Mapping directed networks

Jonathan J. Crofts[†,*], Ernesto Estrada[†,‡], Desmond J. Higham[†] and Alan Taylor[†]

[†]*Department of Mathematics & Statistics, University of Strathclyde, Glasgow, UK.*

[‡]*Department of Physics, University of Strathclyde, Glasgow, UK.*

Abstract

We develop and test a new mapping that can be applied to directed unweighted networks. Although not a "matrix function" in the classical matrix theory sense, this mapping converts an unsymmetric matrix with entries of zero or one into a symmetric real-valued matrix of the same dimension that generally has both positive and negative entries. The mapping is designed to reveal approximate directed bipartite communities within a complex directed network; each such community is formed by two set of nodes S_1 and S_2 such that the connections involving these nodes are predominantly *from* a node in S_1 and *to* a node in S_2. The new mapping is motivated via the concept of *alternating walks* that successively respect and then violate the orientations of the links. Considering the combinatorics of these walks leads us to a matrix that can be neatly expressed via the singular value decomposition of the original adjacency matrix and hyperbolic functions. We argue that this new matrix mapping has advantages over other, exponential-based measures. Its performance is illustrated on synthetic data, and we then show that it is able to reveal meaningful directed bipartite substructure in a network from neuroscience.

1 Background and Notation

Large complex networks can be represented as matrices and studied using the tools of linear algebra. Perhaps most notably, spectral information involving eigenvectors or, more generally, singular

[*]Corresponding author: e-mail: ra.jcro@maths.strath.ac.uk, Tel: +44(0) 141 548 3646, Fax: +44(0) 141 548 3345

vectors, can be used for data mining tasks such as clustering, reordering and discovering various types of substructure [1, 6, 9, 13].

We focus here on the case of an unweighted, directed network of N nodes, with no self-loops. This may be represented by the unsymmetric adjacency matrix $A \in \mathbb{R}^{N \times N}$, where $a_{ij} = 1$ if there is a link from node i to node j, and $a_{ij} = 0$ otherwise.

Quantifying bipartite structure in large complex directed networks has proved to be very informative [6, 11, 15], and our aim here is to consider a specific bipartite pattern that takes account of the orientation of the connections in a directed network. If the set of nodes contains two distinct subsets, S_1 and S_2, such that

- the members of S_1 have very few links between themselves,
- the members of S_2 have very few links between themselves,
- there are many links from members of S_1 to members of S_2, and very few other links in the graph involve the nodes of S_1 and S_2,

then we will say that S_1 and S_2 form an *approximate directed bipartite community*. We are interested in the task of identifying one or more of these communities in a network. We emphasize that this concept has been left deliberately vague in order to acknowledge the fact that real networks are typically noisy—in particular, we do not completely rule out "missing" links from S_1 nodes to S_2 nodes and we also allow the possibility of "spurious" links from S_2 to S_1.

In the next section, we motivate and develop a new mapping that is designed to reveal this type of structure, and test it on a synthetic network. Section 3 gives illustrations that compare the new mapping with the matrix exponential function. In section 4 we describe a method for generating networks to test the significance of bipartite subgraphs and in section 5 we implement these tests on synthetic data. In section 6, we show how meaningful information can be extracted from a network in neuroscience.

2 Motivation and New Mapping

We begin with a definition.

Definition 2.1. *An* alternating walk of length $k-1$ from node i_1 to node i_k *is a list of nodes*

$$i_1, i_2, i_3, \ldots, i_k$$

such that $a_{i_s,i_{s+1}} \neq 0$ *for* s *odd, and* $a_{i_{s+1},i_s} \neq 0$ *for* s *even.*

Loosely, an alternating walk is a traversal that successively follows links in the forward and reverse directions.

From the definition of a matrix product it is immediate that

$$\left(AA^TAA^T\cdots\right)_{ij} \tag{1}$$

with k factors, counts the number of alternating walks of length k from node i to node j.

Suppose now that S_1 and S_2 form an approximate directed bipartite community, as described in section 1. If nodes i and j are both in subset S_1 then there is unlikely to be a link from i to j, but there are likely to be many ways to traverse from i to j by following one link forwards and another link backwards. Hence we expect few alternating walks of length one between i and j but many alternating walks of length two. More generally, we would expect an over-abundance of even length alternating walks and a paucity of odd length alternative walks. Incorporating information about longer walks is an intuitively reasonable way to compensate for possible noise in the network—it smooths out the all-or-nothing issue of whether two nodes are connected. However it is clear that shorter length walks are generally more informative. Hence, motivated by previous work on undirected networks [5, 6], we propose to scale the total number of alternating walks of length k by the factor $1/k!$, and to give negative weight to odd length walks, which produces the mapping

$$f(A) = I - A + \frac{AA^T}{2!} - \frac{AA^TA}{3!} + \frac{AA^TAA^T}{4!} - \cdots \tag{2}$$

In words, the i,j element of $f(A)$ for $i \neq j$ is the difference between the total number of even and odd length alternating walks, with walks of length k scaled by $1/k!$. We have included the identity matrix I in (2) simply for convenience. Using the singular value decomposition (SVD), $A = U\Sigma V^T$,

where $U \in \mathbb{R}^{N\times N}$ is orthogonal, $\Sigma \in \mathbb{R}^{N\times N}$ is diagonal and $V \in \mathbb{R}^{N\times N}$ is orthogonal [8], we have

$$f(A) = I - U\Sigma V^T + \frac{U\Sigma^2 U^T}{2!} - \frac{U\Sigma^3 V^T}{3!} + \frac{U\Sigma^4 U^T}{4!} + \cdots,$$

which can be written

$$f(A) = U\left(I + \frac{\Sigma^2}{2!} + \frac{\Sigma^4}{4!} + \cdots\right)U^T - U\left(\Sigma + \frac{\Sigma^3}{3!} + \frac{\Sigma^5}{5!} + \cdots\right)V^T.$$

This could also be written

$$f(A) = U\cosh(\Sigma)\,U^T - U\sinh(\Sigma)\,V^T, \tag{3}$$

which shows that $f(A)$ may be computed via the SVD. We note that $f(A)$ does not comply with the usual definition of a matrix function in linear algebra [10]. However, it is a well–defined mapping from $\mathbb{R}^{N\times N}$ to $\mathbb{R}^{N\times N}$.

