Supplementary Material for: "A Contrastive Variational Graph Auto-Encoder for Node Clustering"

Nairouz Mrabah[a], Mohamed Bouguessa[a], Riadh Ksantini[b]

[a]*Department of Computer Science, University of Quebec at Montreal, Montreal, Quebec, Canada*
[b]*Department of Computer Science, College of IT, University of Bahrain, Zallaq, Bahrain*

Appendix A. Derivation of $\mathcal{L}^{(pre)}_{ELBO}(X, A)$ in Eq. (9)

We prove that $\mathcal{L}^{(pre)}_{ELBO}(X, A)$ (as stated below) is a lower bound of the input graph $\mathcal{G}$ log-likelihood:

$$\mathcal{L}^{(pre)}_{\text{ELBO}}(X,\, A) = \textstyle\sum_{i,j=1}^{N} \mathbb{E}_{z_i,\, z_j \sim q(.\,|\, X,\, A)}\Big[\log\big(p(a_{ij}^{gen} \,|\, z_i,\, z_j)\big)\Big] - 2\,N\, \textstyle\sum_{i=1}^{N} \text{KL}\big(q(z_i \,|\, X,\, A)\,||\, p(z_i)\big).$$

$$\begin{aligned}
\sum_{i,j=1}^{N} KL\big(q(z_i, z_j|X, A)\,||\, p(z_i, z_j|a_{ij}^{gen})\big) &= -\sum_{i,j=1}^{N} \int_{z_i}\int_{z_j} q(z_i|X,A)\, q(z_j|X,A)\log\Big(\frac{p(z_i, z_j|a_{ij}^{gen})}{q(z_i|X,A)\, q(z_j|X,A)}\Big)\, dz_i\, dz_j,\\
&= -\sum_{i,j=1}^{N} \int_{z_i}\int_{z_j} q(z_i|X,A)\, q(z_j|X,A)\log\Big(\frac{p(z_i, z_j, a_{ij}^{gen})}{q(z_i|X,A)\, q(z_j|X,A)\, p(a_{ij}^{gen})}\Big)\, dz_i\, dz_j,\\
&= -\sum_{i,j=1}^{N} \int_{z_i}\int_{z_j} q(z_i|X,A)\, q(z_j|X,A)\log\Big(\frac{p(z_i, z_j, a_{ij}^{gen})}{q(z_i|X,A)\, q(z_j|X,A)}\Big)\, dz_i\, dz_j\\
&+ \sum_{i,j=1}^{N} \log\big(p(a_{ij}^{gen})\big),\\
&= -\sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)}\Big[\log\Big(\frac{p(z_i, z_j, a_{ij}^{gen})}{q(z_i|X,A)\, q(z_j|X,A)}\Big)\Big] + \sum_{i,j=1}^{N} \log\big(p(a_{ij}^{gen})\big).
\end{aligned}$$

Hence, we can write:

$$\begin{aligned}
\log\big(p(A^{gen})\big) &= \sum_{i,j=1}^{N} \log\big(p(a_{ij}^{gen})\big),\\
&= \sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)}\Big[\log\Big(\frac{p(z_i, z_j, a_{ij}^{gen})}{q(z_i|X,A)\, q(z_j|X,A)}\Big)\Big]\\
&+ \sum_{i,j=1}^{N} KL\big(q(z_i, z_j|X, A)\,||\, p(z_i, z_j|a_{ij}^{gen})\big).
\end{aligned}$$

$$
\begin{aligned}
\text{Since}\ \ \mathcal{L}^{(pre)}_{ELBO}(X,\ A) &= \sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)} \Big[\log\Big(p(a^{gen}_{ij}|z_i, z_j)\Big)\Big] -\ 2\,N \sum_{i=1}^{N} KL\Big(q(z_i|X,A)||p(z_i)\Big),\\
&= \sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)} \Big[\log\Big(p(a^{gen}_{ij}|z_i, z_j)\Big)\Big] + N \sum_{i=1}^{N} \mathbb{E}_{z_i \sim q(.|X,A)} \Big[\log\Big(\frac{p(z_i)}{q(z_i|X,A)}\Big)\Big]\\
&+ N \sum_{j=1}^{N} \mathbb{E}_{z_j \sim q(.|X,A)} \Big[\log\Big(\frac{p(z_j)}{q(z_j|X,A)}\Big)\Big],\\
&= \sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)} \Big[\log\Big(p(a^{gen}_{ij}|z_i, z_j)\Big) + \log\Big(\frac{p(z_i)}{q(z_i|X,A)}\Big) + \log\Big(\frac{p(z_j)}{q(z_j|X,A)}\Big)\Big],\\
&= \sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)} \Big[\log\Big(\frac{p(z_i, z_j, a^{gen}_{ij})}{q(z_i|X,A)\, q(z_j|X,A)}\Big)\Big],
\end{aligned}
$$

then, we can obtain the formulation in Eq. (8):

$$
\log\big(p(A^{gen})\big) = \mathcal{L}^{(pre)}_{ELBO}(X,\ A) + \textstyle\sum_{i,j=1}^{N} KL\big(q(z_i, z_j|X,A)\,||\,p(z_i, z_j|a^{gen}_{ij})\big).
$$

Since $\sum_{i,j=1}^{N} KL\big(q(z_i, z_j|X,A)\,||\,p(z_i, z_j|a^{gen}_{ij})\big) \geqslant 0$, we can see that $\mathcal{L}^{(pre)}_{ELBO}(X,\ A)$ is a lower bound of $\log\big(p(A^{gen})\big)$.

