Partial Multi-Label Learning with Probabilistic Graphical Disambiguation Supplementary Material

Jun-Yi Hang, Min-Ling Zhang*
School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
{hangjy, zhangml}@seu.edu.cn

Appendix A
More Experimental Results for Comparative Studies

Table A.1, A.2 and A.3 report detailed experimental results in terms of *Coverage*, *One-error* and *Hamming loss*, which are not covered in the *Comparative Studies* part of the main body due to page limit. It can be observed that our PARD achieves consistently superior performance to well-established PML approaches.

Table A.1: Predictive performance of each comparing approach (mean±std. deviation) in terms of *Coverage*, where ●/○ indicates whether PARD is significantly superior/inferior to one comparing approach via paired t-test at 0.05 significance level. ↑ (↓) indicates the larger (smaller) the value, the better the performance. Best results are shown in boldface.

Data sets	γ%	*Coverage* ↓						
		FPML	PARVLS	PML-NI	PML-MD	UPML-HL	UPML-RL	PARD
YeastBP		0.437±0.020●	0.537±0.024●	0.265±0.019●	0.306±0.017●	0.301±0.021●	0.394±0.018●	**0.235±0.014**
YeastCC		0.173±0.014●	0.256±0.020●	0.089±0.009●	0.112±0.012●	0.096±0.012●	0.127±0.012●	**0.077±0.009**
YeastMF		0.185±0.013●	0.247±0.015●	0.109±0.012●	0.124±0.013●	0.116±0.016●	0.144±0.011●	**0.094±0.010**
Music_emotion		0.433±0.009●	0.412±0.006●	0.410±0.007●	0.399±0.009	**0.390±0.007**○	0.406±0.009●	0.396±0.008
Music_style		0.218±0.010●	0.209±0.009●	0.198±0.009	0.203±0.009	**0.195±0.008**○	0.207±0.008●	0.199±0.011
corel5k	100	0.343±0.014●	0.449±0.024●	0.372±0.015●	0.340±0.011●	0.324±0.016●	0.314±0.017●	**0.308±0.016**
	150	0.349±0.016●	0.414±0.016●	0.388±0.015●	0.347±0.012●	0.328±0.017	0.322±0.016	**0.321±0.014**
	200	0.351±0.014●	0.423±0.027●	0.404±0.014●	0.352±0.012●	0.345±0.012●	**0.329±0.015**	0.332±0.015
	250	0.356±0.016●	0.432±0.027●	0.414±0.017●	0.355±0.017●	0.349±0.015●	**0.332±0.016**	0.337±0.015
rcv1-s1	100	0.164±0.006●	0.304±0.013●	0.173±0.010●	0.151±0.007●	0.146±0.006●	0.145±0.004●	**0.135±0.006**
	150	0.165±0.007●	0.330±0.024●	0.197±0.009●	0.165±0.006●	0.150±0.006●	0.149±0.005●	**0.140±0.006**
	200	0.172±0.008●	0.350±0.023●	0.214±0.008●	0.173±0.006●	0.156±0.009●	0.155±0.005●	**0.148±0.005**
	250	0.183±0.008●	0.364±0.019●	0.232±0.008●	0.178±0.007●	0.166±0.010●	0.163±0.005●	**0.156±0.006**
Corel16k-s1	100	0.316±0.007●	0.391±0.008●	0.346±0.012●	0.317±0.010●	0.304±0.008●	0.308±0.007●	**0.298±0.009**
	150	0.324±0.007●	0.399±0.011●	0.362±0.011●	0.325±0.009●	0.321±0.008●	0.314±0.008●	**0.307±0.010**
	200	0.336±0.008●	0.427±0.010●	0.375±0.013●	0.330±0.010●	0.348±0.010●	**0.318±0.007**	0.319±0.010
	250	0.347±0.009●	0.470±0.008●	0.387±0.013●	0.338±0.008●	0.349±0.012●	0.325±0.007	**0.322±0.010**
iaprtc12	100	0.328±0.005●	0.352±0.005●	0.354±0.007●	0.334±0.005●	0.345±0.008●	0.349±0.004●	**0.321±0.006**
	150	0.332±0.005●	0.359±0.006●	0.374±0.007●	0.346±0.006●	0.349±0.009●	0.351±0.004●	**0.327±0.006**
	200	0.339±0.005	0.376±0.005●	0.393±0.008●	0.358±0.007●	0.351±0.009●	0.357±0.006●	**0.337±0.007**
	250	0.350±0.007●	0.392±0.006●	0.413±0.008●	0.367±0.007●	0.368±0.010●	0.363±0.007●	**0.346±0.006**
espgame	100	0.365±0.007●	0.398±0.007●	0.404±0.007●	0.371±0.007●	0.372±0.008●	0.391±0.009●	**0.353±0.006**
	150	0.369±0.007●	0.419±0.007●	0.423±0.007●	0.380±0.007●	0.361±0.007●	0.377±0.007●	**0.358±0.006**
	200	0.374±0.007●	0.434±0.006●	0.435±0.008●	0.387±0.007●	0.398±0.010●	0.381±0.006●	**0.362±0.007**
	250	0.384±0.007●	0.451±0.008●	0.449±0.009●	0.389±0.007●	0.388±0.009●	0.388±0.007●	**0.370±0.007**

*Corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table A.2: Predictive performance of each comparing approach (mean±std. deviation) in terms of *One-error*, where •/◦ indicates whether PARD is significantly superior/inferior to one comparing approach via paired t-test at 0.05 significance level. ↑ (↓) indicates the larger (smaller) the value, the better the performance. Best results are shown in boldface.

Data sets	$\gamma\%$	*One-error* ↓						
		FPML	PARVLS	PML-NI	PML-MD	UPML-HL	UPML-RL	PARD
YeastBP		0.833±0.012•	0.977±0.007•	**0.670±0.025**◦	0.749±0.012•	0.690±0.021	0.856±0.012•	0.687±0.014
YeastCC		0.863±0.016•	0.961±0.007•	**0.768±0.023**	0.796±0.018•	0.772±0.017	0.826±0.013•	0.772±0.020
YeastMF		0.933±0.011•	0.983±0.007•	0.838±0.014	0.859±0.014•	0.846±0.010•	0.900±0.014•	**0.837±0.018**
Music_emotion		0.547±0.022•	0.515±0.020•	0.495±0.024•	0.412±0.023	0.407±0.025	0.431±0.018•	**0.403±0.011**
Music_style		0.393±0.018•	0.372±0.020•	0.352±0.017•	0.394±0.020•	0.356±0.017•	0.405±0.017•	**0.337±0.015**
corel5k	100	0.642±0.021•	0.736±0.028•	0.642±0.021•	0.636±0.028•	0.603±0.020•	0.610±0.023•	**0.583±0.023**
	150	0.652±0.017•	0.732±0.015•	0.664±0.022•	0.652±0.020•	0.608±0.021•	0.609±0.021•	**0.589±0.016**
	200	0.656±0.020•	0.747±0.029•	0.681±0.020•	0.662±0.019•	0.607±0.015	0.610±0.017	**0.598±0.019**
	250	0.660±0.018•	0.750±0.027•	0.697±0.021•	0.663±0.024•	0.608±0.018	0.609±0.018	**0.602±0.016**
rcv1-s1	100	0.447±0.017•	0.602±0.030•	0.425±0.022	0.497±0.019•	**0.413±0.025**◦	0.466±0.011•	0.430±0.019
	150	0.458±0.016•	0.609±0.028•	0.449±0.022•	0.513±0.022•	0.428±0.024	0.464±0.015•	**0.424±0.018**
	200	0.463±0.015•	0.622±0.025•	0.466±0.017•	0.524±0.022•	0.434±0.016	0.470±0.014•	**0.433±0.017**
	250	0.472±0.020•	0.623±0.021•	0.495±0.017•	0.535±0.023•	0.440±0.027	0.490±0.022•	**0.435±0.020**
Corel16k-s1	100	0.580±0.015•	0.716±0.017•	0.587±0.015•	0.607±0.017•	0.577±0.020•	0.602±0.009•	**0.567±0.020**
	150	0.589±0.013•	0.715±0.017•	0.610±0.021•	0.617±0.017•	0.582±0.018•	0.605±0.016•	**0.572±0.023**
	200	0.590±0.014	0.720±0.015•	0.624±0.021•	0.624±0.017•	0.604±0.015•	0.609±0.013•	**0.580±0.024**
	250	0.600±0.016	0.721±0.014•	0.635±0.019•	0.631±0.019•	0.603±0.018•	0.611±0.013•	**0.593±0.022**
iaprtc12	100	0.458±0.009•	0.462±0.015•	0.448±0.015•	0.461±0.014•	0.444±0.013•	0.525±0.016•	**0.433±0.020**
	150	0.460±0.009•	0.467±0.014•	0.467±0.017•	0.481±0.016•	0.449±0.009•	0.531±0.010•	**0.438±0.013**
	200	0.463±0.012•	0.474±0.011•	0.485±0.013•	0.494±0.013•	0.444±0.013	0.531±0.013•	**0.443±0.013**
	250	0.469±0.016•	0.485±0.011•	0.507±0.015•	0.503±0.012•	0.463±0.014•	0.540±0.010•	**0.450±0.013**
espgame	100	0.608±0.012•	0.637±0.014•	0.618±0.011•	0.618±0.014•	0.601±0.014•	0.648±0.012•	**0.589±0.014**
	150	0.609±0.010•	0.644±0.014•	0.642±0.012•	0.630±0.013•	0.596±0.013	0.655±0.012•	**0.594±0.013**
	200	0.608±0.011•	0.656±0.015•	0.654±0.009•	0.640±0.011•	0.626±0.016•	0.664±0.011•	**0.598±0.013**
	250	0.611±0.016•	0.678±0.018•	0.677±0.009•	0.652±0.010•	0.613±0.012•	0.672±0.009•	**0.604±0.012**

