SCIENTIFIC REPORTS

OPEN

Finding communities in sparse networks

Abhinav Singh & Mark D. Humphries

Faculty of Life Sciences, University of Manchester.

SUBJECT AREAS:

COMPUTATIONAL SCIENCE

COMPLEX NETWORKS

Received
15 October 2014

Accepted
5 February 2015

Published
6 March 2015

Correspondence and requests for materials should be addressed to A.S. (abhinav.singh@manchester.ac.uk)

Spectral algorithms based on matrix representations of networks are often used to detect communities, but classic spectral methods based on the adjacency matrix and its variants fail in sparse networks. New spectral methods based on non-backtracking random walks have recently been introduced that successfully detect communities in many sparse networks. However, the spectrum of non-backtracking random walks ignores hanging trees in networks that can contain information about their community structure. We introduce the reluctant backtracking operators that explicitly account for hanging trees as they admit a small probability of returning to the immediately previous node, unlike the non-backtracking operators that forbid an immediate return. We show that the reluctant backtracking operators can detect communities in certain sparse networks where the non-backtracking operators cannot, while performing comparably on benchmark stochastic block model networks and real world networks. We also show that the spectrum of the reluctant backtracking operator approximately optimises the standard modularity function. Interestingly, for this family of non- and reluctant-backtracking operators the main determinant of performance on real-world networks is whether or not they are normalised to conserve probability at each node.

Many networks have a modular structure. Social networks contain communities of friends[1–3], collaborators[2], and dolphins[4]; brain networks contain groups of correlated neurons[5,6], circuits of connected groups[7,8], and regions of connected circuits[9]. Similarly modular networks occur across biological domains from protein interaction networks to food webs[10]. This range of applications has driven the dramatic development of "community detection" methods for solving the core problem of finding modules within an arbitrary network[10]. Especially popular are spectral methods based on the eigenvalues and eigenvectors of some matrix representation of the network. These combine speed of execution with considerable information about the network beyond the modular structure[11], including the relative roles of each node[11] and characterisation of the network's dynamical properties[12,13].

Spectral methods can fail for a range of real networks. These methods rely on the eigenvalues falling into two classes, the vast majority – the "bulk" – following a well-defined distribution, and the outliers from that distribution giving information about the community structure. Topological features of a network unrelated to its modules, such as network hub nodes with high degree, can distort this distinction by introducing eigenvalues outside the bulk that mix with those containing information about modules[14–16]. Sparse networks often contain such network hubs and the outlying uninformative eigenvalues cause the breakdown of spectral methods[17]. Unfortunately many real-world networks are sparse (see Table II in Ref. 18 and Table 1 in Ref. 19).

Krzakala et al.[20] proposed a new "non-backtracking" matrix representation of a network that solves this problem: their matrix represents a random walker on the network who cannot immediately return to a node it has just left. The eigenspectrum of this matrix depends on the frequency with which the walker passes through any given node. As the non-backtracking matrix forbids the random walker to return to its immediately previous node, network hubs are not visited disproportionately by this random walker and so the eigenspectrum is not distorted by the presence of hubs in the network. Following this, Newman introduced the closely-related "flow" matrix[21] that conserved the probability for the random walk. Spectral methods applied to these matrices successfully recover modules in sparse networks, down to the theoretical limit for their detection in classes of model networks[20].

However, as noted by Newman[21], these represent an incomplete solution as networks containing trees cannot be handled elegantly. Because the random walker could not escape from such a tree once entered, trees are ignored despite being candidates for separate modules. In this paper we introduce the "reluctant backtracker" approach, which combines the advantages of these new matrix representations by retaining the power of spectral methods for sparse networks with the ability to detect and correctly handle networks with trees. We show that this comes with no penalty for detection performance compared to non-backtracking and flow matrices. Rather, we show that the main difference in performance depends on whether or not such matrix representations are normalised to conserve probability. This finding hints at some deeper difference in network structure than modularity alone.

Non-backtracking and flow matrices

We first outline the non-backtracking[20] and flow matrix[21] approaches to community detection. Both these approaches and ours start from the same representation of the network. Assume an unweighted, undirected, connected network with n vertices and m edges without self loops. We convert the undirected network into a directed network with $2m$ edges by replacing the undirected edge with directed edges in both directions; $j \rightarrow i$ showing the direction of the edge. The binary non-backtracking matrix $\mathbf{B}$ has $2m \times 2m$ elements, each element corresponding to a pair of directed edges in the network. Its elements are given by

$$B_{j\to i,l\to k}=\delta_{il}\left(1-\delta_{jk}\right), \tag{1}$$

which are non-zero only if $B_{j\to i,l\to k}$ corresponds to a directed path from j to k that passes through node i with the restriction that nodes j and k must not be identical, i.e. no backtracking. This matrix encapsulates the biased random walker that is prohibited from returning to its immediately previous node.

Newman modified the non-backtracking matrix by changing the values of its non-zero elements and called it the flow matrix $\mathbf{F}$ in analogy to current flow in an electrical network. Its elements are given by

$$F_{j\to i,l\to k}=\delta_{il}\left(1-\delta_{jk}\right)\frac{1}{d_i-1}, \tag{2}$$

where d_i is the degree of the node i. Consider the random walker that starts from node j and is passing through node i. According to the flow matrix, the random walker can reach any of the $d_i - 1$ nodes except node j with equal probability. The probability of reaching node k from node j passing through node i is $\frac{1}{d_i-1}$, conserving probability at node i. Krzakala et al.[20] and Newman[21] respectively showed that the second leading eigenvector of the non-backtracking and flow matrices is very successful in correctly dividing sparse networks into communities.