Based on this motivation, we would expect $f(A)_{ij}$ to take large positive values when $i, j \in S_1$, and large negative values when $i \in S_1$ and $j \in S_2$.

To test this idea, the picture on the left in Figure 1 shows an adjacency matrix for a 50 node directed network that we constructed where nodes $\{1, 2, \ldots, 10\}$ were made to point to nodes $\{11, 12, \ldots, 25\}$ with independent probability 0.65. Similarly, nodes $\{30, 31, \ldots, 39\}$ point to nodes $\{40, 41, \ldots, 49\}$ with independent probability 0.8, and all other links occur with independent probability 0.05. Hence, there are two approximate directed bipartite communities in the network. In the right of Figure 1 we show a heat map of $f(A)$, and it is clear that the dominant regions of positive and negative values are highlighting the $S_1 \to S_1$ and $S_1 \to S_2$ relationships, respectively, as expected.

We note at this stage that the node ordering in Figure 1 was chosen to make it easy to visualize the results—the communities share contiguous indices. However, it is clear from the derivation, or from the relation $f(PAP^T) = Pf(A)P^T$ for any permutation P, that the same hot/cold values relating two nodes would be preserved under any node reordering. Entirely analogously, we may argue that $f(A^T)$ will have positive entries for $S_2 \to S_2$ relationships and negative for $S_2 \to S_1$. Hence the sum $f(A) + f(A^T)$ should be a useful tool for revealing inter-cluster ($S_1 \to S_1$ and $S_2 \to$

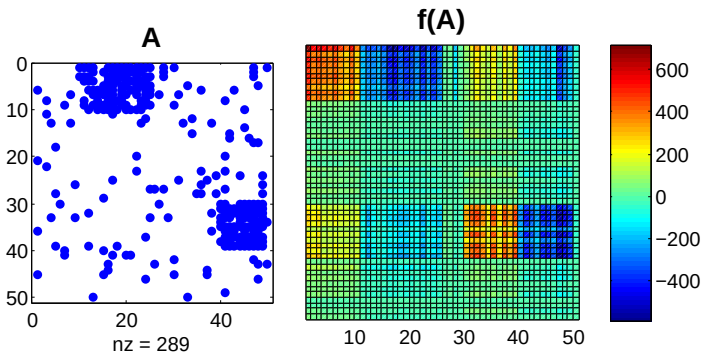

Figure 1: Left: Adjacency matrix. Right: $f(A)$ from (3).

S_2) relationships through positive entries and extra-cluster ($S_1 \to S_2$ and $S_2 \to S_1$) relationships through negative entries. It is straightforward to show that $f(A) + f(A^T)$ is a symmetric matrix, and hence it is amenable to standard clustering techniques, with positively connected clusters representing the common parts of the bipartite communities and negatively-connected clusters representing the disparate parts. We note that the SVD can be used for clustering or reordering this type of symmetric two-signed data into the desired two-by-two checkerboard patterns [9]. Hence, we propose that two separate SVDs may be computed, one to create $f(A) + f(A^T)$ and another to analyse it.

3 Comparison with the Matrix Exponential

In the case of undirected networks, arguments based on the combinatorics of walks between nodes have been used to show that $\exp(A)$ and $\exp(-A)$ can be useful to reveal connectivity patterns [5, 6]. In order to show that the new mapping $f(A) + f(A^T)$ is better suited for pre-processing directed networks, we may consider a hierarchical structure where there are three sets of nodes, S_1, S_2 and S_3, such that

- nodes in S_1 tend to point to nodes in S_2,
- nodes in S_2 tend point to nodes in S_3,
- few of the other possible links are present.

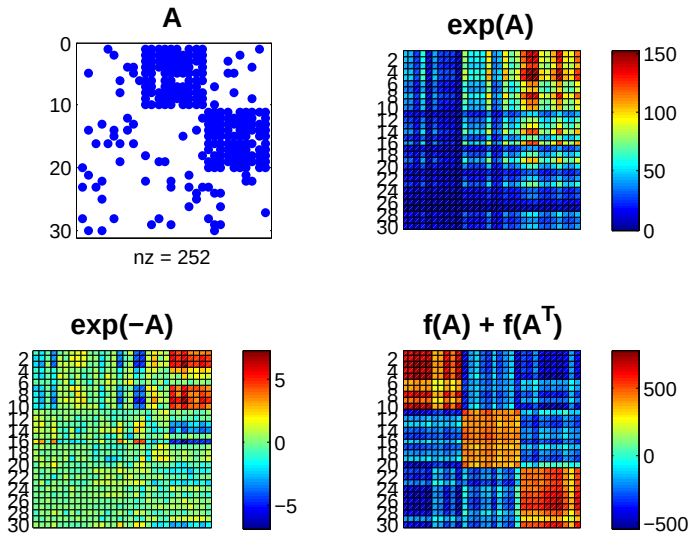

Figure 2: Left: Unsymmetric adjacency matrix A and three different mappings.