The first term $\sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)} \Big[\log\Big(p(a^{gen}_{ij}|z_i, z_j)\Big)\Big]$ of the lower bound $\mathcal{L}^{(pre)}_{ELBO}(X,\ A)$ represents the reconstruction of the adjacency matrix. This term is approximated on the basis of MCMC and the reparameterization trick to ensure a straightforward gradient-based optimization. Given L is the number of Monte Carlo samples, the time complexity for computing this approximation is $\mathcal{O}(dL^2N^2)$ and the obtained expression is:

$$
\begin{aligned}
\sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X,A)} \Big[\log\Big(p(a^{gen}_{ij}|z_i, z_j)\Big)\Big] &\simeq\ \frac{1}{L^2} \sum_{l_1,l_2=1}^{L} \sum_{i,j=1}^{N} \log\Big(p(a^{gen}_{ij}|z^{(l_1)}_i, z^{(l_2)}_j)\Big),\\
&\simeq\ \frac{1}{L^2} \sum_{l_1,l_2=1}^{L} \sum_{i,j=1}^{N} a^{gen}_{ij} \log\Big(\text{Sigmoid}\Big((z^{(l_1)}_i)^T z^{(l_2)}_j\Big)\Big)\\
&+\ \frac{1}{L^2} \sum_{l_1,l_2=1}^{L} \sum_{i,j=1}^{N} (1 - a^{gen}_{ij}) \log\Big(1 - \text{Sigmoid}\Big((z^{(l_1)}_i)^T z^{(l_2)}_j\Big)\Big).
\end{aligned}
$$

The second term $\sum_{i=1}^{N} KL\Big(q(z_i|X,A)\,||\,p(z_i)\Big)$ of the lower bound $\mathcal{L}^{(pre)}_{ELBO}(X,\ A)$ represents a regularization to impose a specific structure on the latent space. Since

$q(z_i|X,A)$ and $p(z_i)$ are two multivariate Gaussian distributions, we can compute their KL divergence analytically with a complexity $\mathcal{O}(dN)$ as follows:

$$\sum_{i=1}^{N} KL\Big(q(z_i|X,A)\,||\,p(z_i)\Big) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{d}\Big(\mu_{z_i}^2[j] + \sigma_{z_i}^2[j] - \log(\sigma_{z_i}^2[j]) - 1\Big).$$

Appendix B. Derivation of $\mathcal{L}_{ELBO}^{(clus)}(X, A)$ in Eq. (16)

We prove that $\mathcal{L}_{ELBO}^{(clus)}(X,A)$ is a lower bound of $\log\big(p(A^{gen})\big)$ such that: $\mathcal{L}_{ELBO}^{(clus)}(X,A) = \mathcal{L}_{ELBO}^{(pre)}(X,A) - 2N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(.|X,A)}\Big[KL\Big(q(c_i|z_i)||p(c_i|z_i)\Big)\Big]$

$$\begin{aligned}
&\log\big(p(A^{gen})\big) = \log\big(\prod_{i,j=1}^{N} p(a_{ij}^{gen})\big),\\
&= \sum_{i,j=1}^{N}\log\big(p(a_{ij}^{gen})\big),\\
&= \sum_{i,j=1}^{N}\log\big(\sum_{c_i}\sum_{c_j}\int_{z_i}\int_{z_j} p(a_{ij}^{gen}, z_i, c_i, z_j, c_j)\,dz_i\,dz_j\big),\\
&= \sum_{i,j=1}^{N}\log\big(\sum_{c_i}\sum_{c_j}\int_{z_i}\int_{z_j} p(a_{ij}^{gen}|z_i,z_j)\,p(c_i|z_i)\,p(c_j|z_j)\,p(z_i)\,p(z_j)\,dz_i\,dz_j\big),\\
&= \sum_{i,j=1}^{N}\log\Big(\sum_{c_i}\sum_{c_j}\int_{z_i}\int_{z_j}\frac{p(a_{ij}^{gen}|z_i,z_j)\,p(c_i|z_i)\,p(c_j|z_j)\,p(z_i)p(z_j)}{q(z_i|X,A)\,q(c_i|z_i)\,q(z_j|X,A)\,q(c_j|z_j)}\,q(z_i|X,A)\,q(c_i|z_i)\,q(z_j|X,A)\,q(c_j|z_j)\,dz_i\,dz_j\Big),\\
&= \sum_{i,j=1}^{N}\log\Big(\mathbb{E}_{\substack{c_i\sim q(.|z_i),\\ c_j\sim q(.|z_j),\\ z_i,z_j\sim q(.|X,A)}}\Big[\frac{p(a_{ij}^{gen}|z_i,z_j)\,p(c_i|z_i)\,p(c_j|z_j)\,p(z_i)\,p(z_j)}{q(z_i|X,A)\,q(c_i|z_i)\,q(z_j|X,A)\,q(c_j|z_j)}\Big]\Big),\\
&\geqslant \sum_{i,j=1}^{N}\mathbb{E}_{\substack{c_i\sim q(.|z_i),\\ c_j\sim q(.|z_j),\\ z_i,z_j\sim q(.|X,A)}}\Big[\log\Big(\frac{p(a_{ij}^{gen}|z_i,z_j)\,p(c_i|z_i)\,p(z_i)\,p(c_j|z_j)\,p(z_j)}{q(c_i|z_i)\,q(z_i|X,A)\,q(c_j|z_j)\,q(z_j|X,A)}\Big)\Big], \quad \text{(Jensen's inequality)}\\
&\geqslant \sum_{i,j=1}^{N}\mathbb{E}_{\substack{c_i\sim q(.|z_i),\\ c_j\sim q(.|z_j),\\ z_i,z_j\sim q(.|X,A)}}\Big[\log\Big(p(a_{ij}^{gen}|z_i,z_j)\Big) + \log\Big(\frac{p(z_i)}{q(z_i|X,A)}\Big) + \log\Big(\frac{p(z_j)}{q(z_j|X,A)}\Big) + \log\Big(\frac{p(c_i|z_i)}{q(c_i|z_i)}\Big) + \log\Big(\frac{p(c_j|z_j)}{q(c_j|z_j)}\Big)\Big],\\
&\geqslant \sum_{i,j=1}^{N}\mathbb{E}_{z_i,z_j\sim q(.|X,A)}\Big[\log\Big(p(a_{ij}^{gen}|z_i,z_j)\Big)\Big] + N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(.|X,A)}\Big[\log\Big(\frac{p(z_i)}{q(z_i|X,A)}\Big)\Big]\\
&+ N\sum_{j=1}^{N}\mathbb{E}_{z_j\sim q(.|X,A)}\Big[\log\Big(\frac{p(z_j)}{q(z_j|X,A)}\Big)\Big] + N\sum_{i=1}^{N}\mathbb{E}_{\substack{c_i\sim q(.|z_i),\\ z_i\sim q(.|X,A)}}\Big[\log\Big(\frac{p(c_i|z_i)}{q(c_i|z_i)}\Big)\Big]
\end{aligned}$$