Results

Reluctant backtracking operators. To solve the problem of detecting communities in the presence of trees, we introduce the idea of a reluctant backtracking random walker that admits a small probability of returning to a node immediately. The reluctance, but not impossibility, of immediately returning to a node mitigates network hub effects on the spectrum of the operators, while allowing the walker to explore and return from hanging trees unlike the non-backtracking operator or flow matrix.

Based on this idea of reluctance, we define two new reluctant backtracking operators $\mathbf{R}$ and $\mathbf{P}$ whose matrix elements are

$$\mathbf{R}: R_{j\to i,l\to k}=\delta_{il}\left(1-\delta_{jk}\right)+\delta_{il}\delta_{jk}\frac{1}{d_j} \tag{3}$$

$$\mathbf{P}: P_{j\to i,l\to k}=\left[\delta_{il}\left(1-\delta_{jk}\right)+\delta_{il}\delta_{jk}\frac{1}{d_j}\right]\frac{1}{d_i-1+\frac{1}{d_j}}, \tag{4}$$

where $R_{j\to i,l\to k}$ and $P_{j\to i,l\to k}$ represents the probability that the random walker shall move from node j to node k with nodes i and l as intermediate nodes. The probability of returning to a node for both operators $\mathbf{R}$ and $\mathbf{P}$ is inversely proportional to the degree of the node, thus discouraging strongly a return to a high degree node.

The operator $\mathbf{R}$ is a reluctant version of the non-backtracking operator $\mathbf{B}$ as it allows the additional probability $\frac{1}{d_j}$ of returning immediately to the node j. The operator $\mathbf{P}$ is a normalised version of the operator $\mathbf{R}$ just like the flow operator $\mathbf{F}$ is a normalised version of the non-backtracking operator $\mathbf{B}$. Similar to the non-backtracking $\mathbf{B}$ and flow $\mathbf{F}$ matrix operators, the new reluctant backtracking operators $\mathbf{R}$ and $\mathbf{P}$ can currently only be applied to undirected networks.

The procedure for detecting the communities is identical for both operators. Given the adjacency matrix of a network, we first generate one of the matrices $\mathbf{R}$ or $\mathbf{P}$. Following Krzakala et al.[20], we calculate its second largest absolute real eigenvalue and the associated eigenvector. The eigenvector has $2m$ elements corresponding to each directed edge in the network. We group the elements of the eigenvector by the group index of the source node of each edge and sum them up to create a new vector that has n elements corresponding to each node in the network. We divide the network into two communities by grouping all nodes with the same sign in that vector: the sign of each element represents the estimate of the reluctant backtracking operators of the node's community.

Communities composed of trees. The indifference of non-backtracking operators towards trees can impair their abilities to detect communities in networks. As an extreme case, consider the network suggested by Newman[21]: a network composed of two binary trees connected at a single node. The non-backtracking operator $\mathbf{B}$ and the flow matrix $\mathbf{F}$ cannot detect communities in such a network, but the reluctant backtracking operators $\mathbf{R}$ and $\mathbf{P}$ do.

We show this using a network composed of two communities A and B where each community is a tree and the two communities are connected by a *single* node. The ratio of the number of nodes in community A and B is denoted by f. The number of nodes in community A is fixed and the number of nodes in community B varies.

Figure 1 shows that when the size of the two communities is comparable ($f \approx 1$), the reluctant operators detect communities perfectly since a random walker will remain within the same community for substantial periods of time. There is a sharp transition in the ability of the reluctant backtracking operators around $f \approx 0.6$ in the network where community A consists of 400 nodes (Figure 1a). When one community becomes much smaller than the other, random walkers keep moving to the larger community from the small community in a short period of time and leads to the loss of performance. The transition point f is dependent on many factors such as the structure of the network, total number of nodes in the network, and the relative sizes of different communities (as illustrated in Figure 1b where community A has 500 nodes, and the transition point is $f \approx 0.48$). Why there is a sharp discontinuity rather than a gradual decline in performance is presently unclear.

Stochastic block model with additional leaves. Networks composed solely of trees are of course very artificial, but we also show that reluctant backtracking operators can detect communities in a more plausible network where the non-backtracking operators fail. Consider a more typical network, created by the classic stochastic block model. The addition or deletion of hanging trees to this network or any other will not affect the eigenspectrum of the non-backtracking operator $\mathbf{B}$. However, the presence of hanging trees can significantly alter the structure of communities in such a network.

