Supporting Information

Karl Rohe, Tai Qin, and Bin Yu 10.1073/pnas.XXXXXXXXXX

A. Political blogs

Of the 1222 blogs in the largest connected component, 549 blogs have at least three incoming edges and three outgoing edges. Of these 549, only six blogs change partitions. This supplementary material discusses these six blogs.

All of the six blogs are labeled as Kerry blogs. Five of these blogs are "dem2gop" blogs that appear to take links from Kerry (i.e. dem) blogs and send links to Bush (i.e. gop) blogs. The final blog in the table (quando.net) is the only "gop2dem" blog, taking more edges from Bush blogs and sending links to Kerry blogs. We visited the blog urls and performed related web searches (Table 2). Many of the sites are now defunct.

[1] labels quando.net as a Kerry blog. Upon closer inspection, this blog hosts a collection of conservative/libertarian bloggers. Perhaps their ambiguity is best illustrated with the following quote, "Face it - the only thing Bush can brag about is his comparative conservative advantage over Kerry. And that's akin to saying a tornado is–comparatively–better at home improvement projects than a hurricane."

Table S1. Of the 549 blogs that have at least three incoming edges and at least three outgoing edges, these are the only six blogs whose receiving cluster (from.cluster) is different from their sending cluster (to.cluster). The numbers to the right of the line are generated using the labels provided in the data set. These numbers reveal that di-sim identifies the nodes with asymmetric relationships between the true blocks. Code available at `https://github.com/karlrohe/disim`

blog url	from.clust 2 to.clust	from.dem	from.gop	to.dem	to.gop
chepooka.com	dem2gop	13	2	1	2
clarified.blogspot.com	dem2gop	4	2	0	6
politics.feedster.com	dem2gop	3	0	13	18
polstate.com	dem2gop	31	7	3	2
shininglight.us	dem2gop	2	2	4	7
qando.net	gop2dem	5	57	14	10

Table S2. After accounting for the fact that the data set appears to mislabel quando.net as a liberal blog, all asymmetric blogs link to blogs of the opposite political leaning.

blog url	label in data set	upon visit
chepooka.com	liberal	unclear, possibly defunct
clarified.blogspot.com	liberal	Kerry supporter
politics.feedster.com	liberal	defunct, evidence for Kerry supporter
polstate.com	liberal	defunct, old twitter feed self-identifies as "pan-partisian"
shininglight.us	liberal	defunct
qando.net	liberal	collection of *conservative* bloggers, see `http://www.qando.net/archives/2004_09.htm`

B. *C. elegans*

Figure S1 plots the singular value of L. Section B.1 discusses the results for $K = 7$ clusters. Section B.2 discusses the results for $K = 5$ clusters. Section B.3 discusses the transformation of the edge weights.

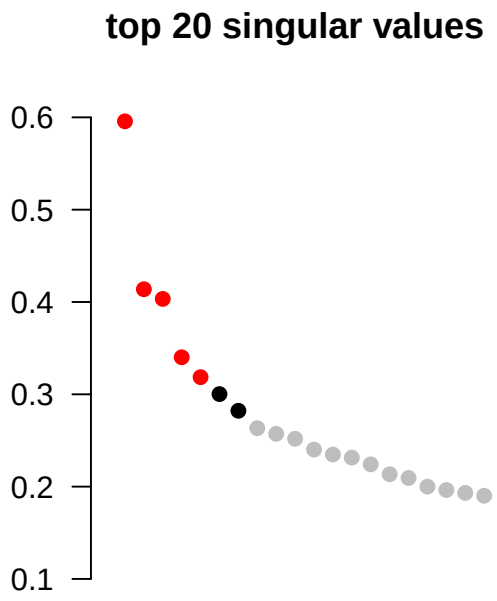

Fig. S1. The first five singular values are colored in red. The next two singular values are black. The rest are grey. There is an elbow after the last black singular value. The analysis in the body of the paper presents results for $K = 7$ because of this elbow. This SI contains the analysis for $K = 5$ to match the analysis in [2]. Data and code available at `https://github.com/karlrohe/disim`.

Results for $K = 7$ clusters. Figure S2 presents three partitions of the nodes. The first two partitions correspond to the sending clusters (on left) and receiving clusters (on right) in DI-SIM.[1] Because the k-means step was run only once, the left and right clusters are comparable. So, the vertical orientation of the clusters is informative because the ith sending cluster from the top sends several edges to the ith receiving cluster from the top.

Each neuron has exactly one line that connects the node's sending cluster to the node's receiving cluster.[1] Darker lines indicate that the neuron moves further between its left and right representations in the singular vectors (as measured by ASYMMETRY SCORE). If there were no co-clustering structure, then all of the lines would be horizontal. However, several lines traverse diagonally, connecting different clusters, and thus indicating non-trivial co-clustering structure. To identify which neurons have a larger ASYMMETRY SCORE, they are written in a slightly larger font in the sending and receiving clusters.

The final partition represented in Figure S2 is represented by the color of the text; this partition corresponds to the communities or modules estimated in [2]. The sensory input to the Response Module (orange) comes from the ventral side of the worm's fan; this module plays an important role in helping the worm physically align with another worm for reproduction. The response module feeds into the Locomotion Module (pink). The locomotion module contains the body-wall motor neurons, helping the worm to move. The R(1-5)A module (green) contains sensory neurons that "promote ventral curling of the tail during mating". The PVV Module (blue) is likely "involved in aspects of male posture during mating". The Insemination Module (yellow) contains neurons that "will take over the male's behavior once the vulva is sensed". The interpretation of the clusters in [2] comes from that paper.

Figure S2 suggests that there is co-clustering structure beyond the standard one-way clustering. In

[1] Some nodes are listed off to the right side of the receiving cluster. These are nodes that do not send any edges, thus they do not have a sending cluster. These nodes are largely motor neurons that control muscles.

[1] The lines in Figure S2 do *not* represent the edges in the graph.

particular, PVV, PVX, PVY, and PVZ neurons are all in large bold font because they have large movement scores. These findings are consistent with the discussion of feedforward circuits in [2]. In particular, Figure 6 in [2] illustrates how these neurons (PVV, PVX, PVY, and PVZ) send most of their edges to neurons in separate clusters.

Results for $K = 5$ clusters.

1. The matrix M for $K = 5$ is given in Figure S4 below. This is the analogue of Figure 3 in the body of the paper.

2. The table in Figure S5 gives the number of nodes with sending cluster u and receving cluster v in element u, v. This is the analogue of the table in Figure 5 of the body of the paper. This table has a feedback score of eight. Under the permutation test described in the paper, only 4.6% of the permuted networks have a smaller feedback score.

3. Figure S6 gives the analogue of Figure S2 (which is also given below). Figure S2 gives the DI-SIM partitions for $K = 7$. Figure S6 gives the DI-SIM partitions for $K = 5$.

Transformation of the edge weights. We investigated three possible transformations of the edge weights: log, square root, and binary (i.e. edge / no edge, which leads to the unweighted graph). Figure S7 gives the histogram of the (non-zero) edge weights and the weights after square root and log transformations.

The results for these three transformations were largely similar. For example, (1) using the novel edge weight transformations to recompute everything up to the feed-forward score and then (2) performing the permutation test that is described in the body of the paper, reveals that both the binary transformation and the square root transformation create statistically significant feed-forward structures (p-values .03 and .002 respectively). However, the transformation is important; with no transformation, DI-SIM struggled to find the hierarchical organization (p-value .15). Moreover, the specific cluster assignments show a dependence on the choice of transformation.

C. Estimating the Stochastic co-Blockmodel with DI-SIM

Throughout, for $x \in \mathbb{R}^d$, $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$, for $M \in \mathbb{R}^{d\times p}$, $\|M\|$ denotes the spectral norm and $\|M\|_F$ denotes the Frobenius norm.

Theorem C.1 bounds the number of nodes that DI-SIM "misclusters". This demonstrates that the co-clusters from DI-SIM estimate both the row- and column-block memberships, one in matrix Y and the other in matrix Z, corresponding to the two types of stochastic equivalence. This implies that the two notions of stochastic equivalence relate to the two sets of singular vectors of L.

Several previous papers have explored the use of spectral tools to aid the estimation of the Stochastic Blockmodel, including [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]; and [15]. The results below build on this previous literature in several ways. Theorem C.1 gives the first statistical estimation results for directed graphs or bipartite graphs with general degree distributions. Because we study a graph that is directed, DI-SIM uses the leading singular vectors of a sparse and asymmetric matrix. As such, the proof required novel extensions of previous proof techniques. These techniques allow the results to also hold for bipartite graphs; previous results for bipartite graphs have only studied computationally intractable techniques, e.g. [16, 17]. For directed graphs and particularly for bipartite graphs, it is not necessarily true that the number of sending clusters should equal the number of receiving clusters. Theorem C.1 below does not presume that the number of sending clusters equals the number of receiving clusters; the theoretical results highlight the statistical price that is paid when they are not equal. Finally, we study a sparse degree corrected model and the theoretical results highlight the importance of the regularization and projection steps in DI-SIM.

Previous theoretical papers that use the non-regularized graph Laplacian all require that the minimum degree grows with the number of nodes (e.g. [7, 12, 15]). However, in many empirical networks, most nodes

have 1, 2, or 3 edges. In these settings, the non-regularized graph Laplacian often has highly localized eigenvectors that are uninformative for estimating large partitions in the graph. Because DI-SIM uses a *regularized* graph Laplacian, the concentration of the singular vectors does not require a growing minimum node degree. Several previous papers have realized the benefits of regularizing the graph Laplacian (e.g. [18, 19, 20, 9, 11, 10]). While the regularized singular vectors concentrate without a growing minimum degree, the weakly connected nodes effect the conclusions through their statistical leverage scores. From the perspective of numerical linear algebra, the leverage scores and the localization of the singular vectors are essential to controlling the algorithmic difficulty of computing the singular vectors [21].

In a diverse set of large empirical networks, the optimal clusters, as judged by a wide variety of graph cut objective functions, are not very large ([22]). To account for this, the results below limit the growth of community sizes by allowing the number of communities to grow with the number of nodes. Previously, [7, 23, 24], and [25] have also studied this high dimensional setting for the undirected Stochastic Blockmodel.

