Appendix

The Information Bottleneck (IB) principle is appealing, since it defines what we mean by a good representation, in terms of the fundamental tradeoff between having a concise representation and one with good predictive power (Tishby and Zaslavsky 2015). In this paper, we propose to model the structural mutual information constraint based on Information Bottleneck theory for node representation in graphs. Specifically, IB principle promotes the learned representations to be maximally informative about the target in the downstream task. Essentially, the Information Bottleneck seeks a trade-off between data fit and model generalization. And based on the IB principle, the learned representation is naturally more robust. In this paper, we give the detailed derivations based on deep VIB (Alemi et al. 2017).

A The proof of Eqn. (6)

$$I_\Theta(\mathbf{w}_{ij}, \mathbf{A}_{ij}) = \int d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\;\; p(\mathbf{w}_{ij}, \mathbf{A}_{ij}) \log \frac{p(\mathbf{w}_{ij}, \mathbf{A}_{ij})}{p(\mathbf{w}_{ij})p(\mathbf{A}_{ij})} \tag{18}$$

$$= \int d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\; p(\mathbf{w}_{ij}, \mathbf{A}_{ij}) \log \frac{p(\mathbf{A}_{ij}|\mathbf{w}_{ij})}{p(\mathbf{A}_{ij})} \tag{19}$$

where $p(\mathbf{A}_{ij}|\mathbf{w}_{ij})$ is fully defined by our encoder and Markov Chain as follows:

$$p(\mathbf{A}_{ij}|\mathbf{w}_{ij}) = \int d\mathcal{G}_i\; p(\mathcal{G}_i, \mathbf{A}_{ij}|\mathbf{w}_{ij}) \tag{20}$$

$$= \int d\mathcal{G}_i\; p(\mathbf{A}_{ij}|\mathcal{G}_i)p(\mathcal{G}_i|\mathbf{w}_{ij}) \tag{21}$$

$$= \int d\mathcal{G}_i\; \frac{p(\mathbf{A}_{ij}|\mathcal{G}_i)p(\mathbf{w}_{ij}|\mathcal{G}_i)p(\mathcal{G}_i)}{p(\mathbf{w}_{ij})} \tag{22}$$

Let $q(\mathbf{A}_{ij}|\mathbf{w}_{ij})$ be a variational approximation to $p(\mathbf{A}_{ij}|\mathbf{w}_{ij})$, using the fact that the Kullback Leibler divergence is always positive, we have,

$$D_{KL}(p(\mathbf{A}_{ij}|\mathbf{w}_{ij})||q(\mathbf{A}_{ij}|\mathbf{w}_{ij})) \geq 0 \tag{23}$$

$$\Longrightarrow \int d\mathbf{A}_{ij}\; p(\mathbf{A}_{ij}|\mathbf{w}_{ij})\; \log p(\mathbf{A}_{ij}|\mathbf{w}_{ij}) \tag{24}$$

$$\geq \int d\mathbf{A}_{ij}\; p(\mathbf{A}_{ij}|\mathbf{w}_{ij}) \log q(\mathbf{A}_{ij}|\mathbf{w}_{ij}) \tag{25}$$

and hence,

$$I_\Theta(\mathbf{w}_{ij}, \mathbf{A}_{ij}) \geq \int d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\; p(\mathbf{w}_{ij}, \mathbf{A}_{ij}) \log \frac{q(\mathbf{A}_{ij}|\mathbf{w}_{ij})}{p(\mathbf{A}_{ij})} \tag{26}$$

$$= \int d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\; p(\mathbf{w}_{ij}, \mathbf{A}_{ij}) \log q(\mathbf{A}_{ij}|\mathbf{w}_{ij}) - \int d\mathbf{A}_{ij}\;\; p(\mathbf{A}_{ij}) \log p(\mathbf{A}_{ij}) \tag{27}$$

$$= \int d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\; p(\mathbf{w}_{ij}, \mathbf{A}_{ij}) \log q(\mathbf{A}_{ij}|\mathbf{w}_{ij}) + H(\mathbf{A}_{ij}) \tag{28}$$

Note that the entropy of target $H(\mathbf{A}_{ij})$ is independent of our optimization procedure and so can be ignored. Focusing on the first term in Eqn (28), we leverage the Markov assumption and rewrite $p(\mathbf{w}_{ij}, \mathbf{A}_{ij})$ as

$$p(\mathbf{w}_{ij}, \mathbf{A}_{ij}) = \int d\mathcal{G}_i\; p(\mathcal{G}_i, \mathbf{w}_{ij}, \mathbf{A}_{ij}) \tag{29}$$

$$= \int d\mathcal{G}_i\, d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\; p(\mathcal{G}_i)p(\mathbf{w}_{ij}|\mathcal{G}_i)p(\mathbf{A}_{ij}|\mathcal{G}_i) \tag{30}$$

which gives us the lower bound on the first term of our objective, i.e.,

$$I_\Theta(\mathbf{w}_{ij}, \mathbf{A}_{ij}) \geq \int d\mathcal{G}_i\, d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\;\; p(\mathcal{G}_i)p(\mathbf{w}_{ij}|\mathcal{G}_i)\; p(\mathbf{A}_{ij}|\mathcal{G}_i) \log q(\mathbf{A}_{ij}|\mathbf{w}_{ij}). \tag{31}$$