$$
\begin{aligned}
\log\big(p(A^{gen})\big) \geqslant & \sum_{i,j=1}^{N} \mathbb{E}_{z_i,z_j\sim q(.|X,A)}\Big[\log\Big(p(a_{ij}^{gen}|z_i,z_j)\Big)\Big] + N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(.|X,A)}\Big[\log\Big(\frac{p(z_i)}{q(z_i|X,A)}\Big)\Big] \\
& + N\sum_{j=1}^{N}\mathbb{E}_{z_j\sim q(.|X,A)}\Big[\log\Big(\frac{p(z_j)}{q(z_j|X,A)}\Big)\Big] + N\sum_{i=1}^{N}\mathbb{E}_{\substack{c_i\sim q(.|z_i),\\ z_i\sim q(.|X,A)}}\Big[\log\Big(\frac{p(c_i|z_i)}{q(c_i|z_i)}\Big)\Big] \\
& + N\sum_{i=1}^{N}\mathbb{E}_{\substack{c_j\sim q(.|z_j),\\ z_j\sim q(.|X,A)}}\Big[\log\Big(\frac{p(c_j|z_j)}{q(c_j|z_j)}\Big)\Big], \\
& \geqslant \sum_{i,j=1}^{N} \mathbb{E}_{z_i,z_j\sim q(.|X,A)}\Big[\log\Big(p(a_{ij}^{gen}|z_i,z_j)\Big)\Big] + 2N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(.|X,A)}\Big[\log\Big(\frac{p(z_i)}{q(z_i|X,A)}\Big)\Big] \\
& + 2N\sum_{i=1}^{N}\mathbb{E}_{\substack{c_i\sim q(.|z_i),\\ z_i\sim q(.|X,A)}}\Big[\log\Big(\frac{p(c_i|z_i)}{q(c_i|z_i)}\Big)\Big], \\
& \geqslant \sum_{i,j=1}^{N} \mathbb{E}_{z_i,z_j\sim q(.|X,A)}\Big[\log\Big(p(a_{ij}^{gen}|z_i,z_j)\Big)\Big] - 2N\sum_{i=1}^{N} KL\Big(q(z_i|X,A)||p(z_i)\Big) \\
& - 2N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(.|X,A)}\Big[KL\Big(q(c_i|z_i)||p(c_i|z_i)\Big)\Big], \\
& \geqslant \mathcal{L}_{ELBO}^{(pre)}(X,\ A) - 2N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(.|X,A)}\Big[KL\Big(q(c_i|z_i)||p(c_i|z_i)\Big)\Big] = \mathcal{L}_{ELBO}^{(clus)}(X,\ A).
\end{aligned}
$$

The clustering term can be approximated on the basis of MCMC and the reparameterization trick. The time complexity to compute this approximation is $\mathcal{O}(LKN)$ and the obtained expression is:

$$
\begin{aligned}
2\,N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(z_i|X,A)}\Big[KL\Big(q(c_i|z_i)||p(c_i|z_i)\Big)\Big] &= 2\,N\sum_{i=1}^{N}\mathbb{E}_{z_i\sim q(z_i|X,A)}\Big[\sum_{c_i} q(c_i|z_i)\log\Big(\frac{q(c_i|z_i)}{p(c_i|z_i)}\Big)\Big], \\
&\simeq 2\,\frac{N}{L}\sum_{l=1}^{L}\sum_{i=1}^{N}\sum_{c_i} q(c_i|z_i^{(l)})\log\Big(\frac{q(c_i|z_i^{(l)})}{p(c_i|z_i^{(l)})}\Big).
\end{aligned}
$$

Appendix C. Derivation of $\mathcal{L}_{CVGAE}$ in Eq. (18)

Theorem 1. Given the design choices for the generative and inference models, we have:

$$
\mathcal{L}_{ELBO}^{(clus)}(X^{pos},\ A^{pos}) \leqslant \mathcal{L}_{\text{CVGAE}} \leqslant \log\big(p(A^{gen})\big),
$$

such that $\mathcal{L}_{\text{CVGAE}} = \mathcal{L}^{(clus)}_{ELBO}(X^{pos},\ A^{pos}) + 2\ \sum_{i=1}^{N}\ KL\Big(q(z_i|X^{pos}, A^{pos}) \parallel q(z_i|X^{neg}, A^{neg})\Big)$.

