Learning Tractable Probabilistic Models from Inconsistent Local Estimates

Shasha Jin, Vasundhara Komaragiri, Tahrima Rahman, and Vibhav Gogate
The University of Texas at Dallas

Appendix

A Proof of Proposition 2

The partial derivative of

$$\lambda_1 \sum_{(X_j,X_k)\in\mathcal{E}} \sum_{x_j,x_k} \mathcal{P}_{jk}(x_j,x_k) \log \mathcal{R}_\Theta(x_j,x_k)$$

w.r.t. $\theta_{x_i,\boldsymbol{u}_i}$ is given by

$$\lambda_1 \sum_{(X_j,X_k)\in\mathcal{E}} \sum_{x_j,x_k} \mathcal{P}_{jk}(x_j,x_k) \left(\frac{\mathcal{R}_\Theta(\boldsymbol{u}_i, X_i=1|x_j,x_k)}{\theta_{x_i,\boldsymbol{u}_i}} - \frac{\mathcal{R}_\Theta(\boldsymbol{u}_i, X_i=0|x_j,x_k)}{1-\theta_{x_i,\boldsymbol{u}_i}} \right) \tag{1}$$

Proof.

$$\frac{\partial \log \mathcal{R}_\theta(x_j,x_k)}{\partial \theta_{x_i,\boldsymbol{u}_i}} = \frac{1}{\mathcal{R}_\theta(x_j,x_k)} \frac{\partial \mathcal{R}_\theta(x_j,x_k)}{\partial \theta_{x_i,\boldsymbol{u}_i}} \tag{2}$$

Using the sum rule of probability theory, $\mathcal{R}_\theta(x_j,x_k)$ equals

$$\begin{aligned}
&= \sum_{x_i,\boldsymbol{u}_i} \mathcal{R}_\theta(x_j,x_k,\boldsymbol{u}_i,x_i) \\
&= \sum_{x_i,\boldsymbol{u}_i} \mathcal{R}_\theta(\boldsymbol{u}_i,x_i)\mathcal{R}_\theta(x_j,x_k,|\boldsymbol{u}_i,x_i) &(3)\\
&= \sum_{x_i,\boldsymbol{u}_i} \mathcal{R}_\theta(x_i|\boldsymbol{u}_i)\mathcal{R}_\theta(\boldsymbol{u}_i)\mathcal{R}_\theta(x_j,x_k|\boldsymbol{u}_i,x_i) &(4)
\end{aligned}$$

Since $\mathcal{R}_\theta(x_i|\boldsymbol{u}_i) = \theta_{x_i,\boldsymbol{u}_i}$ if $X_i=1$ and $\mathcal{R}_\theta(x_i|\boldsymbol{u}_i) = 1-\theta_{x_i,\boldsymbol{u}_i}$ if $X_i=0$, we have:

$$\begin{aligned}
\mathcal{R}_\theta(x_j,x_k) = &\sum_{\boldsymbol{u}_i} \theta_{x_i,\boldsymbol{u}_i}\mathcal{R}_\theta(\boldsymbol{u}_i)\mathcal{R}_\theta(x_j,x_k|\boldsymbol{u}_i,X_i=1) \\
&+\sum_{\boldsymbol{u}_i}(1-\theta_{x_i,\boldsymbol{u}_i})\mathcal{R}_\theta(\boldsymbol{u}_i)\mathcal{R}_\theta(x_j,x_k|\boldsymbol{u}_i,X_i=0)
\end{aligned} \tag{5}$$

Differentiating the Equation above (Eq. (5)) w.r.t. $\theta_{x_i,\boldsymbol{u}_i}$, we get:

$$\begin{aligned}
\frac{\partial \mathcal{R}_\theta(x_j,x_k)}{\partial \theta_{x_i,\boldsymbol{u}_i}} = &\mathcal{R}_\theta(\boldsymbol{u}_i)\mathcal{R}_\theta(x_j,x_k|\boldsymbol{u}_i,X_i=1) \\
&-\mathcal{R}_\theta(\boldsymbol{u}_i)\mathcal{R}_\theta(x_j,x_k|\boldsymbol{u}_i,X_i=0)
\end{aligned} \tag{6}$$

Since

$$\mathcal{R}_\theta(\boldsymbol{u}_i)\mathcal{R}_\theta(x_j,x_k|\boldsymbol{u}_i,X_i=1) = \frac{\mathcal{R}_\theta(x_j,x_k,\boldsymbol{u}_i,X_i=1)}{\theta_{x_i,\boldsymbol{u}_i}} \tag{7}$$

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

and

$$\mathcal{R}_\theta(\boldsymbol{u}_i)\mathcal{R}_\theta(x_j, x_k|\boldsymbol{u}_i, X_i = 0) = \frac{\mathcal{R}_\theta(x_j, x_k, \boldsymbol{u}_i, X_i = 0)}{1 - \theta_{x_i,\boldsymbol{u}_i}} \tag{8}$$

we can rewrite Eq. (6) as:

$$\begin{aligned}\frac{\partial \mathcal{R}_\theta(x_j, x_k)}{\partial \theta_{x_i,\boldsymbol{u}_i}} = \frac{\mathcal{R}_\theta(x_j, x_k, \boldsymbol{u}_i, X_i = 1)}{\theta_{x_i,\boldsymbol{u}_i}} \\ - \frac{\mathcal{R}_\theta(x_j, x_k, \boldsymbol{u}_i, X_i = 0)}{1 - \theta_{x_i,\boldsymbol{u}_i}}\end{aligned} \tag{9}$$

Substituting Eq. (9) in Eq. (2) and using the definition of conditional probability, we get:

$$\begin{aligned}\frac{\partial \log \mathcal{R}_\theta(x_j, x_k)}{\partial \theta_{x_i,\boldsymbol{u}_i}} = \frac{\mathcal{R}_\Theta(\boldsymbol{u}_i, X_i = 1|x_j, x_k)}{\theta_{x_i,\boldsymbol{u}_i}} - \\ \frac{\mathcal{R}_\Theta(\boldsymbol{u}_i, X_i = 0|x_j, x_k)}{1 - \theta_{x_i,\boldsymbol{u}_i}}\end{aligned} \tag{10}$$

□

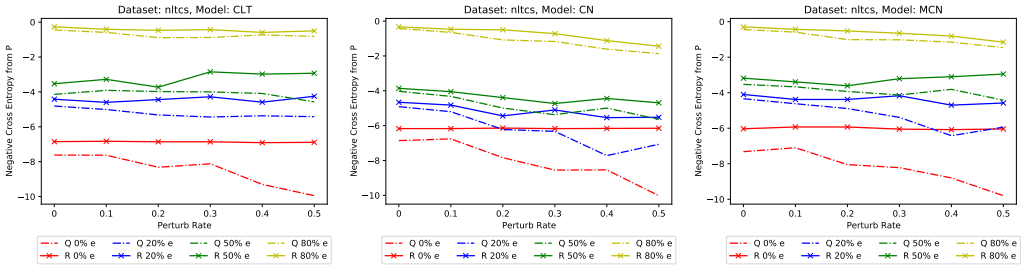
B Additional Experimental Results

Table 1: Discriminative (20% evidence) performance of $\mathcal{Q}$ and $\mathcal{R}$ on three models: CLTs, CNs, MCNs.

