Electronic Transactions on Numerical Analysis.
Volume 37, pp. 337-350, 2010.

ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

MAPPING DIRECTED NETWORKS*

JONATHAN J. CROFTS†, ERNESTO ESTRADA†‡, DESMOND J. HIGHAM†, AND ALAN TAYLOR†

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.

Key words. bipartivity, clustering, communities, exponential, networks, neuroscience, stickiness

AMS subject classifications. 65F60, 05C50

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 vectors, can be used for data mining tasks such as clustering, reordering and discovering various types of substructure [2, 7, 11, 15].

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 [7, 13, 17], 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 network 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 Section 2, 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.

*Received December 4, 2009. Accepted May 4, 2010. Published online November 9, 2010. Recommended by A. Frommer.

†Department of Mathematics & Statistics, University of Strathclyde, Glasgow, UK (jonathan.crofts,ernesto.estrada,d.j.higham,a.taylor@strath.ac.uk)

‡Department of Physics, University of Strathclyde, Glasgow, UK.

ETNA
Kent State University
http://etna.math.kent.edu

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. We emphasize that the nodes and edges that make up an alternating walk need not be distinct.

From the definition of a matrix product it is immediate that

$$\left(AA^TAA^T\cdots\right)_{ij} \tag{2.1}$$

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 alternating 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 [6, 7], 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

$$f(A) = I - A + \frac{AA^T}{2!} - \frac{AA^TA}{3!} + \frac{AA^TAA^T}{4!} - \cdots . \tag{2.2}$$

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 (2.2) 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 [10], we have

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

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

$$f(A) = U\cosh\left(\Sigma\right)U^T - U\sinh\left(\Sigma\right)V^T, \tag{2.3}$$

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 [12]. However, it is a well–defined mapping from $\mathbb{R}^{N\times N}$ to $\mathbb{R}^{N\times N}$.

ETNA
Kent State University
http://etna.math.kent.edu

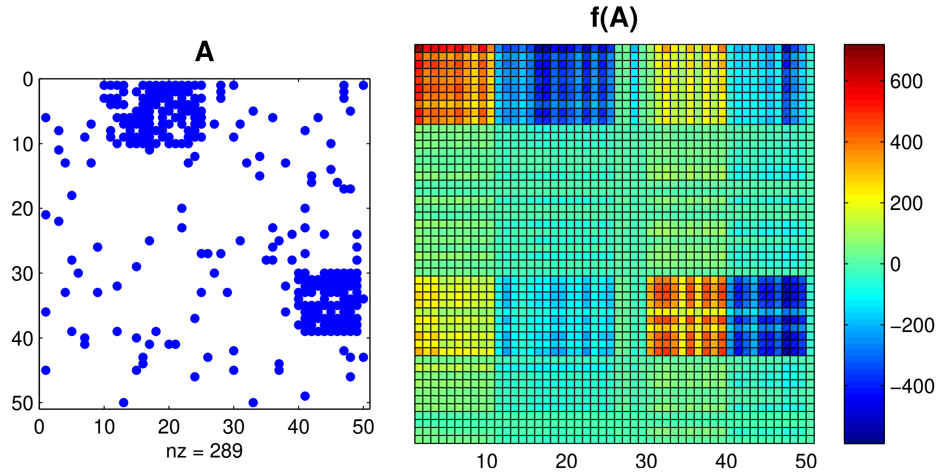

FIG. 2.1. *Left: Adjacency matrix. Right: $f(A)$ from (2.3).*

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 2.1 shows an adjacency matrix for a 50 node directed network that we constructed. Here 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 2.1 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 2.1 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 $S_2 \to S_2$) 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 [11]. Hence, we propose that two separate SVDs may be computed, one to create $f(A) + f(A^T)$ and another to analyze 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 [6, 7]. 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

