LMS J. Comput. Math. 14 (2011) 72–86 © 2011 Authors
doi:10.1112/S1461157009000072

Discovering bipartite substructure in directed networks

Alan Taylor, J. Keith Vass and Desmond J. Higham

Abstract

Bipartivity is an important network concept that can be applied to nodes, edges and communities. Here we focus on directed networks and look for subnetworks made up of two distinct groups of nodes, connected by 'one-way' links. We show that a spectral approach can be used to find hidden substructures of this form. Theoretical support is given for the idealized case where there is limited overlap between subnetworks. Numerical experiments show that the approach is robust to spurious and missing edges. A key application of this work is in the analysis of high-throughput gene expression data, and we give an example where a biologically meaningful directed bipartite subnetwork is found from a cancer microarray dataset.

1. *Motivation*

A bipartite network, or subnetwork, involves objects that may be split into two disjoint groups with connections only occurring across, but not within, the two groups. Sometimes this bipartivity is obvious because the objects naturally fall into two groups. For example, a Hollywood movie network could be constructed with nodes that are either actors or movies, with an edge denoting that an actor appeared in a movie. However, in other cases, bipartite structure might not be immediately apparent and, although there has been interest in deriving measures of the overall level of bipartivity in a network, node or edge [**3**, **6**, **9**], our focus here is on the practical issue of identifying hidden bipartite substructures. More precisely, we are concerned with discovering approximately bipartite subnetworks, because datasets are typically contaminated with missing and spurious edges.

Efforts in this direction have been applied to protein–protein interaction (PPI) networks, which have nodes given by proteins and edges denoting that two proteins have been observed to interact physically [**2**]. It is known that at least some types of protein–protein interaction arise through complementary binding domains, and in this case all proteins that possess one binding domain should interact with all proteins that share the complementary domain [**14**]. In [**11**], an algorithm was developed that aims to find bipartite substructure in PPI networks, thereby offering a route towards the identification of new binding domains using only interaction data. That algorithm uses spectral information — eigenvectors and eigenvalues of the adjacency matrix — and was shown to be robust in the presence of noise. The related approach in [**5**] compares the propensity for even and odd walk lengths between pairs of nodes by forming the negative matrix exponential. This allows the possibility of breaking down the whole network into quasi-bipartite communities.

This work differs from previous studies by considering networks with *directed* edges — a connection from node i to node j does not necessarily have a matching connection from j to i. Our aim is to develop an approach for discovering *directed bipartite substructure*, that is, groups of nodes S_1 and S_2, such that edges point from nodes in S_1 to nodes in S_2. We will show that spectral information is still relevant if we generalize from eigenvectors and

Received 12 March 2009; revised 13 August 2010.

2010 Mathematics Subject Classification 68R10, 92B05.

The third author was supported by Engineering and Physical Sciences Research Council grant EP/E049370/1 and Medical Research Council grant G0601353.

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

eigenvalues to singular values and singular vectors. In Section 2, we develop our theoretical arguments and in Section 3 we test them on synthetically constructed networks, and compute some basic statistical measures. Finally, in Section 4, we look at a larger network arising from cancer microarray data, and show that it is possible to discover biologically relevant directed bipartite subnetworks.

2. *Relevance of the singular value decomposition*

Given a directed network with N nodes, we let the unsymmetric matrix $A \in \mathbb{R}^{N\times N}$ denote the corresponding adjacency matrix, so that $a_{ij} = 1$ if there is a link from i to j and $a_{ij} = 0$ otherwise. To characterize directed bipartite subnetworks, we find it useful to borrow the *lock and key* analogy that was introduced in [11]. We suppose that locks and keys are distributed among the nodes in a network. Each lock and key has a particular colour (red, blue, green, ...) and lock–key matches, corresponding to edges in the network, take place only when the colours agree. Suppose that two sets of nodes, S_1 and S_2, form a bipartite subnetwork, so edges between these nodes only point from nodes in S_1 to nodes in S_2. We may imagine that S_1 consists of all the nodes that possess a certain colour of key, say red, and that S_2 consists of all the nodes that possess the matching red lock. We note that the reference [11] dealt only with undirected edges, whereas this work considers the directed case. So, the concept of locks and keys here is slightly different, and perhaps more natural, and we find that the arguments supporting a spectral algorithm are stronger.

Focussing on this particular red lock–key subnetwork, we may introduce indicator vectors $\mathbf{u}^{\mathrm{red}}, \mathbf{v}^{\mathrm{red}} \in \mathbb{R}^N$ such that

$$(\mathbf{u}^{\mathrm{red}})_i = \begin{cases} 1 & \text{if node } i \text{ has the red key,} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and

$$(\mathbf{v}^{\mathrm{red}})_i = \begin{cases} 1 & \text{if node } i \text{ has the red lock,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

It follows immediately that the edges arising from red key–lock interactions may be characterized through the outer product $\mathbf{u}^{\mathrm{red}}(\mathbf{v}^{\mathrm{red}})^T$. If we let keyred $:= \|\mathbf{u}^{\mathrm{red}}\|_2^2$ and lockred $:= \|\mathbf{v}^{\mathrm{red}}\|_2^2$ denote the total number of red keys and red locks, respectively, then this outer product may be written

$$\sqrt{\text{keyred} \times \text{lockred}}\; \widehat{\mathbf{u}}^{\mathrm{red}}(\widehat{\mathbf{v}}^{\mathrm{red}})^T,$$

where $\widehat{\mathbf{u}}^{\mathrm{red}} := \mathbf{u}^{\mathrm{red}}/\|\mathbf{u}^{\mathrm{red}}\|_2$ and $\widehat{\mathbf{v}}^{\mathrm{red}} := \mathbf{v}^{\mathrm{red}}/\|\mathbf{v}^{\mathrm{red}}\|_2$ are unit vectors. More generally, when all links arise through key–lock interactions, the adjacency matrix for the network may be expanded as

$$\begin{aligned} A = \operatorname{sign}(&\sqrt{\text{keyred} \times \text{lockred}}\; \widehat{\mathbf{u}}^{\mathrm{red}}(\widehat{\mathbf{v}}^{\mathrm{red}})^T + \sqrt{\text{keyblue} \times \text{lockblue}}\; \widehat{\mathbf{u}}^{\mathrm{blue}}(\widehat{\mathbf{v}}^{\mathrm{blue}})^T \\ &+ \sqrt{\text{keygreen} \times \text{lockgreen}}\; \widehat{\mathbf{u}}^{\mathrm{green}}(\widehat{\mathbf{v}}^{\mathrm{green}})^T + \ldots), \end{aligned} \tag{3}$$

where the sign function deals with the possibility of multiple key–lock matches; node i may have both a red and a blue key whilst node j has both a red and a blue lock.