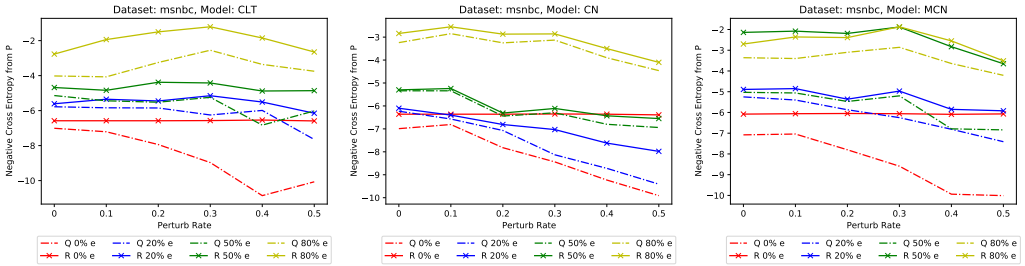
Datasets	#var	Negative cross-entropy with **20% Evidence**					
		CLTs		**CNs**		**MCNs**	
		$\mathcal{Q}$	$\mathcal{R}$	$\mathcal{Q}$	$\mathcal{R}$	$\mathcal{Q}$	$\mathcal{R}$
nltcs	16	-4.81	**-4.42**	-4.91	**-4.66**	-4.34	**-4.10**
msnbc	17	-5.78	**-5.61**	-6.23	**-6.10**	-5.25	**-4.89**
kdd	64	-5.10	**-3.55**	-5.34	**-3.06**	-4.81	**-3.38**
plants	69	-13.63	**-12.82**	-13.32	**-12.88**	-12.59	**-12.02**
audio	100	-42.35	**-40.19**	-40.58	**-37.15**	-37.86	**-35.27**
jester	100	-55.06	**-52.38**	-51.99	**-51.58**	-49.06	**-47.08**
netflix	100	-60.74	**-60.16**	-55.43	**-53.05**	-52.32	**-50.85**
accidents	111	-37.18	**-36.25**	-36.72	**-34.97**	-33.30	**-32.68**
retail	135	-15.12	**-12.42**	-15.05	**-13.02**	-14.14	**-11.65**
pumsb*	163	-34.21	**-32.23**	-31.19	**-28.78**	-29.17	**-28.14**
dna	180	-95.59	**-91.40**	-115.77	**-94.19**	-92.93	**-82.54**
kosarek	190	-16.98	**-12.71**	-13.32	**-12.01**	-11.68	**-9.04**
msweb	294	-13.47	**-11.46**	-14.04	**-11.98**	-12.14	**-11.37**
book	500	-26.51	**-23.10**	-42.92	**-21.25**	-32.43	**-20.19**
movie	500	-86.17	**-66.97**	-67.89	**-59.64**	-59.34	**-41.97**
webkb	839	-182.77	**-155.29**	-173.35	**-144.99**	-162.70	**-131.96**
reuters	889	-124.07	**-99.43**	-132.16	**-105.33**	-128.97	**-102.87**
20newsg	910	-182.35	**-151.92**	-167.36	**-152.8**	-163.80	**-142.50**
bbc	1056	-308.09	**-247.37**	-250.69	**-193.32**	-240.87	**-195.05**
ad	1558	-98.00	**-37.23**	-74.93	**-36.43**	-77.56	**-36.68**
Total AVG		-70.34	**-57.85**	-65.66	**-53.86**	-61.26	**-50.21**

Table 1 has the 20% evidence results to supplement the discriminative performance results presented in the main paper (with 50% and 80% evidences).

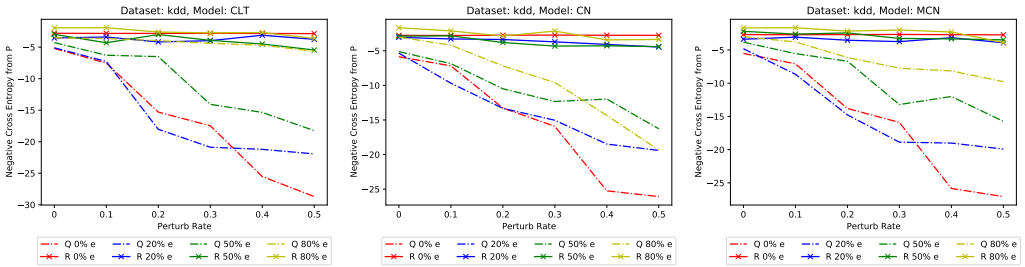
Next, we present the plots for all the 20 datasets. We varied the perturb rate h and standard deviation σ as described in the main paper. Figures 1 – 5 and Figures 6 – 10 show the respective results.



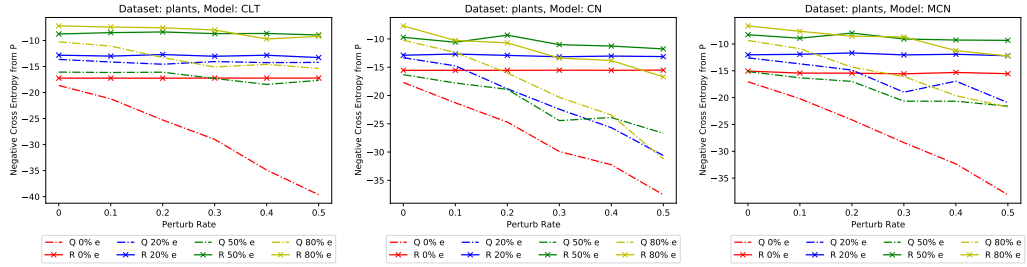
(a) nltcs; CLT (b) nltcs; CN (c) nltcs; MCN