ETNA
Kent State University
http://etna.math.kent.edu

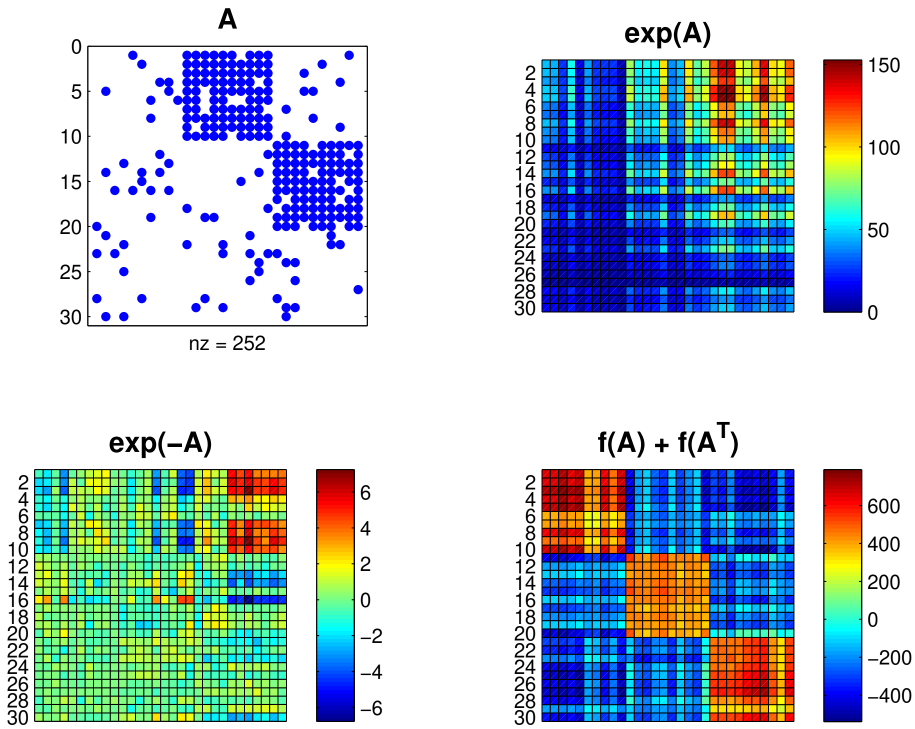

FIG. 3.1. *Left: Unsymmetric adjacency matrix A and three different mappings.*

- 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.

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 pictures in Figure 3.1 shows results for a directed network of 30 nodes where $S_1 = \{1, 2, \dots, 10\}$, $S_2 = \{11, 12, \dots, 20\}$, $S_3 = \{21, 22, \dots, 30\}$. In a similar manner to the network in Figure 2.1, links were chosen probabilistically with a strong bias towards the directed bipartite community connections.

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 pvalue [8] to address the

ETNA
Kent State University
http://etna.math.kent.edu

question "How likely is it that the level of bipartivity identified by the algorithm in a given network 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 [5, 9]. 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 the given network is likely to give a pvalue that indicates statistical significance for the observed pattern. In an attempt to produce a more realistic test, in the next section we develop a new class of directed random networks based on an established modeling principle, 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 stickiness indices, $\theta_{\text{in}}^{[i]}$ and $\theta_{\text{out}}^{[i]}$, that summarize the likelihood of that node having a connection to/from another node in a particular direction. More precisely, we define the probability of a connection from node i to node j as the product

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

This is a natural generalization of the original stickiness model in [18], which was defined for undirected networks. In that case, the stickiness index was justified from a modelling perspective in the context of protein-protein interaction networks.

Our first aim is to choose $\{\theta_{\text{in}}^{[i]}, \theta_{\text{out}}^{[i]}\}$ so that the expected out-degree of node i in the model matches the out-degree of node i in the given network. This requires

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

We may then write

$$\theta_{\text{out}}^{[i]} = \frac{1}{K_1} \sum_{j=1}^{n} a_{ij}, \tag{4.1}$$

for some constant K_1.

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

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

Hence, we may write

$$\theta_{\text{in}}^{[i]} = \frac{1}{K_2} \sum_{j=1}^{n} a_{ji}, \tag{4.2}$$

for some constant K_2.