Population notation. Similar to the Stochastic Blockmodel, let $\mathbf{B}$ be a $k_y \times k_z$ matrix where $\mathbf{B}_{ab} \geq 0$ for all a, b. Under the DC-ScBM,

$$P(A_{ij} = 1) = \theta_i^y \theta_j^z \mathbf{B}_{y_i z_j},$$

where $y_i \in \{1, \ldots, k_y\}$ and $z_j \in \{1, \ldots, k_z\}$ indicate the sending and receiving block memberships of i and j, respectively. The parameters θ_i^y and θ_j^z indicate the propensity of i and j to send and receive edges, respective. To be well defined, $\theta_i^y \theta_j^z \mathbf{B}_{y_i z_j} \in [0, 1]$. Note that parameters θ_i^y and θ_j^z are arbitrary to within a multiplicative constant that is absorbed into $\mathbf{B}$. To make it identifiable, we impose the constraint that within each row block, the summation of θ_i^ys is 1. That is, for each row-block s, $\sum_i \theta_i^y \mathbf{1}(Y_{is} = 1) = 1$. Similarly, for any column-block t, we impose $\sum_j \theta_j^z \mathbf{1}(Z_{jt} = 1) = 1$. Under this constraint, $\mathbf{B}$ has explicit meaning: $\mathbf{B_{st}}$ represents the expected number of links from row-block s to column-block t. Recall that $\mathscr{A} = \mathbb{E}(A)$ is the population version of the adjacency matrix A with

$$\mathscr{A} = \Theta_y Y \mathbf{B} Z^T \Theta_z.$$

Define the matrices

$$\begin{aligned} &\mathscr{O}_{jj} = \textstyle\sum_k \mathscr{A}_{kj} \\ &\mathscr{P}_{ii} = \textstyle\sum_k \mathscr{A}_{ik} \\ &\mathscr{O}_\tau = \mathscr{O} + \tau I, \qquad \mathscr{P}_\tau = \mathscr{P} + \tau I \\ &\mathscr{L} = \mathscr{O}_\tau^{-\frac{1}{2}} \mathscr{A} \mathscr{P}_\tau^{-\frac{1}{2}} \end{aligned} \tag{1}$$

where $\mathscr{O}$ and $\mathscr{P}$ are diagonal matrices. The population graph Laplacian $\mathscr{L}$ has an alternative expression in terms of Y and Z.

Lemma C.1. *(Explicit form for $\mathscr{L}_\tau$) Under the DC-ScBM with parameters $\{\mathbf{B}, Y, Z, \Theta_Y, \Theta_Z\}$, define $\Theta_{Y,\tau} \in \mathbb{R}^{N_r \times N_r} (\Theta_{Z,\tau} \in \mathbb{R}^{N_c \times N_c})$ to be diagonal matrix where*

$$[\Theta_{Y,\tau}]_{ii} = \theta_i^Y \frac{\mathscr{O}_{ii}}{\mathscr{O}_{ii} + \tau} \qquad [\Theta_{Z,\tau}]_{jj} = \theta_j^Z \frac{\mathscr{P}_{jj}}{\mathscr{P}_{jj} + \tau}.$$

Then $\mathscr{L}$ has the following form,

$$\mathscr{L} = \mathscr{O}_\tau^{-\frac{1}{2}} \mathscr{A} \mathscr{P}_\tau^{-\frac{1}{2}} = \Theta_{Y,\tau}^{\frac{1}{2}} Y B_L Z^T \Theta_{Z,\tau}^{\frac{1}{2}},$$

for some matrix $B_L \in \mathbb{R}^{k_y \times k_z}$ that is defined in the proof.

The proof of Lemma C.1 is in Section F.

Definition of misclustered. Rigorous discussions of clustering require careful attention to identifiability. In the ScBM, the *order* of the columns of Y and Z are unidentifiable. This leads to difficulty in defining "misclustered". Theorem C.1 uses the following definition of misclustered that is extended from [7].

By the singular value decomposition, there exist orthonormal matrices $\mathscr{X}_L \in \mathbb{R}^{N_r \times k_y}$ and $\mathscr{X}_R \in \mathbb{R}^{N_c \times k_y}$ and diagonal matrix $\Lambda \in \mathbb{R}^{k_y \times k_y}$ such that

$$\mathscr{L} = \mathscr{X}_L \Lambda \mathscr{X}_R^T.$$

Define $\mathscr{X}_L^*$ and $\mathscr{X}_R^*$ as the row normalized population singular vectors,

$$[\mathscr{X}_L^*]_i = \frac{[\mathscr{X}_L]_i}{||[\mathscr{X}_L]_i||_2}, \quad [\mathscr{X}_R^*]_j = \frac{[\mathscr{X}_R]_j}{||[\mathscr{X}_R]_j||_2}.$$

Unless stated otherwise, we will presume without loss of generality that $k_y \leq k_z$. If $\text{rank}(B) = k_y$, then there exist matrices $\mu^y \in \mathbb{R}^{k_y \times k_y}$ and $\mu^z \in \mathbb{R}^{k_z \times k_y}$ such that $Y\mu^y = \mathscr{X}_L^*$ and $Z\mu^z = \mathscr{X}_R^*$ (implied by Lemma F.1). Moreover, the rows of μ^y are distinct; with a slightly stronger assumption, the rows of μ^z are also distinct. As such, k-means applied to the rows of $\mathscr{X}_L^*$ will reveal the partition in Y. Similarly for μ^z, $\mathscr{X}_R^*$, and Z. As such, DI-SIM applied to the population Laplacian, $\mathscr{L}$, can discover the block structure in the matrices Y and Z.

Let $X_L \in \mathbb{R}^{N_r \times k_y}$ be a matrix whose orthonormal columns are the right singular vectors corresponding to the largest k_y singular values of L. DI-SIM applies k-means (with k_y clusters) to the rows of X_L^*, denoted as $u_1, \ldots, u_{N_r}$. Each row is assigned to one cluster and each cluster has a centroid.

Definition 1. *For $i = 1, \ldots, N_r$, define $c_i^L \in \mathbb{R}^{k_y}$ to be the centroid corresponding to u_i after running $(1+\alpha)$-approximate k-means on $u_1, \ldots, u_{N_r}$ with k_y clusters.*

If c_i^L is closer to some population centroid other than its own, i.e. $y_j \mu^y$ for some $y_j \neq y_i$, then we call node i Y-misclustered. This definition must be slightly complicated by the fact that the coordinates in X_L must first align with the coordinates in $\mathscr{X}_L$. So, the definitions below include an additional rotation matrix $\mathscr{R}_L$.

Definition 2. *The set of nodes Y-misclustered is*

$$\mathscr{M}_y = \left\{ i : \|c_i^L - y_i \mu^y \mathscr{R}_L\|_2 > \|c_i^L - y_j \mu^y \mathscr{R}_L\|_2 \ \text{ for any } \ y_j \neq y_i \right\}, \tag{2}$$

where $\mathscr{R}_L$ is the orthonormal matrix that solves Wahba's problem $\min \|X_L - \mathscr{X}_L \mathscr{R}_L\|_F$, i.e. it is the procrustean transformation.

Defining Z-misclustered, requires defining c_i^R and μ^z analogous to the previous definitions.

Definition 3. *The set of nodes Z-misclustered is*

$$\mathscr{M}_z = \left\{ i : \|c_i^R - z_i \mu^z \mathscr{R}_R\|_2 > \|c_i^R - z_j \mu^z \mathscr{R}_R\|_2 \ \text{ for any } \ z_j \neq z_i \right\}, \tag{3}$$

where $\mathscr{R}_R$ is the orthonormal matrix that solves Wahba's problem $\min \|X_R - \mathscr{X}_R \mathscr{R}_R\|_F$, i.e. it is the procrustean transformation.

Asymptotic performance. Define

$$H = (Y^T \Theta_{Y,\tau} Y)^{1/2} B_L (Z^T \Theta_{Z,\tau} Z)^{1/2}.$$

$H \in \mathbb{R}^{k_y \times k_z}$ shares same top K singular values with the population graph Laplacian $\mathscr{L}$. Define $H_{\cdot j}$ as the jth column of H, and define

$$\gamma_z = \min_{i \neq j} \|H_{\cdot i} - H_{\cdot j}\|_2. \tag{4}$$

When $k_z > k_y$, γ_z controls the additional difficulty in estimating Z.

Define m_y as the minimum row length of $\mathscr{X}_L$. Similarly define m_z as the minimum row length of $\mathscr{X}_R$. That is,

$$m_y = \min_{i=1,..,N_r} |||[\mathscr{X}_L]_i||_2, \qquad m_z = \min_{j=1,..,N_c} |||[\mathscr{X}_R]_j||_2. \tag{5}$$

These are the minimum leverage scores for the matrices $\mathscr{L}\mathscr{L}^T$ and $\mathscr{L}^T\mathscr{L}$.

The next theorem bounds the sizes of the sets of misclustered nodes, $|\mathscr{M}_y|$ and $|\mathscr{M}_z|$.

Theorem C.1. *Suppose $A \in \mathbb{R}^{N_r \times N_c}$ is an adjacency matrix sampled from the Degree-Corrected Stochastic co-Blockmodel with k_y left blocks and k_x right blocks. Let $K = \min\{k_y, k_z\} = k_y$. Define $\mathscr{L}$ as in Equation 1. Define $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ as the K nonzero singular values of $\mathscr{L}$. Let $\mathscr{M}_y$ and $\mathscr{M}_z$ be the sets of Y- and Z-misclustered nodes (Equations 2 and 3) by DI-SIM. Let δ be the minimum expected row and column degree of A, that is $\delta = \min(\min_i \mathscr{O}_{ii}, \min_j \mathscr{P}_{jj})$. Define γ_z, m_y and m_z as in Equations 4 and 5. For any $\epsilon > 0$, if $\delta + \tau > 3\ln(N_r + N_c) + 3\ln(4/\epsilon)$, then with probability at least $1 - \epsilon$,*

$$\frac{\mathscr{M}_y}{N_r} \leq c_0(\alpha)\frac{K\ln(4(N_r + N_c)/\epsilon)}{N_r\lambda_K^2 m_y^2(\delta + \tau)}, \tag{6}$$

$$\frac{\mathscr{M}_z}{N_c} \leq c_1(\alpha)\frac{K\ln(4(N_r + N_c)/\epsilon)}{N_c\lambda_K^2 m_z^2\gamma_z^2(\delta + \tau)}. \tag{7}$$

A proof of Theorem C.1 is broken into three parts. Section E shows that the singular vectors of L converge to the singular vectors of $\mathscr{L}$. Section F shows that DI-SIM applied to $\mathscr{L}$ perfectly recovers the two partitions. Finally Section F uses the convergence of L to show that DI-SIM applied to L returns a clustering that is similar to the clustering obtained from applying DI-SIM to $\mathscr{L}$. The rest of this section and the next section interpret the consequences of this Theorem.