(d) msnbc; CLT (e) msnbc; CN (f) msnbc; MCN

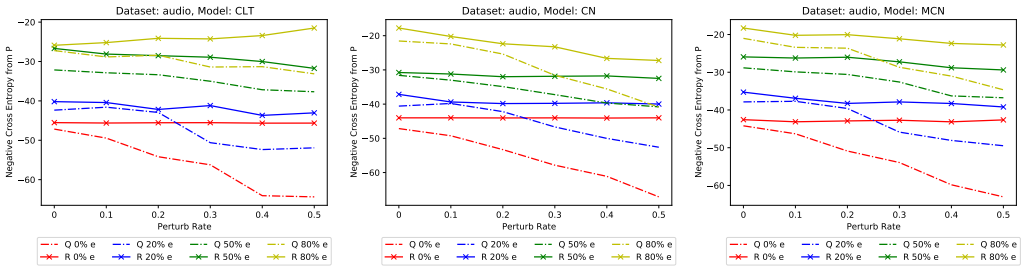


(g) kdd; CLT (h) kdd; CN (i) kdd; MCN

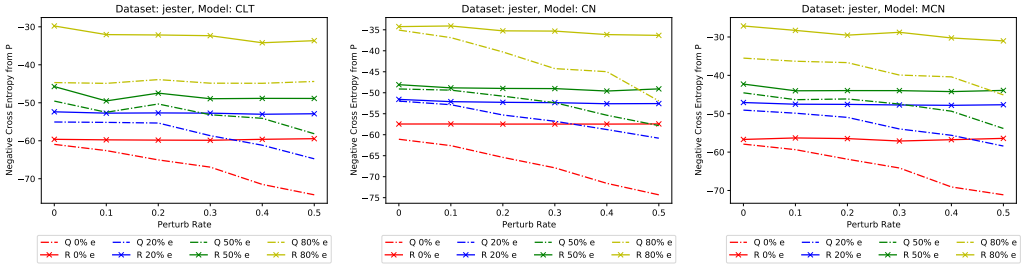


(j) plants; CLT (k) plants; CN (l) plants; MCN

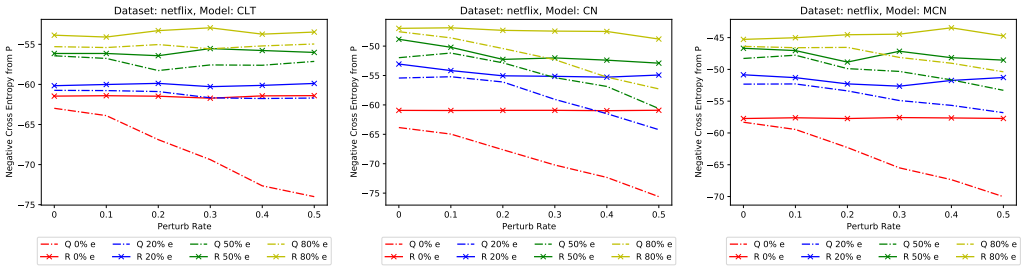
Figure 1: Negative Cross Entropy of $\mathcal{P}$ and $\mathcal{Q}$, $\mathcal{P}$ and $\mathcal{R}$ with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of perturb rate for datasets: nltcs, msnbc, kdd, plants.



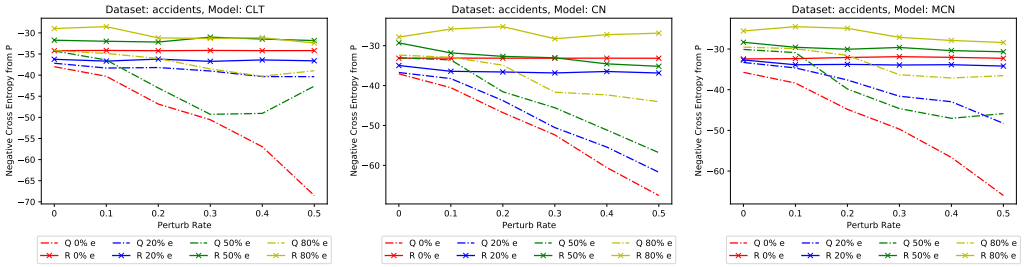
(a) audio; CLT (b) audio; CN (c) audio; MCN



