A. Supplementary Materials

A.1. Proof of Theorems

Theorem 1 *The medians of a $(K-1)$-simplex $\mathcal{S}_{K-1}$ meet at the same point $\mathbf{g}_K$ and they divide each other in the ratio $(K-1):1$.*

Proof. Since the centre of gravity $\mathbf{g}_k$ is a point of each median, for the k^{th} median $[\mathbf{u}_{i^k}^k, \mathbf{g}_k]$, we have the following convex combination

$$\begin{aligned}\mathbf{g}_K &= \frac{1}{K}\sum_{l=1}^{K}\mathbf{u}_{i^l}^l = \frac{1}{K}\Big(\mathbf{u}_{i^k}^k + \sum_{l=1,l\neq k}^{K}\mathbf{u}_{i^l}^l\Big) \\ &= \frac{1}{K}\mathbf{u}_{i^k}^k + \frac{K-1}{K}\mathbf{g}_k\end{aligned} \tag{23}$$

Therefore, $\mathbf{g}_K$ lies on $[\mathbf{u}_{i^k}^k, \mathbf{g}_k]$ and divides it in the ratio $(K-1):1$.

Theorem 2 *By sequentially projecting $\mathcal{S}_{K-1}$, we can generate a series of regular simplexes: $\mathcal{S}_{K-2}$ consisting of $\mathbf{u}_{i^1}^1, \cdots, \mathbf{u}_{i^{K-1}}^{K-1}$ with centre $\mathbf{c}_{K-1} = \mathbf{0}$, $\cdots$, $\mathcal{S}_1$ consisting of $\mathbf{u}_{i^1}^1$ and $\mathbf{u}_{i^2}^2$ with centre $\mathbf{c}_2 = \mathbf{0}$, and $\mathcal{S}_0$ consisting of $\mathbf{u}_{i^1}^1$ with centre $\mathbf{c}_1 = \mathbf{0}$. For radius r_k of $\mathcal{S}_{k-1}$ for any k $(2 \le k \le K)$, we have*

$$r_k = \sqrt{\frac{K(K-1)}{k(k-1)}}r_K, 2 \le k \le K \tag{24}$$

All points in a standard regular simplex $\mathcal{S}_{K-1}$ with $\mathbf{c}_K = \mathbf{0}$ have the following coordinates.

$$\begin{aligned}\mathbf{u}_{i^K}^K &= [0, \cdots, 0, (K-1)r_K] \\ \mathbf{u}_{i^{K-1}}^{K-1} &= [0, \cdots, 0, (K-2)r_{K-1}, -r_K] \\ \mathbf{u}_{i^{K-2}}^{K-2} &= [0, \cdots, 0, (K-3)r_{K-2}, -r_{K-1}, -r_K] \\ \cdots &= \cdots\end{aligned} \tag{25}$$

Proof. By projecting $\mathcal{S}_{K-1}$ into the hyperplane $\mathcal{H}_K$, we can generate a regular simplex of $\mathcal{S}_{K-2}$ consisting of $\mathbf{u}_{i^1}^1, \cdots, \mathbf{u}_{i^{K-1}}^{K-1}$ with centre $\mathbf{c}_{K-1} = \mathbf{0}$. Now, we will derive the relationship between radius r_{K-1} of its inscribed hypersphere and radius r_K of the inscribed hypersphere in $\mathcal{S}_{K-1}$.

Notice that $\mathbf{c}_K$ is the centre of the circumscribed hypersphere in $\mathcal{S}_{K-1}$. The radius R_K of the circumscribed hypersphere of the simplex is equal to the distance between any point and $\mathbf{c}_K$, i.e., $R_K = (K-1)r_K$. We have

$$\|\mathbf{c}_K - \mathbf{u}_{i^{K-1}}^{K-1}\|_2 = \|\mathbf{u}_{i^{K-1}}^{K-1}\|_2 = \|\mathbf{u}_{i^K}^K\|_2 = (K-1)r_K \tag{26}$$

Since line $\overline{\mathbf{c}_K\mathbf{c}_{K-1}}$ is perpendicular to the hyperplane $\mathcal{H}_K$, $\|\mathbf{c}_K - \mathbf{c}_{K-1}\|_2$ is equal to r_K and points $\mathbf{c}_K, \mathbf{c}_{K-1}, \mathbf{u}_{i^{K-1}}^{K-1}$ forms a right triangle. Based on the Pythagoras Theorem, we have

$$\begin{aligned}&\|\mathbf{c}_{K-1} - \mathbf{u}_{i^{K-1}}^{K-1}\|_2 \\ =&\sqrt{\|\mathbf{c}_K - \mathbf{u}_{i^{K-1}}^{K-1}\|_2^2 - \|\mathbf{c}_K - \mathbf{c}_{K-1}\|_2^2} \\ =&\sqrt{((K-1)r_K)^2 - r_K^2} = \sqrt{K^2-2K}r_K\end{aligned} \tag{27}$$

By using the same strategy in Theorem 1, we know $\mathbf{c}_{K-1}$ divides the medians in $\mathcal{S}_{K-2}$ in the ratio $(K-2):1$. We get

$$(K-2)r_{K-1} = \sqrt{K^2-2K}r_K \tag{28}$$

By sequentially projecting $\mathcal{S}_{K-1}$, we can generate a series of regular simplexes: $\mathcal{S}_{K-2}$ with radius r_{K-1} of its inscribed hypersphere, $\cdots$, $\mathcal{S}_1$ with radius r_2, and $\mathcal{S}_0$ with radius r_1.