Then, considering how they represent counts of walks around the network, we can argue that the exponentials of A and $-A$ will have 3-by-3 block structure of the form

$$\exp(A) \approx \begin{bmatrix} 0 & + & + \\ 0 & 0 & + \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \exp(-A) \approx \begin{bmatrix} 0 & - & + \\ 0 & 0 & - \\ 0 & 0 & 0 \end{bmatrix},$$

whereas $f(A) + f(A^T)$ will take the form

$$f(A) + f(A^T) \approx \begin{bmatrix} + & - & 0 \\ - & + & - \\ 0 & - & + \end{bmatrix}.$$

As an illustration, the upper left picture in Figure 2 shows results for a directed network of 30 nodes where $S_1 = \{1, 2, 3, \ldots, 10\}$, $S_2 = \{11, 12, 13, \ldots, 20\}$, $S_3 = \{21, 22, 23, \ldots, 30\}$. In a similar manner to the network in Figure 1, links were chosen probabilistically with a strong bias towards the directed bipartite community connections.

We note in passing that $f(A)$ can be connected to matrix functions of the original adjacency matrix through the relation

$$f(A) = \cosh(B^{1/2}) - \sinh(B^{1/2})B^{-1/2}A,$$

where $B = AA^T$.

In the next two sections we address the issue of judging whether results from the algorithm are significant. On one hand, it is unrealistic to expect that all nodes in a real network can be partitioned into two sets, S_1 and S_2, such that all $S_1 \to S_2$ links are present and no others. On the other hand, simply identifying a pair of nodes i and j such that $a_{ij} = 1$ and $a_{ji} = 0$ is clearly not of interest. We will use the classic notion of a p-value to assess the question "How likely is it that the level of bipartivity identified by the algorithm would arise in an arbitrary network of the same form?" Perhaps the most widely-used random graph classes are the Erdös-Rényi (ER) and Gilbert models [4, 7]. However, it is intuitively clear, and easy to check experimentally, that networks from these classes are extremely unlikely to admit bipartite substructure. Hence, any attempt to fit this type of model to a given network, for example, by matching the expected total in and out degrees, is likely to give a favourable p-value. In an attempt to deal with this, in the next section we develop a new class of directed random networks that are designed to match, in expectation, in and out degrees specified for each node.

4 Directed Stickiness Model

For each node i we define two quantities, $\theta_{\text{in}}^{[i]}$ and $\theta_{\text{out}}^{[i]}$, such that these are a measure of the likelihood of that node having a connection to/from another node in a particular direction. We define the probability of a connection from node i to node j as the product

$$\mathbb{P}(i \to j) = \theta_{\text{out}}^{[i]}\theta_{\text{in}}^{[j]}.$$

Our aim is that the expected out-degree of node i in the model matches the out-degree of node i in the initial data.

This requires

$$\begin{aligned}\sum_{j=1}^{n} a_{ij} &= \mathbb{E}(\text{out-degree of node } i) \\ &= \sum_{j=1}^{n} \theta_{\text{out}}^{[i]} \theta_{\text{in}}^{[j]} \\ &= \theta_{\text{out}}^{[i]} \sum_{j=1}^{n} \theta_{\text{in}}^{[j]}.\end{aligned}$$

But since $\sum_{j=1}^{n} \theta_{\text{in}}^{[j]}$ does not depend on i, this shows that

$$\theta_{\text{out}}^{[i]} \propto \sum_{j=1}^{n} a_{ij}.$$

So let

$$\theta_{\text{out}}^{[i]} = \frac{1}{K_1} \sum_{j=1}^{n} a_{ij}. \tag{4}$$

Similarly we wish the expected in-degree of node i in the model to match the in-degree of node i in the data, giving

$$\begin{aligned}\sum_{j=1}^{n} a_{ji} &= \mathbb{E}(\text{in-degree of node } i) \\ &= \sum_{j=1}^{n} \theta_{\text{out}}^{[j]} \theta_{\text{in}}^{[i]} \\ &= \theta_{\text{in}}^{[i]} \sum_{j=1}^{n} \theta_{\text{out}}^{[j]}.\end{aligned}$$

Since $\sum_{j=1}^{n} \theta_{\text{out}}^{[j]}$ does not depend on i, we have

$$\theta_{\text{in}}^{[i]} \propto \sum_{j=1}^{n} a_{ji},$$

so we let

$$\theta_{\text{in}}^{[i]} = \frac{1}{K_2} \sum_{j=1}^{n} a_{ji}. \tag{5}$$

Having determined their general form, we now wish to find appropriate constants of proportionality, K_1 and K_2, such that the expected in and out-degrees in the model match those of the

initial data. Returning to the out-degree of node i, using Equations 4 and 5, we require

$$\sum_{j=1}^{n} a_{ij} \quad = \quad \left(\frac{1}{K_1}\sum_{j=1}^{n} a_{ij}\right)\left(\frac{1}{K_2}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{kj}\right),$$

which leads to

$$\frac{1}{K_1 K_2}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{kj} = 1. \tag{6}$$

Likewise, considering the in-degree of node i we require

$$\sum_{j=1}^{n} a_{ji} \quad = \quad \left(\frac{1}{K_2}\sum_{j=1}^{n} a_{ji}\right)\left(\frac{1}{K_1}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}\right),$$

which becomes

$$\frac{1}{K_1 K_2}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk} = 1. \tag{7}$$

Since the summations in Equations 6 and 7 involve all entries in the adjacency matrix, they are equivalent, and we arrive at

$$K_1 = K_2 = \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}}. \tag{8}$$

We can now write down an algorithm to produce an instance of such a random graph.