(d) jester; CLT (e) jester; CN (f) jester; MCN

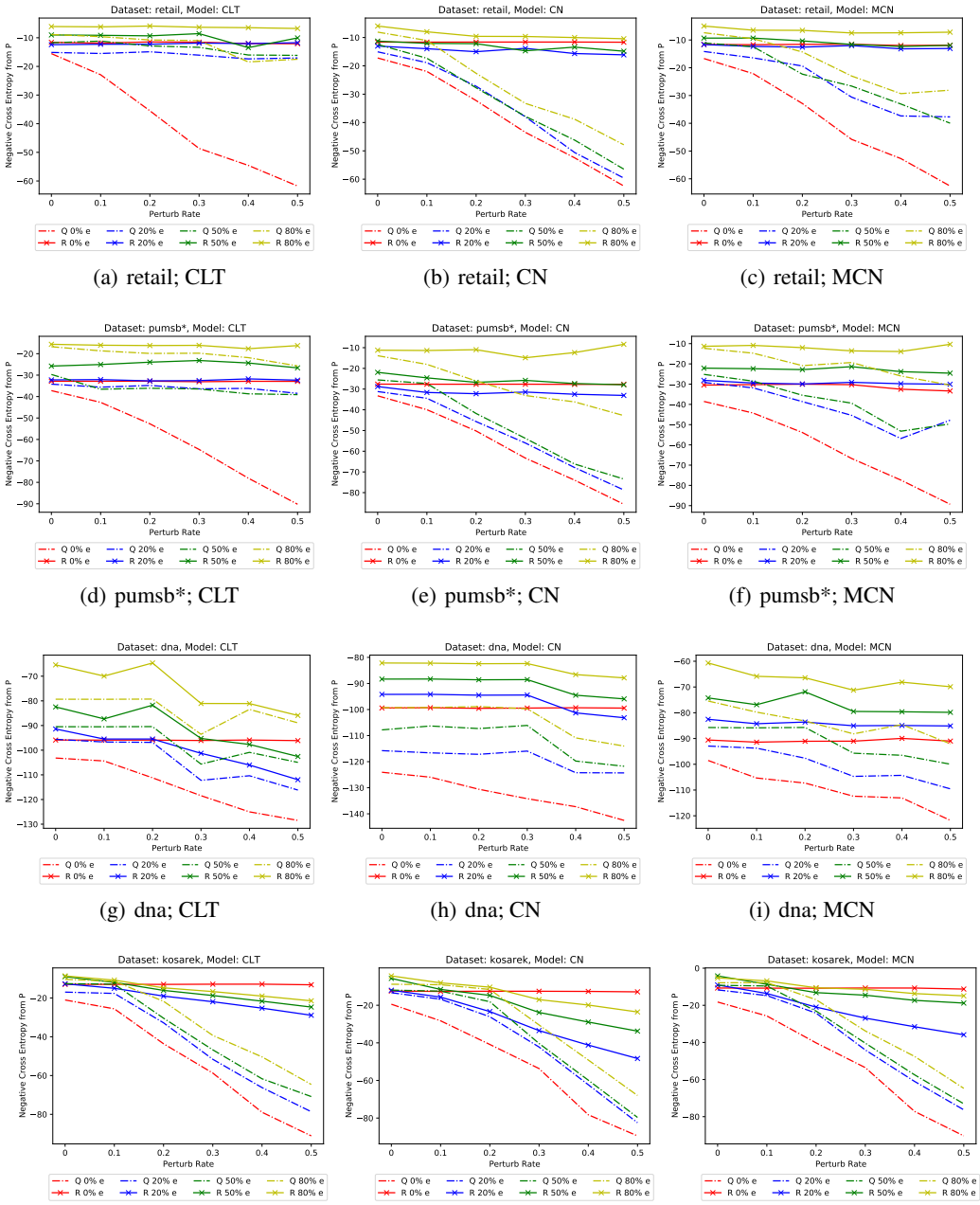


(g) netflix; CLT (h) netflix; CN (i) netflix; MCN



(j) accidents; CLT (k) accidents; CN (l) accidents; MCN

Figure 2: Negative Cross Entropy of $\mathcal{P}$ and $\mathcal{Q}$, $\mathcal{P}$ and $\mathcal{R}$ with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of perturb rate for datasets: audio, jester, netflix, accidents.



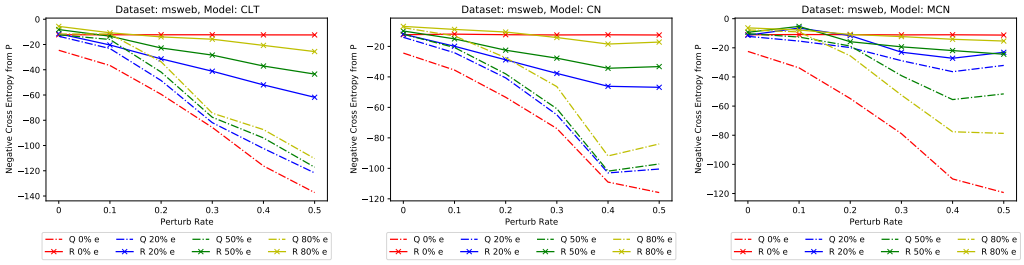
(a) retail; CLT (b) retail; CN (c) retail; MCN

(d) pumsb*; CLT (e) pumsb*; CN (f) pumsb*; MCN

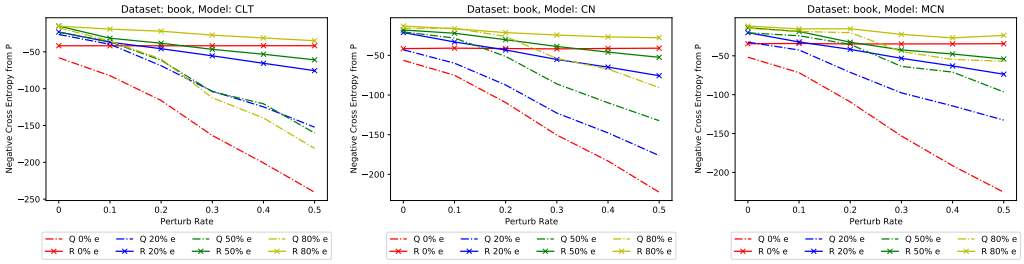
(g) dna; CLT (h) dna; CN (i) dna; MCN

(j) kosarek; CLT (k) kosarek; CN (l) kosarek; MCN

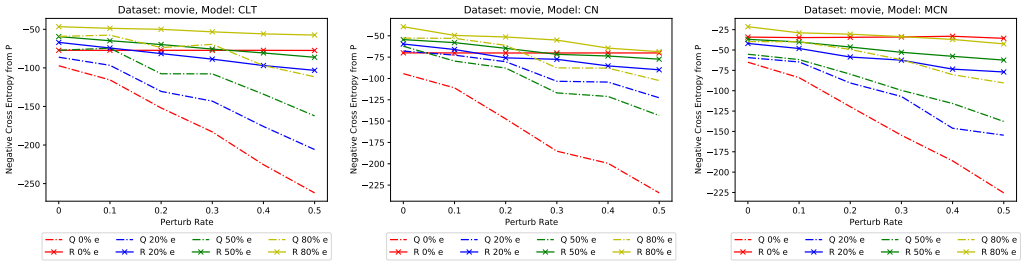
Figure 3: Negative Cross Entropy of $\mathcal{P}$ and $\mathcal{Q}$, $\mathcal{P}$ and $\mathcal{R}$ with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of perturb rate for datasets: retail, pumsb*, dna, kosarek,



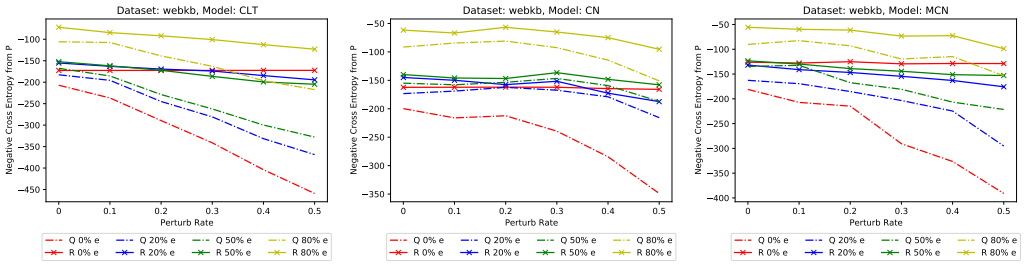
(a) msweb; CLT (b) msweb; CN (c) msweb; MCN