Proof.

Lemma 1. Given the design choices for the generative and inference models, the likelihood of the generated graph structure $A^{gen} = (a^{gen}_{ij})_{1\leq i,j\leq N}$ can be expressed as:

$$\log(A^{gen}) = \sum_{i,j=1}^{N} \text{KL}\Big(q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \parallel p(z_i, z_j, c_i, c_j|a^{gen}_{ij})\Big) + \mathcal{L}^{(clus)}_{ELBO}(X^{pos},\ A^{pos}),$$

$$\begin{aligned}
\text{such that } \mathcal{L}^{(clus)}_{ELBO} = & \sum_{i,j=1}^{N} \mathbb{E}_{z_i,z_j\sim q(.|X^{pos},A^{pos})}\Big[\log\Big(p(a^{gen}_{ij}|z_i, z_j)\Big)\Big] \\
& -\ 2\ N\ \sum_{i=1}^{N} \text{KL}\Big(q(z_i|X^{pos}, A^{pos}) \parallel p(z_i)\Big) \\
& -\ 2\ \mathbb{E}_{z_i\sim q(.|X^{pos},A^{pos})}\Big[\text{KL}\Big(q(c_i|z_i) \parallel p(c_i|z_i)\Big)\Big].
\end{aligned}$$

Proof.

$$\begin{aligned}
&\text{KL}\big(q(z_i,z_j, c_i, c_j|X^{pos}, A^{pos})\big\|p(z_i, z_j, c_i, c_j|a^{gen}_{ij})\big) \\
&\qquad = -\sum_{c_i,c_j}\int_{z_i}\int_{z_j} q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \log\big(\frac{p(z_i, z_j, c_i, c_j|a^{gen}_{ij})}{q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos})}\big)\ dz_i\ dz_j, \\
&\qquad = -\sum_{c_i,c_j}\int_{z_i}\int_{z_j} q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \log\big(\frac{p(z_i, z_j, c_i, c_j, a^{gen}_{ij})}{q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos})}\big)\ dz_i\ dz_j \\
&\qquad + \sum_{c_i,c_j}\int_{z_i}\int_{z_j} q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \log\big(p(a^{gen}_{ij})\big)\ dz_i\ dz_j, \\
&\qquad = -\sum_{c_i,c_j} \mathbb{E}_{z_i,z_j\sim q(.|X^{pos},A^{pos})}\Big[q(c_i|z_i)\ q(c_j|z_j) \log\big(\frac{p(z_i, z_j, c_i, c_j, a^{gen}_{ij})}{q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos})}\big)\Big] + \log\big(p(a^{gen}_{ij})\big).
\end{aligned}$$

Hence, we can write: $\log\big(p(A^{gen})\big) = \log\big(\prod_{i,j=1}^{N} p(a_{ij}^{gen})\big),$

$$= \sum_{i,j=1}^{N} \mathrm{KL}\Big(q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \,\|\, p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\Big)$$

$$+ \sum_{i,j=1}^{N} \sum_{c_i, c_j} \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[q(c_i|z_i)\, q(c_j|z_j) \log\big(\frac{p(z_i, z_j, c_i, c_j, a_{ij}^{gen})}{q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos})}\big) \Big].$$

$$\sum_{c_i, c_j} \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[q(c_i|z_i)\, q(c_j|z_j) \log\big(\frac{p(z_i, z_j, c_i, c_j, a_{ij}^{gen})}{q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos})}\big) \Big]$$

$$= \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[\log\big(p(a_{ij}^{gen} | z_i, z_j)\big) \Big]$$

$$+ \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[\log\big(\frac{p(z_i)}{q(z_i | X^{pos}, A^{pos})}\big) \Big]$$

$$+ \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[\log\big(\frac{p(z_j)}{q(z_j | X^{pos}, A^{pos})}\big) \Big]$$

$$+ \sum_{c_i, c_j} \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[q(c_i|z_i)\, q(c_j|z_j) \log\big(\frac{p(c_i|z_i)}{q(c_i|z_i)}\big) \Big]$$

$$+ \sum_{c_i, c_j} \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[q(c_i|z_i)\, q(c_j|z_j) \log\big(\frac{p(c_j|z_j)}{q(c_j|z_j)}\big) \Big],$$

$$= \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[\log\big(p(a_{ij}^{gen} | z_i, z_j)\big) \Big]$$

$$+ 2\, \mathbb{E}_{z_i \sim q(.|X^{pos}, A^{pos})} \Big[\log\big(\frac{p(z_i)}{q(z_i | X^{pos}, A^{pos})}\big) \Big]$$

$$+ 2 \sum_{c_i} \mathbb{E}_{z_i \sim q(.|X^{pos}, A^{pos})} \Big[q(c_i|z_i) \log\big(\frac{p(c_i|z_i)}{q(c_i|z_i)}\big) \Big]$$

$$= \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})} \Big[\log\big(p(a_{ij}^{gen} | z_i, z_j)\big) \Big]$$