- Input $\deg_{\text{in}}$ and $\deg_{\text{out}}$, vectors of in/out degrees.
- Compute the scaling factor $w = \sqrt{\sum_i \deg_{\text{in}}^{[i]}}$.
- Let $\theta_{\text{in}}^{[i]} = w^{-1}\deg_{\text{in}}^{[i]}$ and $\theta_{\text{out}}^{[i]} = w^{-1}\deg_{\text{out}}^{[i]}$.
- For each pair of nodes i and j, connect i to j with independent probability $\theta_{\text{out}}^{[i]} \times \theta_{\text{in}}^{[j]}$.

We emphasize that this is a natural generalization of the original stickiness model in [16] to the case of directed edges.

5 Statistical Analysis

In order to quantify the likelihood of a given directed bipartite structure arising by chance, we conduct statistical tests based upon a simple measure of bipartivity. Consider a perfectly bipartite

network consisting of two sets S_1 and S_2 containing m_1 and m_2 nodes respectively. Such a network

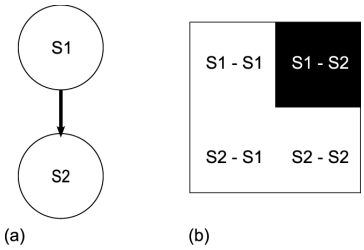

Figure 3: Directed bipartite network: (a) edge structure (b) adjacency matrix

may be represented by an adjacency matrix with a single off-diagonal nonzero block consisting of the edges from S_1 to S_2 as shown in Figure 3. To quantify bipartivity in this subgraph we simply take the ratio of the density of nonzeros in the $S_1 \rightarrow S_2$ block to the density of nonzeros in the remaining L-shaped block plus one (to avoid division by zero) i.e.

$$b = \frac{|S_{12}|/mn}{(|S_{11}| + |S_{21}| + |S_{22}|)/(m^2 + mn + n^2) + 1}.$$

Here $|S_{ij}|$ denotes the number of links from $S_i \rightarrow S_j$ and m, n are the number of nodes in S_1 and S_2 respectively. In the case of perfect directed bipartivity, this measure yields a value of 1. The value decreases as nonzeros are added to the L-shaped block or removed from the $S_1 \rightarrow S_2$ block. This is analagous to adding edges in the "wrong" direction or edges within subsets.

Once a measure of bipartivity has been established for a given subgraph, random graphs with the same expected degree distribution are generated and tested as outlined in the following algorithm;

1. A directed graph with expected degree distribution the same as the original adjacency matrix A is generated.

2. The mapping $f(A) + f(A^T)$ is applied to this matrix and it is reordered according to the first eigenvector of the mapped matrix.

3. A subgraph consisting of sets of the same dimension as S_1 and S_2 is selected and the bipartivity measure is applied to this subgraph.

Once this process has been repeated for a sufficiently large test set, the bipartivity values are plotted

in a histogram. The data is plotted in a quantile-quantile plot, verifying that it fits a log-normal probability distribution. The p-value for the original subgraph may then be computed.

$$p = 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(- \frac{(u-\mu)^2}{2\sigma^2} \right) du$$

where x is the bipartivity measure of our subgraph and μ and σ are the mean and standard deviation respectively of the bipartivity measure on our sample of networks.

5.1 Test case 1

We consider a synthetic network consisting of 100 nodes where a connection between node i and node j occurs with independent probability 0.9 if $i \in \{1, 2, \ldots, 10\}$ and $j \in \{11, 12, \ldots, 20\}$, and with probability 0.3 elsewhere. We then compute the matrix mapping and plot the reordered, mapped matrix to determine what dimension of subgraph to extract as shown in Figure 4. The subgraph comprising the first 20 and last 20 nodes is plotted. 1000 random graphs are then generated by randomly shuffling the in and out degrees of the synthetic network and connecting nodes accordingly. Subgraphs of the correct size are extracted from each of these graphs, their bipartivity measures are computed and the results are plotted in a histogram. The probability distribution of this data was found to be log-normal and thus the p-value for our original subgraph may be computed. For the subgraph obtained from our synthetic network, this process yields a bipartivity score of 0.6138. We measure the significance of this subgraph in two ways. First we compute the quantity $p_1 = p$ as defined at the beginning of this section. Secondly we compute p_2 which is simply the ratio of bipartivity scores larger than that of our subgraph to the sample size. These values will vary depending on the class of graphs from which our samples are drawn. We use four such classes. Standard ER graphs, directed stickiness graphs as defined in section 4 and two variations of the stickiness model. The first of these is 'shuffled stickiness' in which we randomly permute the in and out degree vectors so that node i in the random graph may draw its expected in and out degrees from two different nodes in the original network. The second is 'biased stickiness' where the in and out degrees are assigned in an ordered manner so that the node with the highest in degree will also have the highest out degree and so on. The results of these tests are listed in Table 1.

	p_1	p_2
Erdös-Rényi	0	0/1000
Directed stickiness	2.33×10^{-15}	0/1000
Shuffled stickiness	5.05×10^{-15}	0/1000
Biased stickiness	0	0/1000

Table 1: Significance of subgraph in test case 1 for varying test matrix classes.

5.2 Test case 2

We now consider a test case in which a directed bipartite structure is embedded in a random graph with 100 nodes. In this case, the independent probability of a link between one of nodes 1-20 to one of nodes 21-40 is 0.8, whereas the probability of a link elsewhere is 0.4. Selecting a 40 by 40 subgraph and testing in the same manner as before as shown in Figure 5, a bipartivity score of 0.4083 is obtained. The various p-values are listed in Table 2.

	p_1	p_2
Erdös-Rényi	5.91×10^{-2}	55/1000
Directed stickiness	8.46×10^{-1}	844/1000
Shuffled stickiness	9.03×10^{-1}	896/1000
Biased stickiness	7.00×10^{-2}	57/1000

Table 2: Significance of subgraph in test case 2 for varying test matrix classes.