Stochastic block models provide an easy recipe for constructing networks with specified inter-community and intra-community edge probability. Consider a network of n nodes with two communities. The probability of an edge between nodes a and b is given by

$$P_{ab}=\frac{c_{in}}{n} \quad \text{if a and b belong to same community} \tag{5}$$

$$=\frac{c_{out}}{n} \quad \text{if a and b belong to different communities} \tag{6}$$

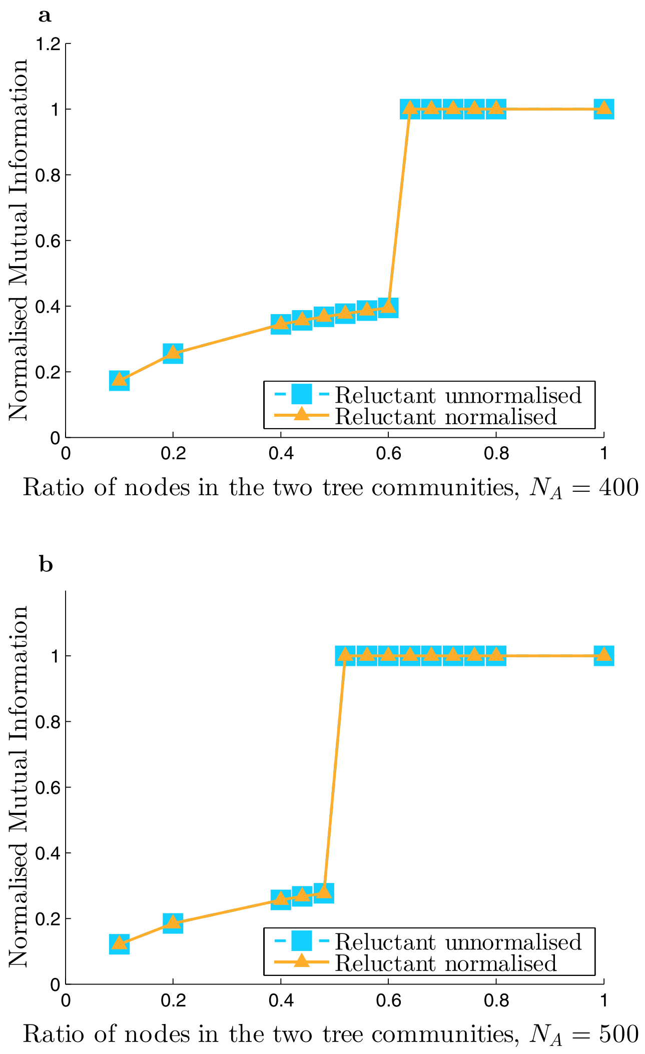

Figure 1 | **Two binary trees connected at one node.** The x-axis shows the number of nodes in community *B* as a fraction *f* of nodes in community *A*. The triangles and squares show the performance of the two operators in detecting communities as measured by the normalised mutual information (NMI): $0 \leq NMI \leq 1$, where $NMI = 1$ means perfect community detection and $NMI = 0$ means random allocation of nodes to communities (see Methods for more details). (a) 400 nodes in community *A*. Number of nodes in community *B* varies from from 40 to 400. (b) 500 nodes in community *A*. Number of nodes in community B varies from 50 to 500.

Let $c = \frac{c_{in} + c_{out}}{2}$ be the average degree of the network and $c_{minus} = \frac{c_{in} - c_{out}}{2}$ denote the degree of mixing between communities in the network. No mixing between the communities implies $c_{minus} = c$ and complete mixing between the two communities implies $c_{minus} = 0$.

We demonstrate the effect of hanging trees by selectively adding leaves to a network based on the stochastic block model. We create a stochastic block model network with two communities, each with 500 nodes, using parameters $c_{in} = 4.8$, $c_{out} = 1.2$. We add one leaf to each node whose number of connections within the community exceeds its connections outside its community by at least 3. This selects the nodes whose degree is greater than the median and whose membership is slightly ambiguous.

Figure 2 shows that the non-backtracking operator **B** does not detect two communities as its spectra has only one real eigenvalue outside the bulk. The additional information provided by the leaves is not available to the non-backtracking operator. On the other hand, Figure 2 shows that the reluctant backtracking operator accounts for the leaves in the network and its second eigenvector successfully detects two communities.

Stochastic block model based networks. The quality of community detection is inversely proportional to the degree of mixing between different communities in a network. The performance of any spectral method falls to chance below a predictable mixing threshold for simulated networks based on the stochastic block model[22–24]. This threshold is defined as the minimum network mixing variable, c_{minus}, where at least one real positive eigenvalue lies outside the bulk, and so some community structure is, in principle, detectable. Below this limit a block-model network becomes spectrally indistinguishable from an Erdös-Rényi random graph and therefore no communities can be reliably detected by spectral methods. Consequently, simulated networks based on the stochastic block model serve as a useful benchmark for testing the performance of different community detection methods. Krzakala et al.[20] showed that the non-backtracking operator **B** can detect communities in sparse networks right down to this theoretical limit where other spectral methods fail.

Figure 3 shows the performance of the four operators **B**, **F**, **R** and **P** on a set of networks based on the stochastic block model with 10^3 nodes with constant average node degree and varying degrees of mixing between communities ($0.1 \leq c_{minus} \leq 3.0$). Both the non-backtracking **B** and flow **F** matrices are able to detect the presence of two communities above chance levels down to the theoretical limit. The reluctant backtracking operator **R**'s performance is comparable to both. Thus the reluctant backtracker **R** accounts for hanging trees in a network, yet there is no penalty for detecting communities down to the theoretical limit.

By contrast, the normalised reluctant backtracker **P** performs worse on average than all other operators, and also has the widest variation in performance. As such, close to the theoretical limit it only occasionally shows above-chance performance.