Table A.3: Predictive performance of each comparing approach (mean±std. deviation) in terms of *Hamming loss*, where •/◦ indicates whether PARD is significantly superior/inferior to one comparing approach via paired t-test at 0.05 significance level. ↑ (↓) indicates the larger (smaller) the value, the better the performance. Best results are shown in boldface.

Data sets	$\gamma\%$	*Hamming loss* ↓						
		FPML	PARVLS	PML-NI	PML-MD	UPML-HL	UPML-RL	PARD
YeastBP		0.025±0.001•	0.025±0.001•	0.026±0.001•	0.026±0.002•	0.025±0.001•	0.026±0.001•	**0.024±0.001**
YeastCC		0.154±0.002•	0.027±0.002•	0.029±0.002•	0.029±0.002•	0.026±0.002•	0.027±0.002•	**0.025±0.002**
YeastMF		**0.025±0.002**	0.026±0.002•	0.031±0.002•	0.028±0.002•	**0.025±0.002**	0.026±0.002•	**0.025±0.002**
Music_emotion		0.217±0.004•	0.214±0.006•	0.212±0.005•	0.218±0.005•	**0.193±0.004**◦	0.210±0.005•	0.196±0.006
Music_style		0.123±0.003•	0.121±0.003•	**0.115±0.003**	0.143±0.004•	**0.115±0.003**	0.125±0.003•	0.117±0.006
corel5k	100	0.117±0.002•	0.134±0.006•	0.129±0.004•	0.117±0.005•	**0.112±0.003**	0.113±0.003•	**0.112±0.003**
	150	0.117±0.003•	0.136±0.006•	0.130±0.004•	0.117±0.003•	**0.112±0.003**	0.113±0.003•	**0.112±0.002**
	200	0.117±0.003•	0.138±0.005•	0.131±0.005•	0.117±0.003•	**0.113±0.003**	**0.113±0.003**	**0.113±0.003**
	250	0.117±0.003•	0.140±0.006•	0.131±0.004•	0.117±0.004•	**0.113±0.003**	**0.113±0.003**	**0.113±0.003**
rcv1-s1	100	0.099±0.002	0.118±0.004•	0.100±0.003	0.106±0.003•	**0.095±0.002**◦	0.108±0.002•	0.098±0.003
	150	0.099±0.002◦	0.123±0.004•	0.103±0.004	0.107±0.003•	**0.096±0.003**◦	0.108±0.002•	0.104±0.003
	200	**0.099±0.002**◦	0.125±0.003•	0.106±0.004•	0.108±0.003•	0.101±0.003	0.107±0.002•	0.103±0.004
	250	**0.101±0.003**◦	0.126±0.003•	0.110±0.005	0.109±0.003	0.125±0.005•	0.109±0.002•	0.108±0.002
Corel16k-s1	100	0.121±0.002•	0.139±0.003•	0.125±0.002•	**0.117±0.002**	**0.117±0.002**	0.118±0.001•	**0.117±0.001**
	150	0.121±0.002•	0.136±0.002•	0.126±0.003•	**0.117±0.002**	**0.117±0.001**	0.118±0.001•	**0.117±0.001**
	200	0.121±0.002•	0.135±0.002•	0.125±0.003•	**0.117±0.002**	**0.117±0.001**	0.118±0.001•	**0.117±0.002**
	250	0.120±0.002•	0.135±0.002•	0.125±0.003•	**0.117±0.002**	**0.117±0.002**	**0.117±0.002**	**0.117±0.001**
iaprtc12	100	0.143±0.002•	0.140±0.002•	0.144±0.002•	0.146±0.002•	0.138±0.002	0.151±0.002•	**0.137±0.002**
	150	0.143±0.002•	0.140±0.002	0.146±0.002•	0.149±0.002•	**0.138±0.003**	0.151±0.002•	0.139±0.002
	200	0.143±0.002•	0.141±0.002	0.148±0.002•	0.150±0.002•	**0.139±0.002**◦	0.151±0.002•	0.141±0.002
	250	0.144±0.002•	**0.143±0.001**◦	0.150±0.003•	0.151±0.002•	0.148±0.003	0.151±0.002•	0.146±0.002
espgame	100	0.133±0.002•	0.135±0.002•	0.141±0.002•	0.130±0.002•	0.131±0.002•	0.132±0.002•	**0.129±0.002**
	150	0.133±0.003•	0.133±0.002•	0.142±0.003•	0.131±0.002•	0.132±0.002•	0.132±0.002•	**0.130±0.002**
	200	0.132±0.002•	0.133±0.002•	0.143±0.002•	**0.131±0.002**	0.132±0.002	0.132±0.002•	**0.131±0.002**
	250	0.132±0.002	0.134±0.002•	0.143±0.003•	0.132±0.002	**0.131±0.002**	0.132±0.002•	**0.131±0.002**