Having determined these general forms, we now wish to find appropriate constants of proportionality, K_1 and K_2. Returning to the out-degree of node i, using (4.1) and (4.2), we require

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

ETNA
Kent State University
http://etna.math.kent.edu

which leads to

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

Considering the in-degree of node i leads to the same condition. We thus arrive at the unique choice

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

We may now outline an algorithm to produce an instance of such a random graph.

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

Of course, for the model to be valid we require all probabilities to be bounded above by one. This can be guaranteed, for example, if the product of the largest in-degree and the largest out-degree is less than the total number of edges in the target network. This constraint was satisfied by the networks that we consider here, and we would not expect it to pose any difficulties in general.

5. Statistical analysis. Having discovered a directed bipartite substructure in a given network, in order to quantify the likelihood of this pattern arising by chance, we must also quantify the level 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 may be represented

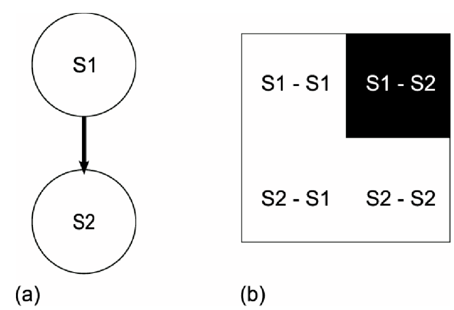

FIG. 5.1. *Directed bipartite network: (a) edge structure (b) adjacency matrix.*

by an adjacency matrix with nonzeros only in an off-diagonal block representing the edges from S_1 to S_2, as shown in Figure 5.1. In practice, there will be some departure from this perfect division, and as our measure of bipartivity we will simply take the ratio of the density of nonzeros in the $S_1 \to S_2$ block to the density of nonzeros in the remaining L-shaped block plus one (to avoid division by zero); that is,

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

ETNA
Kent State University
http://etna.math.kent.edu

Here $|S_{ij}|$ denotes the number of links from $S_i \to S_j$ 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 $S_1 \to S_2$ block. This is analogous to adding edges in the "wrong" direction or edges within subsets.

Once a measure of bipartivity has been chosen for the given subgraph, we may test for significance as follows:

1. Sample a network from an appropriate distribution to give an adjacency matrix $\widehat{A}$.
2. Apply the mapping $f(\widehat{A}) + f(\widehat{A}^T)$ and reorder as for the original network; that is, according to the first eigenvector of the mapped matrix.
3. Use this eigenvector to select the subgraph consisting of sets of the same dimension as S_1 and S_2, and compute the bipartivity measure.

Following a standard hypothesis testing approach, we may then compute a pvalue as the frequency with which the bipartivity measure of the given network exceeds that of a randomly sampled network. This gives one way to answer the question "What is the likelihood that we would observe this level of bipartivity, or higher, in a random network?" Following convention, we will regard a pvalue below 0.05 as an indication of a statistically significant result. Of course, the computed pvalue is dependent on the choice bipartivity measure and the class of random networks used. We tested several variations and present here an indicative summary. We will refer to such a frequency-based pvalue as p_1. In further testing we examined histograms of the sampled bipartivity measures, and, based on quantile-quantile plots, found that a lognormal distribution seemed to be appropriate. For each test, we therefore also compute a second pvalue, p_2, found by fitting a lognormal distribution to the sampled data and using the resulting density function to compute the probability that the given network's bipartivity measures will be exceeded. Both approaches are illustrated in the following subsections.