Eq.(28) can be generalized to the following equation.

$$\begin{aligned}(k-2)r_{k-1} &= \sqrt{k^2-2k}r_k,\ 2 \le k \le K \\ r_1 &= 0\end{aligned} \tag{29}$$

The solution of this recurrence relation is

$$\begin{aligned}r_k &= \sqrt{\frac{K(K-1)}{k(k-1)}}r_K,\ 2 \le k \le K \\ r_1 &= 0\end{aligned} \tag{30}$$

By repeating the same argument for the simplexes $\mathcal{S}_{K-1}$, $\mathcal{S}_{K-1}$, $\cdots$ etc., we get

$$\begin{aligned}\mathbf{u}_{i^K}^K &= [0, \cdots, 0, (K-1)r_K] \\ \mathbf{u}_{i^{K-1}}^{K-1} &= [0, \cdots, 0, (K-2)r_{K-1}, -r_K] \\ \mathbf{u}_{i^{K-2}}^{K-2} &= [0, \cdots, 0, (K-3)r_{K-2}, -r_{K-1}, -r_K] \\ \cdots &= \cdots\end{aligned} \tag{31}$$

The general form of the coordinates of the points are given as follows.

$$\mathbf{u}_{i^k}^k = (k-1)r_k\mathbf{b}_{k-1} - \sum_{l=k+1}^{K} r_l\mathbf{b}_{l-1},\ k = 1, \cdots, K \tag{32}$$

where $\mathbf{b}_1 = \mathbf{0} \in \mathbb{R}^{K-1}$ is an all-zero vector. $\mathbf{b}_k \in \mathbb{R}^{K-1}$ $(1 \le k \le K-1)$ is a one-hot vector with a single value of 1 at the k^{th} component.

Theorem 3 *For any two different points $\mathbf{u}_{i^k}^k$ and $\mathbf{u}_{i^j}^j$ $(1 \le k, j \le K, k \neq j)$ in a standard regular simplex $\mathcal{S}_{K-1}$ with $\mathbf{c}_K = \mathbf{0}$, $\|\mathbf{u}_{i^k}^k - \mathbf{u}_{i^j}^j\|_2^2 = 2K(K-1)r_K^2$.*

Proof. Without loss of generality, we assume that $k > j$. Based on Eqs.(30) and (32), and the orthogonality of the standard unit vectors, we get

$$
\begin{aligned}
&\|\mathbf{u}_{ik}^{k}-\mathbf{u}_{ij}^{j}\|_2^2\\
&=\Big\|(k-1)r_k\mathbf{b}_{k-1}-(j-1)r_j\mathbf{b}_{j-1}+\sum_{l=j+1}^{k}r_l\mathbf{b}_{l-1}\Big\|_2^2\\
&=\Big\|kr_k\mathbf{b}_{k-1}-(j-1)r_j\mathbf{b}_{j-1}+\sum_{l=j+1}^{k-1}r_l\mathbf{b}_{l-1}\Big\|_2^2\\
&=k^2r_k^2+(j-1)^2r_j^2+\sum_{l=j+1}^{k-1}r_l^2\\
&=K(K-1)r_K^2\Big(\frac{k}{k-1}+\frac{j-1}{j}+\sum_{l=j+1}^{k-1}\frac{1}{l(l-1)}\Big)\\
&=K(K-1)r_K^2\Big(\frac{k}{k-1}+\frac{j-1}{j}+\sum_{l=j+1}^{k-1}\frac{1}{l-1}-\frac{1}{l}\Big)\\
&=K(K-1)r_K^2\Big(\frac{k}{k-1}+\frac{j-1}{j}+\frac{1}{j}-\frac{1}{k-1}\Big)\\
&=2K(K-1)r_K^2
\end{aligned}
\tag{33}
$$

Theorem 4 *In a standard regular simplex $\mathcal{S}_{K-1}$ with centre $\mathbf{c}_K=\mathbf{0}$, $\|\mathbf{u}_{ik}^k\|_2^2=(K-1)^2r_K^2$ for any k $(1\le k\le K)$. For any two different points $\mathbf{u}_{ik}^k$ and $\mathbf{u}_{ij}^j$ $(1\le k,j\le K, k\ne j)$, $\mathbf{u}_{ik}^k\cdot\mathbf{u}_{ij}^j=-(K-1)r_K^2$.*