$$- 2\, KL\Big(q(z_i | X^{pos}, A^{pos}) \,\|\, p(z_i)\Big)$$

$$- 2\, \mathbb{E}_{z_i \sim q(.|X^{pos}, A^{pos})} \Big[KL\Big(q(c_i|z_i) \,\|\, p(c_i|z_i)\Big) \Big].$$

$$
\begin{aligned}
\implies \log\big(p(A^{gen})\big) &= \sum_{i,j=1}^{N} KL\Big(q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \,\|\, p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\Big) \\
&+ \sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})}\Big[\log\big(p(a_{ij}^{gen} | z_i, z_j)\big)\Big] \\
&- 2 \sum_{i,j=1}^{N} KL\Big(q(z_i | X^{pos}, A^{pos}) \,\|\, p(z_i)\Big) \\
&- 2 \sum_{i,j=1}^{N} \mathbb{E}_{z_i \sim q(.|X^{pos}, A^{pos})}\Big[KL\Big(q(c_i | z_i) \,\|\, p(c_i | z_i)\Big)\Big], \\
&= \sum_{i,j=1}^{N} KL\Big(q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \,\|\, p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\Big) \\
&+ \sum_{i,j=1}^{N} \mathbb{E}_{z_i, z_j \sim q(.|X^{pos}, A^{pos})}\Big[\log\big(p(a_{ij}^{gen} | z_i, z_j)\big)\Big] \\
&- 2\, N \sum_{i=1}^{N} KL\Big(q(z_i | X^{pos}, A^{pos}) \,\|\, p(z_i)\Big) \\
&- 2\, N \sum_{i=1}^{N} \mathbb{E}_{z_i \sim q(.|X^{pos}, A^{pos})}\Big[KL\Big(q(c_i | z_i) \,\|\, p(c_i | z_i)\Big)\Big], \\
&= \sum_{i,j=1}^{N} KL\Big(q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \,\|\, p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\Big) + \mathcal{L}_{ELBO}^{(clus)}(A^{pos},\, X^{pos}).
\end{aligned}
$$

$\square$

Lemma 2. If $q(z_i, z_j | X^{neg}, A^{neg}) = p(z_i, z_j | a_{ij}^{gen})$,
then $KL\Big(q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \,\|\, p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\Big) \geqslant 2\ KL\Big(q(z_i | X^{pos}, A^{pos}) \,\|\, q(z_i | X^{neg}, A^{neg})\Big)$.

Proof.

$$
\begin{aligned}
&KL\Big(q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \| p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\Big) \\
&\quad= \sum_{c_i, c_j} \int_{z_i} \int_{z_j} q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos})}{p(z_i, z_j, c_i, c_j | a_{ij}^{gen})}\big)\, dz_i\, dz_j, \\
&\quad= \sum_{c_i, c_j} \int_{z_i} \int_{z_j} q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos})\, q(z_i, z_j | X^{neg}, A^{neg})}{p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\, q(z_i, z_j | X^{neg}, A^{neg})}\big)\, dz_i\, dz_j, \\
&\quad= \sum_{c_i, c_j} \int_{z_i} \int_{z_j} q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j | X^{pos}, A^{pos})}{q(z_i, z_j | X^{neg}, A^{neg})}\big)\, dz_i\, dz_j \\
&\quad+ \sum_{c_i, c_j} \int_{z_i} \int_{z_j} q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j | X^{neg}, A^{neg})\, q(c_i | z_i)\, q(c_j | z_j)}{p(z_i, z_j, c_i, c_j | a_{ij}^{gen})}\big)\, dz_i\, dz_j,
\end{aligned}
$$

$$
\begin{aligned}
&KL\Big(q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos})\big\|p(z_i, z_j, c_i, c_j|a_{ij}^{gen})\Big)\\
&\quad= \sum_{c_i,c_j}\int_{z_i}\int_{z_j} q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j|X^{pos}, A^{pos})}{q(z_i, z_j|X^{neg}, A^{neg})}\big)\, dz_i\, dz_j\\
&\quad+ \sum_{c_i,c_j}\int_{z_i}\int_{z_j} q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j|X^{neg}, A^{neg})\; q(c_i|z_i)\; q(c_j|z_j)}{p(z_i, z_j, c_i, c_j|a_{ij}^{gen})}\big)\, dz_i\, dz_j,\\
&\quad= \int_{z_i}\int_{z_j} q(z_i|X^{pos}, A^{pos})\; q(z_j|X^{pos}, A^{pos}) \log\big(\frac{q(z_i|X^{pos}, A^{pos})\; q(z_j|X^{pos}, A^{pos})}{q(z_i|X^{neg}, A^{neg})\; q(z_j|X^{neg}, A^{neg})}\big)\, dz_i\, dz_j\\
&\quad+ \sum_{c_i,c_j}\int_{z_i}\int_{z_j} q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j|X^{neg}, A^{neg})\; q(c_i|z_i)\; q(c_j|z_j)}{p(z_i, z_j|a_{ij}^{gen})\; p(c_i|z_i)\; p(c_j|z_j)}\big)\, dz_i\, dz_j,\\
&\quad= \int_{z_i} q(z_i|X^{pos}, A^{pos}) \log\big(\frac{q(z_i|X^{pos}, A^{pos})}{q(z_i|X^{neg}, A^{neg})}\big)\, dz_i + \int_{z_j} q(z_j|X^{pos}, A^{pos}) \log\big(\frac{q(z_j|X^{pos}, A^{pos})}{q(z_j|X^{neg}, A^{neg})}\big)\, dz_j\\
&\quad+ \underbrace{\sum_{c_i,c_j}\int_{z_i}\int_{z_j} q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \log\big(\frac{q(z_i, z_j|X^{neg}, A^{neg})}{p(z_i, z_j|a_{ij}^{gen})}\big)\, dz_i\, dz_j}_{=0}\\
&\quad+ \sum_{c_i}\int_{z_i} q(z_i, c_i, |X^{pos}, A^{pos}) \log\big(\frac{q(c_i|z_i)}{p(c_i|z_i)}\big)\, dz_i + \sum_{c_j}\int_{z_j} q(z_j, c_j, |X^{pos}, A^{pos}) \log\big(\frac{q(c_j|z_j)}{p(c_j|z_j)}\big)\, dz_j,\\
&\quad= 2\; KL\Big(q(z_i|X^{pos}, A^{pos}) \;\big\|\; q(z_i|X^{neg}, A^{neg})\Big) + \underbrace{2\; N\; \sum_{i=1}^{N} \mathbb{E}_{z_i\sim q(.|X^{pos}, A^{pos})}\Big[KL\Big(q(c_i|z_i) \;\big\|\; p(c_i|z_i)\Big)\Big]}_{\geqslant 0},\\
&\quad\geqslant 2\; KL\Big(q(z_i|X^{pos},\; A^{pos}) \;\big\|\; q(z_i|X^{neg},\; A^{neg})\Big).
\end{aligned}
$$

