Robust Learning of Tractable Probabilistic Models (Supplementary Material)

Rohith Peddi[1] **Tahrima Rahman**[1] **Vibhav Gogate**[1]

[1]The University of Texas at Dallas

1 EXPERIMENTAL RESULTS ON CUTSET NETWORKS: TPMS WITHOUT LATENT VARIABLES

1.1 ROBUST GENERATIVE PERFORMANCE

Tables 1 through 3 report the log-likelihood scores on the test sets $\mathcal{T}$, $\mathcal{T}_a$, and $\mathcal{T}_r$ obtained by cutset networks learned by the algorithm LearnCNet and networks trained by our proposed robust estimation methods. Each of these models were evaluated on three different test sets generated by varying the degree of perturbations ($h = \{1, 3, 5\}$) using the standard model CN. Details are described in section 4.

CN−as and CN−rs have significantly higher log-likelihood scores compared to CNs in all three degrees of perturbations.

1.2 ROBUST PREDICTIVE PERFORMANCE

We report the conditional log-likelihood scores obtained by cutset networks over varying sizes of query and evidence variable sets. We randomly chose {20%, 50%, 80%} variables as query variables and set the remaining variables as evidence variables and computed the conditional probabilities of the query variables given evidence over 200 randomly sampled test points. The average CLL scores for each of the 20 datasets are reported in tables 4 through 12 for varying degrees of corruptions h={1,3,5} to the test data by the standard model. We observe that the robust models consistently have better CLL scores on both adversarial and random perturbations compared to the standard model.

2 EXPERIMENTAL RESULTS ON SUM-PRODUCT NETWORKS: TPMS WITH LATENT VARIABLES

2.1 ROBUST GENERATIVE PERFORMANCE

Tables 13 through 15 report log-likelihood scores on the three different test sets attained by standard SPNs (SPNs) and robust SPNs (SPN−a, SPN−r) trained using our proposed method. Similar to cutset networks, SPN−as and SPN−rs consistently outperform SPNs on robust log-likelihood scores.

2.2 ROBUST PREDICTIVE PERFORMANCE

Tables 16 through 24 report conditional log-likelihood scores of SPNs, SPN−as and SPN−rs on the three test sets $\mathcal{T}$, $\mathcal{T}_a$ and $\mathcal{T}_r$ with variable sizes of evidence sets and degree of corruption. SPN−as and SPN−rs have better robust CLL scores compared to SPNs.
Until now, our experimental results are obtained under the assumption that an adversary has access to the original SPN (both its structure and parameters). In practice, an adversary may only have access to a weaker model. To evaluate the effectiveness of our proposed method under such assumption we conduct a second set of experiments. For each dataset, we learn simple mixtures of tree Bayesian networks using the Chow-Liu algorithm and assume an adversary has access to the mixture models instead of the highly accurate complex SPN. These mixture of tree BNs serve as our example of weak models. Table 25 shows the log-likelihood scores obtained by robust and non-robust SPNs on test data corrupted by these simple SPN. Similar to our previous results, robust SPNs have higher test log-likelihood scores than non-robust SPNs.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Table 1: Generative performance: Test set log-likelihood scores of cutset networks or models without latent variables. $h = 1$: hamming distance threshold. CN: Cutset networks trained on original training data, CN−a: CNs learned from adversarially generated training data by CNs, CN−r: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$			
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r	
nltcs	-6.05	-6.08	-6.06	-12.31	-10.37	-10.66	-10.18	-9.65	-9.77	
msnbc	-6.16	-6.18	-6.17	-12.40	-10.18	-10.43	-10.06	-9.41	-9.51	
kdd	-2.19	-2.43	-2.27	-11.09	-6.48	-7.01	-10.12	-6.59	-7.39	
plants	-13.50	-13.61	-13.56	-35.16	-29.94	-30.68	-25.43	-23.43	-23.81	
baudio	-41.98	-41.97	-41.97	-51.34	-51.00	-51.14	-47.17	-47.00	-47.07	
jester	-55.31	-55.31	-55.31	-64.29	-64.19	-64.24	-59.82	-59.78	-59.80	
bnetflix	-58.71	-58.71	-58.71	-66.26	-66.19	-66.22	-62.91	-62.88	-62.89	
accidents	-30.43	-30.61	-30.54	-62.54	-51.57	-53.14	-45.73	-42.72	-43.32	
tretail	-10.95	-11.48	-11.16	-20.22	-13.88	-14.35	-18.50	-15.18	-15.70	
pumsb_star	-24.24	-24.59	-24.42	-122.89	-86.54	-91.52	-64.63	-55.98	-57.63	
dna	-87.60	-87.82	-87.70	-95.74	-93.52	-93.88	-94.37	-93.08	-93.36	
kosarek	-11.01	-11.43	-11.16	-25.62	-20.39	-21.31	-21.32	-18.43	-19.00	
msweb	-10.04	-10.33	-10.16	-41.27	-35.00	-35.68	-26.05	-25.00	-25.13	
book	-37.35	-37.68	-37.44	-58.74	-52.93	-54.89	-49.36	-47.57	-48.24	
tmovie	-58.20	-58.52	-58.21	-124.66	-117.42	-119.15	-86.10	-83.96	-84.53	
cwebkb	-162.43	-163.04	-162.43	-202.97	-193.70	-196.72	-175.88	-173.92	-174.33	
cr52	-88.63	-90.63	-89.71	-173.80	-156.97	-159.88	-118.61	-115.38	-115.89	
c20ng	-163.07	-165.84	-164.60	-204.58	-191.69	-192.78	-178.52	-175.92	-175.84	
bbc	-261.86	-261.97	-261.89	-271.99	-269.79	-270.12	-269.98	-268.97	-269.21	
ad	-16.88	-18.32	-17.51	-68.40	-55.79	-56.34	-57.35	-53.14	-53.31	
avg.	-57.33	-57.83	-57.55	-86.31	-78.88	-80.01	-71.60	-69.40	-69.79	