Qualitative features of the normalised reluctant operator **P**'s eigenspectrum are potential contributing factors. The maximum eigenvalue of the normalised reluctant operator **P** is always 1, therefore all other eigenvalues are constrained to be less than 1. Additionally, the bound of the bulk eigenvalues is dependent on the average degree of the network, *c* which is held constant at 3 while the degree of mixing c_{minus} varies from 0 to 3. When the number of connections between communities increases due to greater mixing between communities, random walkers associated with the reluctant operators **P** and **R** migrate between communities slightly more easily compared to the non-backtracking operators (**B**, **F**) leading to real eigenvalues being pushed outside the bulk. The fixed bounds on both the bulk and the upper eigenvalue of the normalised reluctant operator **P** suggests a limited range for absorbing these noisy eigenvalues before their magnitude surpasses the second largest real eigenvalue. Thus, close to the theoretical limit where mixing is high, the community structure could become undetectable for **P**. This appears not to be the case for the reluctant operator **R**, as its eigenvalues are unbounded. However, the normalised reluctant operator **P** is seemingly not penalised for this limitation in real-world applications (as we show below in Figure 4). A full understanding of the operator **P**'s performance needs a formal precise analysis of its spectral properties, which is the subject of future work.

Real world networks. Table 1 and Figure 4 compares the effectiveness of the reluctant and non- backtracking matrices on three real world data sets: the Zachary karate club[1], the social network of dolphins in Doubtful Sound[25], and word adjacencies[11]. In Figure 4 we plot the distribution of eigenvalues of each operator, showing that both the non-backtracking (**B**, **F**) and reluctant-backtracking

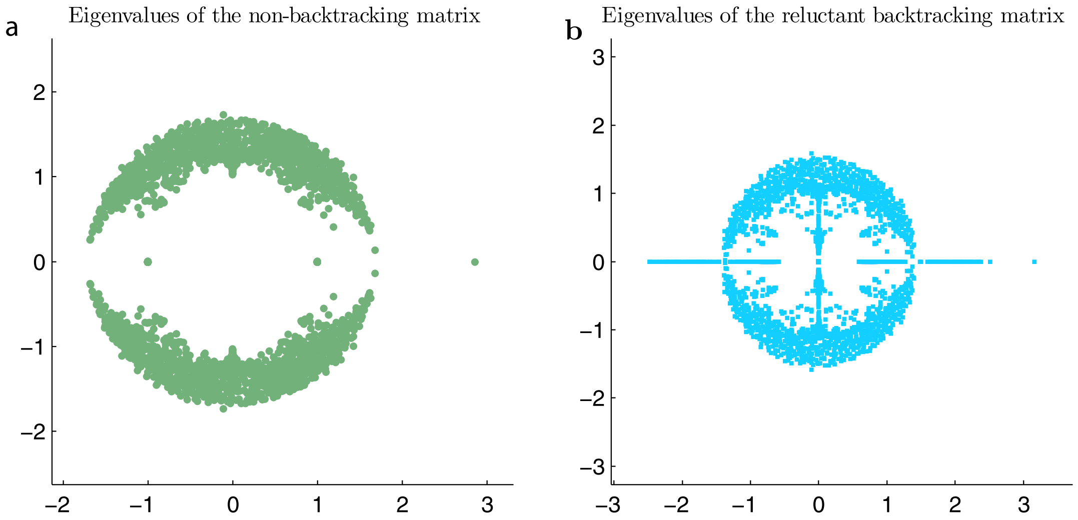

Figure 2 | **Stochastic blockmodel network with additional leaves.** Final network parameters after leaf addition: $n = 1273$, $m = 1818$, $c_{in} = 4.8$, $c_{out} = 1.2$, where n denotes the number of nodes in the network, and m denotes the number of undirected edges in the network. All the random walk operators are square matrices of order $2m$. (a) Eigenvalues of a representative non-backtracking matrix **B**. Note that there is only one real eigenvalue outside the bulk. (b) Eigenvalues of a representative reluctant backtracking matrix **R**.

(**R**, **P**) operators have more than one outlying eigenvalue and can thus detect community structure in these networks. The reluctant backtrackers detect communities comparably to their respective non-backtracking counterparts, and there is no loss of performance when using the reluctant matrices rather than the non-backtracking matrices. Rather, we found that the main difference in performance depended on whether or not the operators are normalised. This is particularly striking for the dolphin social network, for which the normalised operators perform similarly and both markedly better than the unnormalised versions.

Modularity maximisation. Newman[21] showed that the second leading eigenvector of the flow matrix **F** maximises the widely-used modularity function Q[11], connecting the non-backtracking method to the idea of community detection as an optimisation problem. We show that the reluctant backtracking operator **P** also approximately optimises the modularity function Q.

Assume an unweighted undirected network of size n with m edges specified by the adjacency matrix **A**. The modularity function Q is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{d_i d_j}{2m}\right]\delta_{g_i g_j} \tag{7}$$

A_{ij}: presence/absence of edge between nodes i and j.
d_i: degree of node i.
g_i: group membership of node i.
m: number of edges in the network.

