A Supplement

A.1 Comparing batch a-MMSB to MMSB on undirected networks

In Table 1, we compared batch a-MMSB to batch MMSB on benchmark networks [28] to justify our modification to MMSB. We found that a-MMSB converges much faster than MMSB, and achieves a better predictive accuracy. We attribute this to the linear complexity of a-MMSB's per-iteration L-step and the issues discussed in §2.

Table 1: Batch a-MMSB converges faster and achieves a better predictive likelihood than batch MMSB using benchmark networks with varying sparsity [1]. Each network has 400 nodes, 10 communities, 20 nodes with at least 3 overlapping communities and an average degree of nodes, d.

	Heldout log likelihood				**Time (mins)**			
d	5	10	15	20	5	10	15	20
MMSB	-1.49	-1.67	-1.89	-2.03	14	336	612	1128
a-MMSB	-1.06	-1.19	-1.16	-1.67	1	6	12	13

A.2 Optimizing the local variational parameters

For the a-MMSB, optimizing the local variational parameters $\boldsymbol{\phi}$ using coordinate ascent involves iteratively updating $\phi_{a\to b}$ holding $\phi_{a\leftarrow b}$ (and all other variational parameters) fixed, and vice-versa. However, we found that using coordinate ascent to optimize local parameters resulted in poor predictive performance. Instead, we use the algorithm in Figure 5, page 1990 of [2], that iteratively updates $\phi^{t+1}_{a\to b}$ using $\phi^{t}_{a\leftarrow b}$ and $\phi^{t+1}_{a\leftarrow b}$ using $\phi^{t}_{a\to b}$ (and the other variational parameters held constant) until convergence. This is a natural gradient ascent with a step-size of one.

A.3 Comparing sampling methods

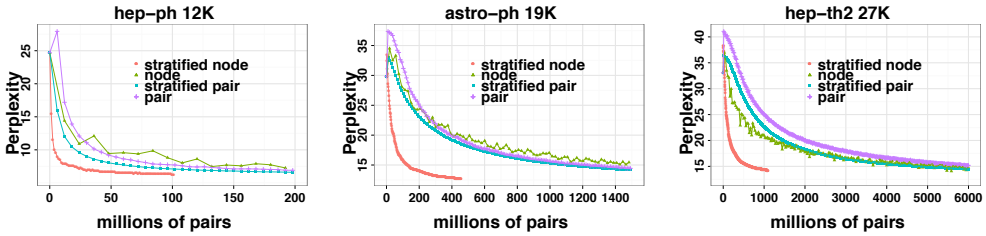

Figure 5: Perplexity vs. number of pairs processed on the hep-ph, astro-ph and hep-th2 datasets.

A.4 Stochastic a-MMSB finds communities as good as batch a-MMSB

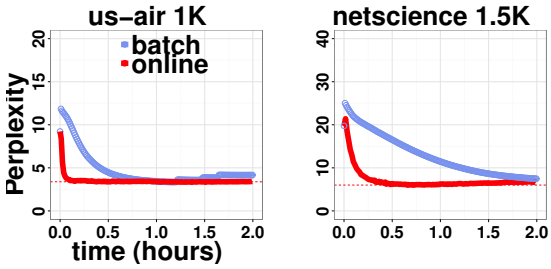

Figure 6: Stochastic a-MMSB (with random pair sampling) scales better than batch a-MMSB on the "netscience" and the "us-air" networks.

A.5 Determining the best number of communities

Figure 7 shows the validation perplexity results on the "relativity" network as K is varied. The best $K = 160$. We analyzed 3 other networks. The best K was 256 for the "hep-th2" network, 15 for the "us-air" network, and 50 for the "netscience" network.