5.1. Test case 1. As a proof of concept, we begin with a synthetic network with known substructure. Here, we have 100 nodes. A connection from node i to j occurs with independent probability 0.9 if $i \in \{1, 2, \ldots, 20\}$ and $j \in \{21, 22, \ldots, 40\}$, 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. The upper pictures in Figure 5.2 show the original adjacency matrix, a heat map of the reordered mapped matrix, and the relevant subgraph comprising the first and last 20 nodes of the relevant eigenvector indices. (In a separate check we found that 85% of the first 20 nodes were in the "correct" set $\{1, 2, \ldots, 20\}$ and, similarly, 85% of the second 20 nodes were in the "correct" set $\{21, 22, \ldots, 40\}$.) We then generated 1000 networks by randomly shuffling the in and out degrees of the given network and connecting nodes according to the algorithm in Section 4. So a node in the random graph typically draws its expected in and out degrees from two different nodes in the original network, but all in and out degrees in the original network are represented in the new graph. The resulting bipartivity measure samples are displayed in a histogram, with the bipartivity value of 0.6138 for the given network indicated as a circle on the x-axis. A lognormal fit to the data is shown, along with a quantile-quantile plot against a lognormal distribution—here data on a straight line is indicative of a good fit [1]. It is clear from the histogram that the given network produces a bipartivity measure deep in the low probability tail of the distribution, and this is reflected in the frequency-based pvalue of $p_1 = 0/1000$ and lognormal version $p_2 = 5.05 \times 10^{-15}$. Table 5.1 summarizes results for other types of randomization. In addition to using the stickiness model from Section 4 to form the "shuffled stickiness" class of networks, we also formed "directed stickiness" networks where node i has the expected in and out degree of the corresponding node in the given network, and "biased stickiness" net-