Following Newman's setting and notation[21], assume that the network is divided into two communities and define the n dimensional group membership vector **s** with elements $s_i \in \{-1, 1\}$ denoting the membership of each node in the network. We define the quadratic form

$$\begin{aligned} &T = \mathbf{u}^T\left(\mathbf{P} - \mathbf{1}\mathbf{1}^T\right)\mathbf{v} \\ &\mathbf{u},\mathbf{v}\ :\ 2m \text{ dimensional unit vectors } 1 = (1,1,1,\ldots)\Big/\sqrt{2m}, \end{aligned} \tag{8}$$

If we make the particular choice $u_{i\to j} = v_{i\to k} = s_i$, meaning that the elements of both vectors **v** and **u** are equal to the group index of the node from which the corresponding edge *emerges*, then

$$\begin{aligned} \mathbf{u}^T\mathbf{P}\mathbf{v} &= \sum_{\substack{\text{edges } j\to i \\ \text{edges } l\to k}} \left[\frac{1}{d_j}\delta_{il}\delta_{jk} + \delta_{il}\left(1-\delta_{jk}\right)\right]\frac{1}{d_i - 1 + \frac{1}{d_j}} s_j s_i \\ &= \sum_j s_j \sum_{ik}\left[\frac{1}{d_i - 1 + \frac{1}{d_j}}\frac{1}{d_j}A_{ik}A_{ij}\delta_{jk} + \frac{1}{d_i - 1 + \frac{1}{d_j}}\left(1-\delta_{jk}\right)A_{ik}A_{ij}\right] s_i \\ &= \sum_j s_j \sum_i A_{ij} s_j \left[\frac{1}{d_i - 1 + \frac{1}{d_j}}\left(\frac{1}{d_j} + d_i - 1\right)\right] \\ &= \sum_j s_j \sum_i A_{ij} s_i \\ &= \mathbf{s}^T\mathbf{A}\mathbf{s} \end{aligned} \tag{9}$$

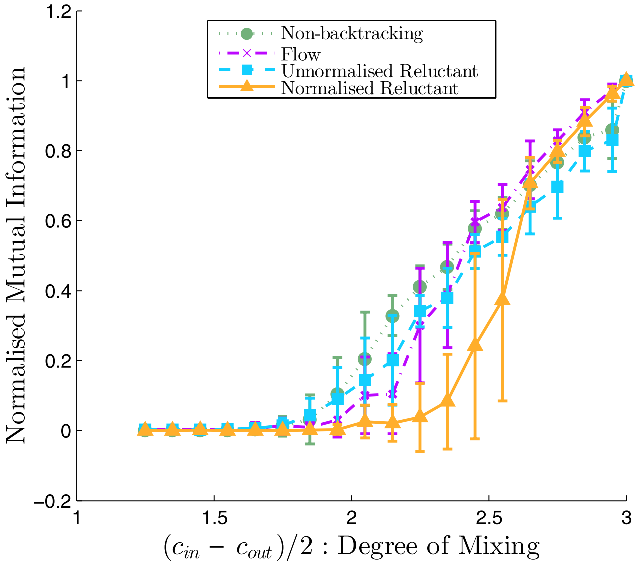

Figure 3 | **Community detection performance on the stochastic block model.** We plot normalised mutual information of the recovered communities compared to the planted communities as a function of the degree of mixing in the block model network (1000 nodes, average degree $c = 3$). Each data point shows the mean and standard deviation of NMI for the different operators as applied to 20 networks with the given mixing parameters.