Table 2: Generative performance: Test set log-likelihood scores of cutset networks or models without latent variables. $h = 3$: hamming distance threshold. CN: Cutset networks trained on original training data, CN−a: CNs learned from adversarially generated training data by CNs, CN−r: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$			
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r	
nltcs	-6.05	-6.36	-6.17	-20.18	-15.88	-16.49	-14.82	-13.59	-13.74	
msnbc	-6.16	-6.20	-6.18	-21.81	-16.80	-17.47	-15.21	-13.99	-14.13	
kdd	-2.19	-2.88	-2.27	-44.77	-36.04	-37.47	-23.42	-19.11	-21.14	
plants	-13.50	-13.72	-13.62	-58.00	-48.74	-49.66	-38.97	-34.88	-35.27	
baudio	-41.98	-42.09	-41.97	-63.18	-58.92	-62.86	-53.18	-51.79	-53.10	
jester	-55.31	-55.34	-55.32	-76.07	-74.46	-75.20	-64.20	-63.80	-63.98	
bnetflix	-58.71	-58.70	-58.70	-75.09	-73.07	-74.56	-66.56	-65.95	-66.43	
accidents	-30.43	-31.80	-31.18	-88.60	-56.49	-60.28	-61.59	-49.15	-50.98	
tretail	-10.95	-11.76	-11.40	-35.23	-19.57	-21.61	-29.42	-23.09	-23.62	
pumsb_star	-24.24	-29.46	-27.28	-232.40	-102.17	-114.24	-100.88	-76.55	-80.75	
dna	-87.60	-89.74	-88.62	-109.12	-99.34	-101.06	-103.41	-97.78	-98.45	
kosarek	-11.01	-11.01	-11.01	-51.91	-50.78	-51.33	-34.81	-34.41	-34.63	
msweb	-10.04	-20.06	-17.33	-69.54	-47.35	-48.32	-46.95	-48.98	-46.72	
book	-37.35	-37.34	-37.34	-76.70	-75.02	-75.61	-63.11	-62.65	-62.85	
tmovie	-58.20	-58.70	-58.37	-184.36	-174.85	-176.03	-112.96	-109.01	-109.62	
cwebkb	-162.43	-162.55	-162.48	-326.57	-322.72	-323.85	-193.89	-193.04	-193.32	
cr52	-88.63	-89.15	-88.94	-268.42	-248.81	-252.28	-151.68	-146.33	-147.38	
c20ng	-163.07	-165.79	-164.51	-408.11	-382.89	-388.35	-197.88	-190.47	-191.67	
bbc	-261.86	-262.61	-262.36	-288.77	-278.96	-280.94	-277.79	-275.59	-275.80	
ad	-16.88	-35.89	-26.01	-152.95	-121.88	-116.87	-84.72	-86.76	-79.27	
Avg.	**-57.33**	-59.56	-58.55	**-132.59**	**-115.24**	**-117.22**	-86.77	**-82.85**	**-83.14**	

Table 3: Generative performance: Test set log-likelihood scores of cutset networks or models without latent variables. $h = 5$: hamming distance threshold. CN: Cutset networks trained on original training data, CN$-$a: CNs learned from adversarially generated training data by CNs, CN$-$r: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN$-$a	CN$-$r	CN	CN$-$a	CN$-$r	CN	CN$-$a	CN$-$r
nltcs	-6.05	-6.39	-6.16	-25.38	-21.14	-21.81	-17.54	-16.80	-16.89
msnbc	-6.16	-6.21	-6.19	-28.68	-21.95	-22.47	-18.92	-17.35	-17.34
kdd	-2.19	-2.35	-2.26	-52.66	-43.99	-45.32	-35.27	-31.34	-32.10
plants	-13.50	-13.82	-13.63	-72.16	-58.08	-61.65	-49.94	-42.80	-44.80
baudio	-41.98	-41.97	-41.97	-71.35	-71.02	-70.94	-57.79	-57.70	-57.68
jester	-55.31	-55.35	-55.34	-84.09	-81.46	-81.71	-67.17	-66.62	-66.67
bnetflix	-58.71	-58.72	-58.70	-81.19	-77.69	-78.82	-69.04	-68.08	-68.43
accidents	-30.43	-31.69	-31.23	-106.02	-65.86	-69.45	-72.83	-56.29	-58.12
tretail	-10.95	-11.00	-10.97	-48.42	-36.65	-39.04	-39.64	-35.34	-36.24
pumsb_star	-24.24	-29.21	-27.28	-271.29	-111.58	-126.03	-122.10	-91.75	-95.80
dna	-87.60	-90.71	-89.19	-121.95	-104.54	-107.37	-110.50	-100.94	-101.89
kosarek	-11.01	-11.03	-11.02	-69.92	-64.78	-66.00	-48.06	-45.57	-46.38
msweb	-10.04	-10.14	-10.09	-91.78	-72.80	-75.34	-63.97	-61.30	-61.45
book	-37.35	-37.34	-37.34	-92.93	-89.81	-90.91	-73.64	-73.10	-73.33
tmovie	-58.20	-58.76	-58.77	-233.61	-222.43	-214.66	-131.36	-126.49	-125.46
cwebkb	-162.43	-163.04	-162.43	-362.88	-354.36	-357.71	-207.46	-200.72	-202.83
cr52	-88.63	-90.71	-89.75	-293.80	-253.27	-260.47	-175.38	-160.28	-163.15
c20ng	-163.07	-165.81	-164.52	-556.33	-518.67	-528.86	-211.09	-200.99	-202.75
bbc	-261.86	-264.97	-262.72	-304.09	-285.92	-290.28	-285.14	-282.64	-282.69
ad	-16.88	-42.73	-29.80	-233.57	-164.26	-168.00	-110.54	-111.97	-104.96
Avg.	**-57.33**	-59.60	-58.47	-160.11	**-136.01**	**-138.84**	-98.37	**-92.40**	**-92.95**

Table 4: Predictive performance: Conditional log-likelihood scores given 80% evidence for models having no latent variables (CNs). $h = 1$: hamming distance threshold. CN: Cutset networks trained on original training data, CN$-$a: CNs learned from adversarially generated training data by CNs, CN$-$r: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