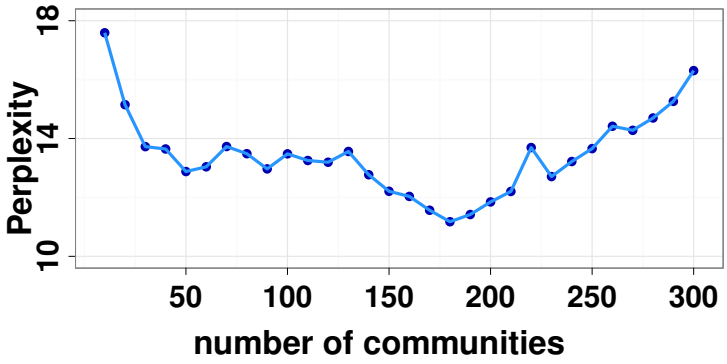

Figure 7: Sensitivity of Stochastic a-MMSB validation perplexity to the number of communities K, on the relativity dataset. The best K is 160.

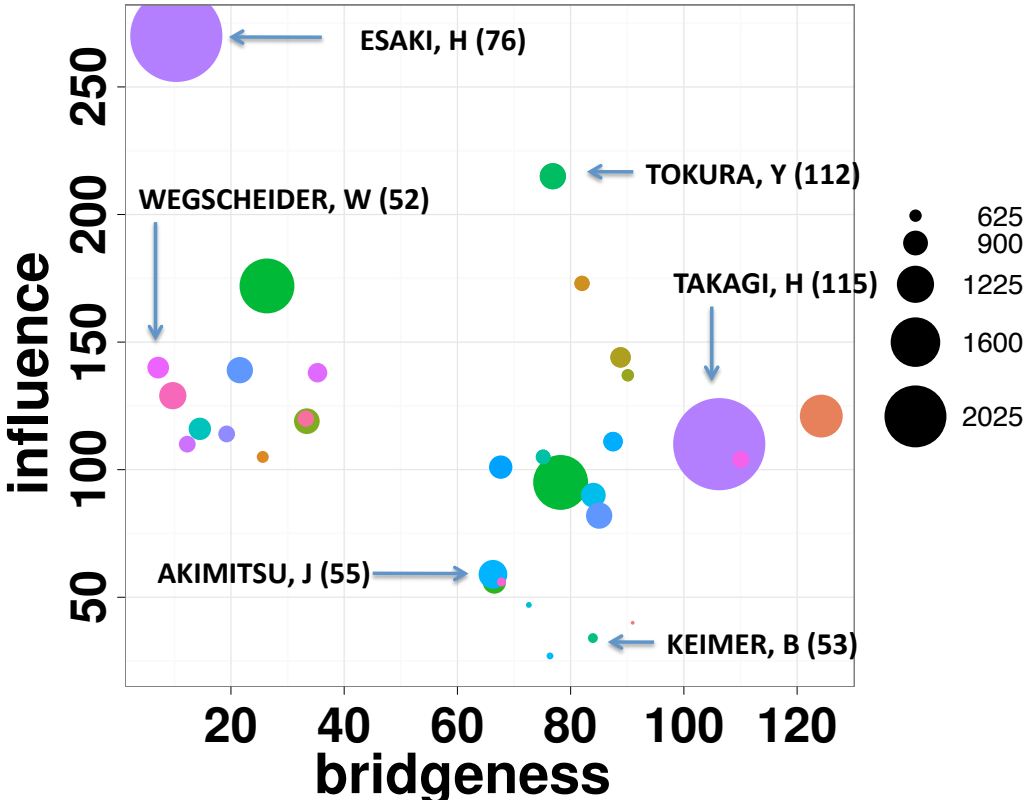

Figure 8: The posterior bridgeness [17] and the posterior influence (centrality within a link community) of top authors (top 20 by influence and top 20 by bridgeness) in the "cond-mat" [26] network is shown. We fit the a-MMSB to the network using Algorithm 1. Each colored circle is the dominant link community of a top author with its size indicating the number of nodes in that community. The names and h-index scores of selected authors is shown.