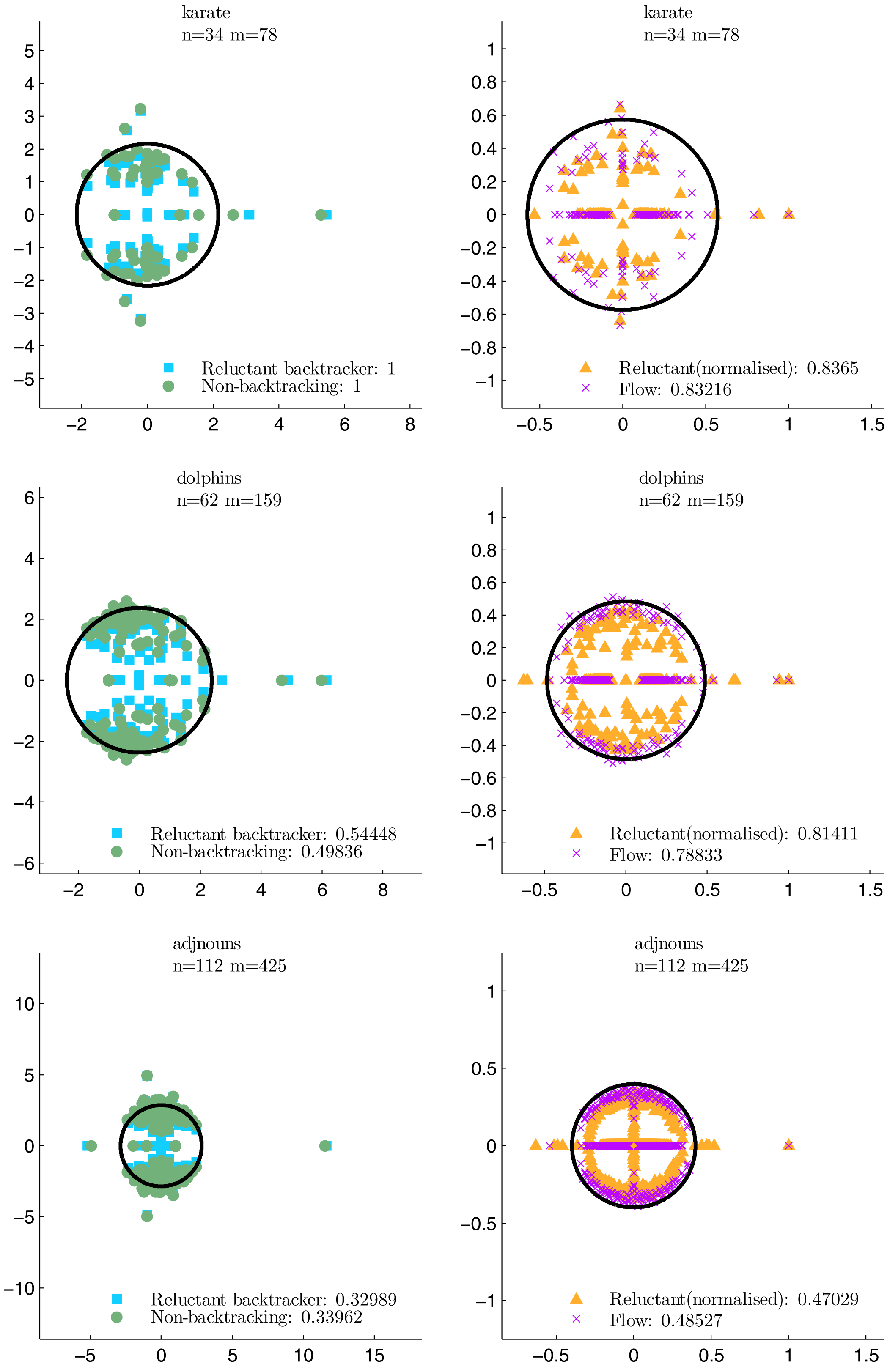

Figure 4 | **Real world performance.** The dots are the eigenvalues of the respective matrices. The black circle is the approximate analytical bound of the bulk eigenvalues for the non-backtracking and flow matrices, respectively $\sqrt{\langle c\rangle}$ (Ref. 20) and $\sqrt{\langle c/(c-1)\rangle/\langle c\rangle}$ (Ref. 21), where c is degree, and $\langle\rangle$ is an average. These bounds were derived for the stochastic block model, so are used here as an heuristic guide for the distribution of eigenvalues resulting from the real-world networks, and computed using their degree distribution. n denotes the number of nodes in the network. m denotes the number of undirected edges in the network. All the random walk operators are square matrices of order $2m$. Values in legends are NMI from Table 1.

Table 1 | Performance (measured as normalised mutual information) of different operators as applied to real datasets

	Reluctant	Non backtracking	Normalised reluctant	Flow
Karate	1	1	0.8365	0.8322
Dolphins	0.5445	0.4984	0.8141	0.7883
Adjnouns	0.3299	0.3396	0.4703	0.4853

Also it follows that

$$\begin{aligned}\mathbf{u}^T\mathbf{1}\mathbf{1}^T\mathbf{v} &= \frac{1}{2m}\sum_{\substack{\text{edges } j\to i\\ \text{edges } l\to k}} s_j s_i = \frac{1}{2m}\sum_{ijkl} A_{ij}A_{kl}s_j s_i \\ &= \frac{1}{2m}\sum_{ji} d_j d_i s_j s_i = \mathbf{s}^T\frac{\mathbf{d}\mathbf{d}^T}{2m}\mathrm{s},\end{aligned} \tag{10}$$

Therefore

$$\begin{aligned}Q &= \frac{1}{2m}\mathbf{u}^T\left(\mathbf{P}-\mathbf{1}\mathbf{1}^T\right)\mathbf{v}, \\ &= \frac{1}{2m}\mathbf{s}^T\left(\mathbf{A}-\frac{\mathbf{d}\mathbf{d}^T}{2m}\right)\mathbf{s}.\end{aligned} \tag{11}$$

Since the normalised reluctant backtracker **P** also optimises the modularity function, our spectral solution coincides with Newman's. We summarise Newman's solution here, refer to Ref. 21 for further details. Solving equation 11 exactly is hard but an approximate solution can be found by standard relaxation techniques. Allow **u** and **v** to independently take any real value rather than only ±1 and apply the constraint that $\mathbf{u}^T\ \mathbf{v} = 2m$. This modified problem can be solved by the method of Lagrange multipliers. We get the following equation by introducing the multiplier λ and differentiating with respect to elements of **u**

$$\left(\mathbf{P}-\mathbf{1}\mathbf{1}^T\right)\mathbf{v}=\lambda\mathbf{v} \tag{12}$$

The leading eigenvector of $\mathbf{P} - \mathbf{1}\mathbf{1}^T$ or the second leading real eigenvector of **P** exactly optimises the relaxed problem. We arrive at the approximate solution of the original unrelaxed problem by setting $s_i = sgn\left(\sum_j v_{i\to j}\right)$, i.e. we sum up all the elements of the eigenvector that emerge from node i and assign $s_i = 1$ if the sum is positive or -1 if it is negative. This is very similar to the algorithm used by Krzakala et al.[20] with the difference that we sum up edges emerging from a node rather the ones incident upon it.

Discussion

We propose a new reluctant backtracking operator to detect communities in sparse networks that accounts for hanging trees. Unlike other recent operators such as the non-backtracking matrix and the flow matrix, the reluctant backtracking operator accounts for the presence of hanging trees in a network and its eigenspectrum is shaped by their presence. We demonstrate the utility of the reluctant backtracking operator by detecting communities in simulated networks where the non-backtracking matrix is unable to do so and also show a comparable ability to detect communities in benchmark simulated and real networks.