ETNA
Kent State University
http://etna.math.kent.edu

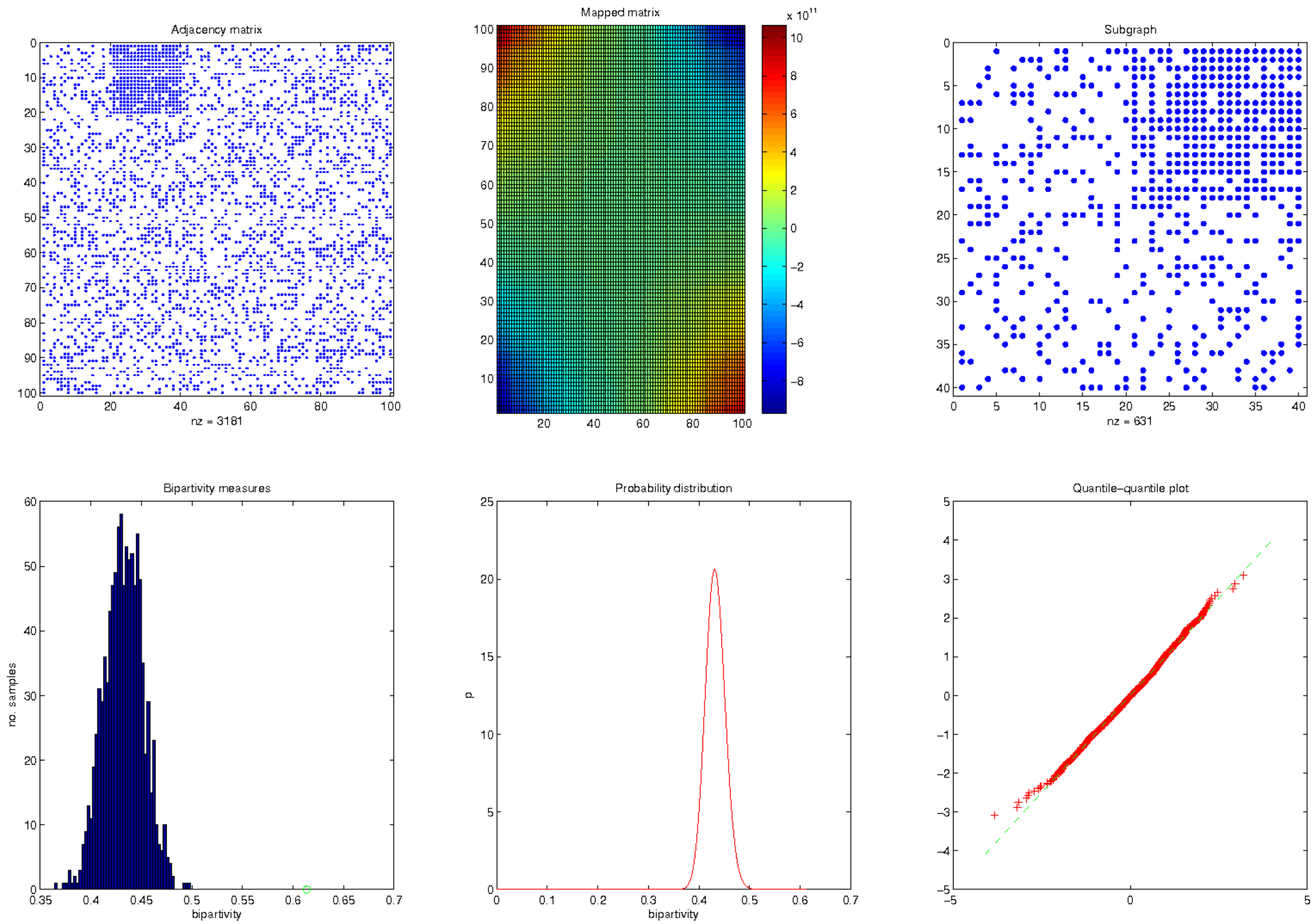

FIG. 5.2. *Results for synthetic network 1: (a) adjacency matrix; (b) mapped matrix; (c) subgraph; (d) histogram of bipartivity values for* 1000 *graphs with same expected degree distribution; (e) fitted probability distribution; and (f) quantile-quantile plot.*

ETNA
Kent State University
http://etna.math.kent.edu

works where node i has expected in and out degrees that match the ith highest over all nodes in the given network. Standard Erdös-Rényi style random graphs with expected number of (directed) edges matching the given network were also used. Intuitively, we would expect the directed stickiness version to be the most likely to preserve any directed bipartivity present in the given network. This is borne out in the results, shown in Table 5.1, although in all cases the pvalue is well below the 0.05 threshold.

TABLE 5.1
Significance of bipartivity substructure discovered in test case 1, for various random network classes.

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

5.2. Test case 2. We now construct a network with a much less well-defined directed bipartite substructure. As before, there are 100 nodes. Now 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, Figure 5.2 changes to Figure 5.3. (We found that 60% of the nodes in the subgraph came from the "correct" sets 1–20 and 21–40.) A bipartivity score of 0.4083 is obtained for this subnetwork. The various pvalues, all of which slightly exceed 0.05, are listed in Table 5.2. We see that for this data set the level of bipartivity discovered by the mapping cannot be regarded as significant. This example gives some indication of the extent of directed bipartivity that we can confidently discover.

TABLE 5.2
Significance of bipartivity substructure discovered in test case 2, for various random network classes.

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

6. Worm brain network. To illustrate the usefulness of the new mapping we analyse two real-world networks[1](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 [14]. To obtain a directed network we removed all gap junctions from the data sets, 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.

[1] The data sets are available at `http://www.biological-networks.org/?page_id=25`.