Proof. (1)

$$
\begin{aligned}
\|\mathbf{u}_{ik}^k\|_2^2=&\Big((k-1)r_k\mathbf{b}_{k-1}-\sum_{l=k+1}^{K}r_l\mathbf{b}_{l-1}\Big)\cdot\\
&\Big((k-1)r_k\mathbf{b}_{k-1}-\sum_{l=k+1}^{K}r_l\mathbf{b}_{l-1}\Big)
\end{aligned}
\tag{34}
$$

By using the orthonormality of the vectors $\{\mathbf{b}_k\}_{k=1}^{K-1}$, we have

$$
\|\mathbf{u}_{ik}^k\|_2^2=(k-1)^2r_k^2+\sum_{l=k+1}^{K}r_l^2 \tag{35}
$$

According to Eq.(30), we get

$$
\begin{aligned}
\|\mathbf{u}_{ik}^k\|_2^2&=\Big\{\frac{(k-1)K(K-1)}{k}+K(K-1)\sum_{l=k+1}^{K}\frac{1}{l(l-1)}\Big\}r_K^2\\
&=\Big\{\frac{k-1}{k}+\frac{1}{k}-\frac{1}{K}\Big\}K(K-1)r_K^2\\
&=(K-1)^2r_K^2
\end{aligned}
\tag{36}
$$

(2) Without loss of generality, we assume that $k>j$.

$$
\begin{aligned}
\mathbf{u}_{ik}^k\cdot\mathbf{u}_{ij}^j=&\Big((k-1)r_k\mathbf{b}_{k-1}-\sum_{l=k+1}^{K}r_l\mathbf{b}_{l-1}\Big)\cdot\\
&\Big((j-1)r_j\mathbf{b}_{j-1}-\sum_{l=j+1}^{K}r_l\mathbf{b}_{l-1}\Big)\\
=&-(k-1)r_k^2+\sum_{l=k+1}^{K}r_l^2\\
=&-\frac{K(K-1)}{k}r_K^2+K(K-1)(\frac{1}{k}-\frac{1}{K})r_K^2\\
=&-(K-1)r_K^2
\end{aligned}
\tag{37}
$$

Therefore, the proof is concluded.

Theorem 5 *The set of phase-type distributions is dense in the field of all positive-valued distributions, namely, it can be used to approximate any positive-valued distribution.*

Proof. Please refer to (O'Cinneide, 1990; 1999) for detailed proof.

Theorem 6 *The estimation with $\bar{F}_{\mathcal{D}}(x_1,\cdots,x_{K-1})=\alpha e^{\mathbf{T}x}\mathbf{D}\mathcal{D}\mathbf{1}$ is more expressive than the one with $\bar{F}(x_1,\cdots,x_{K-1})=\alpha e^{\mathbf{T}x}\mathbf{D}\mathbf{1}$, where $\mathcal{D}=\mathrm{diag}(h(d_1),\cdots,h(d_m))$ is a diagonal matrix to be estimated and h is the sigmoid function.*

Proof. For the original multivariate phase-type distribution $\bar{F}(x_1,\cdots,x_{K-1})=\alpha e^{\mathbf{T}x}\mathbf{D}\mathbf{1}$, by assuming that $\alpha=[\alpha_1,\cdots,\alpha_m]$, we obtain

$$
\begin{aligned}
&\bar{F}(x_1,\cdots,x_{K-1})\\
&=\alpha e^{\mathbf{T}x}\mathbf{D}\mathbf{1}\\
&=[\alpha_1,\cdots,\alpha_m]\begin{bmatrix}e_{11}^{\mathbf{T}x}&\cdots&e_{1m}^{\mathbf{T}x}\\ \vdots&\ddots&\vdots\\ e_{m1}^{\mathbf{T}x}&\cdots&e_{mm}^{\mathbf{T}x}\end{bmatrix}\mathbf{D}_1\cdots\mathbf{D}_{K-1}[1,\cdot,1]^T
\end{aligned}
\tag{38}
$$

where $e_{ij}^{\mathbf{T}x}$ is the element at the i^{th} row and j^{th} column in $e^{\mathbf{T}x}$.

It is obviously that the value of $\bar{F}(x_1,\cdots,x_{K-1})$ is the sum of $n\le m$ columns from vector

$$
[\sum_{i=1}^{m}\alpha_ie_{i1}^{\mathbf{T}x},\cdots,\sum_{i=1}^{m}\alpha_ie_{im}^{\mathbf{T}x}] \tag{39}
$$

where n depends on the number of 1 in the diagonal elements of $\mathbf{D}=\mathbf{D}_1\cdots\mathbf{D}_{K-1}$.

Therefore,

$$
\check{F}=\min_{j=1}^{m}\sum_{i=1}^{m}\alpha_ie_{ij}^{\mathbf{T}x}\le\bar{F}\le\sum_{j=1}^{m}\sum_{i=1}^{m}\alpha_ie_{ij}^{\mathbf{T}x}=\hat{F} \tag{40}
$$

Now, we can see that the value of $\bar{F}$ is one of $\sum_{j=1}^{m}\binom{j}{m}$ states.

