Supplementary Material

Integrating Community and Role Detection in Information Networks

A Detailed derivation of the Gibbs sampling

The collapsed Gibbs sampling relies on the conditional sampling probability of community and role assignment, in which variable θ and π are integrated out. To derive the above conditional sampling probability, we first derive the joint probability of data and membership assignments:

$$
\begin{aligned}
& P(E, Z|\alpha^c, \alpha^r, \xi^1, \xi^2) \\
= & \prod_{(i,j)} P(E_{ij}|Z^c_{ij}, Z^c_{ji}, Z^r_{ij}, Z^r_{ji}, \xi^1, \xi^2) P(Z^c_{ij}|\alpha^c) P(Z^c_{ji}|\alpha^c) P(Z^r_{ij}|\alpha^r) P(Z^r_{ji}|\alpha^r) \\
= & \int \prod_{\delta pq} B_{\delta pq}^{n_{\delta pq+}} (1 - B_{\delta pq})^{n_{\delta pq-}} \prod_{\delta pq} \frac{B_{\delta pq}^{\xi^1 - 1} (1 - B_{\delta pq})^{\xi^2 - 1}}{Beta(\xi^1, \xi^2)} dB \\
& \prod_i \int \prod_a \theta_{ia}^{h_{ia}} \frac{1}{Beta(\alpha^c)} \prod_a \pi_{ia}^{\alpha^c - 1} d\pi_i \\
& \prod_i \int \prod_p \theta_{ip}^{m_{ip}} \frac{1}{Beta(\alpha^r)} \prod_p \theta_{ip}^{\alpha^r - 1} d\theta_i \\
= & \prod_{\delta pq} \frac{Beta(n_{\delta pq+} + \xi^1, n_{\delta pq-} + \xi^2)}{Beta(\xi^1, \xi^2)} \\
& \prod_i \frac{Beta(h_i + \alpha^c)}{Beta(\alpha^c)} \frac{Beta(m_i + \alpha^r)}{Beta(\alpha^r)}
\end{aligned}
\tag{1}
$$

Based on the joint likelihood derived above, we now can derive the conditional sampling probability of membership assignment of given pair of nodes conditioned on the rest of the membership assignments, which is:

$$
\begin{aligned}
& P(Z^c_{ij} = a, Z^c_{ji} = b, Z^r_{ij} = p, Z^r_{ji} = q | E_{ij}, Z_{-ij}, \alpha^r, \alpha^c, \xi^1, \xi^2) \\
\propto & \frac{P(E, Z_{-ij}, Z^c_{ij} = a, Z^c_{ij} = b, Z^r_{ij} = p, Z^r_{ij} = q | \alpha^c, \alpha^r, \xi^1, \xi^2)}{P(E_{-ij}, Z_{-ij}|\alpha^c, \alpha^r, \xi^1, \xi^2)} \\
\propto & \frac{(n^{-ij}_{\delta(a,b)pq+} + \xi^1)^{E_{ij}} (n^{-ij}_{\delta(a,b)pq-} + \xi^2)^{1 - E_{ij}}}{n^{-ij}_{\delta(a,b)pq+} + n^{-ij}_{\delta(a,b)pq-} + \xi^1 + \xi^2} \\
& (h^{-ij}_{ia} + \alpha^c)(h^{-ij}_{jb} + \alpha^c)(m^{-ij}_{ip} + \alpha^r)(m^{-ij}_{jq} + \alpha^r)
\end{aligned}
\tag{2}
$$

B Likelihood and Perplexity

The training/test data likelihood can be useful for convergence monitor and model selection. The training data likelihood $P(E)$ in MMCR can be approximated by the harmonic mean of a set of values of $P(E|Z^{(n)})$ from N samples, where $Z^{(n)}$ is sampled from the posterior $P(Z|E)$. That is:

$$P(E) = \frac{N}{\sum_n^N \frac{1}{P(E|Z^{(n)})}}$$

Where

$$Z^{(n)} \sim P(Z|E)$$

Using the previous derivation in E.q. 1, we can easily compute that:

$$\begin{aligned} &\log P(E|Z^{(n)}) \\ =&\log \prod_{\delta pq} \frac{Beta(n_{\delta pq+} + \xi^1, n_{\delta pq-} + \xi^2)}{Beta(\xi^1, \xi^2)} \\ =&\sum_{\delta pq} \log \mathit{\Gamma}(n_{\delta np+} + \xi^1) + \log \mathit{\Gamma}(n_{\delta np-} + \xi^2) \\ &\quad - \log \mathit{\Gamma}(n_{\delta pq+} + n_{\delta pq-} + \xi^1 + \xi^2) + C \end{aligned} \tag{3}$$

To test the model generality, when training the model, we can hold out a set of links and non-links, and evaluate the held-out likelihood on these links and non-links. The held-out likelihood for a single pair of nodes can be approximated by:

$$\begin{aligned} &P(E_{ij}|E_{\text{observed}}) \\ =&\sum_{pqab} P(E_{ij}|Z^r_{ij} = p, Z^r_{ji} = q, Z^c_{ij} = a, Z^c_{ji} = b, B_{\delta(a,b)pq}) \\ &\quad P(Z^r_{ij} = p|\theta_i) P(Z^c_{ij} = a|\pi_i) P(Z^r_{ji} = q|\theta_j) P(Z^c_{ji} = b|\pi_j) \end{aligned} \tag{4}$$

The perplexity is based on the held-out likelihood, and it is defined as:

$$\text{Perplexity}(E_{\text{test}}) = \exp\left\{ - \frac{\sum_{(i,j)\in E_{\text{test}}} \log P(E_{ij}|E_{\text{observed}})}{|E_{\text{test}}|} \right\} \tag{5}$$