6 Worm Brain Network

To assess the usefulness of the new mapping we consider two real-world networks: (i) the global neuronal network of the nematode (roundworm) *Caenorhabditis elegans*, and (ii) a local subnetwork of 131 frontal neurons of the same organism; see [12] and the website "http://www.biological-networks.org/?page_id2". To obtain a directed network we firstly removed all gap junctions from the data sets, this step is necessary, as experimental techniques used to reconstruct the nervous system of *C. elegans* are unable to infer directionality of such connections. After non-neuronal cells are removed, this results in a local network of 131 neurons and 964 chemical synapses, and a global

network of 191 neurons and 1904 chemical synapses.

Our motivation is that Durbin [3, Figure 8.1] used an ad hoc combinatoric algorithm to search for and display the type of directed bipartite structure that we consider. From a biological viewpoint, this allows us to consider questions such as

> "what is the processing depth from sensory input to motor output, i.e. how many intermediary neurons are there?, and to what extent is the circuitry unidirectional, progressing linearly from input to output?[3]"

In this preliminary work, we are simply using the worm brain network to demonstrate that the new mapping gives a systematic way to discover this type of important structure.

Figure 6 shows how the new mapping can be used to extract useful information from a real network. The upper left plot gives the original adjacency matrix for the local connectivity network. A heat map for $f(A)+f(A^T)$ highlighted certain node pairs as being hot or cold. Applying the SVD to this matrix, and reordering with the first singular vectors to cluster the hot and cold regions [9] produces the lower picture. We see that tight clusters have emerged via contiguous nodes at each end of the new ordering. In the upper right picture, we have picked out the corresponding nodes and plotted the resulting subnetwork. Here the $S_1 \to S_2$ submatrix is respectively, 5, 35 and 9 times more dense than the $S_1 \to S_1$, $S_2 \to S_1$ and $S_2 \to S_2$ subnetworks. Performing a similar analysis on the global network, allows us, in analogous fashion, to pick out two sets of contiguous nodes such that the $S_1 \to S_2$ matrix is respectively, 3, 80 and 27 times more dense than the $S_1 \to S_1$, $S_2 \to S_1$ and $S_2 \to S_2$ subnetworks.

Statistical significance was obtained by fitting a log-normal distribution to the bipartivity scores of 1000 instances of each of the random graph models, as described in the previous section. This was repeated for both the local and global neuronal networks of *C.elegans*. In the case of ER, we found that both patterns were deemed significant and unlikely to arise by chance, see Table 3. However, it is well known that the distribution for both the in and out degrees of the *C.elegans* network deviate significantly from the Poissonian distribution of ER random graphs [17], so perhaps this should not be too surprising. Using the three variants of the stickiness model, we find that the bipartite structure discovered for the local *C. elegans* network is deemed significant only in the case of the 'biased' stickiness model. Note however, that this apparent lack of significance,

	RG model	p_1	p_2
Local			
	Erdos-Renyi	6.81×10^{-5}	0/1000
	Directed stickiness	0.9556	951/1000
	Shuffled stickiness	0.9683	958/1000
	Biased stickiness	0.0468	42/1000
Global			
	Erdos-Renyi	8.47×10^{-12}	0/1000
	Directed stickiness	5.82×10^{-5}	0/1000
	Shuffled stickiness	5.15×10^{-5}	0/1000
	Biased stickiness	8.63×10^{-7}	0/1000

Table 3: Significance of subgraphs found for the local and global networks for *C. elegans* using varying test matrix classes

is due mainly to the fact that, although the $S_1 \to S_2$ matrix has many more connections than the $S_1 \to S_1$, $S_2 \to S_1$ and $S_2 \to S_2$ subnetworks, it is still relatively sparse, thus resulting in a low bipartivity score ($b_{\mathrm{L}} = 0.2645$). For the global network, the connectivity pattern determined remains significant when the differing stickiness models are used ($b_{\mathrm{G}} = 0.6415$ and $p < 0.01$ in all cases; again see Table 3 for the details).

The neuronal classes [1] that were picked out by the algorithm along with a description of their respective functionalities are given in Tables 4 and 5.

For the local neural network of *C. elegans*, neurons contained within S_1 were mainly involved in sensory processes (approximately 65%), whilst those in S_2 involved a mixture of motor neurons and so called 'command' interneurons. Similarly for the global *C. elegans* network, we found that S_1 consisted of a mixture of sensory neurons and nerve ring interneurons, whilst S_2 was made up entirely of command interneurons. Note that in [3], Durbin attempted to vertically order the neuronal classes in the nerve ring of *C. elegans*, in such a way, that as many as possible of the synapses pointed downwards. The resultant ordering placed sensory neurons towards the top, motor neurons towards the bottom, and the remaining interneurons in between. Note that the bipartite structures that we have picked out are in good agreement with the highly directed, hierarchic picture presented by Durbin.

On closer inspection, 60% of neurons contained within S_2 for the local *C. elegans* network, and all neurons belonging to S_2 for the global neural network, were found to belong to a group of neurons

[1] For simplicity we present the combined results for neuronal classes rather than individual cells.