Motivated by the above analysis, we introduce one additional parameter $\mathcal{D}$ to $\bar{F}(x_1,\cdots,x_{K-1})$ and get $\bar{F}_{\mathcal{D}}(x_1,\cdots,x_{K-1})$.

$$\begin{aligned}\bar{F}_{\mathcal{D}}(x_1,\cdots,x_{K-1}) &= \alpha e^{\mathbf{T}x}\mathbf{D}\mathcal{D}\mathbf{1}\\ &= \sum_{j=1}^{m}\Big(\mathbf{D}_{jj}h(d_j)\sum_{i=1}^{m}\alpha_i e_{ij}^{\mathbf{T}x}\Big)\end{aligned} \tag{41}$$

Since Sigmoid is a continuous function with the value in the range of $[0,1]$, $\bar{F}_{\mathcal{D}}(x_1,\cdots,x_{K-1})$ can be treated as a continuous function with the value in the range of $[\check{F},\hat{F}]$, instead of only $\sum_{j=1}^{m}\binom{j}{m}$ possible discrete values in the range of $[\check{F},\hat{F}]$ achieved by the original multivariate phase-type distribution $\bar{F}(x_1,\cdots,x_{K-1})$.

Therefore, the expressiveness of $\bar{F}_{\mathcal{D}}(x_1,\cdots,x_{K-1})$ will be significantly improved.

The following theorem conducts the convergence analysis of our expressive parameter estimation of multivariate phase-type distribution.

Theorem 7 *Given enough iterations I, our expressive parameter estimation $Q_{\mathcal{D}}(x) = 1-\bar{F}(x_1,\cdots,x_{K-1}) = 1-\alpha e^{\mathbf{T}x}\mathbf{D}\mathcal{D}\mathbf{1}$ of multivariate phase-type distribution is able to converge to the true distribution $P(x)$.*

Proof. Let x_i denote x at the i^{th} iteration among total I iterations and $P(x)=\frac{\#(\mathbf{u}\leq[x_i,\cdots,x_i])}{N^k}$ be the true distribution. Based on our estimation, we have

$$\begin{aligned}\mathbb{E}(\mathbf{u}\leq[x_i,\cdots,x_i]) &= -\alpha_i\mathbf{T}_i\mathbf{D}_i\mathcal{D}_i\mathbf{1}\\ \mathbb{V}(\mathbf{u}\leq[x_i,\cdots,x_i]) &= 2\alpha_i\mathbf{T}_i^2\mathbf{D}_i\mathcal{D}_i\mathbf{1}-(\alpha_i\mathbf{T}_i\mathbf{D}_i\mathcal{D}_i\mathbf{1})^2\end{aligned} \tag{42}$$

where $\mathbb{E}$ and $\mathbb{V}$ represent the expectation and variance respectively.

For those samples $\mathbf{u}$ satisfying $\mathbf{u}\leq[x_i,\cdots,x_i]$, we can get its corresponding expectation $\bar{\mathbf{u}}=\mathbb{E}(\cup\mathbf{u},\forall\mathbf{u}\leq[x_i,\cdots,x_i])$, and variance $\sigma_{\mathbf{u}}^2 = \mathbb{V}(\cup\mathbf{u},\forall\mathbf{u}\leq[x_i,\cdots,x_i])$.

In addition, for the true distribution, we get

$$\begin{aligned}\mathbb{E}(\mathbf{u}) &= -\alpha_t\mathbf{T}_t\mathbf{D}_t\mathcal{D}_t\mathbf{1}\\ \mathbb{V}(\mathbf{u}) &= 2\alpha_t\mathbf{T}_t^2\mathbf{D}_t\mathcal{D}_t\mathbf{1}-(\alpha_t\mathbf{T}_t\mathbf{D}_t\mathcal{D}_t\mathbf{1})^2\end{aligned} \tag{43}$$

where subscript t denotes the corresponding terms for the true distribution.

Since $\bar{\mathbf{u}}\in\{\mathbf{u}\}$, it is straightforward to prove

$$\begin{aligned}\mathbb{E}(\bar{\mathbf{u}}) &= \frac{\sum_{i=1}^{I}\mathbb{E}(\mathbf{u}\leq[x_i,\cdots,x_i])}{I}\\ \mathbb{V}(\bar{\mathbf{u}}) &= \frac{1}{I^2}\sum_{i=1}^{I}\mathbb{V}(\mathbf{u}\leq[x_i,\cdots,x_i])\end{aligned} \tag{44}$$