CLL Scores on 20% query, 80% evidence, h = 1									
dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN$-$a	CN$-$r	CN	CN$-$a	CN$-$r	CN	CN$-$a	CN$-$r
nltcs	-1.74	-1.74	-1.74	-3.34	-2.80	-2.86	-3.43	-3.23	-3.29
msnbc	-1.28	-1.28	-1.28	-5.82	-4.08	-4.27	-2.65	-2.33	-2.37
kdd	-0.66	-0.72	-0.67	-9.19	-4.43	-5.05	-3.85	-2.17	-2.51
plants	-4.31	-4.39	-4.37	-17.94	-13.60	-14.13	-11.15	-9.73	-9.96
baudio	-16.61	-16.58	-16.59	-21.58	-21.29	-21.40	-19.69	-19.56	-19.61
jester	-21.20	-21.18	-21.19	-28.71	-28.61	-28.66	-24.17	-24.15	-24.16
bnetflix	-23.55	-24.14	-23.64	-29.33	-26.66	-27.46	-26.68	-25.67	-25.88
accidents	-12.59	-12.67	-12.64	-25.95	-21.51	-22.14	-18.98	-17.73	-18.01
tretail	-3.22	-3.55	-3.36	-9.15	-5.21	-5.49	-7.32	-5.29	-5.44
pumsb_star	-8.98	-9.04	-9.00	-35.93	-25.84	-27.25	-19.71	-17.43	-17.83
dna	-32.29	-32.69	-32.44	-36.62	-34.00	-34.43	-35.05	-34.27	-34.36
kosarek	-4.00	-4.09	-4.01	-13.28	-9.16	-10.03	-7.38	-6.09	-6.39
msweb	-1.53	-1.80	-1.64	-9.71	-5.88	-6.19	-6.17	-5.61	-5.58
book	-15.33	-15.28	-15.28	-21.59	-19.63	-20.35	-19.55	-18.65	-19.05
tmovie	-22.32	-22.13	-22.16	-75.88	-68.80	-70.53	-44.57	-42.17	-42.86
cwebkb	-56.93	-57.16	-56.93	-78.76	-75.06	-76.34	-60.25	-59.78	-59.82
cr52	-28.92	-29.35	-29.13	-48.76	-44.04	-44.83	-35.03	-34.57	-34.67
c20ng	-60.61	-60.73	-60.64	-77.92	-76.10	-76.67	-67.38	-67.01	-67.10
bbc	-94.47	-94.51	-94.48	-96.90	-96.82	-96.84	-97.53	-97.43	-97.47
ad	-6.30	-7.01	-6.56	-17.98	-16.02	-15.79	-20.07	-18.03	-18.18
Avg.	-20.84	-21.00	-20.89	-33.22	-29.98	-30.54	-26.53	-25.55	-25.73

Table 5: Predictive performance: Conditional log-likelihood scores given 80% evidence for models having no latent variables (CNs). $h = 3$: hamming distance threshold. CN: Cutset networks trained on original training data, CN−a: CNs learned from adversarially generated training data by CNs, CN−r: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

CLL Scores on 20% query, 80% evidence, h = 3									
dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-1.74	-1.75	-1.73	-7.46	-5.53	-5.79	-4.95	-4.38	-4.45
msnbc	-1.28	-1.29	-1.28	-9.42	-5.86	-6.25	-4.63	-3.88	-3.96
kdd	-0.66	-0.66	-0.66	-36.02	-34.33	-34.71	-10.75	-9.72	-10.71
plants	-4.31	-4.38	-4.37	-31.90	-23.35	-24.16	-18.87	-15.23	-15.52
baudio	-16.61	-16.56	-16.60	-28.59	-25.51	-28.30	-22.22	-21.36	-22.18
jester	-21.20	-21.07	-21.12	-37.81	-36.25	-36.98	-26.71	-26.39	-26.54
bnetflix	-23.55	-24.56	-23.90	-34.99	-27.88	-29.03	-28.56	-26.44	-26.68
accidents	-12.59	-13.06	-12.86	-31.77	-22.15	-23.28	-25.26	-20.17	-20.86
tretail	-3.22	-3.66	-3.47	-12.93	-6.27	-6.95	-13.09	-9.03	-9.34
pumsb_star	-8.98	-10.50	-9.87	-73.33	-32.69	-36.35	-30.30	-23.57	-24.79
dna	-32.29	-33.20	-32.72	-41.16	-36.56	-37.36	-38.19	-36.18	-36.36
kosarek	-4.00	-4.00	-4.00	-31.12	-30.89	-31.07	-13.64	-13.57	-13.63
msweb	-1.53	-1.55	-1.54	-20.70	-15.68	-16.37	-14.54	-13.23	-13.41
book	-15.33	-15.27	-15.29	-26.76	-24.76	-25.58	-25.07	-24.07	-24.51
tmovie	-22.32	-22.21	-22.14	-119.33	-109.02	-110.43	-61.20	-58.05	-58.57
cwebkb	-56.93	-56.92	-56.93	-191.32	-190.35	-190.82	-67.77	-67.63	-67.70
cr52	-28.92	-29.35	-29.13	-65.57	-57.35	-58.69	-41.54	-39.24	-39.70
c20ng	-60.61	-60.73	-60.64	-187.22	-184.41	-185.31	-71.40	-70.17	-70.53
bbc	-94.47	-94.63	-94.66	-101.02	-100.19	-100.07	-100.54	-100.28	-100.25
ad	-6.30	-17.45	-11.48	-42.80	-31.24	-26.38	-31.37	-35.21	-30.05
Avg.	-20.84	-21.64	-21.22	-56.56	-50.01	-50.69	-32.53	-30.89	-30.99

Table 6: Predictive performance: Conditional log-likelihood scores given 80% evidence for models having no latent variables (CNs). $h = 5$: hamming distance threshold. CN: Cutset networks trained on original training data, CN−a: CNs learned from adversarially generated training data by CNs, CN−r: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