Newman[21] showed that the second leading eigenvector of the flow matrix approximately maximises the modularity function by ensuring conservation of probability at each node. Following a similar argument we also show that the eigenvector of the normalised reluctant backtracking matrix **P** approximately maximises the modularity function.

An interesting future problem is to extend the reluctant backtracking approach to reliably detect more than two communities. Determining the number of communities in a network is a problem by itself and knowing the number of communities in a network can improve the performance of community detection methods[26]. Krzakala et al.[20] suggested a heuristic to determine the number of communities in a given network when using the non-backtracking matrix **B**. They derived an approximate analytical bound for the uninformative eigenvalues lying inside the bulk for sparse stochastic block model networks and found that the number of real-valued eigenvalues lying outside the bulk's radius served as a good heuristic to estimate of the number of modules in model networks. Newman derived a similar bound for the flow matrix **F**[21]. When applied to real-world networks, a further heuristic is to compute these bounds using the mean degree of the real-world network and use them as a guide to the number of modules in that network. We plot these approximated bounds for our sample of real-world networks in Figure 4; we note that, like the flow matrix **F**, the eigenvalue distribution for our normalised reluctant backtracker **P** is particularly well-behaved with respect to the approximated bounds compared to the unnormalised matrices. We leave the determination of the bound for the reluctant operators for future work, as they do not follow simply from those derived for the non-backtracking matrices.

However, because of the approximations involved, the heuristic can fail for real[20] and simulated networks[26], by predicting too many real-valued eigenvalues outside the bulk and thus predicting too many modules. The optimisation of modularity Q by the second eigenvector of both the flow **F** and normalised reluctant-backtracker **P** matrices suggests two further solutions for finding more than two communities. The first solution is a more cautious approach that treats the total number q of real eigenvalues outside the approximated bulk radius as an upper limit for the number of communities in the network[6]. We can identify these communities by first taking each of the $q - 1$ eigenvectors corresponding to the $q - 1$ eigenvalues (remembering that we start from the second eigenvector) and converting them into a length n vector as before – we sum over the eigenvector entries corresponding the same source node. We can then cluster in the $\mathbb{R}^{q-1}$ space defined by these node vectors, using a standard clustering algorithm such as k-means: we cluster for each $k \in [2, q - 1]$, and compute Q for each k, retaining the clustering that maximises Q. The second solution is to apply the iterative bisection algorithm from Ref. 11. We initially divide the network into two communities using the second leading eigenvector of **F** or **P**, then iteratively divide each sub-division using the same algorithm. We compute Q for each sub-division (adjusted to account for the remainder of the network[11]), stopping when $Q \leq 0$.

The difference in performance between the normalised and non-normalised versions of the operators on the real-world networks hints that normalisation is incorporating more information about the network's structure than is available to the unnormalised operator. Normalisation adds information about the degree of the transition node i in the path $j \to i \to k$ to each non-zero element of the matrix of the normalised operators **F** and **P**. By contrast, each path from node $j \to k$ in the non-backtracking matrix **B** has an equal weight of 1 irrespective of the degree of the intermediate node i. This new information affects the eigenspectrum of the normalised operators, and thus likely leads to the observed differences in community detection performance. Precisely how and when this additional information is beneficial for detecting communities is the subject of future work.

Methods

Normalised mutual information. Given a network with two possible partitions of its nodes into communities, normalised mutual information (NMI) quantifies the overlap between these two partitions. NMI serves as a metric to quantify the absolute performance of a community detection method and compare the relative performance of different methods.

Assume a network with N nodes and community partitions $\mathbb{A}$ and $\mathbb{B}$. A_i is the subset of nodes in the network that belong to community i in partition $\mathbb{A}$ and B_j is the subset of nodes in the network that belong to community j in partition $\mathbb{B}$. Let n_A and n_B be the number of communities in the partitions $\mathbb{A}$ and $\mathbb{B}$ respectively (in this paper we have $n_A = n_B = 2$ throughout). The confusion matrix **F** captures the overlap between the two partitions: element F_{ij} counts the number of nodes common to the communities A_i and B_j. Normalised mutual information[27] is defined as

$$NMI(A,B) = \frac{-2\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} F_{ij} ln\left(F_{ij}N/N_iN_j\right)}{\sum_{i=1}^{n_A} N_i ln(N_i/N) + \sum_{j=1}^{n_B} N_j ln\left(N_j/N\right)} \qquad (13)$$

where

$$n_A, n_B : \text{number of communities in partition } \mathbb{A} \text{ and } \mathbb{B}$$

$$N_i, N_j : \text{number of nodes in communities } A_i \text{ and } B_j$$

NMI always lies between 0 and 1; $NMI = 1$ only if the partitions $\mathbb{A}$ and $\mathbb{B}$ are identical and $NMI = 0$ only if the partitions $\mathbb{A}$ and $\mathbb{B}$ are completely independent of each other.

Community detection algorithm and numerical considerations. Given the adjacency matrix of a network, we first generate one of the matrices **R** or **P**. Following Krzakala et al.[20], we calculate its second largest absolute real eigenvalue and the associated eigenvector. The eigenvector has $2m$ elements corresponding to each directed edge in the network. We group the elements of the eigenvector by the group index of the source node of each edge and sum them up to create a new vector that has n elements corresponding to each node in the network. We partition the network into two communities by grouping all nodes that have the same sign; the sign of each element represents the estimate of the reluctant backtracking operators of the node's community.