□

Since maximizing $\mathcal{L}_{ELBO}^{(pre)}(X,\; A)$ during the pretraining phase makes the variational distribution $q(z_i, z_j|X, A)$ approximate the distribution $p(z_i, z_j|a_{ij}^{gen})$, then according to Lemma 2, we have

$$
\begin{aligned}
&KL\Big(q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \;\big\|\; p(z_i, z_j, c_i, c_j|a_{ij}^{gen})\Big) \geqslant 2\; KL\Big(q(z_i|X^{pos}, A^{pos}) \;\big\|\; q(z_i|X^{neg}, A^{neg})\Big),\\
&\implies \sum_{i,j=1}^{N} KL\Big(q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \;\big\|\; p(z_i, z_j, c_i, c_j|a_{ij}^{gen})\Big) \geqslant 2 \sum_{i=1}^{N} KL\Big(q(z_i|X^{pos}, A^{pos}) \;\big\|\; q(z_i|X^{neg}, A^{neg})\Big),\\
&\implies \mathcal{L}_{ELBO}^{(clus)}(X^{pos},\; A^{pos}) + \sum_{i,j=1}^{N} KL\Big(q(z_i, z_j, c_i, c_j|X^{pos}, A^{pos}) \;\big\|\; p(z_i, z_j, c_i, c_j|a_{ij}^{gen})\Big) \geqslant \mathcal{L}_{ELBO}^{(clus)}(X^{pos},\; A^{pos}) + 2 \sum_{i=1}^{N} KL\Big(q(z_i|X^{pos}, A^{pos}) \;\big\|\; q(z_i|X^{neg}, A^{neg})\Big).
\end{aligned}
$$

Based on Lemma 1, we have

$$\log(A^{gen}) = \sum_{i,j=1}^{N} KL\Big(q(z_i, z_j, c_i, c_j | X^{pos}, A^{pos}) \,\|\, p(z_i, z_j, c_i, c_j | a_{ij}^{gen})\Big) + \mathcal{L}_{ELBO}^{(clus)}(X^{pos},\ A^{pos}).$$

Then

$$\log(A^{gen}) \geqslant \mathcal{L}_{ELBO}^{(clus)}(X^{pos},\ A^{pos}) + 2\ \sum_{i=1}^{N}\ KL\Big(q(z_i | X^{pos}, A^{pos}) \,\|\, q(z_i | X^{neg}, A^{neg})\Big).$$

Or

$$2\ \sum_{i=1}^{N}\ KL\Big(q(z_i | X^{pos}, A^{pos}) \,\|\, q(z_i | X^{neg}, A^{neg})\Big) \geqslant 0$$

Then, we conclude

$$\log(A^{gen}) \geqslant \mathcal{L}_{ELBO}^{(clus)}(X^{pos},\ A^{pos}) + 2\ \sum_{i=1}^{N}\ KL\Big(q(z_i | X^{pos}, A^{pos}) \,\|\, q(z_i | X^{neg}, A^{neg})\Big) \geqslant \mathcal{L}_{ELBO}^{(clus)}(X^{pos},\ A^{pos}),$$
$$\implies \log(A^{gen}) \geqslant \mathcal{L}_{\text{CVGAE}} \geqslant \mathcal{L}_{ELBO}^{(clus)}(X^{pos},\ A^{pos}).$$

$\square$

Appendix D. Algorithm for constructing A^{pos}

We provide the full algorithm to build A^{pos}.