CLL Scores on 20% query, 80% evidence, h = 5									
dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-1.74	-1.76	-1.74	-11.56	-9.50	-9.86	-6.23	-5.91	-5.95
msnbc	-1.28	-1.29	-1.28	-12.61	-7.18	-7.62	-5.76	-4.97	-4.97
kdd	-0.66	-0.66	-0.66	-40.49	-39.76	-39.86	-18.86	-18.91	-18.95
plants	-4.31	-4.43	-4.37	-41.78	-28.56	-31.39	-24.33	-18.71	-19.96
baudio	-16.61	-16.60	-16.60	-33.37	-33.03	-32.97	-24.53	-24.47	-24.46
jester	-21.20	-21.06	-21.07	-43.10	-40.70	-40.93	-28.13	-27.72	-27.77
bnetflix	-23.55	-25.08	-23.98	-38.38	-28.29	-30.15	-29.37	-26.87	-27.16
accidents	-12.59	-13.03	-12.88	-36.27	-24.79	-25.81	-29.86	-23.28	-23.99
tretail	-3.22	-3.26	-3.24	-16.99	-12.05	-12.92	-15.67	-12.91	-13.38
pumsb_star	-8.98	-10.60	-9.95	-96.95	-38.27	-43.56	-36.85	-28.19	-29.24
dna	-32.29	-33.58	-32.94	-46.68	-38.87	-40.14	-42.16	-37.82	-38.28
kosarek	-4.00	-4.00	-4.00	-42.81	-39.42	-40.14	-19.57	-18.56	-18.95
msweb	-1.53	-1.55	-1.54	-26.04	-20.42	-21.17	-18.88	-17.46	-17.56
book	-15.33	-15.28	-15.28	-29.89	-27.72	-28.56	-25.80	-24.47	-25.00
tmovie	-22.32	-22.29	-22.12	-172.82	-160.81	-154.84	-73.03	-69.51	-68.04
cwebkb	-56.93	-57.16	-56.93	-220.17	-219.79	-220.52	-76.95	-74.56	-75.33
cr52	-28.92	-29.35	-29.13	-67.34	-59.06	-60.43	-49.98	-45.59	-46.45
c20ng	-60.61	-60.73	-60.97	-257.85	-254.45	-252.20	-77.36	-75.88	-75.47
bbc	-94.47	-95.48	-94.75	-107.79	-103.51	-104.88	-102.92	-102.64	-102.56
ad	-6.30	-21.56	-13.28	-88.68	-57.23	-55.18	-39.49	-44.37	-38.73
Avg.	-20.84	-21.94	-21.34	-71.58	-62.17	-62.66	-37.29	-35.14	-35.11

Table 7: Conditional log-likelihood scores of cutset networks with 50% of the variables set to evidences and corruption size $h = 1$. `CN`: CN learnt from original training data, `CN−a`: CN learnt from adversarially generated training data by `CN`, `CN−r`: CN learnt from data randomly corrupted by `CN`. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `CN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `CN`.

dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-3.82	-3.83	-3.82	-9.65	-7.61	-7.92	-7.49	-7.00	-7.12
msnbc	-4.33	-4.33	-4.33	-9.81	-7.67	-7.92	-7.20	-6.72	-6.80
kdd	-1.76	-1.87	-1.78	-10.21	-5.51	-6.11	-8.66	-5.16	-5.83
plants	-9.61	-9.68	-9.64	-31.26	-25.90	-26.65	-20.83	-18.87	-19.23
baudio	-34.29	-34.26	-34.27	-42.97	-42.59	-42.74	-39.42	-39.25	-39.32
jester	-42.83	-42.81	-42.82	-50.94	-50.84	-50.89	-46.11	-46.08	-46.10
bnetflix	-45.29	-46.12	-45.39	-52.51	-48.87	-49.88	-49.28	-47.95	-48.15
accidents	-21.93	-22.12	-22.04	-52.86	-42.56	-43.96	-36.38	-33.73	-34.28
tretail	-7.77	-8.15	-7.94	-17.17	-10.72	-11.28	-14.95	-11.56	-12.10
pumsb_star	-16.14	-16.37	-16.26	-96.26	-65.64	-69.59	-54.23	-45.69	-47.26
dna	-67.88	-68.71	-68.20	-75.64	-72.17	-72.86	-73.68	-72.17	-72.42
kosarek	-8.41	-8.68	-8.49	-21.35	-16.05	-17.03	-15.29	-13.18	-13.63
msweb	-7.14	-7.45	-7.27	-33.43	-27.26	-27.88	-20.80	-19.93	-20.03
book	-29.82	-30.02	-29.91	-45.68	-39.73	-41.76	-35.88	-34.18	-34.84
tmovie	-41.36	-41.50	-41.35	-107.57	-100.30	-102.19	-74.19	-71.71	-72.49
cwebkb	-125.88	-126.40	-125.91	-155.45	-148.69	-150.97	-133.45	-131.98	-132.32
cr52	-68.01	-69.35	-68.70	-88.07	-87.55	-87.57	-78.14	-78.24	-78.04
c20ng	-128.40	-128.58	-128.43	-151.40	-147.09	-148.44	-139.25	-137.99	-138.29
bbc	-186.91	-187.00	-186.95	-193.91	-193.24	-193.32	-193.95	-193.50	-193.61
ad	-13.63	-14.84	-14.13	-43.34	-34.37	-34.73	-44.67	-40.96	-41.16
Avg.	**-43.26**	-43.60	-43.38	-64.47	**-58.72**	**-59.68**	-54.69	**-52.79**	**-53.15**

Table 8: Conditional log-likelihood scores of cutset networks with 50% of the variables set to evidences and corruption size $h = 3$. `CN`: CN learnt from original training data, `CN−a`: CN learnt from adversarially generated training data by `CN`, `CN−r`: CN learnt from data randomly corrupted by `CN`. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `CN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `CN`.

dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-3.82	-4.05	-3.90	-16.17	-12.03	-12.63	-11.37	-10.11	-10.26
msnbc	-4.33	-4.34	-4.33	-17.09	-12.61	-13.16	-11.50	-10.46	-10.58
kdd	-1.76	-1.76	-1.76	-43.86	-41.47	-42.12	-18.85	-17.03	-18.61
plants	-9.61	-9.80	-9.70	-50.67	-41.50	-42.37	-34.00	-29.51	-29.91
baudio	-34.29	-34.24	-34.29	-54.37	-50.00	-54.01	-44.22	-42.90	-44.16
jester	-42.83	-42.69	-42.74	-61.63	-59.96	-60.72	-49.75	-49.36	-49.54
bnetflix	-45.29	-47.03	-45.96	-60.77	-51.35	-52.93	-52.16	-49.46	-49.71
accidents	-21.93	-23.20	-22.65	-77.83	-48.06	-51.40	-50.90	-39.30	-40.92
tretail	-7.77	-8.31	-8.06	-28.45	-14.38	-16.23	-24.05	-17.87	-18.43
pumsb_star	-16.14	-20.18	-18.54	-195.94	-79.00	-89.34	-82.66	-60.14	-64.05
dna	-67.88	-69.79	-68.80	-86.64	-77.70	-79.34	-80.89	-76.29	-76.81
kosarek	-8.41	-8.41	-8.41	-45.39	-44.91	-45.24	-27.27	-27.13	-27.23
msweb	-7.14	-7.20	-7.17	-60.28	-47.87	-49.58	-39.12	-36.84	-37.19
book	-29.82	-29.92	-29.87	-62.15	-55.17	-57.43	-47.42	-45.19	-46.09
tmovie	-41.36	-41.55	-41.38	-159.81	-147.69	-149.37	-96.51	-92.43	-93.12
cwebkb	-125.88	-126.04	-125.95	-274.90	-271.48	-272.44	-146.05	-145.38	-145.57
cr52	-68.01	-69.37	-68.71	-107.96	-104.02	-105.08	-88.66	-87.10	-87.29
c20ng	-128.40	-128.59	-128.43	-213.50	-206.47	-208.43	-147.78	-145.41	-146.03
bbc	-186.91	-187.38	-187.23	-204.63	-199.52	-200.42	-200.05	-199.00	-198.99
ad	-13.63	-30.97	-21.05	-108.50	-85.30	-78.38	-63.34	-66.95	-58.69
Avg.	**-43.26**	-44.74	-43.95	-96.53	**-82.52**	**-84.03**	-65.83	**-62.39**	**-62.66**