Because $\|\mathscr{X}_L\|_F^2 = K$, the average leverage score $|||[\mathscr{X}_L]_i||_2$ is $\sqrt{K/N_r}$. If the m_y is of the same order, with λ_K and K fixed, then $\frac{\mathscr{M}_y}{N_r}$ goes to zero when $\delta + \tau$ grows faster than $\ln(N_r + N_c)$. In sparse graphs, δ is fixed and so τ must grow with n. To ensure that λ_K remains fixed while τ is growing, it is necessary for the *average* degree to also grow.

In many empirical networks, the vast majority of nodes have very small degrees; this is a regime in which δ is not growing. In such networks, the bounds in Equations Eq. (6) and Eq. (7) are vacuous unless $\tau > 0$. While these equations are upper bounds, the simulations in the appendix show that for sparse networks (i.e. δ small), these bounds align with the performance of DI-SIM. Moreover, the performance of DI-SIM is drastically improves with statistical regularization.

These results highlight the sensitivity to the smallest leverage scores m_y and m_z. When there are excessively small leverage scores, then the bound above can become meaningless. However, a slight modification of DI-SIM that excludes the low leveraged points from the k-means step and the clustering results, obtains a vastly improved bound. If one computes the leading singular vectors and only runs k-means on the with the observations i that satisfy $|||[\mathscr{X}_L]_i||_2 > \eta\sqrt{K/N}$, then the theoretical results are much improved. Denote the nodes misclustered by this procedure as $\mathscr{M}_y^*$. Let there be N^* nodes with $|||[\mathscr{X}_L]_i||_2 > \eta\sqrt{K/N}$. If $N/N^* = O(1)$ and the population eigengap λ_K is not asymptotically diminishing, then

$$\frac{\mathscr{M}_y^*}{N^*} \leq c_2(\alpha)\frac{\ln((N_r + N_c)/\epsilon)}{\eta^2(\delta + \tau)}.$$

The proof mimics the proof of Theorem C.1.

In Theorem C.1, the bound for $\mathscr{M}_z$ exceeds the bound for $\mathscr{M}_y$ because the bound for $\mathscr{M}_z$ contains an additional term γ_z. This asymmetry stems from allowing $k_z \geq k_y$. In fact, if $k_y = k_z$, then γ_z can be removed, making the bounds identical. However, if $k_z > k_y$, then $\text{Rank}(\mathscr{L})$ is at most k_y. So, the singular value decomposition represents the data in k_y dimensions and the k-means steps for both the left and the right clusters are done in k_y dimensions. In estimating Y, there is one dimension in the singular vector

representation for each of the k_y blocks. At the same time, the singular value representation shoehorns the k_z blocks in Z into less than k_z dimensions. So, there is less space to separate each of the k_z clusters, obscuring the estimation of Z.

To further understand the bound in Theorem C.1, define the following toy model.

Definition 4. *The **four parameter ScBM** is an ScBM parameterized by $K \in \mathbb{N}, s \in \mathbb{N}, r \in (0,1)$, and $p \in (0,1)$ such that $p + r \leq 1$. The matrices $Y, Z \in \{0,1\}^{n \times K}$ each contain s ones in each column and $B = pI_K + r\mathbf{1}_K\mathbf{1}_K^T$.*

In the four parameter ScBM, there are K left- and right-blocks each with s nodes and the node partitions in Y and Z are not necessarily related. If $y_i = z_j$, then $P(i \to j) = p + r$. Otherwise, $P(i \to j) = r$.

Corollary C.1. *Assume the four parameter ScBM, with same number of rows and columns, and r, p fixed and K growing with $N = Ks$. Since δ is growing with n, set $\tau = 0$. Then,*

$$\lambda_K = \frac{1}{K(r/p) + 1},$$

where λ_K is the Kth largest singular value of $\mathscr{L}$. Moreover,

$$N^{-1}(|\mathscr{M}_y| + |\mathscr{M}_z|) = O_p\left(\frac{K^2 \log N}{N}\right).$$

The proportion of nodes that are misclustered converges to zero, as long as number of clusters $K = o(\sqrt{N/\log N})$.

The proof of Corollary C.1 is contained in Section F.

These theoretical results are novel because they give the first statistical estimation results results for directed graphs or bipartite graphs with general degree distributions. Moreover, because DI-SIM uses the leading singular vectors of a sparse and asymmetric matrix, the proof required novel extensions of previous proof techniques. These techniques extend the spectral results to bipartite graphs; previous results for bipartite graphs have only studied computationally intractable techniques, e.g. [16, 17]. Another novelty comes from the fact that the results do not presume that the number of sending clusters is equal to the number of receiving clusters and the theoretical results highlight the difficulties presented when they are not equal. Finally, because we study a sparse degree corrected model, the theoretical results highlight the importance of the regularization and projection steps in DI-SIM; the concentration results do not depend on the minimum node degree. Instead, the weakly connected nodes affect the conclusions through their statistical leverage scores in the observed graph Laplacian. From the perspective of numerical linear algebra, the leverage scores are essential to controlling the algorithmic difficulty of computing the singular vectors [21].

D. Simulation

The theoretical results of Theorem C.1 identify (1) the expected node degree and (2) the spectral gap as essential parameters that control the clustering performance of DI-SIM. The simulations investigate DI-SIM's non-asymptotic sensitivity to these quantities under the four parameter Stochastic Co-Blockmodel (Definition 4). Moreover, the simulations investigate the performance under the model without degree correction and with degree correction.

Both simulations use $k = 5$ blocks for both Y and Z. Each of the five blocks contains 400 nodes. So, $n = 2000$. When the model is degree corrected, $\theta_1, \ldots, \theta_n$ are iid with $\theta_i \overset{d}{=} \sqrt{Z + .169}$ where $Z \sim$ exponential(1). The addition of .169 ensures that $\mathbb{E}(\theta_i) \approx 1$ and thus the expected degrees are unchanged between the degree corrected model and the model without degree correction.

In the first simulation, the expected node degree is represented on the horizontal axis; the out of block probability r and the in block probability $p+r$ change in a way that keeps the spectral gap of $\mathscr{L}$ fixed across the horizontal axis. In the second simulation, the spectral gap is represented on the horizontal axis; the probabilities p and r change so that the expected degree $pk+rn$ remains fixed at twenty. In both simulations, the partition matrices Y and Z are sampled independently and uniformly over the set of matrices with $s=400$ and $k=5$.

To design the parameter settings of p and r, note that the population graph Laplacian $\mathscr{L}$ is a rank k matrix. So, its $k+1$ eigenvalue is $\lambda_{k+1}=0$ and the spectral gap is $\lambda_k-\lambda_{k+1}=\lambda_k$. Corollary C.1 says that the kth eigenvalue of $\mathscr{L}$ for $\tau=0$ is

$$\lambda_k=\frac{1}{k(r/p)+1}.$$

To keep the spectral gap λ_k fixed, it is equivalent to keeping r/p fixed.

We use the k-means++ algorithm (**(author?)** [26], **(author?)** [27]) with ten initializations. Only the results for Y-misclustered (Definition 2) are reported. Code is provided at http://www.stat.wisc.edu/~karlrohe/.

Simulation 1. This simulation investigates the sensitivity of DI-SIM to a diminishing number of edges. Figure S8 displays the simulation results for a sequence of nine equally spaced values of the expected degree between 5 and 16. To decrease the variability of the plot, each simulation was run twenty times; only the average is displayed. The solid line corresponds to setting the regularization parameter equal to zero ($\tau=0$). The line with longer dashes represents $\tau=1$. The line with small dashes represents the average degree, $\tau=\frac{1}{n}\sum_i P_{ii}$.

Figure S8 demonstrates two things. First, the number of misclustered nodes increases as the expected degree goes to zero. Second, regularization decreases the number of misclustered nodes for small values of the expected degree.

Simulation 2. This simulation investigates the sensitivity of DI-SIM to a diminishing spectral gap λ_k. Figure S8 displays the simulation results for a sequence of nine equally spaced values of the spectral gap, between .3 and .6. In each simulation, the expected degree is held constant at twenty. To decrease the variability, each simulation was run twenty times; only the average is displayed. The solid line corresponds to setting the regularization parameter equal to zero ($\tau=0$). The line with longer dashes represents $\tau=1$. The line with small dashes represents the average degree, $\tau=\frac{1}{n}\sum_i P_{ii}$.

Figure S8 demonstrates two things. First, the number of misclustered nodes increases as the spectral gap goes to zero. Second, regularization yields slight benefits when the spectral gap is small and the model is degree corrected.

E. Convergence of Singular Vectors

The classical spectral clustering algorithm above can be divided into two steps: (1) find the eigendecomposition of L and (2) run k-means. Several previous papers have studied the estimation performance of the classical spectral clustering algorithm under a standard social network model. However, due to the asymmetry of A, previous proof techniques can not be directly applied to study the singular vectors for DI-SIM. In this analysis, we (a) symmetrize the graph Laplacian, (b) apply modern matrix concentration techniques to this symmetrized version of the graph Laplacian, and (c) apply an updated version of the Davis-Kahn theorem to bound the distance between the singular spaces of the empirical and population Laplacian.

For simplicity, from now on let L denote the regularized graph Laplacian.

Define the symmetrized version of L and $\mathscr{L}$ as

$$\tilde{L}=\begin{pmatrix}0 & L\\ L^T & 0\end{pmatrix},\qquad \tilde{\mathscr{L}}=\begin{pmatrix}0 & \mathscr{L}\\ \mathscr{L}^T & 0\end{pmatrix}.$$

The next theorem gives a sharp bound between $\tilde{L}$ and $\tilde{\mathscr{L}}$.