The sign function in (3) is not needed if we make the following assumption.

ASSUMPTION A. Each node has at most one key and one lock.

Note that this assumption permits a node to possess a key of one colour and a lock of another colour, or a lock and a key of the same colour. A second important consequence of Assumption A is that the key indicator vectors $\{\mathbf{u}^{\mathrm{red}}, \mathbf{u}^{\mathrm{blue}}, \mathbf{u}^{\mathrm{green}}, \ldots\}$ form an orthogonal set and the lock indicator vectors $\{\mathbf{v}^{\mathrm{red}}, \mathbf{v}^{\mathrm{blue}}, \mathbf{v}^{\mathrm{green}}, \ldots\}$ form an orthogonal set. In this case,

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

we see that the expansion (3) has the same form as the singular value decomposition (SVD) [7]

$$A = \sum_{k=1}^{N} \sigma_k \mathbf{u}^{[k]} \mathbf{v}^{[k]}, \tag{4}$$

where $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_N \geqslant 0$ are the singular values of A, and $\{\mathbf{u}^{[k]}\}_{k=1}^{N}$ and $\{\mathbf{v}^{[k]}\}_{k=1}^{N}$ are the corresponding left and right singular vectors, respectively. We conclude that under Assumption A the SVD can be used to discover bipartite subgraphs — the square of the singular value, σ_k^2, indicates the product of the number of locks and keys of the kth colour, the non-zero entries of $\mathbf{u}^{[k]}$ give the key locations and the non-zero entries of $\mathbf{v}^{[k]}$ give the lock locations.

We show now that there is a complementary way to motivate the use of the SVD. This approach, which is based on the ideas in [11] that were used for undirected networks, also goes some way towards allowing for false negatives among the edges. Under Assumption A, suppose that node i does not possess the red key. Then multiplying the ith row of the adjacency matrix into the red lock indicator vector will give a value zero; there will be no matches in the inner product. On the other hand, if node i possesses the red key then multiplying the ith row of the adjacency matrix into the red lock indicator vector will count the number of red locks in existence — each red lock will take part in one non-zero term. Suppose now that there are some 'errors' in the network in the form of missing edges. More precisely, suppose that only a fixed proportion $\theta \in (0,1)$ of the red key–lock matches are recorded as edges. Then generalizing the argument above we have

$$(A\mathbf{v}^{[k]})_i = \sum_{j=1}^{N} a_{ij} v_j^{[k]} = \begin{cases} \theta \, \text{lockred} & \text{if node } i \text{ has the red key,} \\ 0 & \text{otherwise,} \end{cases}$$

which may be written

$$A\widehat{\mathbf{v}}^{\text{red}} = \theta\sqrt{\text{lockred} \ \times \text{keyred}}\,\widehat{\mathbf{u}}^{\text{red}}. \tag{5}$$

Similarly, we find that

$$A^T\widehat{\mathbf{u}}^{\text{red}} = \theta\sqrt{\text{lockred} \ \times \text{keyred}}\,\widehat{\mathbf{v}}^{\text{red}}. \tag{6}$$

The relations (5) and (6) show that $\widehat{\mathbf{u}}^{\text{red}}$ and $\widehat{\mathbf{v}}^{\text{red}}$ correspond to left and right singular vectors of A, respectively, with singular value $\theta\sqrt{\text{lockred} \ \times \text{keyred}}$. Of course, when $\theta = 1$, we recover the singular value expression $\sqrt{\text{lockred} \ \times \text{keyred}}$ that we derived earlier via the argument involving rank one outer products. However, it is worth noting that (5) and (6) require Assumption A to hold only for nodes with red keys or locks. The other nodes in the network could be connected in any way. So, the SVD will reveal isolated substructure hidden within any complex network.

In summary, we have shown that if a directed network can be broken down into isolated or non-overlapping bipartite subnetworks, even when a fixed proportion of edges are missing, then the left and right singular vectors reveal which nodes take part in which subnetwork, and the singular values tell us how many nodes are involved. Hence, in practice, to reveal this type of substructure in a given network, where the nodes will typically be labelled in a manner that hides the bipartivity, the singular values $\sigma_1, \sigma_2, \ldots$ can be taken in order, and the indices of extremal components in the corresponding left and right singular vectors used to identify candidate nodes.

Eigenvectors and, more generally, singular vectors, enjoy important variational properties, and hence the information that they convey tends to be robust to the presence of noise. This has been confirmed experimentally; see for example [1, 4, 8, 10, 12, 17]. In particular, for the case of undirected edges, it was shown in [11] that the SVD can find approximate bipartite subgraphs in both synthetic and real networks. In the next section, we test the robustness of the SVD in the directed network setting.

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

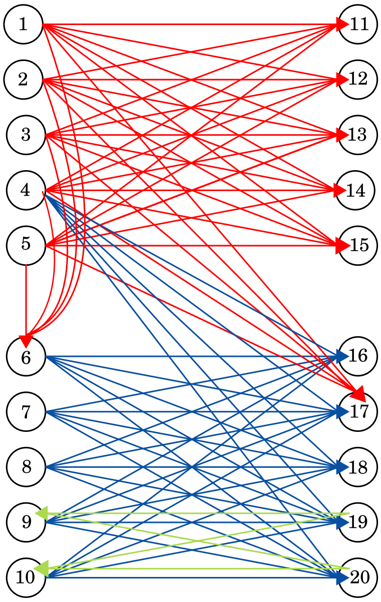

FIGURE 1. *Example with three lock-and-key types.*

3. *Exploratory tests*

3.1. *Detailed example*

The directed network in Figure 1 does not completely satisfy Assumption A from Section 2. The colour coding of the arrows in Figure 1 is designed to emphasize the lock and key distribution. Nodes 1–5 have the red key and nodes 6, 11–15 and 17 the corresponding red lock. Nodes 4 and 6–10 have the blue key and nodes 16–20 the corresponding blue lock. Finally, nodes 19 and 20 have the green key, while nodes 9 and 10 have the green lock. We note that this node ordering was chosen simply to make the output easier to interpret — results from the SVD are invariant under symmetric row and column permutations.