(d) book; CLT (e) book; CN (f) book; MCN

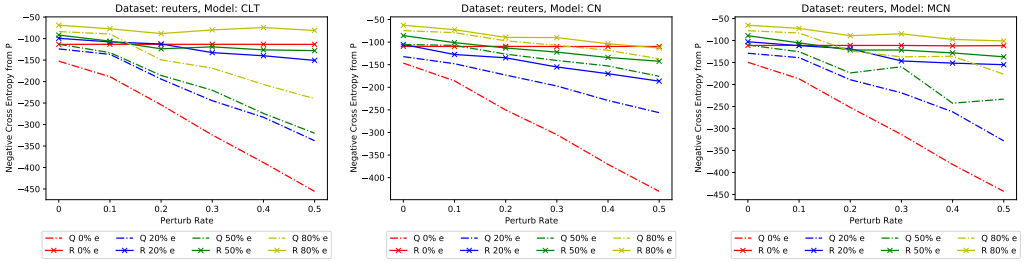


(g) movie; CLT (h) movie; CN (i) movie; MCN

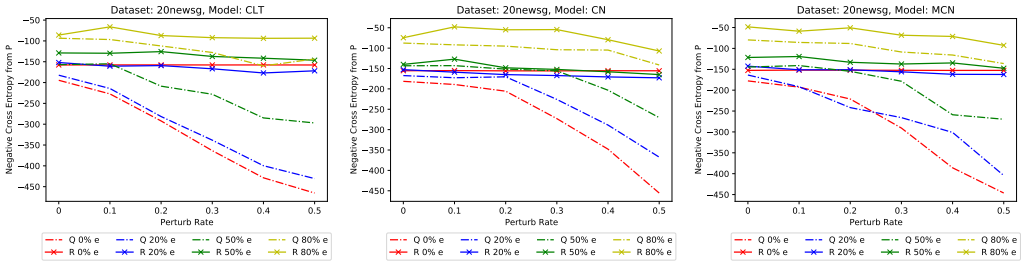


(j) webkb; CLT (k) webkb; CN (l) webkb; MCN

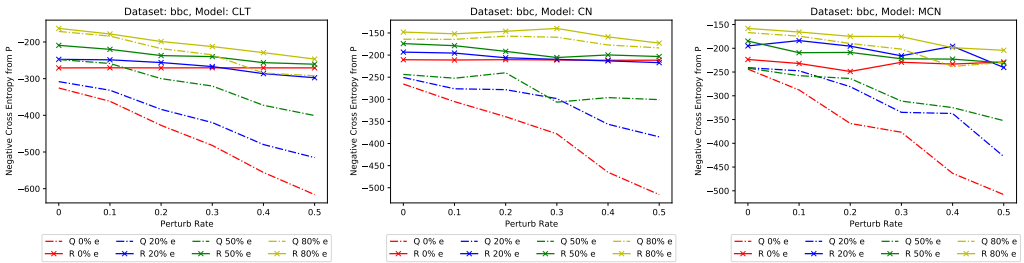
Figure 4: Negative Cross Entropy of $\mathcal{P}$ and $\mathcal{Q}$, $\mathcal{P}$ and $\mathcal{R}$ with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of perturb rate for: msweb, book, movie, webkb.



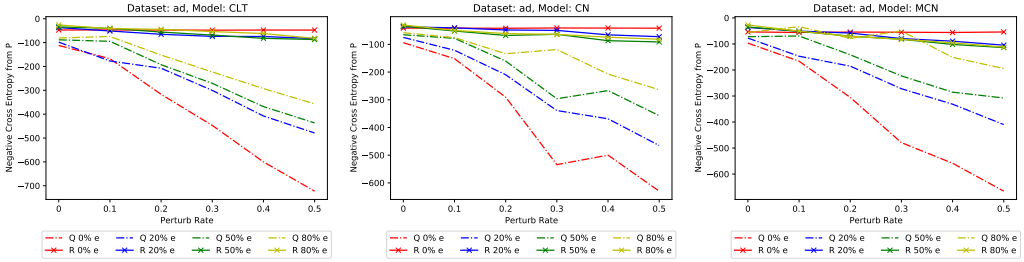
(a) reuters; CLT (b) reuters; CN (c) reuters; MCN



(d) 20newsg; CLT (e) 20newsg; CN (f) 20newsg; MCN

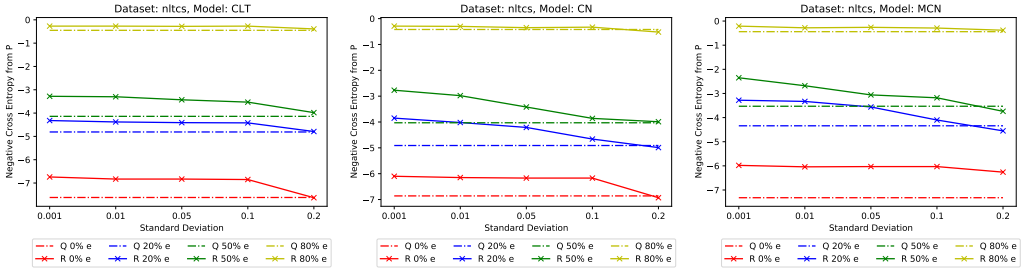


(g) bbc; CLT (h) bbc; CN (i) bbc; MCN

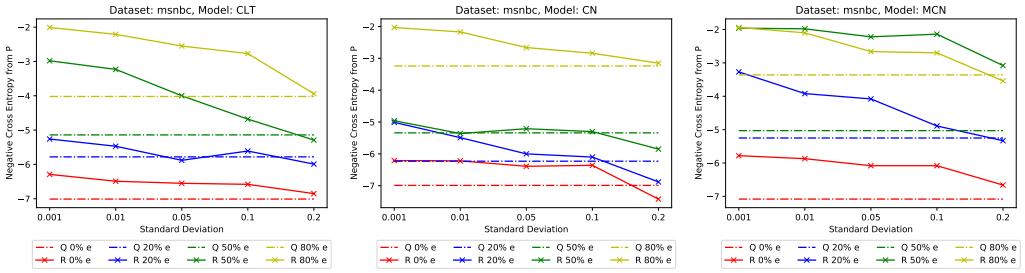


(j) ad; CLT (k) ad; CN (l) ad; MCN

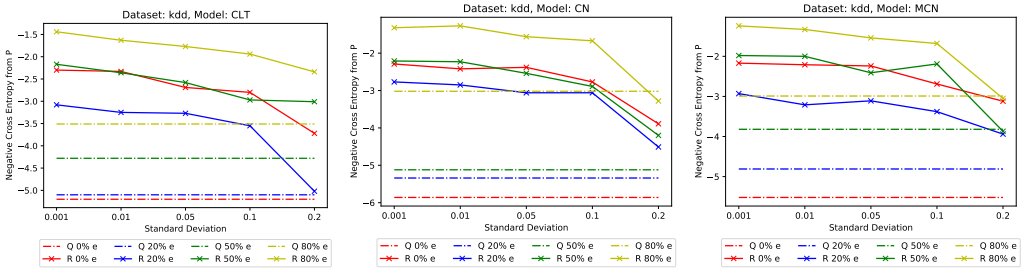
Figure 5: Negative Cross Entropy of $\mathcal{P}$ and $\mathcal{Q}$, $\mathcal{P}$ and $\mathcal{R}$ with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of perturb rate for datasets: reuters, 20newsg, bbc, ad.