Table 9: Conditional log-likelihood scores of cutset networks with 50% of the variables set to evidences and corruption size $h = 5$. CN: CN learnt from original training data, CN−a: CN learnt from adversarially generated training data by CN, CN−r: CN learnt from data randomly corrupted by CN. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-3.82	-4.09	-3.91	-21.04	-16.83	-17.50	-13.60	-13.00	-13.05
msnbc	-4.33	-4.34	-4.33	-23.05	-16.67	-17.14	-14.05	-12.85	-12.81
kdd	-1.76	-1.76	-1.76	-48.72	-46.43	-47.02	-29.87	-29.58	-29.69
plants	-9.61	-9.89	-9.72	-61.56	-47.99	-51.22	-42.68	-35.87	-37.68
baudio	-34.29	-34.29	-34.29	-62.18	-61.80	-61.72	-48.57	-48.49	-48.47
jester	-42.83	-42.69	-42.70	-68.38	-65.77	-66.02	-52.34	-51.82	-51.87
bnetflix	-45.29	-48.01	-46.13	-65.87	-52.34	-55.00	-54.12	-50.58	-50.97
accidents	-21.93	-23.10	-22.70	-94.10	-56.40	-59.60	-60.96	-45.67	-47.30
tretail	-7.77	-7.82	-7.79	-40.71	-30.12	-32.32	-31.53	-27.24	-28.11
pumsb_star	-16.14	-20.07	-18.56	-230.84	-88.93	-100.93	-96.36	-70.60	-73.96
dna	-67.88	-70.55	-69.24	-97.68	-82.01	-84.64	-87.61	-79.06	-79.96
kosarek	-8.41	-8.44	-8.43	-62.99	-57.95	-59.13	-38.12	-35.96	-36.67
msweb	-7.14	-7.23	-7.19	-81.78	-61.83	-64.44	-53.11	-50.23	-50.41
book	-29.82	-30.02	-29.91	-58.41	-55.28	-57.13	-46.80	-45.95	-46.75
tmovie	-41.36	-41.65	-41.62	-217.39	-203.58	-197.10	-112.39	-107.70	-105.93
cwebkb	-125.88	-126.40	-125.91	-309.95	-304.74	-307.12	-158.53	-153.76	-155.27
cr52	-68.01	-69.38	-68.71	-125.26	-117.23	-119.33	-99.35	-95.95	-96.66
c20ng	-128.40	-128.59	-129.19	-244.74	-236.29	-233.75	-156.50	-152.78	-152.04
bbc	-186.91	-188.82	-187.43	-214.46	-204.51	-206.93	-205.69	-204.12	-204.31
ad	-13.63	-36.89	-24.78	-185.34	-127.66	-128.53	-82.75	-86.35	-78.85
Avg.	**-43.26**	-45.20	-44.22	-115.72	**-96.72**	**-98.33**	-74.25	**-69.88**	**-70.04**

Table 10: Predictive performance: Conditional log-likelihood scores given 20% evidence for models having no latent variables (CNs). $h = 1$: hamming distance threshold. CN: Cutset networks trained on original training data, CN−a: CNs learned from adversarially generated training data by CNs, CN−r: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by CN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by CN.

CLL Scores on 80% query, 20% evidence, h = 1									
dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-4.82	-4.87	-4.85	-10.88	-8.87	-9.16	-8.79	-8.34	-8.43
msnbc	-5.84	-5.85	-5.85	-11.82	-9.65	-9.90	-9.27	-8.78	-8.87
kdd	-2.13	-2.31	-2.16	-10.72	-6.09	-6.64	-9.92	-6.34	-7.04
plants	-13.06	-13.15	-13.11	-35.55	-30.17	-30.93	-25.84	-23.80	-24.20
baudio	-41.40	-41.36	-41.38	-50.36	-49.97	-50.12	-46.69	-46.50	-46.58
jester	-54.24	-54.22	-54.23	-62.58	-62.48	-62.53	-57.99	-57.96	-57.97
bnetflix	-57.02	-58.31	-57.30	-64.73	-61.36	-62.19	-61.61	-60.55	-60.58
accidents	-28.81	-29.04	-28.95	-60.54	-49.86	-51.38	-43.84	-41.17	-41.72
tretail	-9.88	-10.36	-10.07	-19.86	-13.35	-13.90	-17.99	-14.37	-14.92
pumsb_star	-21.94	-22.25	-22.10	-117.20	-81.32	-86.19	-65.25	-55.58	-57.37
dna	-84.33	-85.17	-84.65	-92.21	-88.76	-89.43	-90.94	-89.21	-89.52
kosarek	-9.97	-10.42	-10.13	-23.58	-18.23	-19.15	-19.14	-16.29	-16.87
msweb	-9.03	-9.34	-9.17	-38.62	-32.12	-32.80	-23.90	-22.91	-23.02
book	-38.18	-38.32	-38.21	-52.69	-46.79	-48.78	-43.48	-41.77	-42.42
tmovie	-52.74	-53.09	-52.79	-127.08	-119.88	-121.74	-92.64	-90.19	-90.93
cwebkb	-157.13	-157.67	-157.09	-194.24	-185.35	-188.27	-166.40	-164.51	-164.90
cr52	-84.85	-86.53	-85.74	-95.89	-95.65	-95.40	-93.42	-93.57	-93.18
c20ng	-151.87	-152.05	-151.88	-166.12	-164.59	-165.20	-159.59	-159.09	-159.18
bbc	-247.05	-247.20	-247.11	-256.04	-255.21	-255.34	-256.94	-256.23	-256.42
ad	-17.16	-18.53	-17.76	-64.71	-52.62	-53.14	-56.27	-51.94	-52.13
Avg.	-54.57	-55.00	-54.73	-77.77	-71.62	-72.61	-67.50	-65.46	-65.81

Table 11: Predictive performance: Conditional log-likelihood scores given 20% evidence for models having no latent variables (CNs). $h = 3$: hamming distance threshold. `CN`: Cutset networks trained on original training data, `CN−a`: CNs learned from adversarially generated training data by `CN`s, `CN−r`: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `CN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `CN`.