By utilizing Chebyshev's inequality, for any real number $\varepsilon>0$, we have

$$\begin{aligned}&P\big(|\bar{\mathbf{u}}-\mathbb{E}(\mathbf{u})|\geq\varepsilon\big)\\ &= \int_{|\bar{\mathbf{u}}-\mathbb{E}(\mathbf{u})|\geq\varepsilon} f(u)du\\ &\leq \int_{|\bar{\mathbf{u}}-\mathbb{E}(\mathbf{u})|\geq\varepsilon}\frac{|\bar{\mathbf{u}}-\mathbb{E}(\mathbf{u})|^2}{\epsilon^2}f(u)du\\ &\leq \frac{1}{\epsilon^2}\int|\bar{\mathbf{u}}-\mathbb{E}(\mathbf{u})|^2 f(u)du\\ &= \frac{1}{\varepsilon^2 I^2}\sum_{i=1}^{I}\mathbb{V}(\mathbf{u}\leq[x_i,\cdots,x_i])\\ &= \frac{\mathbb{V}(\mathbf{u})}{\varepsilon^2 I}\end{aligned} \tag{45}$$

By taking the limit on both sides at the same time, we get

$$\lim_{I\to\infty}P\big(|\bar{\mathbf{u}}-\mathbb{E}(\mathbf{u})|\geq\varepsilon\big)=\lim_{I\to\infty}\frac{\mathbb{V}(\mathbf{u})}{\varepsilon^2 I}=0 \tag{46}$$

Again, by employing Chebyshev's inequality, for any real number $\psi>0$, we get

$$\begin{aligned}&P\big(|\mathbb{E}(\sigma_{\mathbf{u}}^2)-\mathbb{E}(\mathbb{V}(\mathbf{u}))|\geq\psi\big)\\ &= \int_{|\mathbb{E}(\sigma_{\mathbf{u}}^2)-\mathbb{E}(\mathbb{V}(\mathbf{u}))|\geq\psi} f(u)du\\ &\leq \int_{|\mathbb{E}(\sigma_{\mathbf{u}}^2)-\mathbb{E}(\mathbb{V}(\mathbf{u}))|\geq\psi}\frac{|\mathbb{E}(\sigma_{\mathbf{u}}^2)-\mathbb{E}(\mathbb{V}(\mathbf{u}))|^2}{\psi^2}f(u)du\\ &\leq \frac{1}{\psi^2}\int|\frac{I-1}{I}\sigma_{\mathbf{u}}^2-\mathbb{V}(\mathbf{u})|^2 f(u)du\\ &= \frac{\mathbb{V}(\mathbf{u})^2}{\psi^2 I^2}\end{aligned} \tag{47}$$

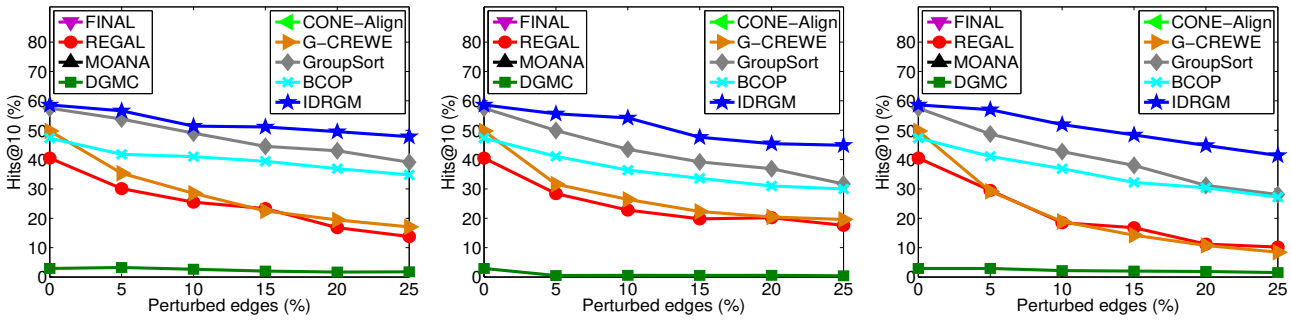
In addition, we have

$$\lim_{I\to\infty}P\big(|\mathbb{E}(\sigma_{\mathbf{u}}^2)-\mathbb{E}(\mathbb{V}(\mathbf{u}))|\geq\psi\big)=\lim_{I\to\infty}\frac{\mathbb{V}(\mathbf{u})^2}{\psi^2 I^2}=0 \tag{48}$$

Therefore, the proof is concluded.