Theorem E.1. *(Concentration of L) Let G be a random graph, with independent edges and* $pr(v_i \sim v_j) = p_{ij}$. *Let* δ *be the minimum expected row and column degree of G, that is* $\delta = \min(\min_i \mathscr{O}_{ii}, \min_j \mathscr{P}_{jj})$. *For any* $\epsilon > 0$, *if* $\delta + \tau > 3\ln(N_r + N_c) + 3\ln(4/\epsilon)$, *then with probability at least* $1 - \epsilon$,

$$\|\tilde{L} - \tilde{\mathscr{L}}\| \leq 4\sqrt{\frac{3\ln(4(N_r + N_c)/\epsilon)}{\delta + \tau}}. \quad [8]$$

Proof. Let $C = \mathscr{P}_\tau^{-\frac{1}{2}} A \mathscr{O}_\tau^{-\frac{1}{2}}$ and define $\tilde{C}$ in the same way as $\tilde{L}$. Then $\|\tilde{L} - \tilde{\mathscr{L}}\| \leq \|\tilde{C} - \tilde{\mathscr{L}}\| + \|\tilde{L} - \tilde{C}\|$. We bound the two terms separately.

For the first term, we apply the following concentration inequality for matrices, see for example **(author?)** [28].

Lemma E.1. *Let* $X_1, X_2, ..., X_m$ *be independent random* $N \times N$ *Hermitian matrices. Moreover, assume that* $\|X_i - \mathbb{E}(X_i)\| \leq M$ *for all i, and* $v^2 = \|\sum var(X_i)\|$. *Let* $X = \sum X_i$. *Then for any* $a > 0$,

$$pr(\|X - \mathbb{E}(X)\| \geq a) \leq 2N \exp\Big(- \frac{a^2}{2v^2 + 2Ma/3}\Big).$$

Let E^{ij} be the matrix with 1 in the i, j and j, i positions and 0 everywhere else. Let $p_{ij} = \mathscr{A}_{ij}$. To use this inequality, express $\tilde{C} - \tilde{\mathscr{L}}$ as the sum of the matrices $Y_{i,m+j}$,

$$Y_{i,m+j} = \frac{1}{\sqrt{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}}(A_{ij} - p_{ij})E^{i,m+j}, i = 1, ..., m, j = 1, ..., n.$$

Note that

$$\|\tilde{C} - \tilde{\mathscr{L}}\| = \|\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{i,m+j}\|,$$

and

$$\|Y_{i,m+j}\| \leq \frac{1}{\sqrt{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}} \leq (\delta + \tau)^{-1}.$$

Moreover,

$$\mathbb{E}[Y_{i,m+j}] = 0 \quad and \quad \mathbb{E}[Y_{i,m+j}^2] = \frac{1}{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}(p_{ij} - p_{ij}^2)(E^{ii} + E^{m+j,m+j}).$$

Then,

$$\begin{aligned}
v^2 &= \|\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}[Y_{i,m+j}^2]\| = \|\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}(p_{ij} - p_{ij}^2)(E^{ii} + E^{m+j,m+j})\| \\
&= \|\sum_{i=1}^{m}[\sum_{j=1}^{n}\frac{1}{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}(p_{ij} - p_{ij}^2)]E^{ii} + \sum_{j=1}^{n}[\sum_{i=1}^{m}\frac{1}{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}(p_{ij} - p_{ij}^2)]E^{m+j,m+j}\| \\
&= \max\Big\{\max_{i=1,...,m}(\sum_{j=1}^{n}\frac{1}{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}(p_{ij} - p_{ij}^2)), \max_{j=1,...,n}(\sum_{i=1}^{m}\frac{1}{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}(p_{ij} - p_{ij}^2))\Big\} \\
&\leq \max\Big\{\max_{i=1,...,m}\frac{1}{\delta + \tau}\sum_{j=1}^{n}\frac{p_{ij}}{\mathscr{O}_{ii} + \tau}, \max_{j=1,...,n}\frac{1}{\delta + \tau}\sum_{i=1}^{m}\frac{p_{ij}}{\mathscr{P}_{jj} + \tau}\Big\} \\
&= (\delta + \tau)^{-1}.
\end{aligned}$$

Take

$$a = \sqrt{\frac{3\ln(4(N_r + N_c)/\epsilon)}{\delta + \tau}}.$$

By assumption, $\delta+\tau > 3\ln(N_r+N_c)+3\ln(4/\epsilon)$. So $a<1$. Applying Lemma E.1,

$$\begin{aligned} pr(\|\tilde{C}-\tilde{\mathscr{L}}\|\geq a) &\leq 2(N_r+N_c)\exp\left(-\frac{\frac{3\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}}{2/(\delta+\tau)+2a/[3(\delta+\tau)]}\right) \\ &\leq 2N\exp(-\frac{3\ln(4(N_r+N_c)/\epsilon)}{3}) \\ &\leq \epsilon/2. \end{aligned}$$

For the second term $\|\tilde{L}-\tilde{C}\|$, define

$$D_\tau=\begin{pmatrix} O_\tau & 0 \\ 0 & P_\tau \end{pmatrix}, \qquad \mathscr{D}_\tau=\begin{pmatrix} \mathscr{O}_\tau & 0 \\ 0 & \mathscr{P}_\tau \end{pmatrix}, \quad D=D_0, \text{ and } \mathscr{D}=\mathscr{D}_0.$$

Apply the two sided concentration inequality for each i, $1\leq i\leq N_r+N_c$, (see for example **(author?)** [29, chap. 2])

$$pr(|D_{ii}-\mathscr{D}_{ii}|\geq\lambda)\leq\exp\{-\frac{\lambda^2}{2\mathscr{D}_{ii}}\}+\exp\{-\frac{\lambda^2}{2\mathscr{D}_{ii}+\frac{2}{3}\lambda}\}.$$

Let $\lambda=a(\mathscr{D}_{ii}+\tau)$, where a is as before.

$$\begin{aligned} pr\Big(|D_{ii}-\mathscr{D}_{ii}|\geq a(\mathscr{D}_{ii}+\tau)\Big) &\leq \exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)^2}{2\mathscr{D}_{ii}}\}+\exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)^2}{2\mathscr{D}_{ii}+\frac{2}{3}a(\mathscr{D}_{ii}+\tau)}\} \\ &\leq 2\exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)^2}{(2+\frac{2}{3}a)(\mathscr{D}_{ii}+\tau)}\} \\ &\leq 2\exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)}{3}\} \\ &\leq 2\exp\{-\ln(4(N_r+N_c)/\epsilon)\frac{(\mathscr{D}_{ii}+\tau)}{\delta+\tau}\} \\ &\leq 2\exp\{-\ln(4(N_r+N_c)/\epsilon)\} \\ &\leq \epsilon/2(N_r+N_c). \end{aligned}$$

Because

$$\|\mathscr{D}_\tau^{-\frac{1}{2}}D_\tau^{\frac{1}{2}}-I\|=max_i\left|\sqrt{\frac{D_{ii}+\tau}{\mathscr{D}_{ii}+\tau}}-1\right|\leq max_i\left|\frac{D_{ii}+\tau}{\mathscr{D}_{ii}+\tau}-1\right|,$$

It follows that

$$\begin{aligned} pr(\|\mathscr{D}_\tau^{-\frac{1}{2}}D_\tau^{\frac{1}{2}}-I\|\geq a) &\leq pr(max_i\left|\frac{D_{ii}+\tau}{\mathscr{D}_{ii}+\tau}-1\right|\geq a) \\ &\leq pr(\cup_i\{|(D_{ii}+\tau)-(\mathscr{D}_{ii}+\tau)|\geq a(\mathscr{D}_{ii}+\tau)\}) \\ &\leq \epsilon/2. \end{aligned}$$

Note that $\|\tilde{L}_\tau\|\leq 1$. Therefore, with probability at least $1-\epsilon/2$,

$$\begin{aligned} \|\tilde{L}_\tau-C\| &= \|D_\tau^{-\frac{1}{2}}\tilde{A}D_\tau^{-\frac{1}{2}}-\mathscr{D}_\tau^{-\frac{1}{2}}\tilde{A}\mathscr{D}_\tau^{-\frac{1}{2}}\| \\ &= \|\tilde{L}_\tau-\mathscr{D}_\tau^{-\frac{1}{2}}D_\tau^{\frac{1}{2}}\tilde{L}_\tau D_\tau^{\frac{1}{2}}\mathscr{D}_\tau^{-\frac{1}{2}}\| \\ &= \|(I-\mathscr{D}_\tau^{-\frac{1}{2}}D_\tau^{\frac{1}{2}})\tilde{L}_\tau D_\tau^{\frac{1}{2}}\mathscr{D}_\tau^{-\frac{1}{2}}+\tilde{L}_\tau(I-D_\tau^{\frac{1}{2}}\mathscr{D}_\tau^{-\frac{1}{2}})\| \\ &\leq \|\mathscr{D}_\tau^{-\frac{1}{2}}D_\tau^{\frac{1}{2}}-I\|\|\mathscr{D}_\tau^{-\frac{1}{2}}D_\tau^{\frac{1}{2}}\|+\|\mathscr{D}_\tau^{-\frac{1}{2}}D_\tau^{\frac{1}{2}}-I\| \\ \leq a^2+2a. \end{aligned}$$

Combining the two parts yields

$$\|\tilde{L}_\tau - \tilde{\mathscr{L}_\tau}\| \leq a^2 + 3a \leq 4a,$$

with probability at least $1-\epsilon$. □

The next theorem bounds the difference between the empirical and population singular vectors in terms of the Frobenius norm.

Theorem E.2. *(Concentration of Singular Space) Let A be the adjacency matrix generated from the DC-ScBM with parameters $\{\mathbf{B}, Y, Z, \Theta_Y, \Theta_Z\}$. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$ be the positive singular values of $\mathscr{L}_\tau$.*

Let $X_L(X_R)$ and $\mathscr{X}_L(\mathscr{X}_R)$ contain the top K left(right) singular vectors of L_τ and $\mathscr{L}_\tau$ respectively. For any $\epsilon > 0$ and sufficiently large N_r and N_r, if $\delta > 3\ln(N_r + N_c) + 3\ln(4/\epsilon)$, then with probability at least $1-\epsilon$

$$\|X_L - \mathscr{X}_L\mathscr{R}_L\|_F \leq \frac{8\sqrt{6}}{\lambda_K}\sqrt{\frac{K\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}} \quad [9]$$

$$\text{and } \|X_R - \mathscr{X}_R\mathscr{R}_R\|_F \leq \frac{8\sqrt{6}}{\lambda_K}\sqrt{\frac{K\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}}, \quad [10]$$

for some orthogonal matrices $\mathscr{R}_L, \mathscr{R}_R \in \mathbb{R}^{K\times K}$.

