications* 421.2 (2007). Special Issue in honor of Miroslav Fiedler, pp. 284–305. ISSN: 0024-3795. URL: 10.1016/j.laa.2006.07.020.

[13] Fan Chung, Linyuan Lu, and Van Vu. "Spectra of random graphs with given expected degrees". In: *Proceedings of the National Academy of Sciences* 100.11 (2003), pp. 6313–6318. ISSN: 0027-8424. DOI: 10.1073/pnas.0937490100.

[14] Essam El Seidy, Salah Eldin Hussein, and Atef AboElkher. "Spectra of some simple graphs". In: *Mathematical Theory and Modeling* 5 (2 2015), pp. 115–121. DOI: 10.14419/ijamr.v5i2.6106.

[15] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2009. ISBN: 9780980232714.

[16] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv:1609.04747* (2016). URL: https://arxiv.org/abs/1609.04747.

[17] Yurii Nesterov. "A method for unconstrained convex minimization problem with the rate of convergence O (1/k^ 2)". In: *Doklady AN USSR*. Vol. 269. 1983, pp. 543–547.

[18] Albert-László Barabási. "Scale-free networks: a decade and beyond". In: *science* 325.5939 (2009), pp. 412–413. DOI: 10.1126/science.1173299.

[19] Vito Latora and Massimo Marchiori. "Efficient behavior of small-world networks". In: *Phys. Rev. Lett.* 87 (19 Oct. 2001), p. 198701. DOI: 10.1103/PhysRevLett.87.198701.

[20] Linton C Freeman. "A set of measures of centrality based on betweenness". In: *Sociometry* (1977), pp. 35–41. DOI: 10.2307/3033543.

[21] Britta Ruhnau. "Eigenvector-centrality—a node-centrality?" In: *Social networks* 22.4 (2000), pp. 357–365. URL: 10.1016/S0378-8733(00)00031-9.

[22] Ernesto Estrada and Naomichi Hatano. "Communicability in complex networks". In: *Physical Review E* 77.3 (2008), p. 036111. DOI: 10.1103/PhysRevE.77.036111.

[23] Dan Chen et al. "Complex network comparison based on communicability sequence entropy". In: *Physical Review E* 98.1 (2018), p. 012319. DOI: 10.1103/PhysRevE.98.012319.

[24] Jure Leskovec, Kevin J Lang, and Michael Mahoney. "Empirical comparison of algorithms for network community detection". In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 631–640.

[25] Daniel P Buxhoeveden and Manuel F Casanova. "The minicolumn hypothesis in neuroscience". In: *Brain* 125.5 (2002), pp. 935–951. DOI: 10.1093/brain/awf110.

[26] Rafael Yuste. "From the neuron doctrine to neural networks". In: *Nature reviews neuroscience* 16.8 (2015), p. 487. DOI: 10.1038/nrn3962.