A.6 Detailed derivation of Stochastic a-MMSB

The joint probability of the data and latent variables in a-MMSB is given by

$$p(\mathbf{y}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\beta}|\alpha, \boldsymbol{\eta}) = \prod_k p(\beta_k|\eta) \prod_p p(\pi_a|\alpha) \prod_{a,b} p(y_{ab}|z_{a\to b}, z_{a\leftarrow b}, \boldsymbol{\beta})\, p(z_{a\to b}|\pi_a)\, p(z_{a\leftarrow b}|\pi_b) \tag{13}$$

For the a-MMSB, the variational distributions are

$$q(z_{a\to b} = k) = \phi_{a\to b,k}; \quad q(\pi_a) = \text{Dirichlet}(\pi_a; \gamma_p); \quad q(\beta_k) = \text{Beta}(\beta_k; \lambda_k). \tag{14}$$

The posterior over link community assignments $\boldsymbol{z}$ is parameterized by the *per-interaction memberships* $\boldsymbol{\phi}$, the node community distributions $\boldsymbol{\pi}$ by the *community memberships* $\boldsymbol{\gamma}$ and the link probability $\boldsymbol{\beta}$ by the *community strengths* $\boldsymbol{\lambda}$. Notice that $\boldsymbol{\lambda}$ is of dimension $K \times 2$, and $\boldsymbol{\gamma}$ is of dimension $N \times K$.

Following [29], we bound the log likelihood of the network adjacency matrix.

$$\log p(\mathbf{y}|\alpha, \boldsymbol{\beta}) \geq \mathcal{L}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \triangleq \mathbb{E}_q[\log p(\mathbf{y}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\beta}|\alpha, \boldsymbol{\eta})] - \mathbb{E}_q[\log q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{z})]. \tag{15}$$

Consider drawing a node pair (a, b) at random from a population distribution $g(a, b)$ over the $M = N(N-1)/2$ node pairs. We can rewrite Eq. 4 as the sum over its global terms and an expectation over the local terms.

$$\begin{aligned}
\mathcal{L} = &\textstyle\sum_k \mathbb{E}_q[\log p(\beta_k|\eta_k)] - \sum_k \mathbb{E}_q[\log q(\beta_k|\lambda_k)] \\
&+ \textstyle\sum_n \mathbb{E}_q[\log p(\pi_n|\alpha)] - \sum_n \mathbb{E}_q[\log q(\pi_n|\gamma_n)] \\
&+ \mathbb{E}_g[\tfrac{1}{g(a,b)} \mathbb{E}_q[\log p(y_{ab}|z_{a\to b}, z_{a\leftarrow b}, \boldsymbol{\beta})]] \\
&+ \mathbb{E}_g[\tfrac{1}{g(a,b)} \mathbb{E}_q[\log p(z_{a\to b}|\pi_a)] + \tfrac{1}{g(a,b)} \mathbb{E}_q[\log p(z_{a\leftarrow b}|\pi_b)]] \\
&- \mathbb{E}_g[\tfrac{1}{g(a,b)} \mathbb{E}_q[\log q(z_{a\to b}|\phi_{a\to b})] - \tfrac{1}{g(a,b)} \mathbb{E}_q[\log q(z_{a\leftarrow b}|\phi_{a\leftarrow b})]]
\end{aligned} \tag{16}$$

Evaluating the rewritten Eq. 16 gives a noisy but unbiased estimate of the ELBO in Eq. 4. Expanding the terms in Eq. 16 (see page 2008 of [2]) and taking the derivative with respect to the global variational parameters we find the following noisy gradients for $\gamma_{a,k}$ and $\lambda_{k,j}$, where $j = 0$ or $j = 1$.