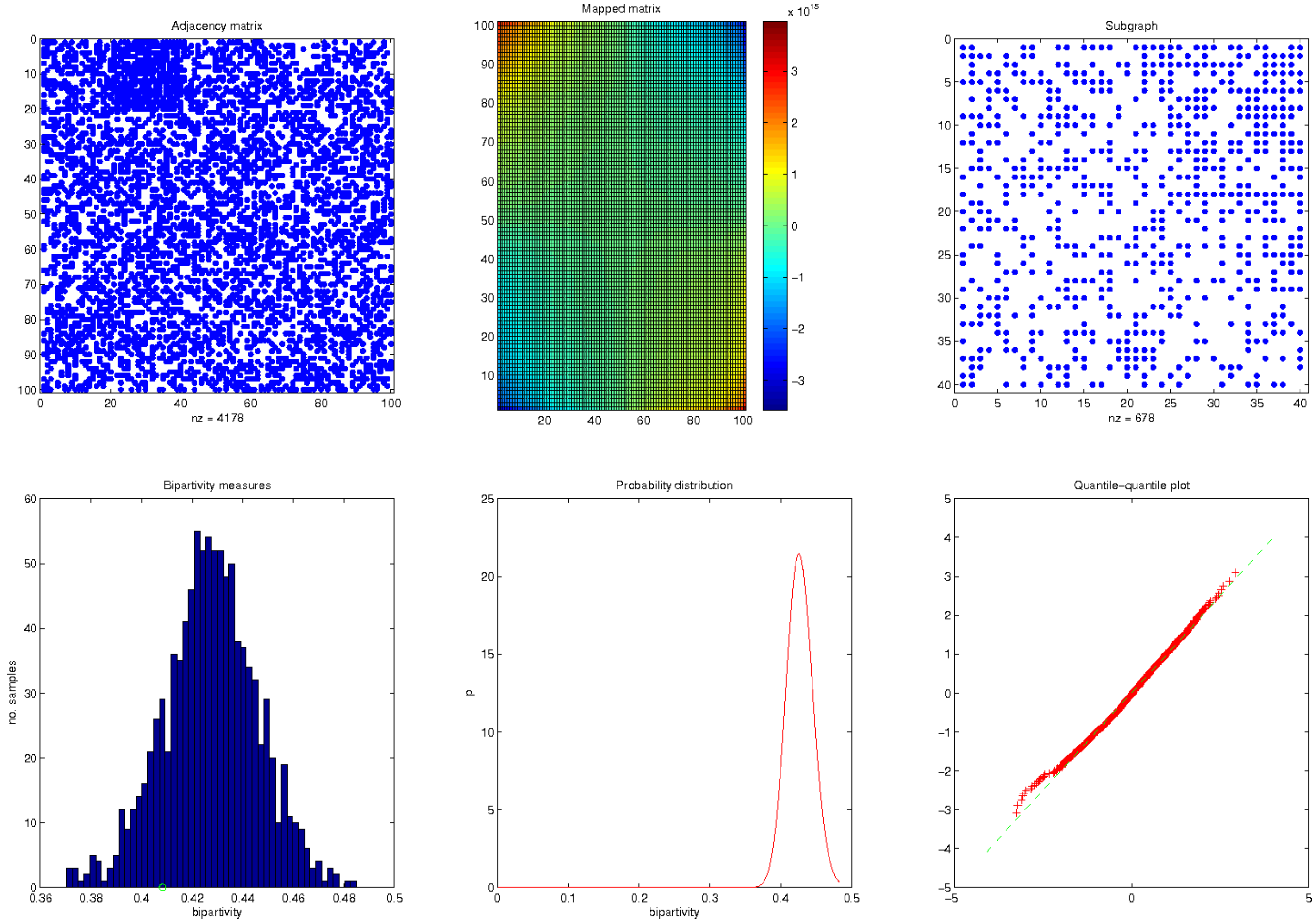

FIG. 5.3. *Results for synthetic network 2: (a) adjacency matrix; (b) mapped matrix; (c) subgraph; (d) histogram of bipartivity values for* 1000 *graphs with same expected degree distribution; (e) fitted probability distribution; and (f) quantile-quantile plot.*

.