Algorithm 1 Function Υ: constructs A^{pos}

Input: Input adjacency matrix: A, Clustering assignments: (p_{ij}), Set of reliable nodes: $\Theta(t)$.
Output: Clustering-oriented graph structure: A^{pos}.
$\Pi \leftarrow [$1-NN$(\Omega_j, \Theta) \mid j \in \{1, ..., K\}]$ $\triangleright$ 1-NN$(\Omega_j,\, \Theta)$ returns the index of the nearest neighbor of Ω_j among the set $\Theta(t)$.
$A^{pos} \leftarrow A$
for i in $\Theta(t)$ **do**
    $k_1 \leftarrow$ arg $\max_k (p_{ik})$
    $l \leftarrow \Pi[k_1]$
    if arg $\max_k (p_{ik}) =$ arg $\max_k (p_{lk})$ **then**
        $a_{il}^{pos} \leftarrow 1$
    end if
    for j in $\Theta(t) \cap$ indices$(A,\, i)$ **do** $\triangleright$ indices$(A,\, i)$ returns the indices of the nodes connected to the i^{th} node.
        if arg $\max_k (p_{ik}) \neq$ arg $\max_k (p_{jk})$ **then**
            $a_{ij}^{pos} \leftarrow 0$
        end if
    end for
end for
Return A^{pos}

Appendix E. Algorithm for constructing A^{gen}

We provide the full algorithm to build the self-supervisory signal A^{gen} for the decoding process.

Algorithm 2 Function Ψ: constructs A^{gen}

Input: Input adjacency matrix: A, Clustering assignments: (p_{ij}), Set of reliable nodes: $\Theta(t)$.
Output: Self-supervisory signal: A^{gen}.
$\Pi \leftarrow [\text{1-NN}(\Omega_j, \Theta) \mid j \in \{1, ..., K\}]$ $\triangleright$ 1-NN(Ω_j, Θ) returns the index of the nearest neighbor of Ω_j among the set $\Theta(t)$.
$A^{gen} \leftarrow A$
for i in $\Theta(t)$ **do**
  $k_1 \leftarrow \arg\max_k (p_{ik})$
  $l \leftarrow \Pi[k_1]$
  if $\arg\max_k (p_{ik}) = \arg\max_k (p_{lk})$ **then**
    $a_{il}^{gen} \leftarrow 1$
  end if
end for
Return A^{gen}

Appendix F. Description of the datasets

We used six datasets for our experimental protocol: Cora [1], Citeseer [1], Pubmed [1], Wiki [2], ACM [3], and DBLP [3]. We provide the statistics of these datasets in Table F.1. A detailed description is provided as follows:

- **Cora**: is a network that contains $2,708$ publications. Two papers are connected if one of them cites the other one. The number of links of this citation network is equal to $5,429$. The features of each node are computed on the basis of a vector of words with 0/1 value from the dictionary.

- **Citeseer**: is a network that contains $3,327$ publications. Two papers are connected if one of them cites the other one. The number of links of this citation network is equal to $4,732$. The features of each node are computed based on 0/1-valued vector of words from the dictionary.

- **Pubmed**: is a network that contains $19,717$ publications. Two papers are connected if one of them cites the other one. The number of links of this citation network is equal to $44,338$. The features of each node are computed based on a TF/IDF-weighted vector of words from the dictionary.

- **Wiki**: is a network containing $2,405$ Wikipedia documents. Two documents are connected if one of them cites the other one. The number of links of this citation network is equal to $17,981$. The features of each node are computed based on a TF/IDF-weighted vector of words from the dictionary.

- **ACM**: is a network containing $3,025$ papers in the fields of database, wireless communication, and data mining. These papers are published in different conferences such as MobiCOMM, VLDB, SIGMOD, KDD, and SIGCOMM. Two papers are connected if they have common authors. The features of each node are computed based on the bag-of-words of keywords.

- **DBLP**: is a network containing $4,057$ authors from different fields of computer science such as machine learning, information retrieval, database, and data mining. Two authors are connected if they have at least one paper in common. The features of each node are computed based on the bag-of-words of keywords.

Table F.1: Dataset statistics.

Dataset	**Cora**	**Citeseer**	**Pubmed**	**Wiki**	**ACM**	**DBLP**
Number of nodes	2,708	3,327	19,717	2,405	3,025	4,057
Number of edges	5,429	4,732	44,338	17,981	26,256	7,056
Number of features	1,433	3,703	500	4,973	1,870	334
Number of classes	7	6	3	17	3	4

Appendix G. Evaluation metrics

The metrics Λ_{FR} [4] and Λ_{FD} [4] assess the level of Feature Randomness (FR) and Feature Drift (FD), respectively. We use the metric AU [5], which computes the number of Active Units in the latent space, to estimate the level of Posterior Collapse. The first metric (i.e., Λ_{FR}) is defined by the cosine of the angle between the gradient of the pseudo-supervised loss computed based on pseudo-labels and the gradient of the corresponding supervised loss computed based on the true labels. Let L_{clus} be the clustering loss, P is the clustering assignment matrix, and Q is the matrix of the true clustering assignments. The equation defining Λ_{FR} is described as follows:

$$\Lambda_{FR} = cos\left(\frac{\partial L_{clus}(P)}{\partial W}, \frac{\partial L_{clus}(Q)}{\partial W}\right). \tag{G.1}$$

The values of Λ_{FR} are within the range $[0, 1]$. Higher values of Λ_{FR} correspond to fewer randomnesses in the optimization process. The second metric (i.e., Λ_{FD}) is defined by the cosine of the angle between the gradient of the general purpose self-supervised loss (i.e., node level, proximity level, graph level) and the task-specific self-supervised loss (i.e., clustering level). This metric captures the level of competition between these two self-supervised losses to assess the deviation from the clustering objective. In our case, we use this metric to evaluate the competition between the gradients of two cross-entropy losses: Vanilla reconstruction of the original graph A defined by $L_{bce}(A)$ and construction of the clustering-oriented graph A^{gen} defined by $L_{bce}(A^{gen})$. The equation defining Λ_{FD} is described as follows:

$$\Lambda_{FD} = cos\left(\frac{\partial L_{bce}(A)}{\partial W}, \frac{\partial L_{bce}(A^{gen})}{\partial W}\right). \tag{G.2}$$