CLL Scores on 80% query, 20% evidence, h = 3									
dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-4.82	-5.11	-4.94	-18.47	-14.24	-14.86	-13.17	-11.90	-12.05
msnbc	-5.84	-5.86	-5.85	-20.79	-16.11	-16.71	-14.39	-13.27	-13.41
kdd	-2.13	-2.17	-2.13	-44.45	-42.09	-42.71	-22.57	-20.58	-22.29
plants	-13.06	-13.29	-13.18	-57.29	-47.67	-48.64	-40.22	-35.54	-35.98
baudio	-41.40	-41.32	-41.40	-62.42	-57.91	-62.06	-52.03	-50.61	-51.97
jester	-54.24	-54.15	-54.17	-74.15	-72.43	-73.22	-62.10	-61.69	-61.88
bnetflix	-57.02	-59.41	-58.01	-73.71	-64.24	-65.69	-64.81	-62.36	-62.42
accidents	-28.81	-30.43	-29.71	-86.00	-55.10	-58.72	-60.09	-47.94	-49.66
tretail	-9.88	-10.63	-10.30	-33.73	-18.06	-20.04	-28.69	-22.09	-22.69
pumsb_star	-21.94	-26.69	-24.76	-220.39	-94.93	-106.42	-99.82	-73.69	-78.23
dna	-84.33	-86.46	-85.35	-105.18	-95.66	-97.33	-99.72	-94.21	-94.86
kosarek	-9.97	-9.97	-9.97	-49.97	-49.47	-49.82	-32.12	-31.97	-32.08
msweb	-9.03	-9.11	-9.07	-66.20	-53.66	-55.40	-44.77	-42.45	-42.81
book	-38.18	-38.24	-38.19	-70.37	-63.38	-65.61	-56.63	-54.28	-55.22
tmovie	-52.74	-53.22	-52.91	-184.56	-172.14	-173.92	-117.48	-113.28	-114.00
cwebkb	-157.13	-157.28	-157.19	-319.80	-316.08	-317.18	-182.85	-182.03	-182.29
cr52	-84.85	-86.56	-85.75	-114.44	-111.35	-111.94	-105.37	-103.73	-103.88
c20ng	-151.87	-152.05	-151.88	-180.53	-178.16	-179.06	-169.41	-168.37	-168.72
bbc	-247.05	-247.69	-247.56	-270.45	-264.42	-265.33	-264.35	-262.92	-262.96
ad	-17.16	-36.00	-26.25	-148.77	-118.60	-113.54	-82.62	-84.49	-76.88
Avg.	-54.57	-56.28	-55.43	-110.08	-95.29	-96.91	-80.66	-76.87	-77.21

Table 12: Predictive performance: Conditional log-likelihood scores given 20% evidence for models having no latent variables (CNs). $h = 5$: hamming distance threshold. `CN`: Cutset networks trained on original training data, `CN−a`: CNs learned from adversarially generated training data by `CN`s, `CN−r`: trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `CN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `CN`.

CLL Scores on 80% query, 20% evidence, h = 5									
dataset	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
	CN	CN−a	CN−r	CN	CN−a	CN−r	CN	CN−a	CN−r
nltcs	-4.55	-4.82	-4.64	-23.00	-18.67	-19.37	-15.26	-14.60	-14.67
msnbc	-5.48	-5.50	-5.49	-27.04	-20.33	-20.83	-17.09	-15.74	-15.70
kdd	-2.13	-2.13	-2.13	-52.61	-50.27	-50.87	-34.94	-34.44	-34.63
plants	-13.06	-13.39	-13.21	-71.02	-56.49	-60.04	-50.17	-42.95	-44.90
baudio	-41.40	-41.40	-41.40	-70.63	-70.25	-70.17	-56.99	-56.91	-56.90
jester	-54.24	-54.16	-54.16	-81.86	-79.14	-79.40	-64.87	-64.34	-64.40
bnetflix	-57.02	-60.64	-58.25	-79.82	-65.69	-68.28	-67.00	-63.64	-63.77
accidents	-28.81	-30.30	-29.77	-103.31	-64.32	-67.78	-71.42	-55.17	-56.99
tretail	-9.88	-9.93	-9.90	-46.95	-35.08	-37.47	-38.78	-34.19	-35.14
pumsb_star	-21.94	-26.58	-24.81	-258.73	-105.61	-118.99	-114.87	-85.59	-89.40
dna	-84.33	-87.40	-85.90	-117.66	-100.74	-103.50	-106.96	-97.35	-98.34
kosarek	-9.97	-10.00	-9.99	-67.91	-62.75	-63.95	-44.69	-42.26	-43.04
msweb	-9.03	-9.14	-9.09	-88.44	-68.35	-70.99	-60.89	-57.96	-58.15
book	-38.18	-38.32	-38.21	-68.89	-66.05	-67.87	-62.68	-61.13	-62.24
tmovie	-52.74	-53.30	-53.26	-245.35	-231.23	-224.70	-135.60	-130.64	-128.84
cwebkb	-157.13	-157.67	-157.09	-356.89	-348.96	-352.15	-197.67	-191.62	-193.50
cr52	-84.85	-86.56	-85.75	-114.44	-111.35	-111.94	-105.37	-103.73	-103.88
c20ng	-151.87	-152.05	-151.88	-180.53	-178.16	-179.06	-169.41	-168.37	-168.72
bbc	-247.05	-250.06	-247.91	-285.53	-271.26	-274.82	-271.61	-269.76	-269.77
ad	-17.16	-42.85	-30.04	-229.32	-161.28	-164.75	-106.82	-108.15	-101.14
Avg.	-54.54	-56.81	-55.64	-128.50	-108.30	-110.35	-89.65	-84.93	-85.21