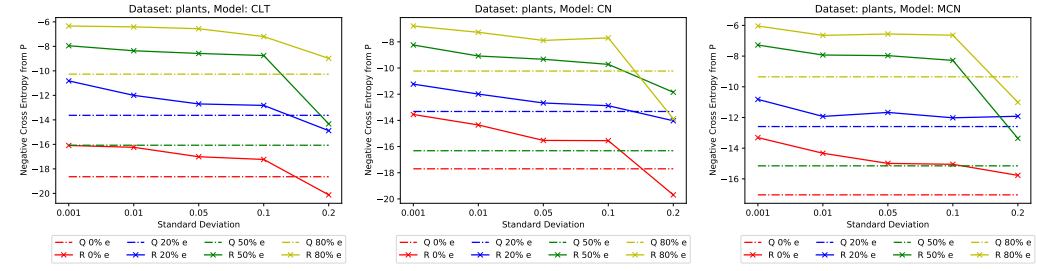
(a) nltcs; CLT (b) nltcs; CN (c) nltcs; MCN



(d) msnbc; CLT (e) msnbc; CN (f) msnbc; MCN

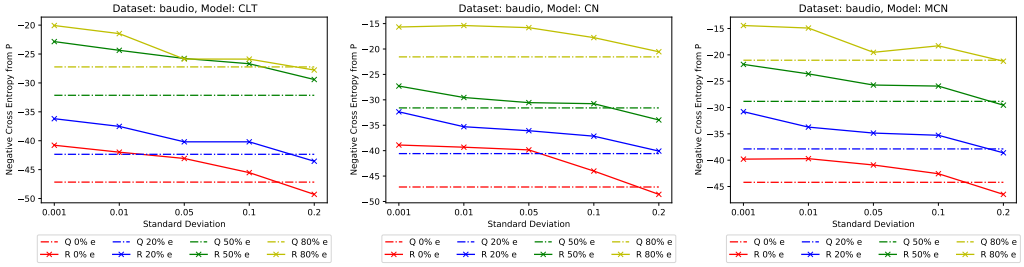


(g) kdd; CLT (h) kdd; CN (i) kdd; MCN

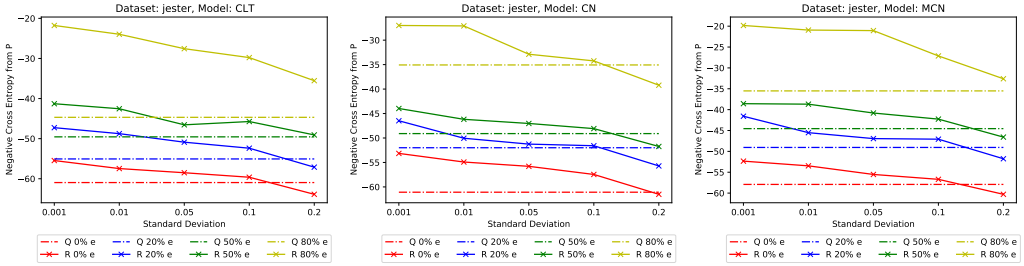


(j) plants; CLT (k) plants; CN (l) plants; MCN

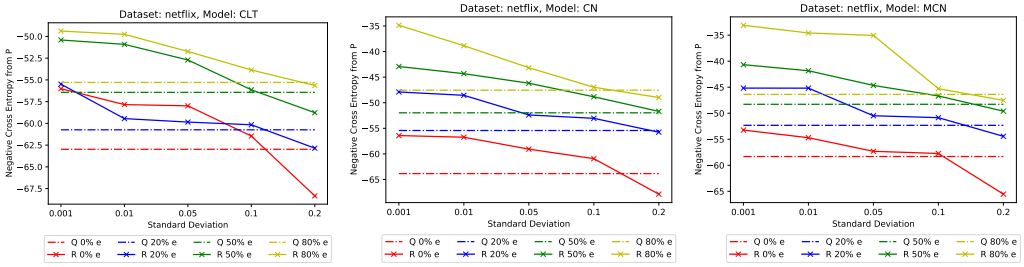
Figure 6: Negative Cross Entropy between $\mathcal{P}$ and $\mathcal{Q}$, and between $\mathcal{P}$ and $\mathcal{R}$, with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of standard deviation σ (of Gaussian noise that is applied to the local statistics $\mathcal{P}_{jk}(X_j, X_k)$) for datasets: nltcs, msnbc, kdd, plants.



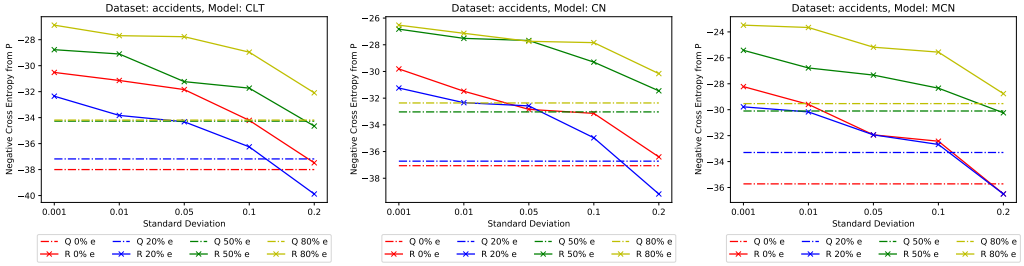
(a) audio; CLT (b) audio; CN (c) audio; MCN