The non-trivial singular values are, to two digits, 6.5, 4.7, 2.0 and 0.7, which is consistent with our expectation that the first, second and third pairs of singular vectors should correspond to the red, blue and green groups.

3.1.1. *First left and right singular vectors.* Figure 2 shows the components of the first left and right singular vectors of the adjacency matrix in increasing order. On the horizontal axis are the indices, so, for example, the most negative left and right singular vector entries correspond to nodes 4 and 17, respectively.

For the first left singular vector, $\mathbf{u}^{[1]}$, there are two main groups of nodes with components away from zero, one with values around -0.34 and the other with values around -0.23. There is also an outlying vertex at around -0.5. The group at height -0.34 involves nodes 1, 2, 3, 5, which, as we see from Figure 1, share the red key. The outlier is node 4. This is the only other red key node, but it also has the blue key.

The first right singular vector, $\mathbf{v}^{[1]}$, splits up the network in a similar fashion. Nodes 6, 11, 12, 13, 14 form a clear group and we see from Figure 1 that they share the red lock. The outlier is node 17. This is the only other red lock node, but it also has the blue lock. We note that node 6 differs from its neighbours in Figure 2 in that it also has a blue key, but the left singular vector has not been affected by this — node 6 is placed at the same height as the purely red

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

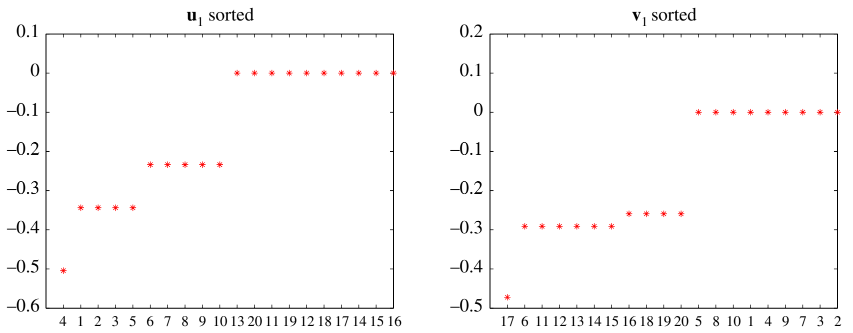

FIGURE 2. *First left and right singular vectors for the network in Figure 1.*

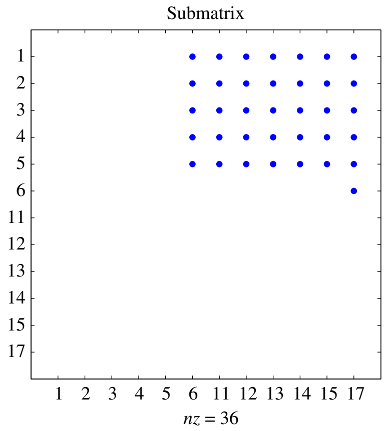

FIGURE 3. *Subgraph showing red lock–key interactions arising from nodes* 4, 1, 2, 3, 5 *and* 17, 6, 11, 12, 13, 14, 15 *taken from* $\mathbf{u}^{[1]}$ *and* $\mathbf{v}^{[1]}$ *in Figure 2.*

lock nodes. This is consistent with the theory in Section 2, where Assumption A does not rule out the case of a node having a key and a lock of different colours.

In Figure 3, we show the adjacency matrix for the subgraph created by nodes 4, 1, 2, 3, 5, 6, taken from the left-hand end of $\mathbf{u}^{[1]}$ up to the cut-off from -0.34 to -0.23, and nodes 17, 6, 11, 12, 13, 14, 15, taken from the left-hand end of $\mathbf{v}^{[1]}$ up to the corresponding cut-off. We see that there is a clear two-by-two block checkerboard structure, corresponding to a directed bipartite subgraph, with node 6 having an extra link to its red lock colleague (arising from the separate blue lock–key connection).

Similarly, Figure 4 reveals the blue lock–key interactions by plotting the adjacency matrix arising from nodes 6, 7, 8, 9, 10 and 16, 18, 19, 20 that were grouped together above the main cut-offs in Figure 2. Here node 17 is missing from the blue lock group; this makes sense because it is also a lock member of the larger red lock–key group. In Figure 4, we also see that the green lock–key group of 9, 10 and 19, 20 appears as a block in the adjacency matrix. However, from Figure 2 this appears to be a coincidental feature caused by the fact that these nodes are also part of the blue subgraph — the components of nodes 9, 10 in $\mathbf{u}^{[1]}$ and 19, 20 in $\mathbf{v}^{[1]}$ are not visually distinguishable from their blue key and lock neighbours, 6, 7, 8 and 16, 18, respectively.

Overall, the first left and right singular vectors have done an excellent job of sorting out the key and lock nodes, respectively, for the red group. They also made a reasonable delineation of

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

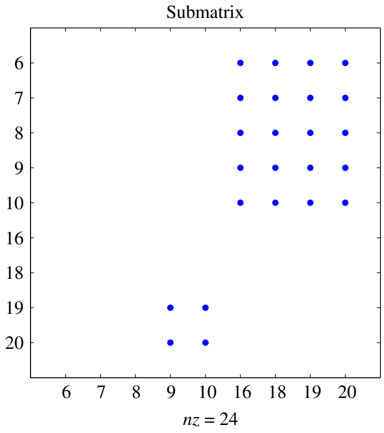

FIGURE 4. *Subgraph showing blue lock–key interactions arising from nodes* 6, 7, 8, 9, 10 *and* 16, 18, 19, 20 *taken from* $\mathbf{u}^{[1]}$ *and* $\mathbf{v}^{[1]}$ *in Figure* 2.