Table 13: Generative performance: Test set log-likelihood scores of models having latent variables. $h = 1$: hamming distance threshold. SPN: SPN trained on original training data, SPN−a: SPN trained on the adversarially generated training data by SPN, SPN−r: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by SPN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by SPN.

	h=1								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-6.02	-6.9	-6.41	-11.51	-8.37	-8.52	-10.98	-8.36	-8.57
msnbc	-6.06	-8.76	-7.02	-12.55	-9.27	-8.86	-11.97	-9.81	-9.32
kdd-2k	-2.13	-3.88	-2.79	-11.18	-7.17	-10.48	-10.24	-8.79	-8.28
plants	-13.52	-14.17	-13.8	-21.78	-18.01	-18.2	-21.18	-17.83	-18.28
jester	-52.88	-53.28	-53.12	-56.5	-55.15	-54.97	-55.48	-55.05	-55.03
audio	-40.04	-41.12	-40.5	-45.31	-43.46	-43.03	-43.85	-43.66	-43.39
netflix	-56.85	-57.98	-57.31	-61.02	-59.37	-58.87	-59.47	-59.58	-59.1
accidents	-35.74	-35.82	-35.72	-43.09	-39.46	-39.89	-42.08	-39.33	-39.66
retail	-10.9	-11.41	-11.13	-18.52	-14.64	-16.13	-16.98	-15.2	-15.41
pumsb-star	-30.92	-31.38	-31.01	-40.06	-36.12	-36.54	-39.1	-36.15	-36.72
dna	-96.95	-97.38	-97.37	-100.01	-99.15	-99.43	-99.29	-99.01	-99.11
kosarek	-11.01	-11.8	-11.45	-18.88	-17.38	-17.41	-18.45	-16.8	-17.13
msweb	-10.04	-10.73	-10.3	-21.57	-16.83	-17.58	-20.9	-16.56	-17.19
book	-34.94	-35.59	-35.33	-42.59	-39.6	-39.83	-41.6	-40.42	-40.53
each-movie	-53.33	-54.47	-53.83	-77.64	-69.13	-70.72	-76.91	-68.62	-70.27
web-kb	-159.21	-159.53	-159.94	-171.15	-165.91	-166.12	-168.87	-165.88	-166.51
reuters-52	-90.64	-91.82	-91.66	-111.19	-102.55	-102.85	-108.41	-103.1	-105.85
20ng	-155.47	-155.33	-155.84	-169.11	-160.32	-162.19	-165.26	-160.53	-161.85
bbc	-250.75	-266.75	-253.7	-259.35	-275.39	-260.35	-258.37	-274.78	-260.54
ad	-32.16	-40.89	-35.22	-44.64	-50.5	-45.34	-43.48	-50.47	-45.18

Table 14: Generative performance: Test set log-likelihood scores of models having latent variables. $h = 3$: hamming distance threshold. SPN: SPN trained on original training data, SPN−a: SPN trained on the adversarially generated training data by SPN, SPN−r: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by SPN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by SPN.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-6.02	-9.04	-6.68	-18.68	-10.54	-11.01	-18.16	-10.57	-11.03
msnbc	-6.06	-13.41	-7.08	-21.09	-8.15	-9.9	-19.9	-10.22	-11.96
kdd-2k	-2.13	-7.82	-2.84	-25.58	-15.25	-17.45	-22.48	-18.52	-18.7
plants	-13.52	-16.04	-14.46	-37.47	-23.83	-25.06	-35.67	-23.83	-24.86
jester	-52.88	-54.91	-53.79	-62.04	-57.93	-57.43	-60.12	-58.72	-58.44
audio	-40.04	-43.52	-41.35	-52.71	-47.58	-46.68	-50.38	-49.06	-47.85
netflix	-56.85	-61.56	-58.01	-67.18	-61.43	-61.12	-64.22	-64.65	-62.16
accidents	-35.74	-36.8	-35.9	-56.8	-44.7	-45.68	-53.9	-44.58	-45.49
retail	-10.9	-12.89	-11.42	-31.09	-22.9	-22.29	-28.16	-23.01	-22.95
pumsb-star	-30.92	-32.85	-31.91	-56.97	-43.54	-45.2	-55.23	-43.6	-45.13
dna	-96.95	-97.97	-97.42	-104.87	-102.28	-103.05	-103.58	-102.08	-102.36
kosarek	-11.01	-13.61	-12.01	-33.68	-26.01	-27.47	-32.58	-25.89	-26.82
msweb	-10.04	-12.59	-10.91	-43.35	-26.2	-27.57	-42.35	-26.19	-27.38
book	-34.94	-37.11	-36.02	-55.85	-50.6	-51.39	-54.2	-50.67	-50.82
each-movie	-53.33	-57.82	-55.13	-124.37	-86.04	-96.13	-121.93	-87.12	-95.49
web-kb	-159.21	-160.88	-160.75	-191.55	-175.45	-177.97	-185.88	-176.41	-179.1
reuters-52	-90.64	-93.93	-92.78	-146.1	-119.42	-120.03	-138.25	-119.44	-124.32
20ng	-155.47	-156.29	-156.01	-190.14	-168.54	-171.26	-182.15	-169.29	-171.12
bbc	-250.75	-261.12	-254.68	-275.56	-280.13	-272.46	-273.37	-279.1	-272.4
ad	-32.16	-33.49	-32.11	-68.55	-58.71	-59.88	-65.74	-58.27	-59.97