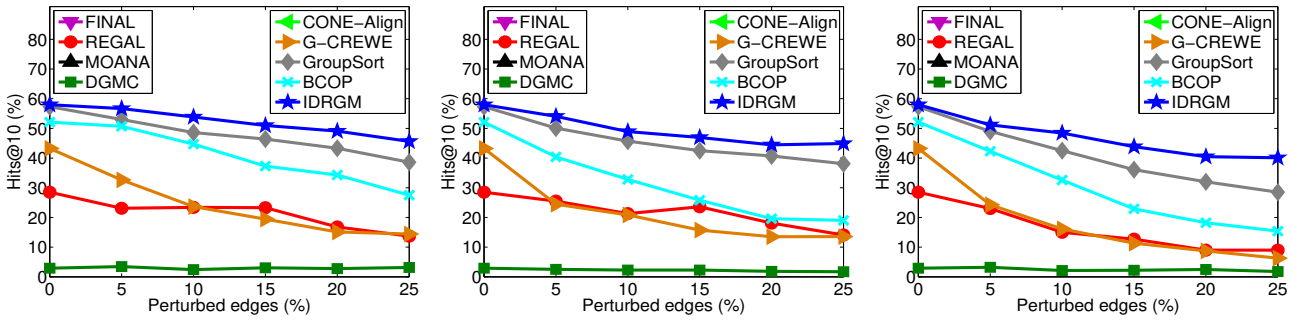
A.2. Additional Experiments

In this section, we use another measure $Hits@10$ to evaluate our model and other competitors on graph matching. Notice that three graph matching methods of FINAL, MOANA, and CONE-Align perform one-to-one matching among multiple nodes in different graphs, i.e., they find only one node as



(a) RND Attack (b) NEA Attack (c) GMA Attack

Figure 9: AS with varying perturbed edges and $Hits@10$ (%)



(a) RND Attack (b) NEA Attack (c) GMA Attack

Figure 10: AS with varying perturbed edges and $Hits@10$ (%)

a matching in each of other graphs for a node in a graph. Thus, the $Hits@10$ scores are the same as the $Hits@1$ values for these three methods. Therefore, we do not plot the $Hits@10$ curves of these three algorithms in Figures 9-11. The $Hits@1$ values of these three methods have been reported in Figures 3-5.

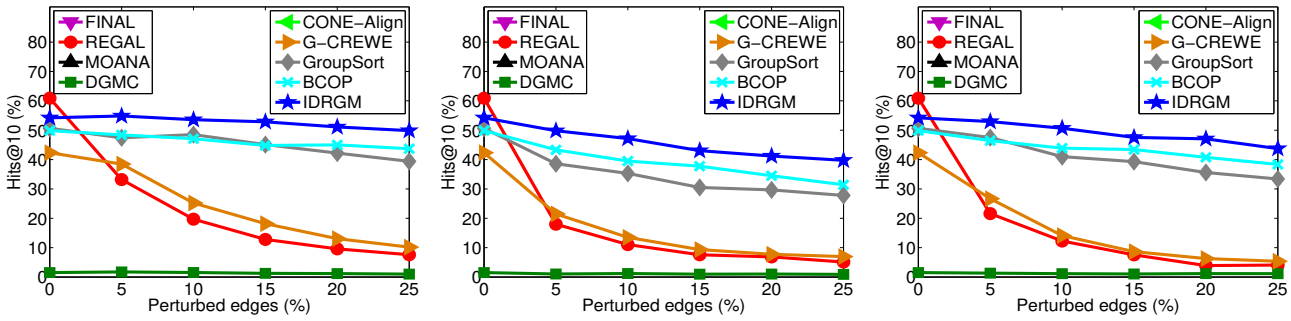
Figures 9-11 presents the quality of graph matching by nine different algorithms under varying strengths of adversarial attacks. Similar trends are observed for the graph matching quality comparison: IDRGM achieves the largest $Hits@10$ values (>41.4%, >40.1%, and >39.8%) on three datasets of AS, CAIDA, DBLP respectively, which are obviously better than all other methods. Especially, as shown in Figure 9, compared to the best competitors among nine graph matching algorithms, the $Hits@10$ scores achieved by IDRGM averagely achieves 19.3% increase. This demonstrates that the integration of the simplex detection technique for tackling the inter-graph dispersion attacks and the distribution estimation and node separation methods for handling the intra-graph assembly attacks is able to make the graph matching results achieved by our integrated defense model robust to various adversarial attacks.

Table 3 presents the ablation study results with $Hits@10$ as evaluation metric of graph matching on three datasets by four variants of our IDRGM model. The number of perturbed edges is fixed to 5%. It is observed that the complete IDRGM achieves the highest $Hits@10$ ($>$ 55.6%) on AS, ($>$ 51.2%) over CAIDA, and ($>$ 49.8%) on DBLP, which are obviously better than other versions. Compared with IDRGM-A, IDRGM-D performs better in all experiments. Concretely, there are three critical reasons for high accuracy of IDRGM: IDRGM-D is able to tackle inter-graph dispersion attacks only. IDRGM-A can handle intra-graph assembly attacks only. IDRGM-N fail to utilize any defense techniques for defending the graph matching models against adversarial attacks.

A.3. Parameter Sensitivity

In this section, we conduct more experiments to validate the sensitivity of various parameters in the phase-type distribution estimation and our integrated defense model.