(d) jester; CLT (e) jester; CN (f) jester; MCN

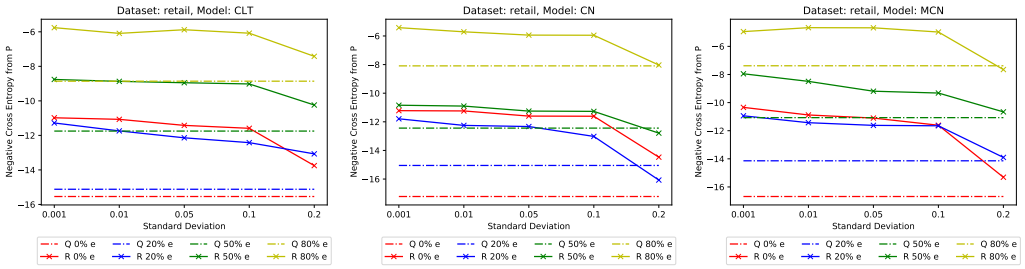


(g) netflix; CLT (h) netflix; CN (i) netflix; MCN

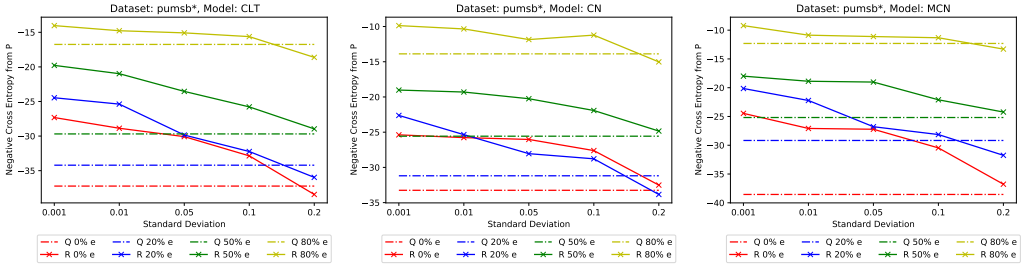


(j) accidents; CLT (k) accidents; CN (l) accidents; MCN

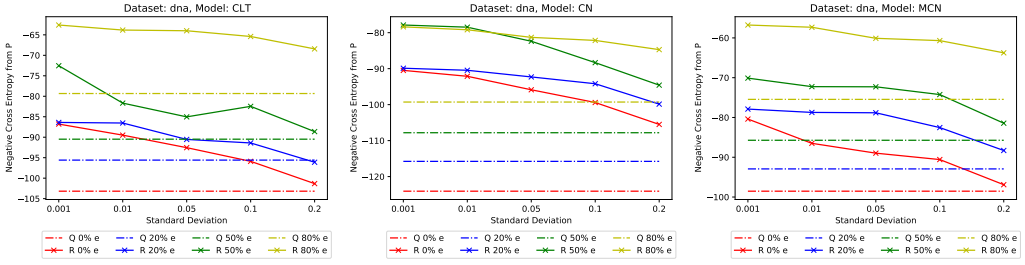
Figure 7: Negative Cross Entropy between $\mathcal{P}$ and $\mathcal{Q}$, and between $\mathcal{P}$ and $\mathcal{R}$, with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of standard deviation σ (of Gaussian noise that is applied to the local statistics $\mathcal{P}_{jk}(X_j, X_k)$) for datasets: audio, jester, netflix, accidents.



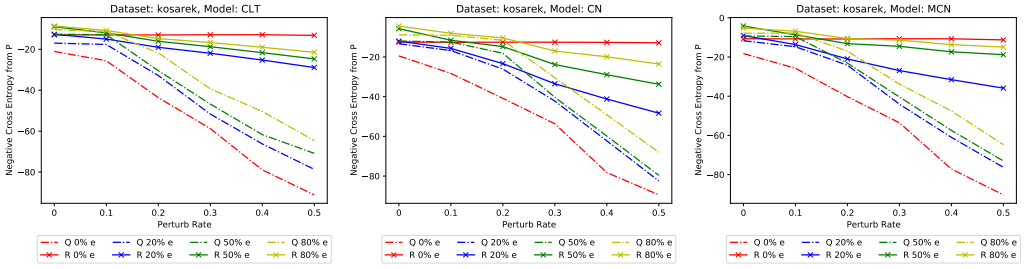
(a) retail; CLT (b) retail; CN (c) retail; MCN



(d) pumsb*; CLT (e) pumsb*; CN (f) pumsb*; MCN

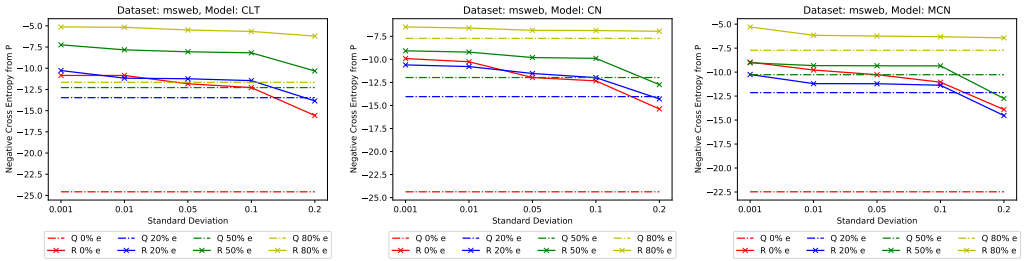


(g) dna; CLT (h) dna; CN (i) dna; MCN

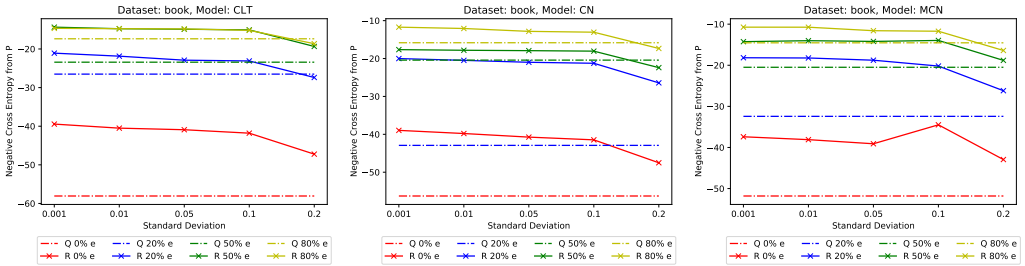


(j) kosarek; CLT (k) kosarek; CN (l) kosarek; MCN

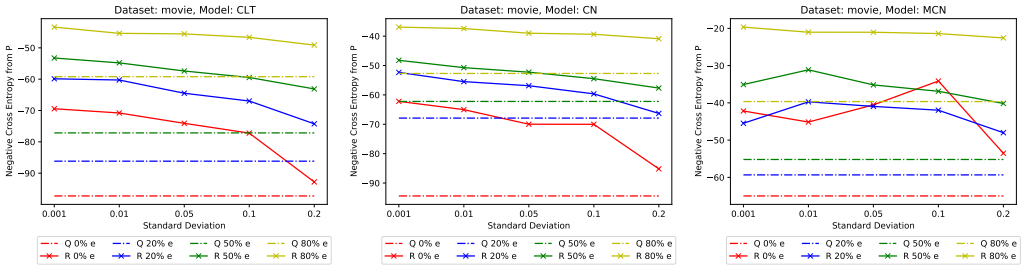
Figure 8: Negative Cross Entropy between $\mathcal{P}$ and $\mathcal{Q}$, and between $\mathcal{P}$ and $\mathcal{R}$, with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of standard deviation σ (of Gaussian noise that is applied to the local statistics $\mathcal{P}_{jk}(X_j, X_k)$) for datasets: retail, pumsb*, dna, kosarek.