	Neuronal Class	Description
S_1	OLL	Head sensory neuron
	URY	Head sensory neuron
	IL2	Head sensory neuron
	RIH	Ring interneuron
	ASH	Amphids; sensory neuron
	RIM	Ring motor neuron
	RIV	Ring motor/interneuron
	CEP	Head sensory neuron
	AVH	Interneuron
	ADL	Amphids; sensory neuron
S_2	SMD	Ring motor neuron
	RME	Ring motor neuron
	RMD	Ring motor neuron
	AVB	Command interneuron
	AVA	Command interneuron
	AVE	Command interneuron
	AVD	Command interneuron

Table 4: Neuronal class and type for bipartite subgraph found in the local network of 131 frontal neurons of *C. elegans*.

termed the *lateral ganglion* which are known to be highly interconnected with both sensory and motor neurons - particularly those motor neurons in the ventral cord. Indeed, it has been suggested that the lateral ganglion is the principal pathway between sensory and motor components of the nematode *C. elegans* [2]. In addition, the neuronal classes AVA, AVB, AVD and AVE, which were picked out both in the local and global networks, have been previously identified as 'hub' or 'centre' neurons that are essential for normal biological function [14]. For example, it is well known that both AVA and AVB neurons are necessary for normal coordinated movement.

7 Conclusions

This paper addresses the problem of determining *directed bipartite structures* within complex networks via a new matrix mapping. Initial tests on a network from neuroscience show that the new mapping can be used to successfully infer biologically relevant information using only the network topology. We emphasise the importance of choosing the correct random graph model by comparing

	Neuronal Class	Description
S_1	DVA	Interneuron
	FLP	Sensory neuron
	DVC	Ring interneuron
	PVP	Interneuron
	ADL	Amphids; sensory neuron
	AIM	Ring interneuron
	ADE	Anterior deirid; sensory neuron
	ASH	Amphids; sensory neuron
	AQR	Sensory neuron
	ADA	Ring interneuron
	AVM	Sensory neuron
S_2	AVA	Command interneuron
	AVB	Command interneuron
	AVD	Command interneuron
	AVE	Command interneuron

Table 5: Neuronal class and type for bipartite subgraph found in the global network of 191 neurons of the *C. elegans*.

statistics for several such models. In particular, we see that the 'significance' or 'non-significance' of the determined connectivity patterns can be extremely sensitive to the class of random matrices chosen. In future work in this area we plan to develop automated algorithms for discovering and quantifying the quality of approximate directed bipartite communities and to test these ideas on further real life data sets.

Acknowledgements

We are very grateful to Markus Kaiser for kindly providing the directed connectivity data and for valuable feedback.

References

[1] D. Barash. Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation. *Bioinformatics*, 20:1861–1869, 2004.

[2] N. Chaterjee and S. Sinha. Understanding the mind of a worm: hierarchical network structure underlying nervous system function in *C. elegans*. *Progress in Brain Research*, 168:145–153, 2008.

[3] Richard Michael Durbin. *Studies on the Development and Organisation of the Nervous System of Caenorhabditis Elegans*. PhD thesis, University of Cambridge, 1987.

[4] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[5] E. Estrada and N. Hatano. Communicability in complex networks. *Physical Review E*, 77:036111, 2008.

[6] E. Estrada, D. J. Higham, and N. Hatano. Communicability and multipartite structures in complex networks at negative absolute temperatures. *Physical Review E*, 78:026102, 2008.

[7] A. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–1144, 1959.

[8] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, third edition, 1996.

[9] D.J. Higham, G. Kalna, and J.K. Vass. Spectral analysis of two-signed microarray expression data. *IMA Mathematical Medicine and Biology*, 24:131–148, 2007.

[10] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[11] Petter Holme, Fredrik Liljeros, Christofer R. Edling, and Beom Jun Kim. Network bipartivity. *Physical Review E*, 68:056107, 2003.

[12] M. Kaiser and C. C. Hilgetag. Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS Computational Biology*, 2:e95, 2006.

[13] C. Kamp and K. Christensen. Spectral analysis of protein-protein interactions in Drosophila melanogaster. *Physical Review E*, 71:041911, 2005.

[14] S. Morita, K. Oshio, Y. Osana, Y. Funabashi, K. Oka, and K. Kawamura. Geometrical structure of the neuronal network of *Caenorhabditis elegans*. *Physica A*, 298:553–561, 2001.

[15] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert. A lock-and-key model for protein-protein interactions. *Bioinformatics*, 2:2012–2019, 2006.

[16] N. Pržulj, D. G. Corneil, and I. Jurisica. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics*, 22:974–980, 2006.

[17] M. Reigl, U. Alon, and D. B. Chklovskii. Search for computational modules in the *C. elegans* brain. *BMC Biology*, 2:25, 2004.