Impact of boundary parameter X. Figure 12 (a) presents the impact of boundary parameter X in phase-type distribution estimation with X between 0.2 and 1. It is observed that the $Hits@1$ values are stable with varying X. Namely, our expressive estimation of phase-type distribution is insensitive to X when normalizing the node embedding vectors into a bounded range $[0, X]$. This demonstrates that our



(a) RND Attack (b) NEA Attack (c) GMA Attack

Figure 11: AS with varying perturbed edges and $Hits@10$ (%)

Table 3: $Hits@10$ of IDRGM variants with 5% perturbed edges

Dataset	**AS**			**CAIDA**			**DBLP**		
Attack Model	**RND**	**NEA**	**GMA**	**RND**	**NEA**	**GMA**	**RND**	**NEA**	**GMA**
IDRGM-D	53.3	53.7	53.8	55.8	51.5	50.4	53.4	47.7	51.2
IDRGM-A	53.4	51.6	53.9	55.6	51.8	50.3	53.5	48.7	51.7
IDRGM-N	50.0	48.9	50.0	51.5	48.6	46.5	51.3	47.1	49.6
IDRGM	**56.6**	**55.6**	**57.0**	**56.7**	**54.1**	**51.2**	**54.8**	**49.8**	**53.0**

IDRGM model can always result in the good parameter estimation for phase-type distribution, no matter which normalization bound is selected.

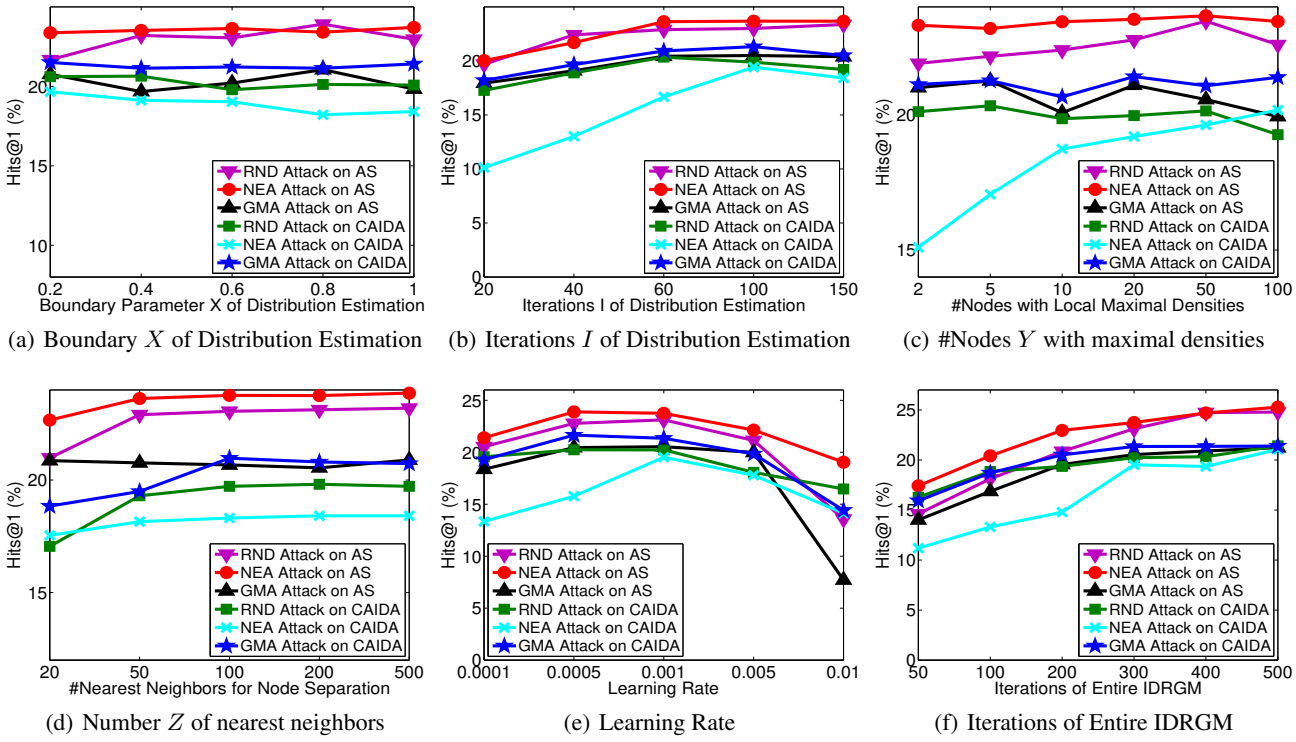
Sensitivity of iterations I. Figure 12 (b) shows the impact of iterations I in phase-type distribution estimation over two groups of datasets by varying I from 20 to 150. We have witnessed the performance curves initially increase quickly and then become stable or even drop when I continuously increases. Initially, a large I can help partition the bounded range $[0, X]$ into more intervals to derive more accurate parameter estimation. Later on, when I continues to increase and goes beyond some thresholds, too many intervals may lead to each interval with very small size, i.e., very few samples. Therefore, inadequate samples at each iteration may result in the performance loss of parameter estimation.

Influence of number Y. Figure 12 (c) exhibits the sensitivity of number Y of nodes with local maximal densities with Y between 2 and 100. The performance curves oscillates up and down. This demonstrates that there must exist the optimal Y that makes the performance of our node separation method be maximally optimized. On one hand, too small Y may not ensure the nodes separated into a wide enough space, such that the similar neighbors may still affect the matching of the perturbed nodes. On the other hand, too large Y may push the similar neighbors of a perturbed node close to another perturbed node, such that the matching performance of another perturbed node is significantly affected.