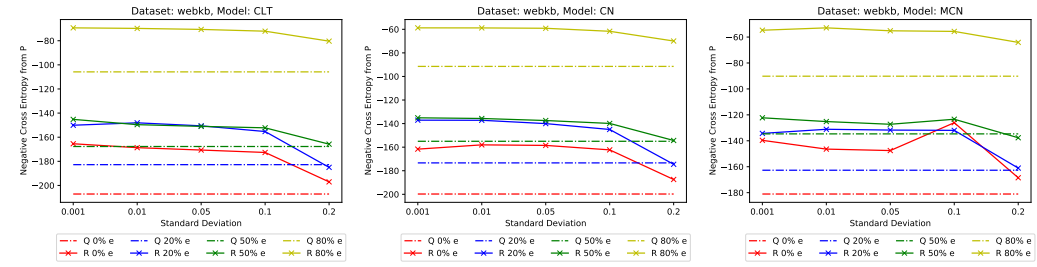
(a) msweb; CLT (b) msweb; CN (c) msweb; MCN



(d) book; CLT (e) book; CN (f) book; MCN

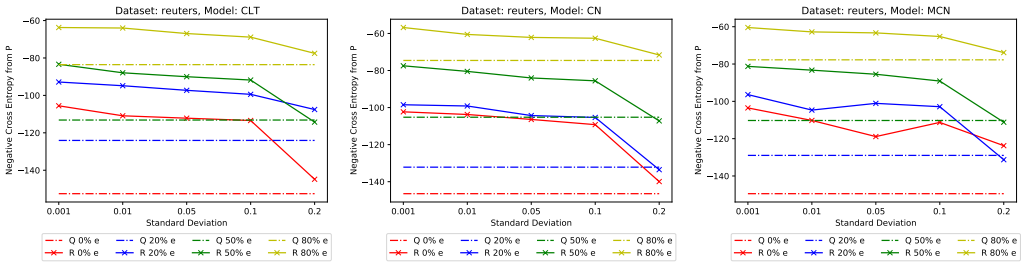


(g) movie; CLT (h) movie; CN (i) movie; MCN

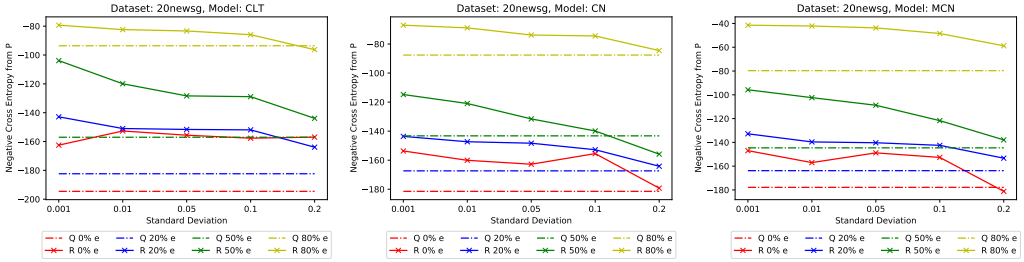


(j) webkb; CLT (k) webkb; CN (l) webkb; MCN

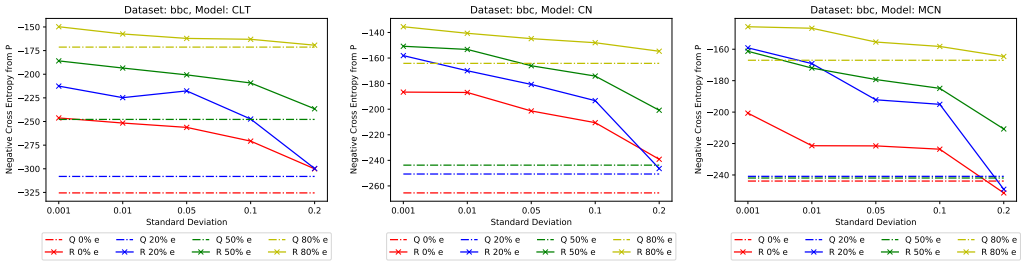
Figure 9: Negative Cross Entropy between $\mathcal{P}$ and $\mathcal{Q}$, and between $\mathcal{P}$ and $\mathcal{R}$, with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of standard deviation σ (of Gaussian noise that is applied to the local statistics $\mathcal{P}_{jk}(X_j, X_k)$) for datasets: msweb, book, movie, webkb.



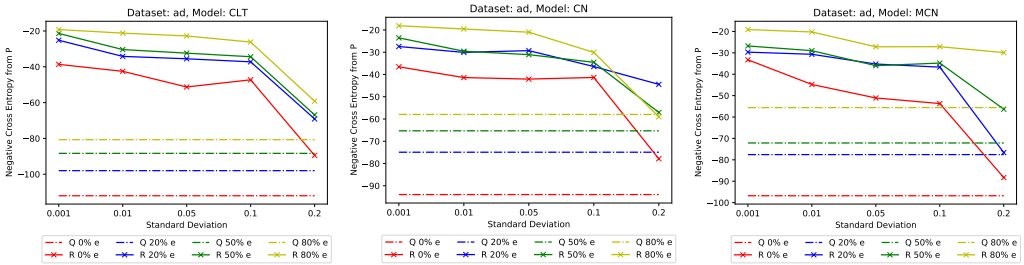
(a) reuters; CLT (b) reuters; CN (c) reuters; MCN



(d) 20newsg; CLT (e) 20newsg; CN (f) 20newsg; MCN



(g) bbc; CLT (h) bbc; CN (i) bbc; MCN



(j) ad; CLT (k) ad; CN (l) ad; MCN

Figure 10: Negative Cross Entropy between $\mathcal{P}$ and $\mathcal{Q}$, and between $\mathcal{P}$ and $\mathcal{R}$, with evidence of 0%, 20%, 50%, and 80% on three different models: CLT, CN, and MCN, as a function of standard deviation σ (of Gaussian noise that is applied to the local statistics $\mathcal{P}_{jk}(X_j, X_k)$) for datasets: reuters, 20newsg, bbc, ad.