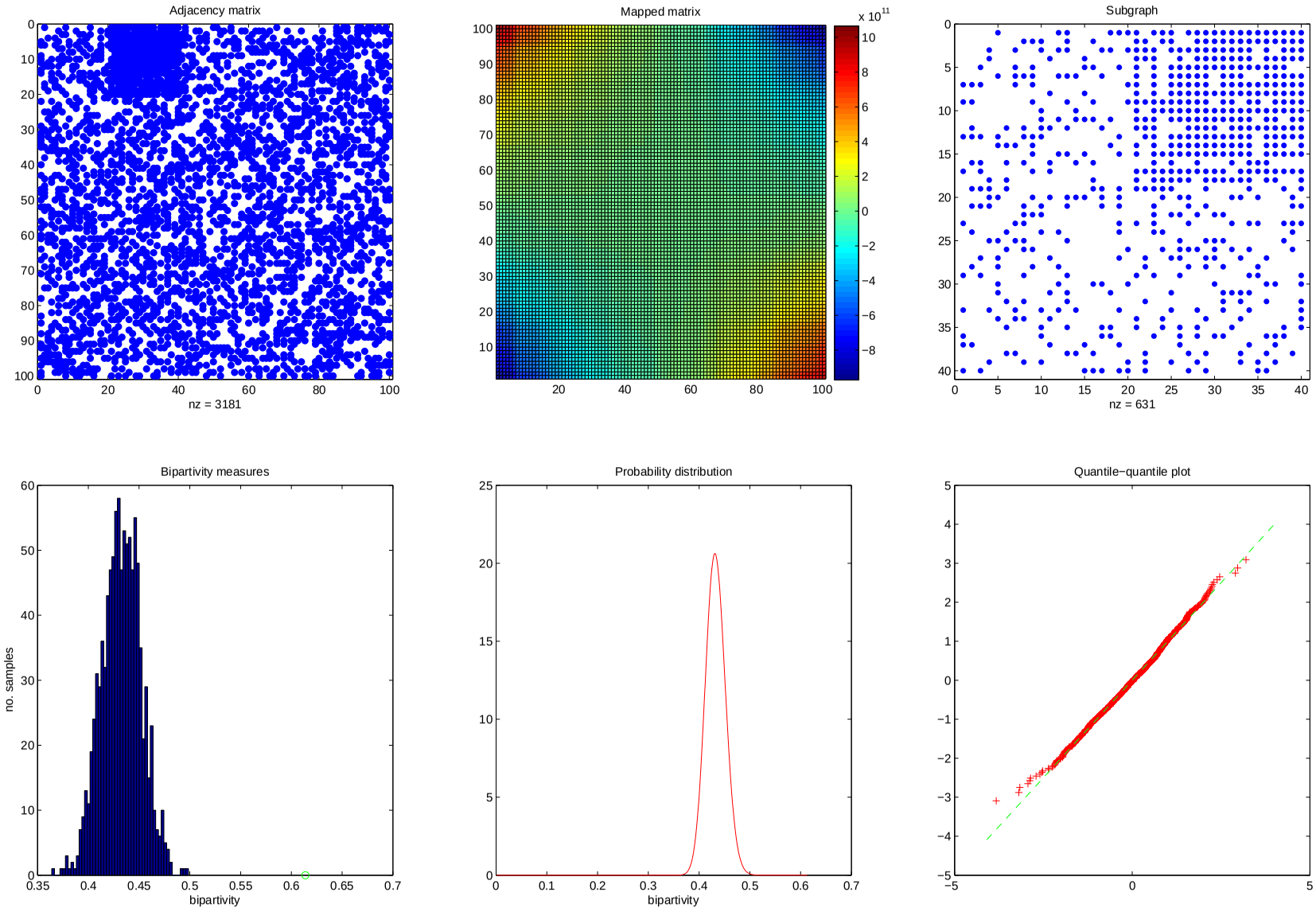

Figure 4: Results for synthetic network 1: (a) adjacency matrix; (b) mapped matrix; (c) subgraph; (d) histogram of bipartivity values for 1000 graphs with same expected degree distribution; (e) fitted probability distribution; and (f) quantile-quantile plot.