Proof. Define

$$\tilde{\mathscr{X}} = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathscr{X}_L \\ \mathscr{X}_R \end{pmatrix}.$$

A simple calculation shows that $\tilde{\mathscr{X}} \in \mathbb{R}^{(N_r+N_c)\times K}$ contains the top K eigenvectors of $\tilde{L}$ corresponding to its top K eigenvalues.

We apply an improved version of Davis Kahn theorem from [30]. By a slightly modified proof of Lemma 5.1 in **(author?)** [30], it can be shown that

$$\|\tilde{X}\tilde{X}^T - \tilde{\mathscr{X}}\tilde{\mathscr{X}}^T\|_F \leq \frac{\sqrt{2K}}{\lambda_K}\|\tilde{L}_\tau - \tilde{\mathscr{L}_\tau}\|.$$

Combining it with Theorem E.1 and its assumptions,

$$\|\tilde{X}\tilde{X}^T - \tilde{\mathscr{X}}\tilde{\mathscr{X}}^T\|_F \leq \frac{4\sqrt{6}}{\lambda_K}\sqrt{\frac{K\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}},$$

with probability at lease $1-\epsilon$. By definition of $\tilde{\mathscr{X}}$ and $\tilde{X}$,

$$\begin{aligned}\|\tilde{X}\tilde{X}^T - \tilde{\mathscr{X}}\tilde{\mathscr{X}}^T\|_F &= \left\| \begin{pmatrix} \frac{1}{2}(X_LX_L^T - \mathscr{X}_L\mathscr{X}_L^T) & \frac{1}{2}(X_LX_R^T - \mathscr{X}_L\mathscr{X}_R^T) \\ \frac{1}{2}(X_RX_L^T - \mathscr{X}_R\mathscr{X}_L^T) & \frac{1}{2}(X_RX_R^T - \mathscr{X}_R\mathscr{X}_R) \end{pmatrix} \right\|_F \\ &\geq \frac{1}{2}\|X_LX_L^T - \mathscr{X}_L\mathscr{X}_L^T\|_F \\ &\geq \frac{1}{2}\|X_L - \mathscr{X}_L\mathscr{R}_L\|_F.\end{aligned}$$

Similarly $\|\tilde{X}\tilde{X}^T - \tilde{\mathscr{X}}\tilde{\mathscr{X}}^T\|_F \geq \frac{1}{2}\|X_R - \mathscr{X}_R\mathscr{R}_R\|_F$. This proves the above theorem. □

F. Clustering

To rigorously discuss the asymptotic estimation properties of DI-SIM, the next subsections examine the behavior of DI-SIM applied to a population version of the graph Laplacian $\mathscr{L}$, and compare this to DI-SIM applied to the observed graph Laplacian L.

The population version of DI-SIM. This subsection shows that DI-SIM applied to $\mathscr{L}$ can perfectly identify the blocks in the Stochastic co-Blockmodel. Recall DI-SIM applied to L.

1. Find the left singular vectors $X_L \in \mathbb{R}^{N_r \times k_y}$.
2. Normalize each row of X_L to have unit length. Denote the normalized rows of X_L as $u_1, \ldots, u_{N_r} \in \mathbb{R}^{k_y}$ with and $\|u_i\|_2 = 1$.
3. Run $(1+\alpha)$-approximate k-means on $u_1, \ldots, u_{N_r}$ with k_y clusters.
4. Repeat steps (a), (b), and (c) for the the right singular vectors $X_R \in \mathbb{R}^{N_c \times k_y}$ with k_z clusters.

k-means clusters points $u_1, \ldots, u_n$ in Euclidean space by optimizing the following objective function ([31]),

$$\min_{\{m_1,\ldots,m_{k_y}\} \subset \mathbb{R}^{k_y}} \sum_i \min_g \|u_i - m_g\|_2^2. \qquad [11]$$

Define the *centroids* as the arguments $m_1^*, \ldots, m_{k_y}^*$ that optimize (11). Finding $m_1^*, \ldots, m_{k_y}^*$ is NP-hard. DI-SIM uses a linear time algorithm, $(1+\alpha)$-approximate k-means (**(author?)** [26]). That is, the algorithm computes $\hat{m}_1, \ldots, \hat{m}_{k_y}$ such that

$$\sum_i \min_g \|u_i - \hat{m}_g\|_2^2 \leq (1+\alpha) \sum_i \min_g \|u_i - m_g^*\|_2^2.$$

To study DI-SIM applied to $\mathscr{L}$, Lemma C.1 gives an explicit form as a function of the parameters of the DC-ScBM. Recall that $\mathscr{A} = E(A)$ and under the DC-ScBM,

$$\mathscr{A} = \Theta_y Y \mathbf{B} Z^T \Theta_z,$$

where $Y \in \{0,1\}^{N_r \times k_y}, Z \in \{0,1\}^{N_c \times k_z}$, and $B \in [0,1]^{k_y \times k_z}$. Assume that $k_y \leq k_z$, without loss of generality. Moreover, recall that the regularized population versions of O, P, and L defined in Equation Eq. (1).

The following proves Lemma C.1.

Proof. Define $O_B \in \mathbb{R}^{k_y \times k_y}$ as a diagonal matrix whose (s,s)'th element is $[O_B]_{ss} = \sum_t B_{st}$. Similarly define $P_B \in \mathbb{R}^{k_z \times k_z}$ as a diagonal matrix whose (t,t)'th element is $[P_B]_{tt} = \sum_s B_{st}$. A couple lines of algebra shows that $[O_B]_{ss}$ is the total expected out-degrees of row nodes from block s and that $\mathscr{O}_{ii} = \theta_i^Y [O_B]_{y_i y_i}$. Similarly $[P_B]_{tt}$ is the total expected in-degrees of column nodes from block t and that $\mathscr{P}_{jj} = \theta_j^Z [P_B]_{z_j z_j}$.

Recall that $\mathscr{O}_{ii} = \theta_i^Y [P_B]_{y_i y_i}$ and $\mathscr{P}_{jj} = \theta_j^Z [O_B]_{z_j z_j}$. In addition,

$$[\Theta_{Y,\tau}]_{ii} = \theta_i^Y \frac{\mathscr{O}_{ii}}{\mathscr{O}_{ii} + \tau} \quad \text{and} \quad [\Theta_{Z,\tau}]_{jj} = \theta_j^Z \frac{\mathscr{P}_{jj}}{\mathscr{P}_{jj} + \tau}.$$

The ij'th element of $\mathscr{L}_\tau$ is

$$[\mathscr{L}]_{ij} = \frac{\mathscr{A}_{ij}}{\sqrt{(\mathscr{O}_{ii}+\tau)(\mathscr{P}_{jj}+\tau)}} = \frac{\theta_i^Y \theta_j^Z B_{y_i z_j}}{\sqrt{\mathscr{O}_{ii}\mathscr{P}_{jj}}} \sqrt{\frac{\mathscr{O}_{ii}}{\mathscr{O}_{ii}+\tau} \frac{\mathscr{P}_{jj}}{\mathscr{P}_{jj}+\tau}} = \frac{B_{z_i z_j}}{\sqrt{[P_B]_{y_i}[O_B]_{z_j}}} \sqrt{[\Theta_{Y,\tau}]_{ii}[\Theta_{Z,\tau}]_{jj}}.$$

Hence,

$$\mathscr{L} = \Theta_{Y,\tau}^{\frac{1}{2}} Z B_L Z^T \Theta_{Z,\tau}^{\frac{1}{2}},$$

where B_L is defined as

$$B_L = O_B^{-1/2} B P_B^{-1/2}. \qquad [12]$$

□

Recall that $\mathscr{A} = \Theta_Y Y \mathbf{B} Z^T \Theta_Z$. Lemma C.1 demonstrates that $\mathscr{L}$ has a similarly simple form that separates the block-related information (B_L) and node specific information (Θ_Y and Θ_Z).

Assume that $rank(B_L) = K, 0 < K = k_y \leq k_z$. Recall $H = (Y^T\Theta_{Y,\tau}Y)^{\frac{1}{2}} B_L (Z^T\Theta_{Z,\tau}Z)^{\frac{1}{2}}$. Singular value decomposition of H gives

$$H = U\Lambda V^T.$$

where $U \in \mathbb{R}^{k_y \times K}$ and $V \in \mathbb{R}^{k_z \times K}$ contain the left and right singular vectors of H corresponding to the singular values of H ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$) in the diagonal of $\Lambda \in \mathbb{R}^{K \times K}$. The proof of the next lemma shows that H and $\mathscr{L}$ share the same nonzero singular values.

The next lemma gives the explicit form of the left and right population singular vectors and further shows that their normalized versions are block constant.

Theorem F.1. *(Singular value decomposition for $\mathscr{L}$) Under the DC-ScBM with parameters $\{\mathbf{B}, Y, Z, \Theta_Y, \Theta_Z\}$, Let $\mathscr{X}_L \in \mathbb{R}^{N_r \times K}(\mathscr{X}_R \in \mathbb{R}^{N_c \times K})$ contain the left/right singular vectors of $\mathscr{L}_\tau$. Define $\mathscr{X}_L^*/\mathscr{X}_R^*$to be the row-normalized $\mathscr{X}_L/\mathscr{X}_R$. Then*

1. $\mathscr{X}_L = \Theta_{Y,\tau}^{\frac{1}{2}} Y (Y^T\Theta_{Y,\tau}Y)^{-\frac{1}{2}} U$,
2. $\mathscr{X}_R = \Theta_{Z,\tau}^{\frac{1}{2}} Z (Z^T\Theta_{Z,\tau}Z)^{-\frac{1}{2}} V$.
3. $\mathscr{X}_L^* = YU$, $Y_i \neq Y_j \Leftrightarrow Y_iU \neq Y_jU$.
4. $\mathscr{X}_R^* = ZV^*$, *where* $V_j^* = V_j/\|V_j\|_2$.

Proof. Recall that $H = (Y^T\Theta_{Y,\tau}Y)^{\frac{1}{2}} B_L (Z^T\Theta_{Z,\tau}Z)^{\frac{1}{2}}$ and singular value decompositon of H gives $H = U\Lambda V^T$.

Define $\mathscr{X}_L = \Theta_{Y,\tau}^{\frac{1}{2}} Y (Y^T\Theta_{Y,\tau}Y)^{-\frac{1}{2}} U$, and $\mathscr{X}_R = \Theta_{Z,\tau}^{\frac{1}{2}} Z (Z^T\Theta_{Z,\tau}Z)^{-\frac{1}{2}} V$. It is easy to check that $\mathscr{X}_L^T\mathscr{X}_L = I$ and $\mathscr{X}_R^T\mathscr{X}_R = I$.