Impact of number Z. Figure 12 (d) presents the impact of number Z of nearest neighbors for node separation over two datasets. We have observed that the performance initially raises when the number Z increases. Intuitively, a larger Z can help separate mode nodes into a wide space, such that the interference from the similar neighbors of the perturbed nodes is significantly reduced. However, when Z continues to increase and goes beyond a certain threshold, the performance curves become stable. A rational guess is that after the perturbed nodes and their neighbors have been already separated at a certain threshold, our IDRGM model is able to generate a good graph matching result. When Z continuously increases, this does not affect the performance of graph matching any more.

Sensitivity of learning rate. Figure 12 (e) shows the impact of learning rate in our IDRGM model by varying it from 0.0001 to 0.01. the $Hits@1$ values have concave curves when increasing learning rate. A too small learning rate may result in a long training process that could get stuck, whereas a too large learning rate may result in learning a sub-optimal set of weights too fast or an unstable training process. Thus, this demonstrates that there must exist the optimal learning rate that makes the performance of our IDRGM model be maximally optimized.

Convergence study. Figure 12 (f) exhibits the convergence of our IDRGM model for resilient graph matching. As we can see, the $Hits@1$ values keep increasing when we iteratively perform the defense task. The method converges



(a) Boundary X of Distribution Estimation (b) Iterations I of Distribution Estimation (c) #Nodes Y with maximal densities

(d) Number Z of nearest neighbors (e) Learning Rate (f) Iterations of Entire IDRGM

Figure 12: $Hits@1$ (%) with varying parameters

when the numbers of iterations go beyond some thresholds. We have observed that most curves on two datasets converge within 200-300 iterations. This verifies the efficiency of our IDRGM model to combat with two kinds of adversarial attacks.

A.4. Experimental Details

Environment. Our experiments were conducted on a compute server running on Red Hat Enterprise Linux 7.2 with 2 CPUs of Intel Xeon E5-2650 v4 (at 2.66 GHz) and 8 GPUs of NVIDIA GeForce GTX 2080 Ti (with 11GB of GDDR6 on a 352-bit memory bus and memory bandwidth in the neighborhood of 620GB/s), 256GB of RAM, and 1TB of HDD. Overall, our experiments took about 4 days in a shared resource setting. We expect that a consumer-grade single-GPU machine (e.g., with a 1080 Ti GPU) could complete our full set of experiments in around 8 days, if its full resources were dedicated. The codes were implemented in Python 3.7.3 and PyTorch 1.0.14. We also employ Numpy 1.16.4 and Scipy 1.3.0 in the implementation. Since the datasets used are all public datasets and the hyperparameter settings are explicitly described, our experiments can be easily reproduced on top of a GPU server.

Implementation. For random attack model, we add the noisy edges to the datasets with different levels of noisy data, say 5% (i.e., 0.05), by randomly adding or removing edges with the half noise level respectively, say 2.5%. For other two attack models of NEA [1] and GMA [2], we used the open-source implementation and default parameter settings by the original authors for our experiments. For six graph matching methods of FINAL [3], REGAL [4], MOANA [5], DGMC [6], CONE-Align [7] and G-CREWE [8], two Lipschitz-bound neural architectures of GroupSort [9] and BCOP [10], we also utilized the same model architecture as the official implementation provided by the original authors and used the same perturbed graphs to validate the robustness of these graph learning models in all experiments.

For our integrated defense model for resilient graph matching, we performed hyperparameter selection by performing a parameter sweep on boundary parameter $X \in \{0.2, 0.4, 0.6, 0.8, 1\}$ in phase-type distribu-

[1] https://www.in.tum.de/daml/node-embedding-attack/
[2] https://github.com/DMML-AU/GMA
[3] https://github.com/sizhang92/FINAL-network-alignment-KDD16
[4] https://github.com/GemsLab/REGAL
[5] https://github.com/sizhang92/Multilevel-network-alignment-Moana-
[6] https://github.com/rusty1s/deep-graph-matching-consensus
[7] https://github.com/GemsLab/CONE-Align
[8] https://github.com/cruiseresearchgroup/G-CREWE
[9] https://github.com/cemanil/LNets
[10] https://github.com/ColinQiyangLi/LConvNet

tion estimation, iterations $I \in \{20, 40, 60, 100, 150\}$ in phase-type distribution estimation number $Y \in \{2, 5, 10, 20, 50, 100\}$ of nodes with local maximal densities, number $Z \in \{20, 50, 100, 200, 500\}$ of nearest neighbors for node separation, and learning rate $\in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$. We select the best parameters over 50 iterations of training and evaluate the model at test time. After the hyperparameter selection, the model was trained for 500 iterations, with a batch size of 512, and a learning rate of 0.001.