We sincerely thank all reviewers for their valuable comments. Our responses to the comments are listed below.

To Reviewer #1: Q1. About Eq.12 and reference [19] (q1 under **Correctness**).

$$
\begin{aligned}
p(A|Z) &= \xi \prod_{A_{ij}=1} \exp(f(C_i, C_j)) \prod_{A_{ij}=0} \exp(1 - f(C_i, C_j)) \\
&\approx \prod_{A_{ij}=1} \frac{\exp(f(C_i,C_j))}{\exp(f(C_i,C_j))+\exp(1-f(C_i,C_j))} \prod_{A_{ij}=0} \frac{\exp(1-f(C_i,C_j))}{\exp(f(C_i,C_j))+\exp(1-f(C_i,C_j))} \\
&= \prod_{A_{ij}=1} \frac{1}{1+\exp(1-2f(C_i,C_j))} \prod_{A_{ij}=0} \frac{\exp(1-2f(C_i,C_j))}{1+\exp(1-2f(C_i,C_j))} \\
&= \prod_{A_{ij}=1} \sigma(C_i^\top C_j) \prod_{A_{ij}=0} (1 - \sigma(C_i^\top C_j))) \quad \textit{consider} \;\; f(C_i, C_j) = \tfrac{1}{2} C_j^\top C_j + \tfrac{1}{2}
\end{aligned}
\tag{1}
$$

where $\sigma(\cdot)$ is the logistic sigmoid function and the last line in the above equation is the one used in [19]. Z does not appear on the RHS of the equation, as Z coincides with cluster memberships C (line 117 in the paper). This point can be verified since both $z_i \in \Delta_{K-1}$ and $C_i \in \Delta_{K-1}$ are the $(K-1)$-simplex.

Q2. About Claim 4.1 and ELBO (q2 under **Correctness**). If we understand it correctly, "the approximate posterior in the reconstruction term" refers to the KL term in ELBO. If so, we do consider the effects of the KL term in the proof. As in line 107 of the paper, we regard $C_i \sim \text{Dir}(\beta)$, which is a Dirichlet posterior with a Dirichlet prior $\text{Dir}(\alpha)$. In other words, our proof starts with the effects of the KL term. We set the distance metric to be $f(C_i, C_j) = 1 - \text{MSE}(C_i, C_j)$ (line 140 in the paper) in the proof of Claim 4.1.

Q3. Does Claim 4.1 rely on a specific distribution (q1 under **Weaknesses**)? A quick answer is yes, as our proof relies on the fact $C_i \sim \text{Dir}(\beta)$. As for Gaussian variables (the case in VGAE), we are not sure if the claim still holds.

Q4. Laplace approximation towards Dirichlet (q3 under **Weaknesses**). In section 3, we used $\mathcal{N}(z_i'; \mu^0, \sigma^0)$ to approximate the Dirichlet distribution $\text{Dir}(z_i; \phi)$, where $z_i = \text{softmax}(z_i')$ is based on softmax Laplace approximation. We agree sometimes it might be hard to follow and we shall add the details in the revision.

To Reviewer #2: Q1. About Heatts and $q_\phi(\cdot)$ (q1 under **Feedback**). The parameters of the variational approximation $q_\phi(\cdot)$ are determined by Eq.7 of the DGVAE framework. Putting Heatts into the framework, we derive $\mu^0, \sigma^0 = \text{Heatts}_\phi(A, X)$, as Heatts is a variant of GNN.

Q2. About DGVAE and DGAE (q1 under **Weaknesses**). We agree non-variational version performs better in most cases than its variational one, as observed by many other references [8][19]. In our opinion, the reasons are: (1) component collapsing, a good optimization of KL term results in a bad reconstruction term; (2) inappropriate priors, we use the same priors through the experiments. As for the theoretical results (Claim 4.1), we think it holds for DGAE, if we consider the Dirichlet posterior is only determined by the likelihood, i.e., without priors.

Q3. About the limitation of balanced cluster sizes (q2 under **Weaknesses**). We agree balanced cluster sizes could be a limitation in some cases. However, Claim 4.1 is based on asymptotic analysis. Moreover, the optimization is also affected by the Dirichlet priors, which provides our model with flexibility.

To Reviewer #3: Q1. About the interpretation of Dirichlet factors (q1 under **Weaknesses**). We agree that the Dirichlet approach does not lead to a better explanation compared to other clustering approaches, e.g., spherical k-means. Compared to the traditional graph variational auto-encoders which focus on Gaussian distributions, Dirichlet approach leads to a better explanation since its latent variables can be interpreted as cluster memberships.

Q2. About comparison of SOTA (q2 under **Weaknesses**). We report the experimental results of VGAE + spherical K-means (VGAE&S), for Pubmed, VGAE&S achieves 61.2%, 16.7% and 61.1% on ACC, NMI and F1 respectively; for Citeseer, VGAE&S gets 46.8%, 21.7% and 44.7% respectively; for Wiki, VGAE&S gets 27.8%, 24.3% and 19.5% respectively. The results show our method outperforms VGAE&S on all metrics. In addition, we would like to highlight that our method does not rely on any outsourcing, e.g., spherical K-means or GMM, and cluster memberships can be derived directly from the latent Z, while previous approaches such as VGAE do not have this capacity.

Q3. About the three contributions (q3 under **Weaknesses**). Yes, we agree Heatts is the key component that leads to the good experimental results. However, the good results are not the sole purpose of this work. More precisely, this work wants to figure out a clear way to understand how the graphs are generated and how we can improve the design of graph VAEs analytically. In this vein, we first describe the DGVAE framework (the 1st contribution); we then analyse the framework and find that the optimization of the framework favors *low pass* of GNN (the 2nd contribution); based on that, we propose Heatts. Without the 1st and 2nd contributions, the design of Heatts may reduce to trial-and-error.

Q4. About the corners of simplex (q1 under **Feedback**). Yes, it is imposed by the prior assumption and by optimizing the reconstruction term. At the initial stage of our model (randomly initialized parameters), the data may appear anywhere in the simplex. As the optimization goes, the Dirichlet posterior $C_i \sim \text{Dir}(\beta)$ would be "dragged" towards $\sum_k \beta_k \to 0$ (Lemma 4.2), which makes the data sampled from this distribution lie at the corners of the simplex. For the stability, the right figure of Fig. 3 provides an intuitive way to understand the process. As shown, the learning curve is smooth and drops quickly, indicating our approach is stable. The sampling strategy is illustrated in Eq.8.