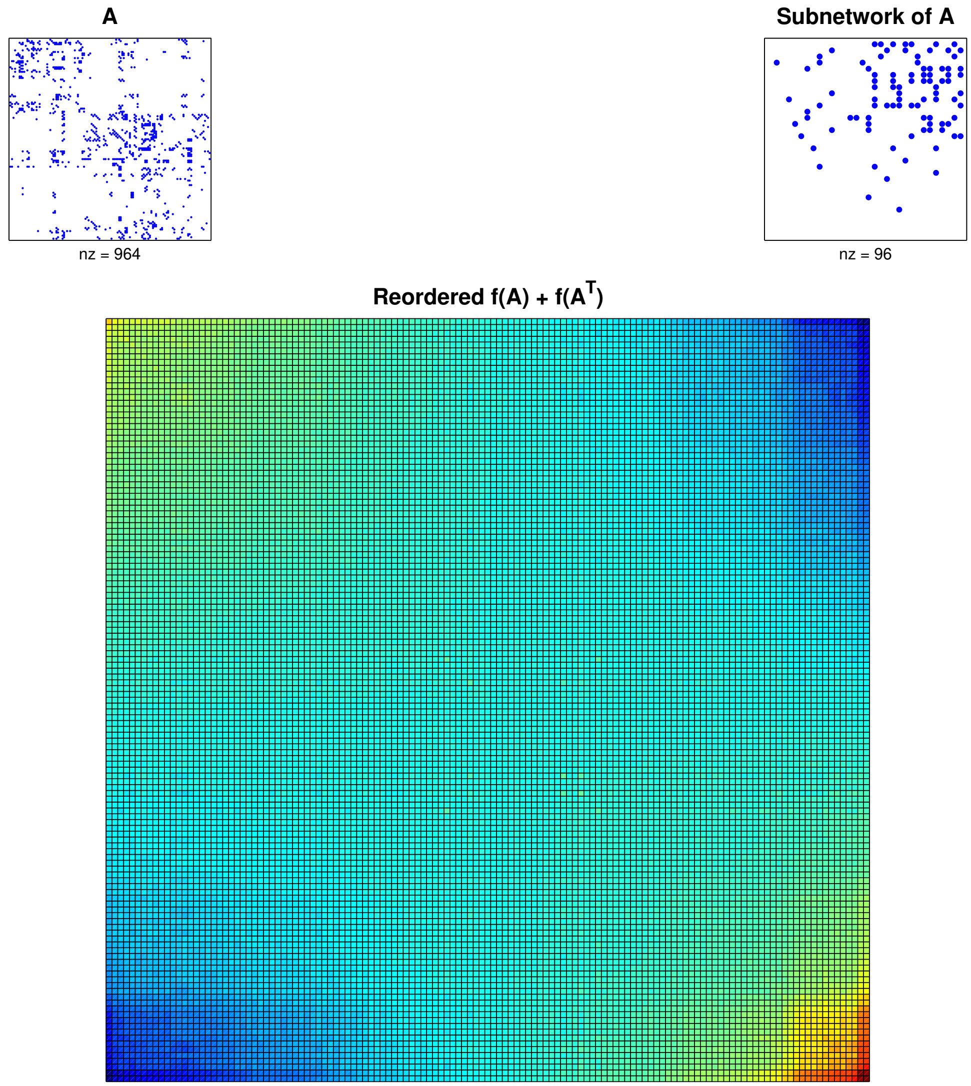

FIG. 6.1. *Upper left: worm neural network with* 131 *nodes. Lower: reordered version of* $f(A) + f(A^T)$*. Upper right: subnetwork of* 32 *nodes obtained from the reordering.*

Our motivation is that Durbin [4, 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?* [4]

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.

The upper plot in Figure 6.1 shows 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 spectrally reordering to reveal the hot and cold

ETNA
Kent State University
http://etna.math.kent.edu

regions 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.

TABLE 6.1
Significance of subgraphs found for the local and global networks for C. elegans using varying test matrix classes.

	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

Using the methods developed in the previous sections, we tested for statistical significance and obtained the results shown in Table 6.1. We see that for the smaller, local network, the bipartite structure that we discovered is deemed significant in the case of the "biased stickiness" and Erdös-Rényi models, but not for the more demanding stickiness versions. This inconsistency may be due to the fact that although the $S_1 \to S_2$ submatrix has many more connections than the $S_1 \to S_1$, $S_2 \to S_1$ and $S_2 \to S_2$ submatrices, it is still relatively sparse, thus resulting in a low bipartivity score of 0.2645. For the global network, the connectivity pattern is significant under all stickiness models. The bipartivity measure is 0.6415 and all pvalues are below 0.01.

In this example we are fortunate that in addition to statistical significance testing, we can validate the results against known biological information.

The neuronal classes[2] that were picked out by the algorithm along with a description of their respective functionalities are given in Tables 6.2 and 6.3.