If the network has less than 500 nodes, we calculated all the eigenvalues and eigenvectors using the eig function in MATLAB based on the QR algorithm because it is feasible to quickly calculate on a desktop computer all the eigenvalues and eigenvectors for networks where the number of edges, m is within one order of 1000. When the network is large and the number of edges becomes greater than 10000, it became impractical to quickly calculate all the eigenvalues and eigenvectors and we resort to a different approach since our community detection algorithm does not require us to know all the eigenvalues to estimate the community structure of the network. If the network was larger than 500 nodes, we employed a heuristic to find the second largest real eigenvalue by magnitude. We first calculated the largest 50 eigenvalues by absolute value and the associated eigenvectors using the eigs function in MATLAB that is suited for sparse matrices and is based on the implicitly restarted Arnoldi iteration method[28]. We then selected the eigenvalues whose complex part was less than 0.5×10^{-4} to allow for the inexactness of the eigenvalue algorithms and from these finally chose the eigenvalue with the second highest absolute value and its associated eigenvector. The number of eigenvalues that need to be calculated before a real eigenvalue is found is mostly dependent on the degree of mixing in the network rather than the number of edges or nodes in the network. If the communities in the network are strongly mixed then the real eigenvalues will be buried deep within the bulk even if the network has few nodes and the real eigenvalues will be detached from the bulk in a large network if the communities are weakly mixed.

1. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
2. Girvan, M. & Newman, M. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).
3. Newman, M. Modularity and community structure in networks. *P. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
4. Lusseau, D. & Newman, M. Identifying the role that animals play in their social networks. *P. Roy. Soc. Lond. B Bio.* **271**, S477–S481 (2004).
5. Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G. & Buzsáki, G. Organization of cell assemblies in the hippocampus. *Nature* **424**, 552–556 (2003).
6. Humphries, M. Spike-train communities: finding groups of similar spike trains. *J. Neurosci.* **31**, 2321–2336 (2011).
7. Binzegger, T., Douglas, R. J. & Martin, K. A. C. A quantitative map of the circuit of cat primary visual cortex. *J. Neurosci.* **24**, 8441–8453 (2004).
8. Perin, R., Berger, T. K. & Markram, H. A synaptic organizing principle for cortical neuronal groups. *P. Natl. Acad. Sci. USA* **108**, 5419–5424 (2011).
9. Meunier, D., Lambiotte, R., Fornito, A., Ersche, K. D. & Bullmore, E. T. Hierarchical modularity in human brain functional networks. *Front. Neuroinform.* **3**, 37; DOI:10.3389/neuro.11.037.2009 (2009).
10. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
11. Newman, M. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).
12. Rajan, K. & Abbott, L. Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.* **97**, 188104 (2006).
13. Zhou, Q., Jin, T. & Zhao, H. Correlation between eigenvalue spectra and dynamics of neural networks. *Neural Comput.* **21**, 2931–2941 (2009).
14. Farkas, I. J., Derenyi, I., Barabasi, A. L. & Vicsek, T. Spectra of real-world graphs: beyond the semicircle law. *Phys. Rev. E* **64**, 026704 (2001).
15. Goh, K. I., Kahng, B. & Kim, D. Spectra and eigenvectors of scale-free networks. *Phys. Rev. E* **64**, 051903 (2001).
16. Nadakuditi, R. R. & Newman, M. Spectra of random graphs with arbitrary expected degrees. *Phys. Rev. E* **87**, 012803 (2013).
17. Zhang, P., Krzakala, F., Reichardt, J. & Zdeborová, L. Comparative study for inference of hidden classes in stochastic block models. *J. Stat. Mech-Theory. E.* **2012**, P12021; DOI:10.1088/1742-5468/2012/12/P12021 (2012).
18. Newman, M. The structure and function of complex networks. *SIAM Review* **45**, 167–256 (2003).
19. Humphries, M. & Gurney, K. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS One* **3**, e0002051; DOI: 10.1371/journal.pone.0002051 (2008).
20. Krzakala, F. *et al.* Spectral redemption in clustering sparse networks. *P. Natl. Acad. Sci. USA* **110**, 20935–20940 (2013).
21. Newman, M. Spectral community detection in sparse networks. arXiv:1308.6494 (2013).
22. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**, 066106 (2011).
23. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Lett.* **107**, 065701 (2011).
24. Mossel, E., Neeman, J. & Sly, A. Stochastic block models and reconstruction. arXiv:1202.1499 (2012).
25. Lusseau, D. *et al.* The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **54**, 396–405 (2003).
26. Darst, R. K., Nussinov, Z. & Fortunato, S. Improving the performance of algorithms to find communities in networks. *Phys. Rev. E* **89**, 032809 (2014).
27. Danon, L., Diaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *J. Stat. Mech-Theory. E.* **2005**, P09008; DOI:10.1088/1742-5468/2005/09/P09008 (2005).
28. Lehoucq, R. B. & Sorensen, D. C. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM J. Matrix Anal. Appl.* **17**, 789–821 (1996).

Acknowledgments

This work was supported by a Medical Research Council Senior non-Clinical Fellowship award to M.D.H.

Author contributions

A.S. and M.H. designed the study. A.S. analysed the data and prepared figures. A.S. and M.H. wrote the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Singh, A. & Humphries, M.D. Finding communities in sparse networks. *Sci. Rep.* **5**, 8828; DOI:10.1038/srep08828 (2015).