The values of Λ_{FD} are within the range $[0, 1]$. Higher values for Λ_{FD} correspond to less competition between the two self-supervised tasks. Finally, the PC problem is associated with an increasing number of hidden units of the embedded space becoming inactive and unable to capture any useful information from the input signal. This problem is caused by the regularization term $KL\big(q(z_i \,|\, X,\, A) \,||\, p(z_i)\big)$, which pushes the latent variables to follow a standard multivariate Gaussian distribution without any consideration of the input data. More precisely, the hidden units that does not participate into optimizing the pseudo-supervised or self-supervised objectives of the ELBO will be limited to minimizing the regularization term (i.e., they are considered as inactive units). According to [5], we can quantify the activity of a hidden unit using the metric AU. The AU metric of a hidden unit u is described as follow:

$$AU(u) = \text{COV}_{X,A}\left(\mathbb{E}_{u \sim q(u \,|\, X,\, A)}\Big[u\Big]\right). \tag{G.3}$$

Appendix H. Hyperparameters specification

We divide the hyperparameters of our model into two categories. The first category consists of fixed hyperparameters that are independent of the processed dataset. This category is presented in Table H.2. Among this category, we have hyperparameters related to the architecture, the pretraining phase, and the clustering phase. In particular, we pretrain our model for 200 iterations similar to previous VGAE models such as $\mathcal{N}$-VGAE [6] and GMM-VGAE [7]. For the Monte Carlo estimation, we set the number of samples to one ($L = 1$) following the previous methods [6, 7, 8, 9, 10, 11, 12]. The second category consists of two data-dependent hyperparameters M and α. This category is presented in Table H.3. We fix M and α from the ranges $[10,\ 20,\ 30,\ 40,\ 50]$ and $[0.1,\ 0.15,\ 0.2,\ 0.25,\ 0.3,\ 0.35,\ 0.4]$, respectively, based on grid search.

Table H.2: Fixed hyperparameters for all datasets.

Level	Parameter	Value
Architecture	Dimension of the first GCN layer	32
	Dimension of the second GCN layer	16
Pretraining phase	Optimizer	Adam
	Learning rate	0.01
	Number of iterations	200
	Number of Monte Carlo samples	1
Clustering phase	Optimizer	Adam
	Learning rate	0.001
	Number of Monte Carlo samples	1

Table H.3: Specific hyperparameters for each dataset.

Parameter	Cora	Citeseer	Pubmed	Wiki	ACM	DBLP
Updating indicator M	20	10	20	10	30	20
Confidence threshold α	0.25	0.3	0.3	0.15	0.3	0.3

Appendix I. Software and Hardware environments

We provide the experimental settings of our software and hardware environments in Table I.4.

Table I.4: Experimental settings of our software and hardware environments.

Software	Operating System	Ubuntu 18.04.5 LTS
	Python	3.8.8
	PyTorch	1.7.0
	Sklearn	0.24.1
Hardware	CPU model	Intel(R) Xeon(R) CPU E5-2620 V4 @ 2.10GHz
	Number of CPUs	32
	GPU model	GeForce RTX 2080 Ti
	GPU memory	11 GB
	Number of GPUs	2
	RAM	132 GB

References

[1] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, T. Eliassi-Rad, Collective classification in network data, AI magazine 29 (3) (2008) 93–93.

[2] C. Yang, Z. Liu, D. Zhao, M. Sun, E. Chang, Network representation learning with rich text information, in: IJCAI, 2015, pp. 2111–2117.

[3] D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu, P. Cui, Structural deep clustering network, in: WWW, 2020, pp. 1400–1410.

[4] N. Mrabah, M. Bouguessa, R. Ksantini, Adversarial deep embedded clustering: on a better trade-off between feature randomness and feature drift, TKDE 34 (4) (2022) 1603–1617.

[5] Y. Burda, R. B. Grosse, R. Salakhutdinov, Importance weighted autoencoders, in: ICLR, 2016.

[6] T. N. Kipf, M. Welling, Variational graph auto-encoders, in: NeurIPS workshop, Vol. 32, 2016, pp. 1–3.

[7] B. Hui, P. Zhu, Q. Hu, Collaborative graph convolutional networks: Unsupervised learning meets semi-supervised learning, in: AAAI, Vol. 34, 2020, pp. 4215–4222.

[8] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, J. M. Tomczak, Hyperspherical variational auto-encoders, in: UAI, Vol. 2, 2018, pp. 856–865.

[9] J. Li, J. Yu, J. Li, H. Zhang, K. Zhao, Y. Rong, H. Cheng, J. Huang, Dirichlet graph variational autoencoder, NeurIPS 33 (2020) 5274–5283.

[10] A. Hasanzadeh, E. Hajiramezanali, K. Narayanan, N. Duffield, M. Zhou, X. Qian, Semi-implicit graph variational auto-encoders, in: NeurIPS, Vol. 32, 2019, pp. 10712–10723.

[11] A. Grover, A. Zweig, S. Ermon, Graphite: Iterative generative modeling of graphs, in: ICML, Vol. 97, 2019, pp. 2434–2444.

[12] J. Li, J. Li, Y. Liu, J. Yu, Y. Li, H. Cheng, Deconvolutional networks on graph data, in: A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), NeurIPS, Vol. 34, 2021, pp. 21019–21030.