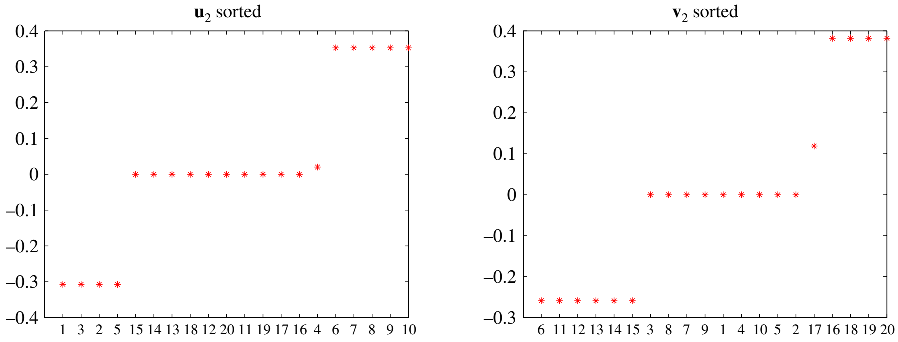

FIGURE 5. *Second left and right singular vectors for the network in Figure* 1.

the blue group, but the ambiguous nodes that shared red and blue characteristics were placed next to their red colleagues, which form the dominant group.

3.1.2. *Second left and right singular vectors.* Figure 5 shows the second left and right singular vectors. In $\mathbf{u}^{[2]}$, we see that nodes 1, 2, 3, 5 are grouped together. These are four out of the five red key nodes. Node 4, which also has the blue key, has been given a positive value, in line with the positivity of nodes 6, 7, 8, 9, 10, which complete the blue key group and are classified together. In $\mathbf{v}^{[2]}$, we see nodes 6, 11, 12, 13, 14, 15 grouped together. These are all red lock nodes, with the exception of node 17, which has been given a positive value in line with the other blue lock nodes 16, 18, 19, 20 that are classified together.

So, overall, the second singular vectors also give information about both red and blue groups, but they favour the second-largest, blue, group.

3.1.3. *Third left and right singular vectors.* Figure 6 shows the third left and right singular vectors. In this case, we see that the green keys 19, 20 and the green locks 9, 10 have been picked out unambiguously.

3.2. *Parameterized example*

Our second experiment tests the robustness of the SVD approach when spurious and missing edges contaminate the directed bipartivity. We constructed a network of 50 nodes with

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

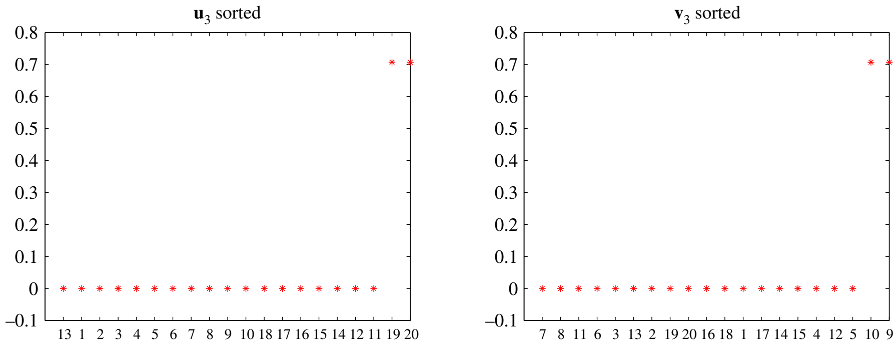

FIGURE 6. *Third left and right singular vectors for the network in Figure 1.*

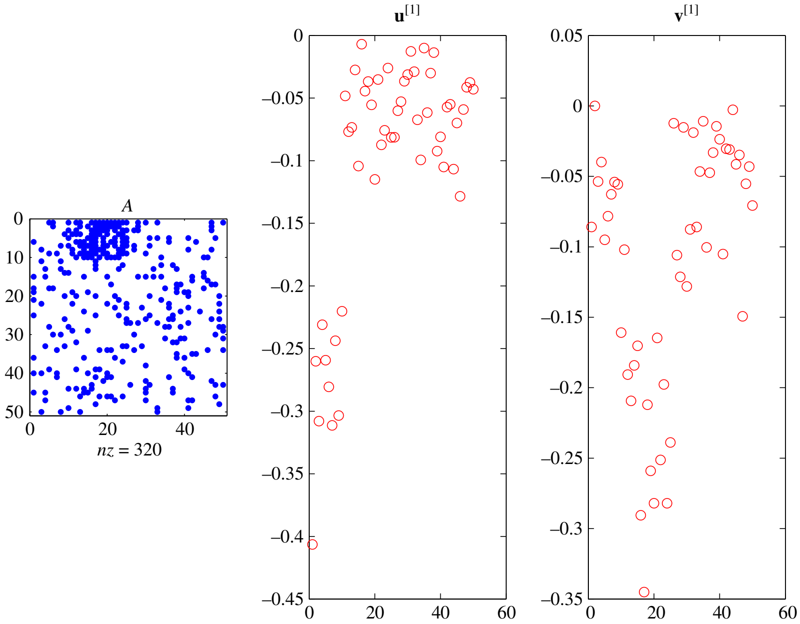

FIGURE 7. *Left: adjacency matrix. Middle: first left singular vector. Right: first right singular vector. Here $p_2 = 0.1$ for the non-bipartite connectivity probability.*

nodes 1–10 forming group S_1 and nodes 11–25 forming group S_2. We formed links independently at random such that the probability of a link from node i to node j is given by 0.6 if $i \in S_1$ and $j \in S_2$, and p_2 otherwise.

To the left in Figure 7, we show the adjacency matrix that arose for $p_2 = 0.1$. We see that the S_1-to-S_2 block forms a dense patch, but there is a significant amount of non-bipartite 'noise'. In fact, there are 94 S_1-to-S_2 edges and 226 others. In the centre and left of Figure 7, we show the first left and right singular vectors. It is clear that the key group, S_1, is picked out by $\mathbf{u}^{[1]}$ and the lock group, S_2, is picked out by $\mathbf{v}^{[1]}$; in each case the group members appear sequentially, taking the extreme values in the vector.