Appendix B
Derivation of The Variational Lower Bound

The variational lower bound of the log-likelihood (i.e. Eq. (2) in the main body) is derived as follows

$$\begin{aligned}
\log p_\theta(\mathbf{s}|\mathbf{x}) &= \log \int p_\theta(\mathbf{s},\mathbf{y}|\mathbf{x}) d\mathbf{y} \\
&= \log \int p_\theta(\mathbf{s}|\mathbf{x},\mathbf{y}) p_\theta(\mathbf{y}|\mathbf{x}) d\mathbf{y} \\
&= \log \int q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s}) \frac{p_\theta(\mathbf{s}|\mathbf{x},\mathbf{y}) p_\theta(\mathbf{y}|\mathbf{x})}{q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s})} d\mathbf{y} \\
&\geq \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s})}[\log \frac{p_\theta(\mathbf{s}|\mathbf{x},\mathbf{y}) p_\theta(\mathbf{y}|\mathbf{x})}{q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s})}] \\
&= \mathcal{L}(\mathbf{x},\mathbf{s};\theta,\phi).
\end{aligned}$$

Appendix C
Derivation of The KL-Divergence Term's Closed-Form Solution

With mean-field approximation technique, closed-form solution of the KL-divergence term can be derived as follows

$$\begin{aligned}
& KL[q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s})||p_\theta(\mathbf{y}|\mathbf{x})] \\
&= \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s})}[\log q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s}) - \log p_\theta(\mathbf{y}|\mathbf{x})] \\
&= \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s})}[\sum_{k=1}^{t} \log q_\phi(y_k|\mathbf{x},\mathbf{s}) - \sum_{k=1}^{t} \log p_\theta(y_k|\mathbf{x})] \\
&= \sum_{k=1}^{t} \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{x},\mathbf{s})}[\log \frac{q_\phi(y_k|\mathbf{x},\mathbf{s})}{p_\theta(y_k|\mathbf{x})}] \\
&= \sum_{k=1}^{t} \mathbb{E}_{q_\phi(y_k|\mathbf{x},\mathbf{s})}[\log \frac{q_\phi(y_k|\mathbf{x},\mathbf{s})}{p_\theta(y_k|\mathbf{x})}] \\
&= \sum_{k=1}^{t} KL[q_\phi(y_k|\mathbf{x},\mathbf{s})||p_\theta(y_k|\mathbf{x})] \\
&= \sum_{k=1}^{t} p_\phi^{y_k} \log \frac{p_\phi^{y_k}}{p_\theta^{y_k}} + (1 - p_\phi^{y_k}) \log \frac{1 - p_\phi^{y_k}}{1 - p_\theta^{y_k}}.
\end{aligned}$$

Althouth mean-field approximation would restrict model capacity, it is a routine in VAE-related literatures [17] to make graphical model tractable. A natural direction for future work is to investigate whether it is possible to compute the KL-divergence term without mean-field approximation.