Following [2, 14], the stochastic gradients for the nodes memberships of a and b, and the community strengths $\boldsymbol{\lambda}$ are

$$\frac{\partial \mathcal{L}}{\partial \gamma_{a,j}} = \sum_{k=1}^{K} \frac{\partial \mathbb{E}_q[\log \pi_{a,k}]}{\partial \gamma_{a,j}} \left(\frac{1}{g(a,b)} \phi_{a\to b,k} + \alpha_k - \gamma_{a,k}\right) \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_{b,j}} = \sum_{k=1}^{K} \frac{\partial \mathbb{E}_q[\log \pi_{b,k}]}{\partial \gamma_{b,j}} \left(\frac{1}{g(a,b)} \phi_{a\leftarrow b,k} + \alpha_k - \gamma_{b,k}\right) \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_{k,j}} = \sum_{i=0}^{1} \frac{\partial \mathbb{E}_q[\log \beta_{k,i}]}{\partial \lambda_{k,j}} \left(\frac{1}{g(a,b)} \phi_{a\to b,k} \phi_{a\leftarrow b,k} y_{ab,i} + \eta_{k,i} - \lambda_{k,i}\right) \tag{19}$$

Note that for nodes x that have not been sampled during the iteration, we have the following gradient.

$$\frac{\partial \mathcal{L}}{\partial \gamma_{x,j}} = \sum_{k=1}^{K} \frac{\partial \mathbb{E}_q[\log \pi_{x,k}]}{\partial \gamma_{x,j}} (\alpha_k - \gamma_{x,k}) \tag{20}$$

We now substitute $\mathbb{E}_q[\log \beta_{k,i}]$, $\mathbb{E}_q[\log \pi_{a,k}]$ and $\mathbb{E}_q[\log \pi_{x,k}]$ with the derivative of the log-normalizer of the respective exponential distributions. For example, since $q(\beta_k|\lambda_k)$ is a Beta distribution and is in the exponential family, we can write $\mathbb{E}_q[\log \beta_{k,i}]$ as the derivative of the log partition function of $q(\beta_k|\lambda_k)$, $a(\lambda_k)$.

$$\mathbb{E}_q[\log \beta_{k,i}] = \frac{\partial a(\lambda_k)}{\partial \lambda_{k,i}} \tag{21}$$

For any exponential family $p(x|\eta)$,

$$p(x|\eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\}$$

$$\log p(x|\eta) = \log h(x) + \eta^T t(x) - a(\eta) \tag{22}$$

So we have,

$$\frac{\partial^2 \log p(x|\eta)}{\partial \eta_j \eta_i} = -\frac{\partial^2 a(\eta)}{\partial \eta_j \eta_i} \tag{23}$$

Substituting the derivative of the log-normalizer of the exponential distributions in Eq. 17 and Eq. 19, we have

$$\frac{\partial \mathcal{L}}{\partial \gamma_{a,j}} = \sum_k -\frac{\partial^2 \log q(\pi_a|\gamma_a)}{\partial \gamma_{a,j}\gamma_{a,k}}(\frac{1}{g(a,b)}\phi_{a\to b,k} + \alpha_k - \gamma_{a,k}) \tag{24}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_{k,j}} = \sum_{i=0}^{1} \frac{\partial^2 \log q(\beta_k|\lambda_k)}{\partial \lambda_{k,j}\partial \lambda_{k,i}}(\frac{1}{g(a,b)}\phi_{a\to b,k}\phi_{a\leftarrow b,k}y_{ab,i} + \eta_{k,i} - \lambda_{k,i}) \tag{25}$$

Following [15], we multiply the gradients in Eq. 24 and Eq. 25 by the inverse of the Fisher information matrix of q which gives us the noisy natural gradients in Eq. 26 and Eq. 27.

A similar derivation gives us the noisy natural gradients in the case of set-based sampling (see §3).

$$\partial \gamma^t_{a,k} = \alpha_k + \tfrac{1}{h(t)} \textstyle\sum_{(a,b)\in s_t} \phi^t_{a\to b,k} - \gamma^{t-1}_{a,k} \tag{26}$$

$$\partial \lambda^t_{k,i} = \eta_{k,i} + \tfrac{1}{h(t)} \textstyle\sum_{(a,b)\in s_t} \phi_{a\to b,k} \cdot \phi_{a\leftarrow b,k} \cdot y_{ab,i} - \lambda^{t-1}_{k,i}, \tag{27}$$

where $y_{ab,0} = y_{ab}$, and $y_{ab,1} = 1 - y_{ab}$.