In Figure 8, we increase p_2 to 0.3. Now there are 665 edges outside the S_1-to-S_2 class. In this case, we are at the extremes of the noise level that the SVD can tolerate. In $\mathbf{u}^{[1]}$, the 10 nodes in group S_1 appear in positions 1, 2, 3, 4, 6, 7, 10, 11, 13, 20 as we search through the components with largest to smallest absolute value. Similarly, in $\mathbf{v}^{[1]}$, the 15 nodes in group S_2

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

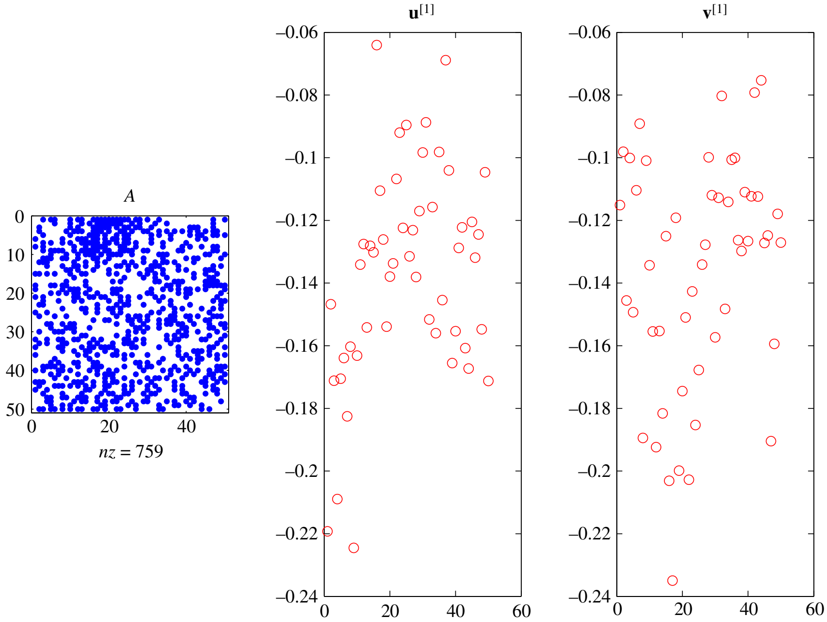

FIGURE 8. *Left: adjacency matrix. Middle: first left singular vector. Right: first right singular vector. Here $p_2 = 0.3$ for the non-bipartite connectivity probability.*

appear in positions 1, 2, 3, 4, 5, 8, 9, 10, 11, 14, 15, 16, 20, 29, 31. So, the dominant singular vectors do not reproduce perfectly the key and lock groups, although at this high level of noise it may be argued that the groups are not clearly defined.

3.3. *Statistical testing*

Having established that the left and right singular vectors may be useful in identifying members of directed bipartite communities, we would like to assess the success with which nodes are classified into subgroups. To this end, we construct an adjacency matrix consisting of 100 nodes with sets S_1 and S_2 comprising nodes 1–10 and 11–20, respectively. We connect pairs of nodes i and j independently at random with probability p_1 if $i \in S_1$ and $j \in S_2$ and with probability p_2 otherwise.

With an adjacency matrix set up in this fashion, we may compute the SVD and examine the components of the first left and right singular vectors. If the SVD perfectly organizes nodes into the appropriate subgroups, then nodes 1–10 should correspond to the ten components with highest absolute value in $\mathbf{u}_1$ and nodes 11–20 should correspond to the ten components with highest absolute value in $\mathbf{v}_1$. We fix p_1 and vary p_2 from 0 to p_1, generating several instances of the synthetic network described above for each value of p_2. The SVD of each adjacency matrix is calculated and the proportion of correctly identified nodes in the first ten positions of the relevant singular vector is recorded in each instance.

In Figure 9, we show a plot of the mean proportion of correctly identified nodes for $p_1 = 0.9$ and p_2 varying from 0 to 0.9 in increments of 0.01. The blue line shows the mean proportion of ‘key’ nodes correctly identified and the red line corresponds to the mean proportion of ‘lock’ nodes correctly identified. In this case, the probability of false negatives in the region of bipartite connectivity is 0.1, and the SVD is reasonably tolerant of false positives elsewhere in the adjacency matrix. The singular vectors correctly identify more than half of the correct

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

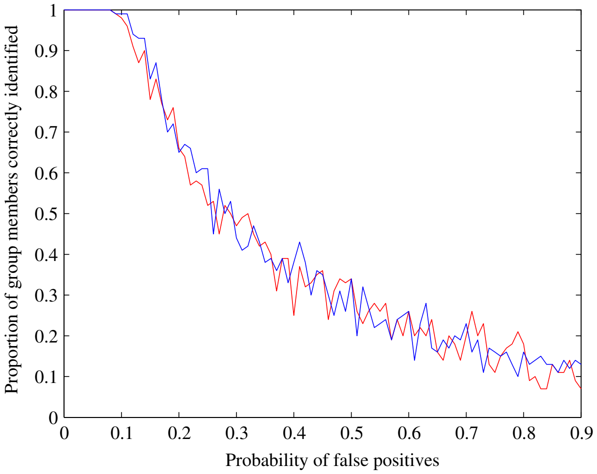

FIGURE 9. *Proportions of 'key' nodes (red) and 'lock' nodes (blue) correctly identified by first singular vectors for* $p_1 = 0.9$ *with varying levels of 'noise'.*

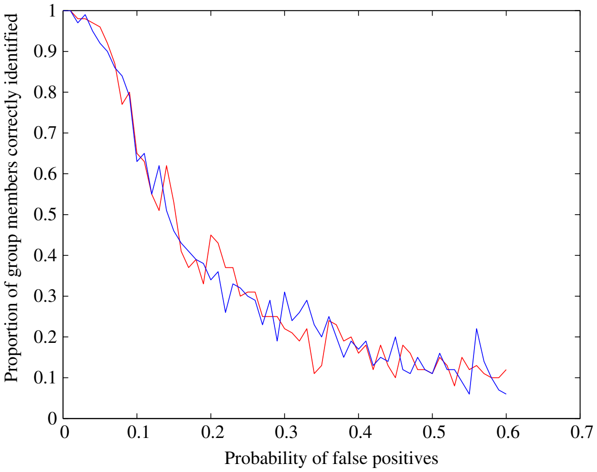

FIGURE 10. *Proportions of 'key' nodes (red) and 'lock' nodes (blue) correctly identified by first singular vectors for* $p_1 = 0.6$ *with varying levels of 'noise'.*

nodes until the probability of a false positive in a given position in the adjacency matrix reaches around 0.3.