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 [4], Durbin attempted to display the neuronal classes in the nerve ring of *C. elegans* vertically, in such a way that as many of the synapses as possible pointed downwards. The resultant ordering placed sensory neurons towards the top, motor neurons towards the bottom, and the remaining interneurons in between. Overall, the bipartite structures that we have picked out are in good agreement with the highly directed, hierarchical 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

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

TABLE 6.2
Neuronal class and type for bipartite subgraph found in the local network of 131 *frontal neurons of C. elegans.*

	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 6.3
Neuronal class and type for bipartite subgraph found in the global network of 191 *neurons of the C. elegans.*

	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

a group of neurons 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* [3]. 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 'center' neurons that are essential for normal biological function [16]. 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 and validating them statistically. Initial tests on a network from neuroscience show that the new mapping can be used to infer biologically relevant information using only the network topology. We also found that the statistical significance of the connectivity patterns can be extremely sensitive to the class of random matrices chosen for comparison. In future work in this area we plan to develop automated algorithms for discovering and quantifying the presence 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. JJC and DJH were supported through MRC project grant G0601353. DJH was supported through EPSRC project grant GR/S62383/01. AT was supported through an EPSRC 'Bridging the Gaps' grant.

REFERENCES

[1] B. C. ARNOLD, N. BALAKRISHNAN, AND H. N. NAGARAJA, *A First Course in Order Statistics*, SIAM, Philadelphia, 2008.
[2] D. BARASH, *Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation*, Bioinformatics, 20 (2004), pp. 1861–1869.
[3] 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 (2008), pp. 145–153.
[4] R. M. DURBIN, *Studies on the Development and Organisation of the Nervous System of Caenorhabditis Elegans*, Ph.D. thesis, University of Cambridge, MRC Laboratory of Molecular Biology, 1987.
[5] P. ERDÖS AND A. RÉNYI, *On random graphs*, Publ. Math. Debrecen, 6 (1959), pp. 290–297.
[6] E. ESTRADA AND N. HATANO, *Communicability in complex networks*, Phys. Rev. E, 77 (2008), 036111 (12 pages).
[7] E. ESTRADA, D. J. HIGHAM, AND N. HATANO, *Communicability and multipartite structures in complex networks at negative absolute temperatures*, Phys. Rev. E, 78 (2008), 026102 (7 pages).
[8] W. J. EWENS AND G. R. GRANT, *Statistical Methods in Bioinformatics: An Introduction*, Springer, Berlin, 2001.
[9] E. N. GILBERT, *Random graphs*, Ann. Math. Statist., 30 (1959), pp. 1141–1144.
[10] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, third ed., 1996.
[11] D. HIGHAM, G. KALNA, AND J. VASS, *Spectral analysis of two-signed microarray expression data*, Math. Med. Biol., 24 (2007), pp. 131–148.
[12] N. J. HIGHAM, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.
[13] P. HOLME, F. LILJEROS, C. R. EDLING, AND B. J. KIM, *Network bipartivity*, Phys. Rev. E, 68 (2003), 056107 (12 pages).
[14] M. KAISER AND C. C. HILGETAG, *Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems*, PLoS Computat. Biol., 2 (2006), e95 (11 pages).
[15] C. KAMP AND K. CHRISTENSEN, *Spectral analysis of protein-protein interactions in Drosophila melanogaster*, Phys. Rev. E, 71 (2005), 041911 (8 pages).
[16] S. MORITA, K. OSHIO, Y. OSANA, Y. FUNABASHI, K. OKA, AND K. KAWAMURA, *Geometrical structure of the neuronal network of Caenorhabditis elegans*, Phys. A, 298 (2001), pp. 553–561.
[17] J. L. MORRISON, R. BREITLING, D. J. HIGHAM, AND D. R. GILBERT, *A lock-and-key model for protein-protein interactions*, Bioinformatics, 2 (2006), pp. 2012–2019.
[18] N. PRŽULJ AND D. J. HIGHAM, *Modelling protein-protein interaction networks via a stickiness index*, J. Roy. Soc. Interface, 3 (2006), pp. 711–716.