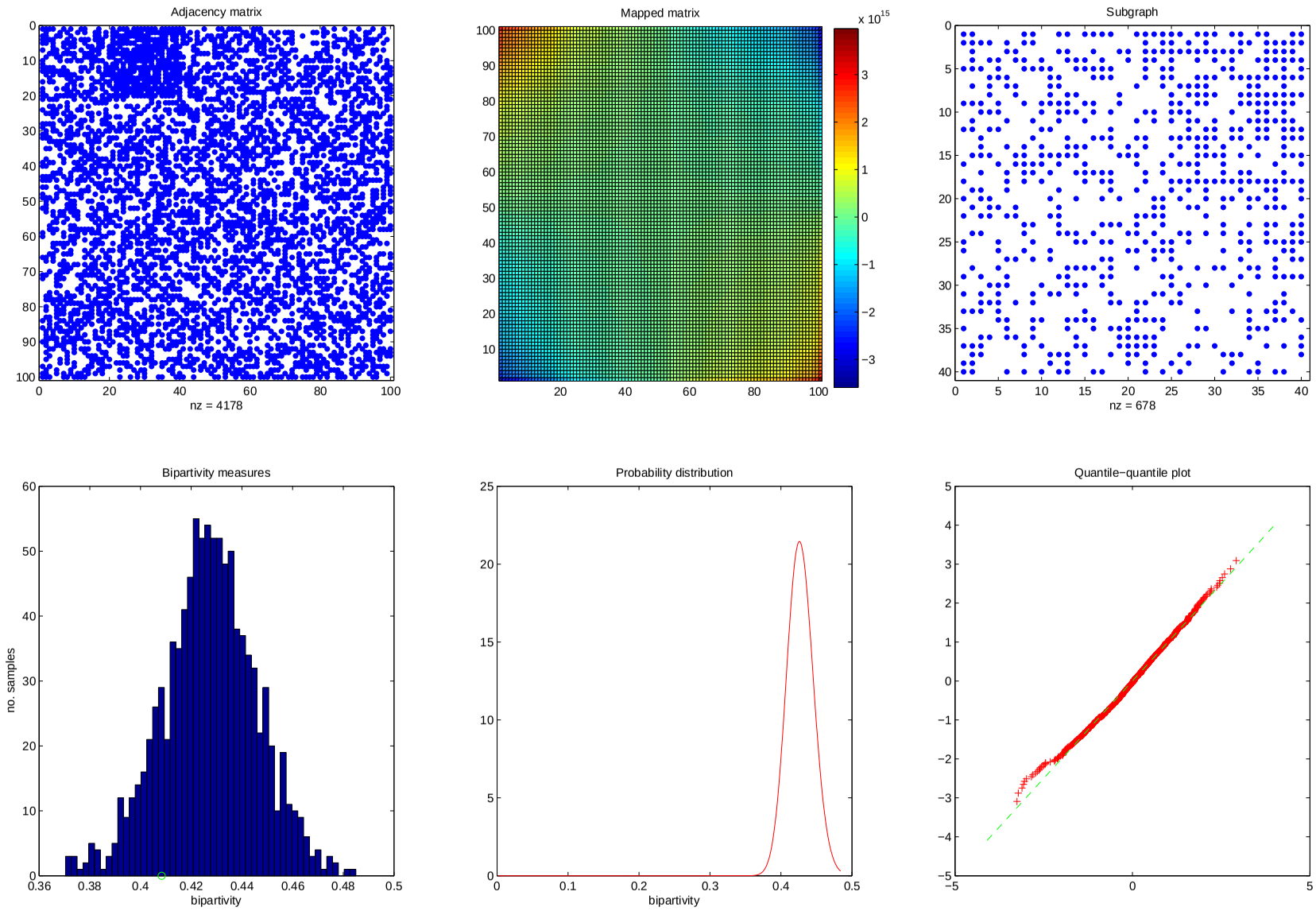

Figure 5: Results for synthetic network 2: (a) adjacency matrix; (b) mapped matrix; (c) subgraph; (d) histogram of bipartivity values for 1000 graphs with same expected degree distribution; (e) fitted probability distribution; and (f) quantile-quantile plot.

.

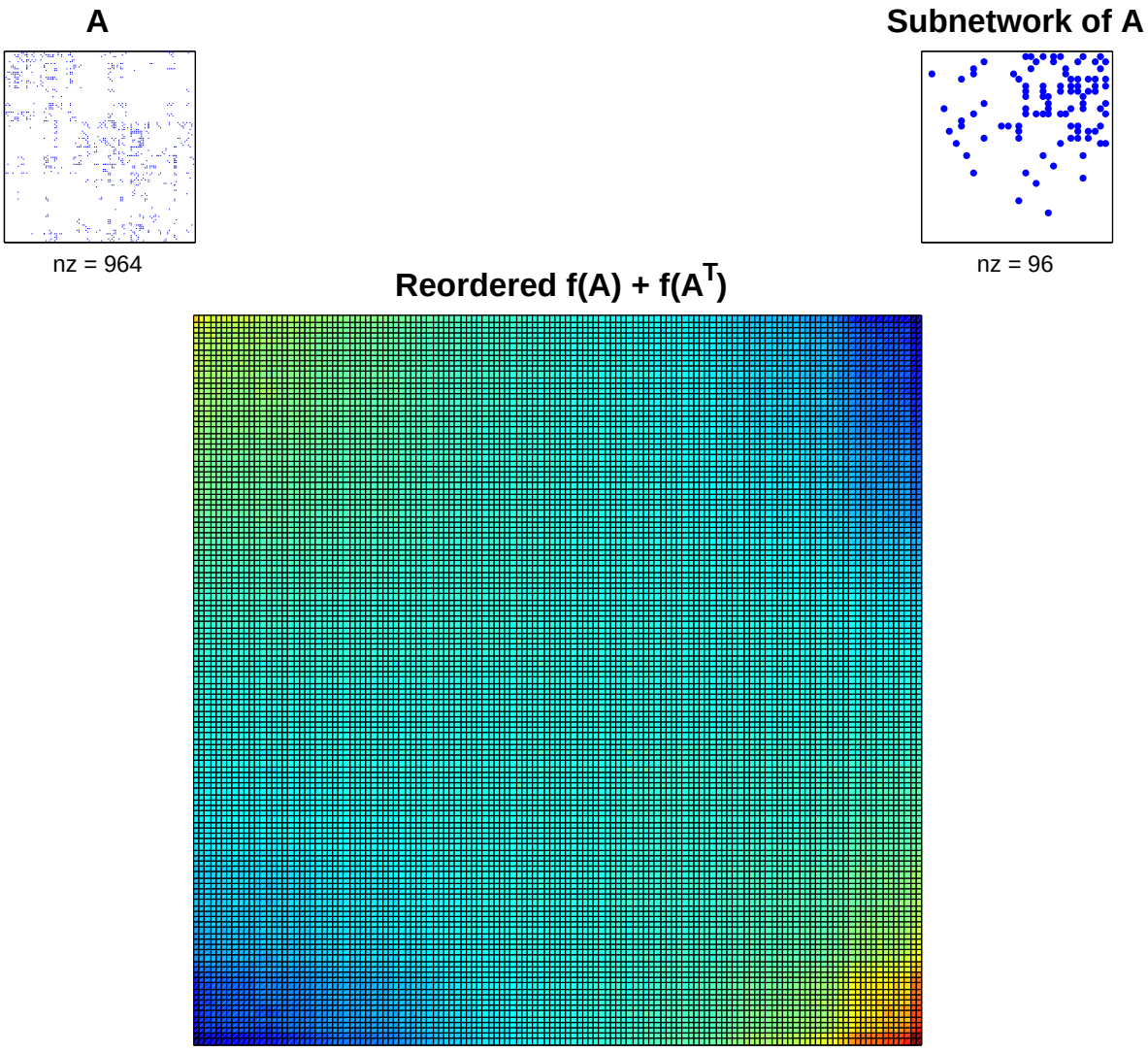

Figure 6: Upper left: worm neural network with 131 nodes. Lower: reordered version of $f(A) + f(A^T)$. Upper right: subnetwork of 32 nodes obtained from the reordering.