Figure 10 shows a plot of the same type with $p_1 = 0.6$. In this case, the singular vectors successfully identify more than half of the correct nodes until the probability of a false positive in a given position in the adjacency matrix reaches around 0.15.

By varying p_1 from 0 to 1 and carrying out the procedure described above, we can obtain three-dimensional plots of the proportions of locks and keys correctly recovered by the SVD for varying probabilities of false negatives and false positives. Figure 11 shows the proportion of 'key' nodes correctly recovered for varying values of p_1 and p_2. The data is also plotted as a heat map in Figure 12. Similarly, Figures 13 and 14 show the proportion of 'lock' nodes correctly recovered as a surface plot and a heat map, respectively. By inspection, we see that the left and right singular vectors perform very similarly in their identification of 'lock' and 'key' nodes. Approximately half the correct nodes are identified if the probability of a false

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

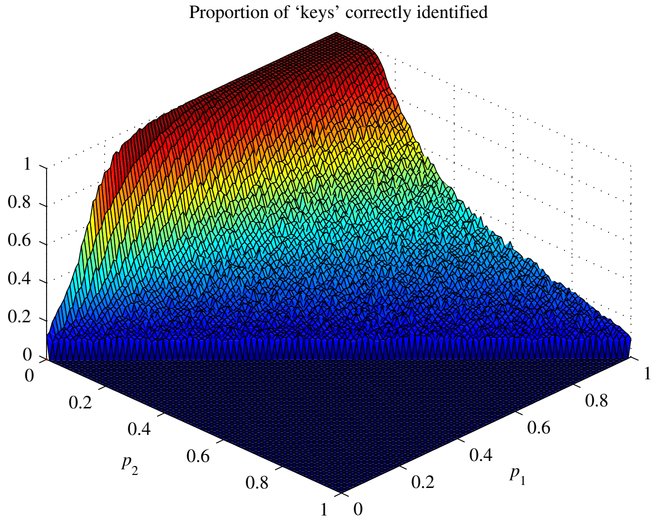

Figure 11. *Proportion of 'key' nodes correctly identified by first left singular vector for varying levels of 'noise'.*

positive is around a third of the probability of a 'true positive', and all the correct nodes are identified if every true connection is present and the probability of a false positive is less than 0.2.

It is important to stress that the success of the SVD in identifying members of directed bipartite subgroups is dependent upon a number of factors, including the proportion of 'noise' in the data (that is, the frequency of false positives and false negatives), the number of lock-and-key pairs (including overlapping group membership) and the size of a directed bipartite community relative to the dataset as a whole.

Having quantified the success of the SVD in a number of test cases, we now apply our method to a dataset from genomics and attempt to uncover meaningful communities.

4. *Application to cancer microarray data*

In this section, we consider a network arising through gene expression. Cancer microarray data from [**15**] was treated with the classification method developed in [**16**]. More precisely, we selected 133 genes related to the oncogene p53, and computed the 'plus–minus' network. Here an edge between nodes i and j indicates that when gene i expresses significantly above its usual level, gene j generally expresses significantly below its usual level. This produced a directed network with 133 nodes and 558 edges, whose adjacency matrix is shown in Figure 15. Discovering directed bipartite subgraphs in this setting is of major biological interest, as it reveals a pair of gene groups such that over-expression in one group is associated with under-expression in the other.

The singular values for this adjacency matrix are plotted in Figure 16. We see that the largest singular values are around 8. The first left and right singular vectors were found to produce an approximately bipartite subnetwork where edges crossed between groups in both directions, a feature that suggests that there is a significant component of symmetry in the network. Since we are interested here in directional information, we focus on the second left and right singular vectors, which are shown in Figure 17.

Keeping in mind the typical network size suggested by the singular values and considering the natural break points in the singular vector components, we chose the indices from four extreme

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

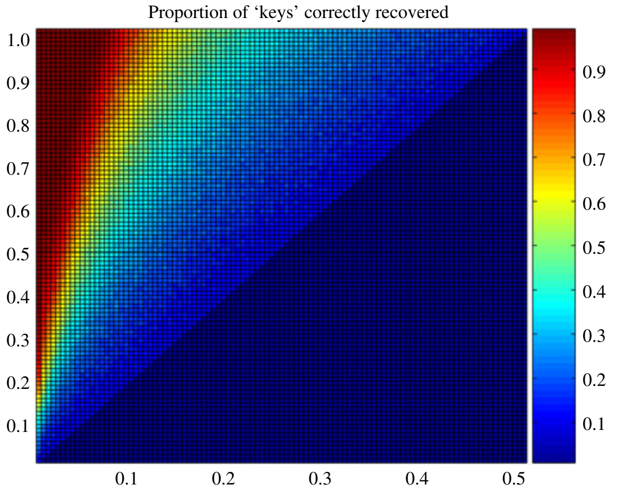

FIGURE 12. *Proportion of 'key' nodes correctly identified as a heat map.*

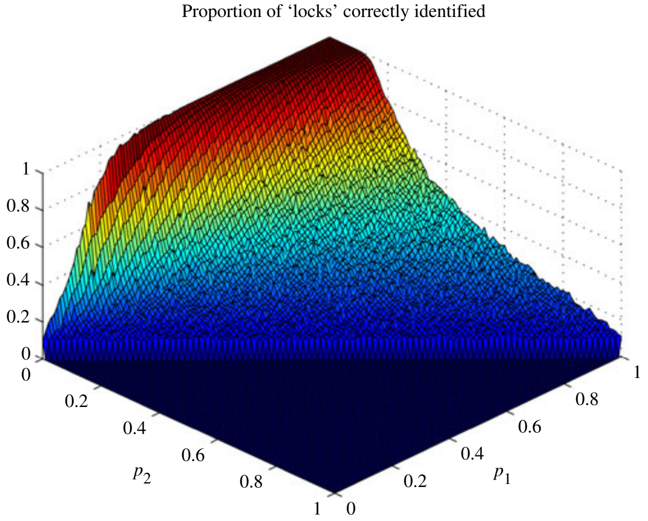

FIGURE 13. *Proportion of 'lock' nodes correctly identified by first right singular vector for varying levels of 'noise'.*