B The proof of Eqn. (7)

$$I_\Theta(\mathbf{w}_{ij}, \mathcal{G}_i) = \int d\mathbf{w}_{ij}\, d\mathcal{G}_i\; p(\mathbf{w}_{ij}, \mathcal{G}_i) \log \frac{p(\mathbf{w}_{ij}|\mathcal{G}_i)}{p(\mathbf{w}_{ij})}, \tag{32}$$

$$= \int d\mathbf{w}_{ij}\, d\mathcal{G}_i\; p(\mathbf{w}_{ij}, \mathcal{G}_i) \log p(\mathbf{w}_{ij}|\mathcal{G}_i) - \int d\mathbf{w}_{ij}\, d\mathcal{G}_i\; p(\mathbf{w}_{ij}, \mathcal{G}_i) \log p(\mathbf{w}_{ij}). \tag{33}$$

Let $r(\mathbf{w}_{ij})$ be prior distribution. Then,

$$D_{KL}(p(\mathbf{w}_{ij})||r(\mathbf{w}_{ij})) \geq 0 \tag{34}$$

$$\Longrightarrow \int d\mathbf{w}_{ij}\; p(\mathbf{w}_{ij}) \log p(\mathbf{w}_{ij}) \geq \int d\mathbf{w}_{ij}\; p(\mathbf{w}_{ij}) \log r(\mathbf{w}_{ij}) \tag{35}$$

So far, we have derived the Eqn. (7),

$$I_\Theta(\mathbf{w}_{ij}, \mathcal{G}_i) \leq \int d\mathbf{w}_{ij}\, d\mathcal{G}_i\;\; p(\mathcal{G}_i)p(\mathbf{w}_{ij}|\mathcal{G}_i) \log \frac{p(\mathbf{w}_{ij}|\mathcal{G}_i)}{r(\mathbf{w}_{ij})}.$$

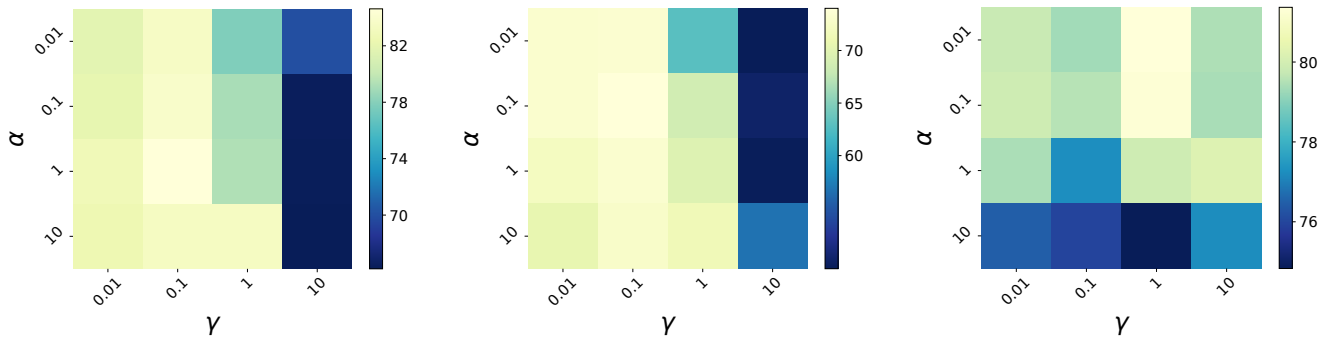
C The proof of Eqn. (9)

$$\zeta = -\int d\mathcal{G}_i\, d\mathbf{w}_{ij}\, d\mathbf{A}_{ij}\; p(\mathcal{G}_i)p(\mathbf{w}_{ij}|\mathcal{G}_i)p(\mathbf{A}_{ij}|\mathcal{G}_i) \log q(\mathbf{A}_{ij}|\mathbf{w}_{ij}) + \beta \int d\mathbf{w}_{ij}\, d\mathcal{G}_i\;\; p(\mathcal{G}_i)p(\mathbf{w}_{ij}|\mathcal{G}_i) \log \frac{p(\mathbf{w}_{ij}|\mathcal{G}_i)}{r(\mathbf{w}_{ij})} \tag{36}$$

We can approximate $p(\mathbf{A}_{ij}, \mathcal{G}_i) = p(\mathcal{G}_i)p(\mathbf{A}_{ij}|\mathcal{G}_i)$ using the empirical data distribution $p(\mathbf{A}_{ij}, \mathcal{G}_i) = \frac{1}{n^2}\sum_n \delta_{\mathbf{A}_n}(\mathbf{A}_{ij})\; \delta_{\mathcal{G}_n}(\mathcal{G}_i)$, then

$$\zeta \approx \frac{1}{n}\sum_n - \int d\mathbf{w}_{ij}\; p(\mathbf{w}_{ij}|\mathcal{G}_n) \log q(\mathbf{A}_n|\mathbf{w}_{ij}) + \beta \int d\mathbf{w}_{ij}\; p(\mathbf{w}_{ij}|\mathcal{G}_n) \log \frac{p(\mathbf{w}_{ij}|\mathcal{G}_n)}{r(\mathbf{w}_{ij})} \tag{37}$$

$$\approx \frac{1}{n^2}\sum_n \mathbb{E}\left[-\log q(\mathbf{A}_n|\mathbf{w}_{ij})\right] + \beta\, D_{KL}\left[p(\mathbf{w}_{ij}|\mathcal{G}_n)||r(\mathbf{w}_{ij})\right] \tag{38}$$



Cora Citeseer Pubmed

Figure A: Heat map about the hyper-parameters sensitivity of α and γ.

D Experimental results

Here, we visualize the results of different hyperparameter settings via 2D heat map. Figure A shows the corresponding results that are same as the Figure 2 in the main paper.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop*, 1–5.