On the other hand,

$$\mathscr{X}_L \Lambda \mathscr{X}_R^T = \Theta_{Y,\tau}^{\frac{1}{2}} Y B_L Z^T \Theta_{Z,\tau}^{\frac{1}{2}} = \mathscr{L}.$$

Hence, $\lambda_s, s = 1, ..., r$ are $\mathscr{L}_\tau$'s nonzero singular values and $\mathscr{X}_L/\mathscr{X}_R$ contains $\mathscr{L}_\tau$'s left/right singular vectors corresponding to its nonzero singular values.

Let $\mathscr{X}_L^i$ denote the i'th row of $\mathscr{X}_L$. For part (c), notice that

$$\|\mathscr{X}_L^i\|_2 = \left(\frac{[\Theta_{Y,\tau}]_{ii}}{[Y^T\Theta_{Y,\tau}Y]_{y_iy_i}}\right)^{\frac{1}{2}}.$$

So,

$$[\mathscr{X}_L^*]^i = \frac{\mathscr{X}_L^i}{\|\mathscr{X}_L^i\|_2} = Y_iU.$$

Therefore, $\mathscr{X}_L^* = YU$. For (d), notice that

$$\|\mathscr{X}_R^j\|_2 = \left(\frac{[\Theta_{Z,\tau}]_{jj}\|V_{Z_j}\|^2}{[Z^T\Theta_{Z,\tau}Z]_{z_jz_j}}\right)^{\frac{1}{2}}.$$

Hence,

$$[\mathscr{X}_R^*]^j = \frac{\mathscr{X}_R^j}{\|\mathscr{X}_R^j\|_2} = Z_jV^*.$$

□

Comparing the population and observed clusters. The first part of the section proves the bound of misclustering rate for row nodes.

Clustering for Y**.**

Proof. Recall that the set of misclustered row nodes is defined as:

$$\mathscr{M}_y = \left\{ i : \|c_i^L - y_i \mu^y \mathscr{R}_L\|_2 > \|c_i^L - y_j \mu^y \mathscr{R}_L\|_2 \ \text{ for any } \ y_j \neq y_i \right\}.$$

Let $\mathcal{C}_i$ denote $y_i \mu^y$. Note that Theorem F.1 implies that the population centroid corresponding to the i'th row of $\mathscr{X}_L^*$ is

$$\mathcal{C}_i = y_i \mu^y = y_i U.$$

Since all population centroids are of unit length and are orthogonal to each other, a simple calculation gives a sufficient condition for one observed centroid to be closest to the population centroid:

$$\|c_i^L \mathscr{R}_L^T - \mathcal{C}_i^L\|_2 < 1/\sqrt{2} \Rightarrow \|c_i^L \mathscr{R}_L^T - \mathcal{C}_i^L\|_2 < \|c_i^L \mathscr{R}_L^T - \mathcal{C}_j^L\|_2, \quad \forall j \neq i.$$

Define the following set of nodes that do not satisfy the sufficient condition,

$$\mathscr{B}_y = \{ i : \|c_i^L \mathscr{R}_L^T - \mathcal{C}_i^L\|_2 \geq 1/\sqrt{2} \}.$$

The mis-clustered nodes $\mathscr{M}_y \subset \mathscr{B}_y$.

Define $C_L \in \mathbb{R}^{N_r \times K}$, where the i'th row of C_L is c_i^L, the observed centroid of node i from the $(1+\alpha)$-approximate k-means. Define $M_L \in \mathbb{R}^{N_r \times K}$ to be the global solution of k-means. By definition,

$$\|X_L^* - C_L\|_F \leq (1+\alpha)\|X_L^* - M_L\|_F \leq (1+\alpha)\|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F.$$

Further, by the triangle inequality,

$$\|C_L - YU\mathscr{R}_L\|_F = \|C_L - \mathscr{X}_L^* \mathscr{R}_L\|_F \leq \|X_L^* - C_L\|_F + \|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F \leq (2+\alpha)\|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F.$$

Thus,

$$\begin{aligned}
\frac{|\mathscr{M}_y|}{N_r} \leq \frac{|\mathscr{B}_y|}{N_r} &= \frac{1}{N_r} \sum_{i \in \mathscr{B}_y} 1 \\
&\leq \frac{2}{N_r} \sum_{i \in \mathscr{B}_y} \|c_i^L \mathscr{R}_L^T - \mathcal{C}_i^L\|_2^2 \\
&= \frac{2}{N_r} \|C_L - YU\mathscr{R}_L\|_F^2 \\
&\leq \frac{2(2+\alpha)^2}{N_r} \|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F^2 \\
&\leq \frac{8(2+\alpha)^2}{N_r m_y^2} \|X_L - \mathscr{X}_L \mathscr{R}_L\|_F^2.
\end{aligned}$$

The last inequality is due to the following fact.

Lemma F.1. *For two non-zero vectors v_1, v_2 of the same dimension, we have*

$$\left\| \frac{v_1}{\|v_1\|_2} - \frac{v_2}{\|v_2\|_2} \right\|_2 \leq 2 \frac{\|v_1 - v_2\|_2}{max(\|v_1\|_2, \|v_2\|_2)}.$$

By Theorem E.2, we have, with probability at least $1 - \epsilon$,

$$\frac{|\mathscr{M}_y|}{N_r} \leq c_0(\alpha) \frac{K \ln(4(N_r + N_c)/\epsilon)}{N_r \lambda_K^2 m_y^2 (\delta + \tau)}.$$

□

The second part proves the bound of the misclustering rate for column nodes.

Clustering for Z**.** Because $k_y \leq k_z$, it is slightly more challenging to bound $\mathscr{M}_z$.

Proof. Recall that $H = (Y^T\Theta_{Y,\tau}Y)^{\frac{1}{2}}B_L(Z^T\Theta_{Z,\tau}Z)^{\frac{1}{2}}$ and $H = U\Lambda V^T$. Left multiply by $\Lambda^{-1}U^T$, we have

$$V = H^T U\Lambda^{-1}.$$

Hence

$$\|V_i - V_j\|_2 \geq \frac{1}{\lambda_1}\|H_{\cdot i}U - H_{\cdot j}U\|_2 \geq \|H_{\cdot i} - H_{\cdot j}\|_2.$$

The second inequality is due to the facts that $\lambda_1 \leq 1$ and U is an orthogonal matrix. Recall that

$$\gamma_z = \min_{i\neq j}\|H_{\cdot i} - H_{\cdot j}\|_2 + (1-\kappa),$$

where $\kappa = \max_{i,j}\|V_i\|_2/\|V_j\|_2$. We have that, $\forall i \neq j$,

$$\|V_i^* - V_j^*\|_2 \geq \gamma_z.$$

This is because

$$\begin{aligned}
\|V_i^* - V_j^*\|_2 &= \|\frac{V_i - V_j}{\|V_j\|_2} + V_i(\frac{1}{\|V_i\|_2} - \frac{1}{\|V_j\|_2})\|_2 \\
&\geq \|V_i - V_j\|_2 + 1 - \frac{\|V_i\|_2}{\|V_j\|_2} \\
&\geq \|H_{\cdot i} - H_{\cdot j}\|_2 + (1-\kappa) \\
&\geq \gamma_z.
\end{aligned}$$

Recall that the set of misclustered row nodes is defined as:

$$\mathscr{M}_z = \left\{i : \|c_i^R - z_i\mu^z\mathscr{R}_R\|_2 > \|c_i^R - z_j\mu^z\mathscr{R}_R\|_2 \ \text{ for any } \ z_j \neq z_i\right\}.$$

Let $\mathcal{C}_i^R$ denote $z_i\mu^z$. Note that Theorem F.1 implies that the population centroid corresponding to the i'th row of $\mathscr{X}_R^*$ is

$$\mathcal{C}_i^R = z_i\mu^z = Z_iV^*.$$

Define the following set of column nodes,

$$\mathscr{B}_z = \{i : \|c_i^R\mathscr{R}_R^T - \mathcal{C}_i^R\|_2 \geq \gamma_z/2\}.$$

It is straightforward to show that $\mathscr{M}_z \in \mathscr{B}_z$.

Define $C_R \in \mathbb{R}^{N_c\times K}$, where the i'th row of M is c_i^R, the observed centroid of column node i from $(1+\alpha)$-approximate k-means. Define $M_R \in \mathbb{R}^{N_r\times K}$ to be the global solution of k-means. By definition, we have

$$\|X_R^* - C_R\|_F \leq (1+\alpha)\|X_R^* - M_R\|_F \leq (1+\alpha)\|X_R^* - \mathscr{X}_R^*\mathscr{R}_R\|_F.$$

Further, by the triangle inequality,

$$\|C_R - ZV^*\mathscr{R}_R\|_F = \|C_R - \mathscr{X}_R^*\mathscr{R}_R\|_F \leq \|X_R^* - C_R\|_F + \|X_R^* - \mathscr{X}_R^*\mathscr{R}_R\|_F \leq (2+\alpha)\|X_R^* - \mathscr{X}_R^*\mathscr{R}_R\|_F.$$

Putting all of these pieces together,

$$\begin{aligned}\frac{|\mathscr{M}_z|}{N_c} &\leq \frac{|\mathscr{B}_z|}{N_c} = \frac{1}{N_c}\sum_{i\in\mathscr{B}_z} 1 \\ &\leq \frac{4}{N_c\gamma_z^2}\sum_{i\in\mathscr{B}_y}\|c_i^R\mathscr{R}_L^R - \mathcal{C}_i^R\|_2^2 \\ &= \frac{4}{N_c\gamma_z^2}\|C_R - ZV^*\mathscr{R}_R\|_F^2 \\ &\leq \frac{4(2+\alpha)^2}{N_c\gamma_z^2}\|X_R^* - \mathscr{X}_R^*\mathscr{R}_R\|_F^2 \\ &\leq \frac{16(2+\alpha)^2}{N_c\gamma_z^2 m_z^2}\|X_R - \mathscr{X}_R\mathscr{R}_R\|_F^2.\end{aligned}$$

By Theorem E.2, we have with probability at least $1-\epsilon$,

$$\frac{|\mathscr{M}_z|}{N_c} \leq c_1(\alpha)\frac{K\ln(4(N_r+N_c))/\epsilon)}{N_r\lambda_K^2 m_z^2\gamma_z^2(\delta+\tau)}.$$

□

The following is a proof of Corollary C.1.