components of $\mathbf{u}^{[2]}$ and seven extreme components of $\mathbf{v}^{[2]}$. This produced the subnetwork shown in Figure 18. We see that there is a high degree of directed bipartivity. Denoting the first four nodes in this subnetwork as group S_1 and the remaining seven nodes as group S_2, twenty-five out of the possible twenty-eight S_1-to-S_2 connections are present, but none of the other S_1-to-S_1, S_2-to-S_1 or S_2-to-S_2 connections.

The subnetwork in Figure 18 was produced using only the network data. Because of the high level of interest in p53, there is extra biological information available, which can be used to justify the relevance of the subnetwork. Details of the genes are given in Table 1. Looking at these genes, BTG2, CCNG2 and FHL1 all appear to have inhibitory effects on growth or cell division, while the role of HLA-F with respect to growth is unclear. KIF15, CDC20, PRC1, CCNB2, KIF20A and NEK2 all seem to be involved in the cell-division process and, in most cases, inhibiting the genes seems to prevent growth. So, it appears that we have identified two

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

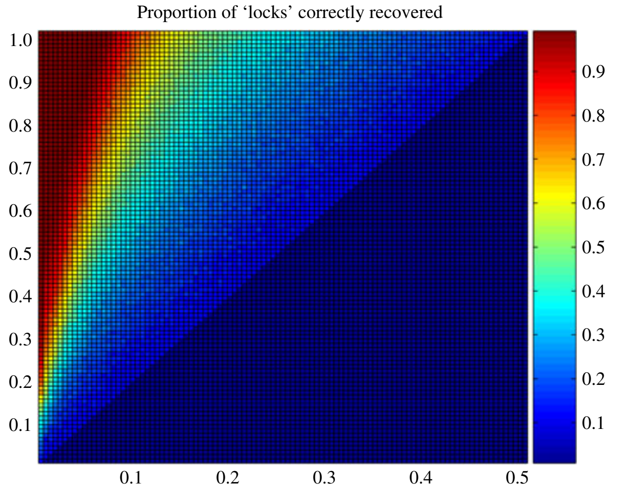

FIGURE 14. *Proportion of 'lock' nodes correctly identified as a heat map.*

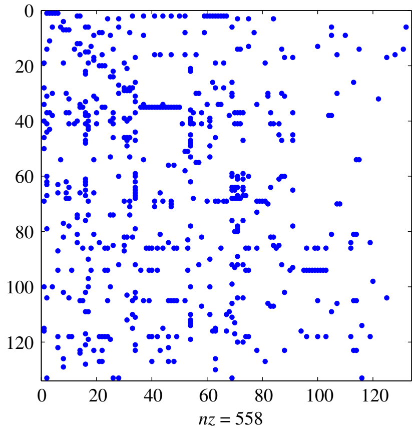

FIGURE 15. *Adjacency matrix for a 'plus–minus' network of genes relating to the oncogene* p53.

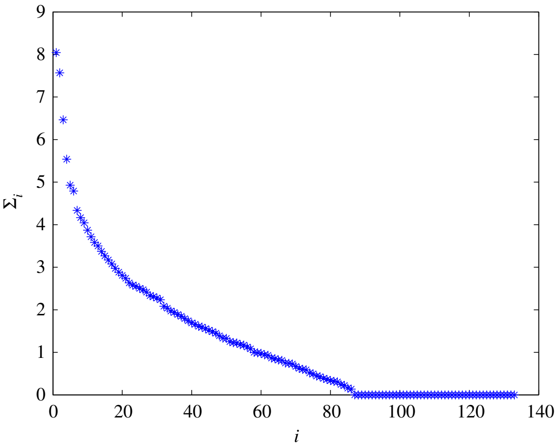

FIGURE 16. *Singular values of the adjacency matrix from Figure* 15.

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

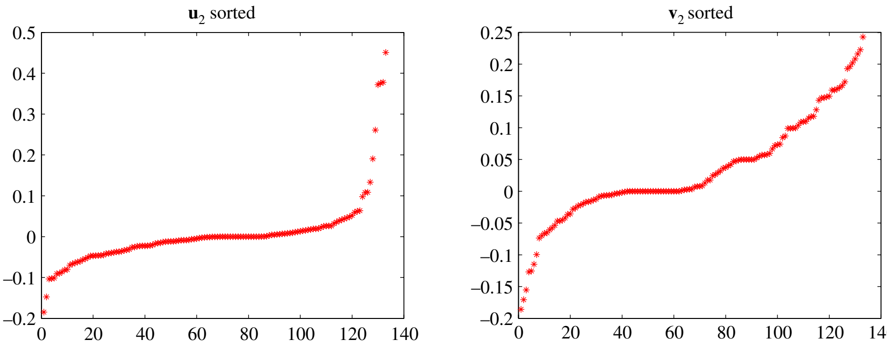

FIGURE 17. *Second left and right singular vectors of the adjacency matrix from Figure 15.*

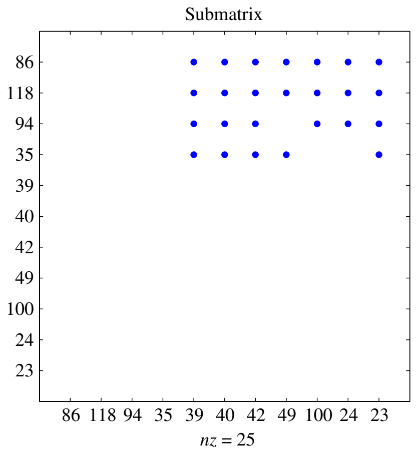

FIGURE 18. *Adjacency matrix for the subgraph of* 11 *nodes discovered via the second singular vectors.*

groups of genes whose products have opposite functions; whilst we make no detailed biological interpretation here, it seems reasonable that their mutually exclusive expression patterns are consistent with growth promotion and inhibition being correlated, with one of these groups being switched on when the other is off.

TABLE 1. *Details of genes corresponding to the groups in Figure 18. The two groups are separated by a double horizontal line.*

Probe set ID	Gene ID	Description
201235_s_at	BTG2	BTG family, member 2
202769_at	CCNG2	Cyclin G2
201540_at	FHL1	Four and a half LIM domains 1
221978_at	HLA-F	Major histocompatibility complex, class I, F
219306_at	KIF15	Kinesin family member 15
202870_s_at	CDC20	CDC20 cell division cycle 20 homologue (*S. cerevisiae*)
218009_s_at	PRC1	Protein regulator of cytokinesis 1
204822_at	TTK	TTK protein kinase
202705_at	CCNB2	Cyclin B2
218755_at	KIF20A	Kinesin family member 20A
204641_at	NEK2	NIMA (never in mitosis gene a)-related kinase 2

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

Statistical testing to assess the significance of the bipartite subgraph identified in Figure 18 was carried out as follows; see [13] for a more detailed discussion of this methodology. The rows and columns of the original adjacency matrix were independently permuted and the SVD of the resulting 'shuffled' matrix was computed. A subgraph of the same size as the one described above was formed from the 4 and 7 extremal entries in $\mathbf{u}^{[2]}$ and $\mathbf{v}^{[2]}$, respectively, and a measure of bipartivity was recorded by taking the ratio of the density of non-zeros in the upper right matrix block (that is, the region of directed bipartite connectivity) to the density of the remaining L-shaped block (plus 1 to avoid division by zero). The experiment was repeated 1000 times and in no instance did the bipartivity measure of the submatrix from a shuffled network exceed that of the initial submatrix. This indicates that the degree of bipartivity seen in Figure 18 is unlikely to have arisen 'by chance'.

Overall, we believe that these initial results show a proof of principle for the SVD as a tool for discovering directed bipartite communities. In the context of analysing high-throughput expression data its main utility, of course, will lie in the case where directed bipartite patterns are found that involve gene groups for which annotational information is missing or only partially known. In this way, the algorithm could suggest putative functional and cause-and-effect relationships that may direct more specific experiments. More generally, for any set of unsymmetric interaction data this new method for detecting directed bipartite community structure offers a useful tool for highlighting meaningful information.

References

1. D. Barash, 'Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation', *Bioinformatics* 20 (2004) 1861–1869.
2. E. de Silva and M. P. H. Stumpf, 'Complex networks and simple models in biology', *J. R. Soc. Interface* 2 (2005) 419–430.
3. E. Estrada, 'Protein bipartivity and essentiality in the yeast protein–protein interaction network', *J. Proteome Res.* 5 (2006) 2177–2184.
4. E. Estrada and N. Hatano, 'Communicability in complex networks', *Phys. Rev. E* 77 (2008) 036111.
5. 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.
6. E. Estrada and J. Rodríguez-Velázquez, 'Spectral measures of bipartivity in complex networks', *Phys. Rev. E* 72 (2005) 046105.
7. G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd edn (Johns Hopkins University Press, Baltimore, 1996).
8. P. Grindrod and M. Kibble, 'Review of uses of network and graph theory concepts within proteomics', *Expert Rev. Proteom.* 1 (2004) 229–238.
9. P. Holme, F. Liljeros, C. R. Edling and B. J. Kim, 'Network bipartivity', *Phys. Rev. E* 68 (2003) 056107.
10. Y. Hu and J. A. Scott, HSL_MC73: a fast multilevel Fiedler and profile reduction code, RAL-TR-2003-36, Numerical Analysis Group, Computational Science and Engineering Department, Rutherford Appleton Laboratory, 2003.
11. J. L. Morrison, R. Breitling, D. J. Higham and D. R. Gilbert, 'A lock-and-key model for protein–protein interactions', *Bioinformatics* 2 (2006) 2012–2019.
12. A. Spence, Z. Stoyanov and J. K. Vass, 'The sensitivity of spectral clustering applied to gene expression data', *Proceedings of the 1st International Conference on Bioinformatics and Biomedical Engineering*, 2007, 1343–1346.
13. A. J. Taylor, 'Computational tools for complex networks', PhD Thesis, University of Strathclyde, 2009.
14. A. Thomas, R. Cannings, N. A. M. Monk and C. Cannings, 'On the structure of protein–protein interaction networks', *Biochem. Soc. Trans.* 31 (2003) 1491–1496.
15. P. J. Valk, R. G. Verhaak, M. A. Beijen, C. A. Erpelinck, S. B. van Waalwijk van Doorn-Khosrovani, J. M. Boer, H. B. Beverloo, M. J. Moorhouse, P. J. van der Spek, B. Lüwenberg and R. Delwel, 'Prognostically useful gene-expression profiles in acute myeloid leukemia', *New Engl. J. Med.* 16 (2004) 1617–1628.
16. J. K. Vass, D. J. Higham, X. Mao and D. Crowther, 'New controls of TCA-cycle genes revealed in networks built by discretization or correlation', Technical Report, Department of Mathematics, University of Strathclyde, Glasgow, UK, 2009, 10.
17. C. Walshaw and M. Cross, 'JOSTLE: parallel multilevel graph-partitioning software – an overview', *Mesh partitioning techniques and domain decomposition techniques* (ed. F. Magoules; Saxe-Coburg Publications, Stirling, UK, 2007) 27–58.

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press

Alan Taylor
Department of Mathematics and Statistics
University of Strathclyde
Glasgow, G1 1XH
United Kingdom

ta.atay@maths.strath.ac.uk

Desmond J. Higham
Department of Mathematics and Statistics
University of Strathclyde
Glasgow, G1 1XH
United Kingdom

djh@maths.strath.ac.uk

J. Keith Vass
Translational Medical Research Collaboration,
The Sir James Black Centre
University of Dundee
Dundee, DD1 5EH
United Kingdom

j.k.vass@dundee.ac.uk

https://doi.org/10.1112/S1461157009000072 Published online by Cambridge University Press