Proof. Under the four parameter ScBM, presume that $\theta_i = 1/s$ for all i. From the proof of Theorem F.1, $\mathscr{L}$ has the same singular values as

$$H = (Y^T\Theta_{Y,\tau=0}Y)^{\frac{1}{2}}B_L(Z^T\Theta_{Z,\tau=0}Z)^{\frac{1}{2}} = B_L = O_B^{-\frac{1}{2}}BP_B^{-\frac{1}{2}} = \frac{1}{s^2(Kr+p)}(s^2pI_K + s^2r\mathbf{1}_K\mathbf{1}_K^T).$$

By inspection, the constant vector is an eigenvector of this matrix. It has eigenvalue

$$\lambda_1 = \frac{p+Kr}{Kr+p} = 1.$$

Any vector orthogonal to a constant vector is also an eigenvector. These eigenvectors have eigenvalue

$$\lambda_k = \frac{p}{Kr+p} = \frac{1}{K(r/p)+1}.$$

The result follows from using $m_y^2 = K/n$ (see discussion after Theorem C.1) and $\delta \propto N$. □

G. References

1. Adamic LA, Glance N (2005) The political blogosphere and the 2004 us election: divided they blog in *Proceedings of the 3rd international workshop on Link discovery*. (ACM), pp. 36–43.
2. Jarrell TA et al. (2012) The connectome of a decision-making neural network. *Science* 337(6093):437–444.
3. McSherry F (2001) Spectral partitioning of random graphs in *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*. (IEEE), pp. 529–537.
4. Dasgupta A, Hopcroft JE, McSherry F (2004) Spectral analysis of random graphs with skewed degree distributions in *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*. (IEEE), pp. 602–610.
5. Coja-Oghlan A, Lanka A (2009) Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics* 23(4):1682–1714.
6. Ames BP, Vavasis SA (2010) Convex optimization for the planted k-disjoint-clique problem. *arXiv preprint arXiv:1008.2814*.
7. Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* 39(4):1878–1915.
8. Sussman DL, Tang M, Fishkind DE, Priebe CE (2012) A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* 107(499):1119–1128.
9. Chaudhuri K, Graham FC, Tsiatas A (2012) Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research-Proceedings Track* 23:35–1.
10. Joseph A, Yu B (2014) Impact of regularization on spectral clustering. *arXiv preprint arXiv:1312.1733*.
11. Qin T, Rohe K (2013) Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Advances in Neural Information Processing Systems*.
12. Sarkar P, Bickel PJ (2013) Role of normalization in spectral clustering for stochastic blockmodels. *arXiv preprint arXiv:1310.1495*.
13. Krzakala F et al. (2013) Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences* 110(52):20935–20940.
14. Jin J (2015) Fast community detection by score. *The Annals of Statistics* 43(1):57–89.
15. Lei J, Rinaldo A (2015) Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* 43(1):215–237.
16. Flynn CJ, Perry PO (2012) Consistent biclustering. *arXiv preprint arXiv:1206.6927*.

17. Choi D, Wolfe P (2014) Co-clustering separately exchangeable network data. *Annals of Statistics*.
18. Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: Bringing order to the web.
19. Andersen R, Chung F, Lang K (2006) Local graph partitioning using pagerank vectors in *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*. (IEEE), pp. 475–486.
20. Amini AA, Chen A, Bickel PJ, Levina E, et al. (2013) Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics* 41(4):2097–2122.
21. Mahoney MW (2011) Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning* 3(2):123–224.
22. Leskovec J, Lang K, Dasgupta A, Mahoney M (2008) Statistical properties of community structure in large social and information networks in *Proceeding of the 17th international conference on World Wide Web*. (ACM), pp. 695–704.
23. Choi D, Wolfe P, Airoldi E (2012) Stochastic blockmodels with growing number of classes. *Biometrica (in press)*.
24. Rohe K, Qin T, Fan H (2012) The highest dimensional stochastic blockmodel with a regularized estimator. *arXiv preprint arXiv:1206.2380*.
25. Bhattacharyya S, Bickel PJ (2014) Community detection in networks using graph distance. *arXiv preprint arXiv:1401.3915*.
26. Kumar A, Sabharwal Y, Sen S (2004) A simple linear time (1+ ε)-approximation algorithm for geometric k-means clustering in any dimensions in *Proceedings-Annual Symposium on Foundations of Computer Science*. (IEEE), pp. 454–462.
27. Borchers HW (2012) [r] k-means++. `https://stat.ethz.ch/pipermail/r-help/2012-January/300051.html`.
28. Chung F, Radcliffe M (2011) On the spectra of general random graphs. *the electronic journal of combinatorics* 18(P215):1.
29. Chung FRK, Lu L (2006) *Complex graphs and networks*. (American Mathematical Soc.) No. 107.
30. Lei J, Rinaldo A (2013) Consistency of spectral clustering in sparse stochastic block models. *arXiv preprint arXiv:1312.2050*.
31. Steinhaus H (1956) Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci* 1:801–804.

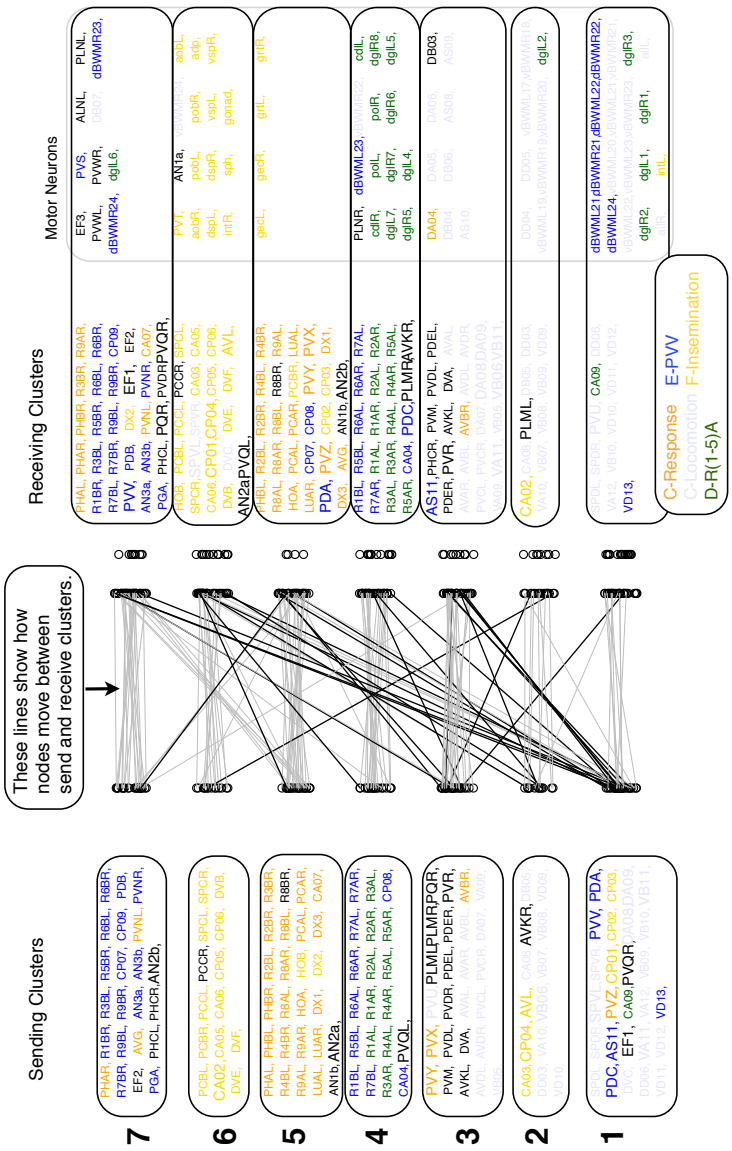

Fig. S2. Results for $K = 7$. Cluster 1 in Figure S3 corresponds to the bottom cluster in this figure, cluster 2 corresponds to the next cluster up, and so on. Data and code available at `https://github.com/karlrohe/disim`.

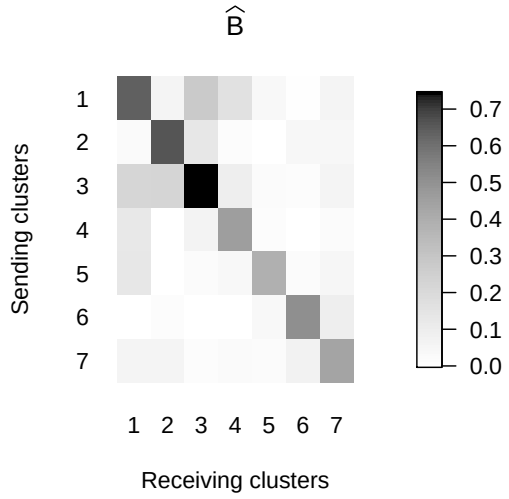

Fig. S3. Results for $K = 7$. Element u, v is darker when there are more (unweighted) edge from block u to block v. A strong diagonal in this matrix suggests that if an edge comes from a node in sending block u, then it probably goes to a node in receiving block u. Data and code available at `https://github.com/karlrohe/disim`.

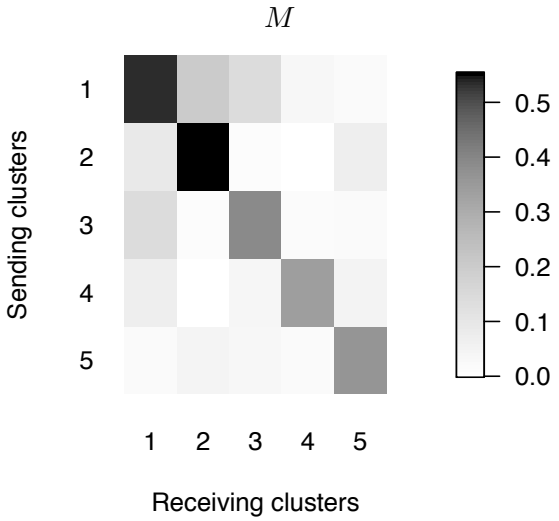

Fig. S4. This is the analogue of Figure 3 in the body of the paper, but for the DI-SIM results with $K = 5$. Element u, v is darker when there are stronger connections from block u to block v. A strong diagonal in this matrix suggests that most connections stay within the same block.

	1	2	3	4	5
1	15	5	2	8	2
2	2	21	1	1	4
3	1	1	20	1	3
4	.	.	.	17	1
5	.	.	3	1	38

Fig. S5. This is the analogue of the table in Figure 5 of the body of the paper. This table gives the results of DI-SIM with $K = 5$. Element u, v is the number of nodes with sending cluster u and receiving cluster v. Only nodes that both send and receive edges are counted.

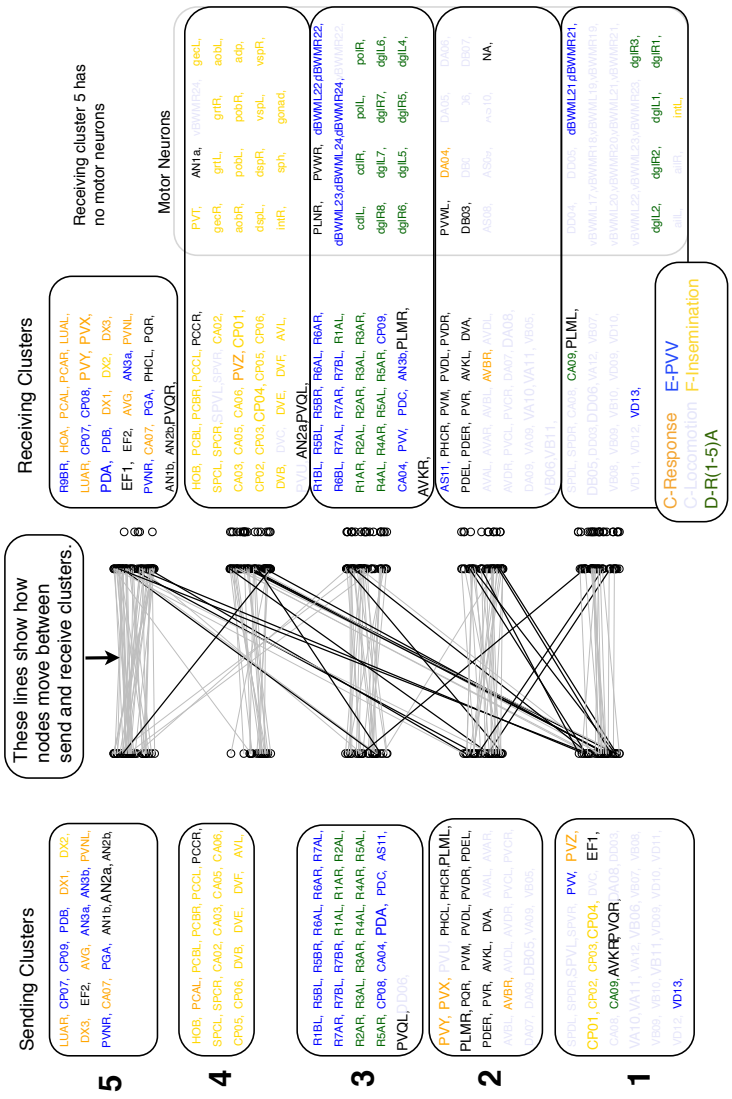

Fig. S6. This is the analogue of Figure S2, but for $K = 5$. Cluster 1 in Figure S4 corresponds to the bottom cluster in this figure, cluster 2 corresponds to the next cluster up, and so on. Data and code available at `https://github.com/karlrohe/disim`.

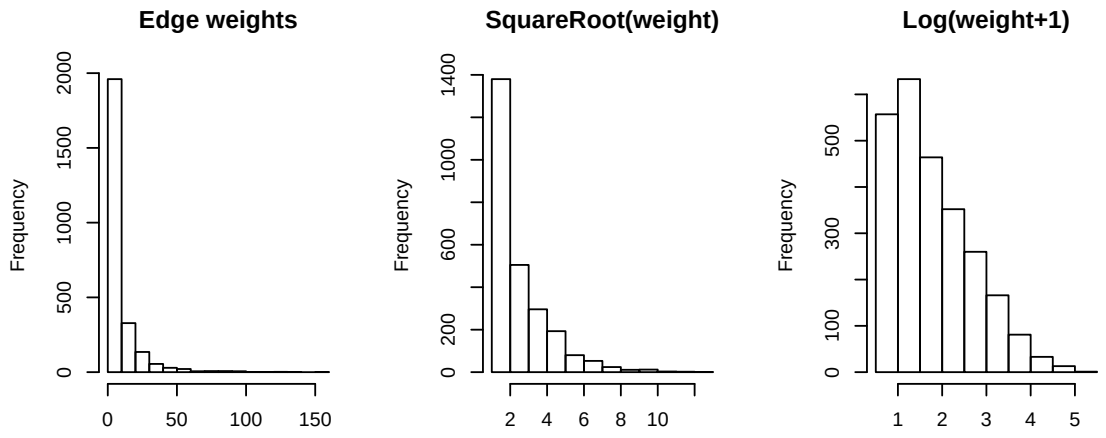

Fig. S7. These three histograms display the nonzero edge weights. In the left panel, there is no transformation. In the center panel, the weights have been transformed with a square root. In the right panel, each edge weight $w \in \mathbb{R}$ is transformed to be $\log(w+1)$. Data and code available at `https://github.com/karlrohe/disim`.

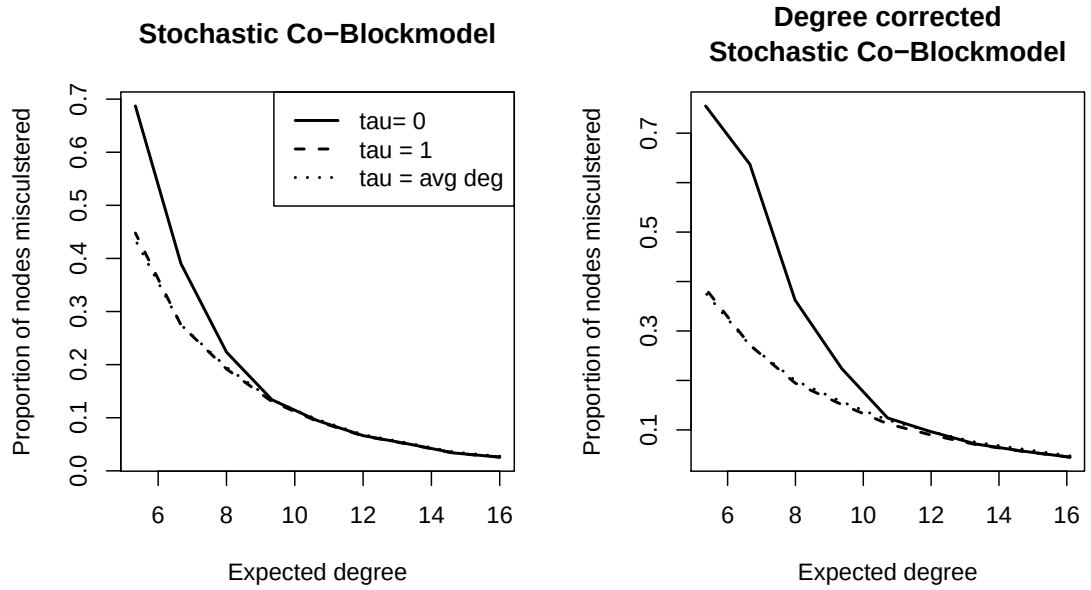

Fig. S8. In the simulation on the left, the data comes from the four parameter Stochastic Co-Blockmodel. On the right, the data comes from the same model, but with degree correction. The θ_i parameters have expectation one. In both models, $k = 5$ and $s = 400$. The probabilities p and r vary such that $p = 5r$, keeping the spectral gap fixed at $\lambda_k = 1/2$. This simulation shows that for small expected degree, regularization decreases the proportion of nodes that are misclustered. Moreover, the benefits of regularization are more pronounced under the degree corrected model.

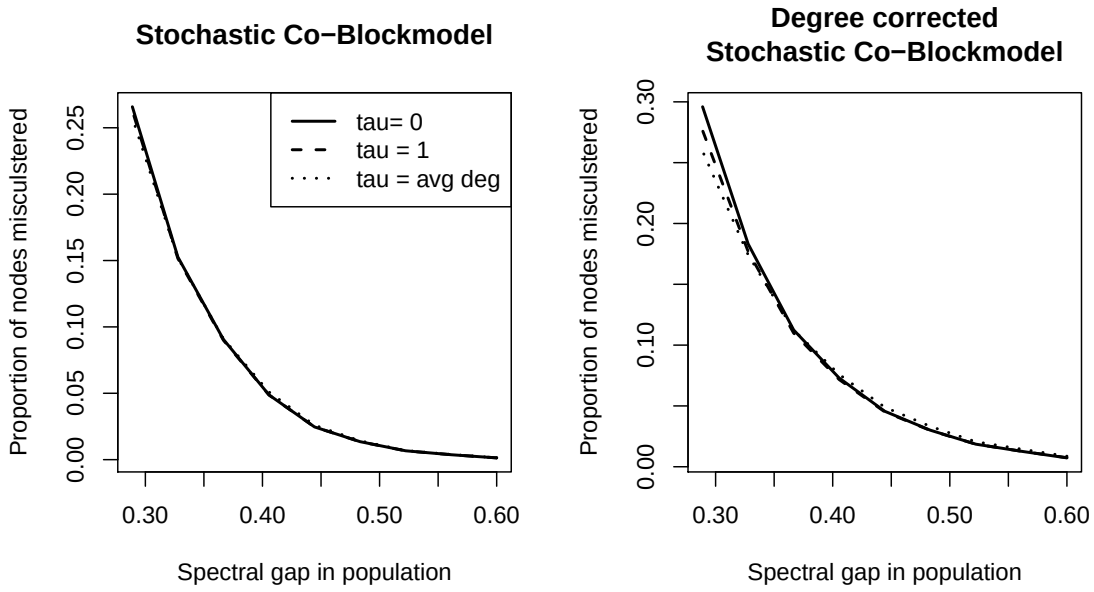

Fig. S9. In the simulation on the left, the data comes from the four parameter Stochastic Co-Blockmodel. On the right, the data comes from the same model, but with degree correction. The θ_i parameters have expectation one. In both models, $k = 5$ and $s = 400$. The spectral gap, displayed on the horizontal axis, changes because the probabilities p and r change. The values of p and r vary in a way that keeps the expected degree fixed at twenty for all simulations. Without degree correction, the three separate lines are difficult to distinguish because they are nearly identical. Under the degree corrected model, regularization improves performance when the spectral gap is small.