Table 15: Generative performance: Test set log-likelihood scores of models having latent variables. $h = 5$: hamming distance threshold. SPN: SPN trained on original training data, SPN−a: SPN trained on the adversarially generated training data by SPN, SPN−r: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by SPN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by SPN.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-6.02	-10.75	-6.71	-23.22	-11.23	-11.36	-22.44	-11.19	-11.42
msnbc	-6.06	-17.85	-7.04	-26.06	-10.1	-11.96	-25.18	-11.53	-13.26
kdd-2k	-2.13	-11.61	-2.85	-37.35	-25.87	-23.43	-33.58	-26.99	-25.34
plants	-13.52	-17.9	-14.68	-52.31	-28.35	-29.77	-49.19	-28.38	-29.58
jester	-52.88	-57.18	-54.19	-66.42	-60.4	-59.58	-64.08	-62.21	-61.24
audio	-40.04	-46.23	-41.81	-58.56	-50.09	-49.33	-55.78	-53.87	-51.31
netflix	-56.85	-65.28	-58.49	-72.06	-63.12	-62.82	-68.63	-68.99	-64.54
accidents	-35.74	-38.93	-36.64	-69.76	-49.22	-50.42	-65.38	-49.19	-50.26
retail	-10.9	-14.65	-11.52	-41.9	-28.78	-30.02	-39.16	-29.56	-30.19
pumsb-star	-30.92	-34.77	-32.47	-73.13	-49.94	-51.59	-71.02	-49.84	-51.47
dna	-96.95	-98.37	-97.54	-108.87	-104.71	-106.29	-107.44	-104.48	-105.36
kosarek	-11.01	-15.5	-11.71	-47.45	-34.45	-33.3	-45.64	-34.14	-33.77
msweb	-10.04	-14.63	-10.64	-64.78	-35.18	-36.66	-63.55	-34.74	-36.62
book	-34.94	-39.21	-36.44	-68.06	-59.48	-58.26	-65.97	-60.28	-59.1
each-movie	-53.33	-63.87	-55.88	-165.06	-109.16	-101.77	-159.01	-108.74	-103.39
web-kb	-159.21	-162.14	-161.1	-208.77	-186.24	-186.78	-201.04	-185.39	-188.32
reuters-52	-90.64	-96.41	-93.4	-173.71	-123.88	-131.36	-161.36	-130.86	-137.45
20ng	-155.47	-158.2	-156.61	-207.68	-178.39	-179.1	-196.91	-178.44	-179.6
bbc	-250.75	-257.18	-254.13	-291.12	-284.1	-281.15	-288.14	-283.67	-281.78
ad	-32.16	-35.63	-32.63	-91.52	-73.72	-75.81	-87.53	-74.1	-75.19

Table 16: Predictive performance: Conditional log-likelihood scores given 20% evidence for models having latent variables (SPNs). $h = 1$: hamming distance threshold. SPN: SPN trained original training data, SPN−a: SPN trained on the adversarially generated training data by SPN, SPN−r: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by SPN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by SPN.

	h=1								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-4.82	-5.61	-5.18	-10.21	-7.02	-7.23	-9.62	-6.98	-7.23
msnbc	-4.3	-6.99	-5.26	-10.79	-7.51	-7.09	-10.22	-8.05	-7.56
kdd-2k	-2.0	-2.46	-2.05	-2.59	-3.42	-3.07	-3.96	-4.54	-3.9
plants	-9.53	-10.06	-9.75	-15.0	-12.74	-12.46	-14.92	-12.54	-13.21
jester	-40.61	-40.99	-40.83	-44.0	-42.6	-42.47	-43.02	-42.72	-42.54
audio	-31.44	-32.47	-31.83	-36.22	-34.44	-34.07	-34.86	-34.56	-34.29
netflix	-44.42	-45.52	-44.87	-48.4	-46.79	-46.31	-46.76	-46.93	-46.45
accidents	-26.94	-26.93	-26.89	-31.68	-28.81	-29.39	-31.42	-29.12	-30.01
retail	-5.69	-6.17	-5.9	-13.31	-9.38	-10.9	-11.41	-9.4	-9.77
pumsb-star	-23.19	-23.58	-23.25	-30.47	-27.12	-27.51	-29.7	-27.63	-27.67
dna	-77.25	-77.67	-77.67	-80.19	-79.32	-79.62	-79.44	-79.13	-79.22
kosarek	-4.46	-5.25	-4.88	-12.32	-10.82	-10.84	-11.87	-10.24	-10.55
msweb	-3.06	-3.77	-3.38	-13.71	-9.85	-10.05	-13.42	-9.53	-9.89
book	-26.92	-27.56	-27.32	-34.23	-30.99	-31.53	-33.18	-31.84	-32.14
each-movie	-37.31	-38.59	-37.86	-59.23	-47.49	-53.81	-58.48	-52.09	-53.44
web-kb	-127.08	-127.35	-127.66	-137.33	-132.64	-132.62	-135.4	-132.69	-133.17
reuters-52	-71.59	-72.67	-72.61	-91.68	-82.24	-83.4	-87.92	-82.76	-85.19
20ng	-123.39	-123.16	-123.75	-136.7	-127.81	-129.8	-132.45	-127.76	-129.12
bbc	-163.8	-178.49	-166.5	-172.29	-187.09	-173.03	-171.32	-186.35	-173.23
ad	-23.71	-30.6	-26.1	-35.32	-39.39	-35.42	-33.48	-38.59	-34.49

Table 17: Predictive performance: Conditional log-likelihood scores given 20% evidence for models having latent variables (SPNs). $h = 3$: hamming distance threshold. `SPN`: SPN trained original training data, `SPN−a`: SPN trained on the adversarially generated training data by `SPN`, `SPN−r`: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `SPN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `SPN`.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-4.82	-7.33	-5.36	-16.26	-8.61	-9.01	-15.92	-8.68	-9.13
msnbc	-4.3	-11.63	-5.31	-19.33	-6.37	-8.13	-18.13	-8.43	-10.19
kdd-2k	-2.0	-5.8	-2.06	-14.53	-8.07	-11.86	-13.95	-11.03	-13.17
plants	-9.53	-11.68	-10.31	-25.22	-16.25	-16.91	-26.76	-17.23	-18.12
jester	-40.61	-42.53	-41.47	-49.22	-45.27	-44.75	-47.35	-46.08	-45.67
audio	-31.44	-34.72	-32.66	-42.69	-37.9	-37.0	-40.54	-39.53	-38.02
netflix	-44.42	-48.97	-45.53	-54.21	-48.57	-48.32	-51.06	-51.62	-49.21
accidents	-26.94	-27.72	-26.98	-39.79	-31.94	-32.62	-41.54	-33.27	-34.5
retail	-5.69	-7.59	-6.15	-25.86	-17.51	-17.02	-22.07	-17.01	-17.01
pumsb-star	-23.19	-24.64	-23.99	-43.75	-32.55	-34.08	-42.26	-33.08	-34.21
dna	-77.25	-78.17	-77.68	-84.76	-82.12	-82.94	-83.37	-81.79	-82.05
kosarek	-4.46	-6.97	-5.48	-27.11	-19.35	-20.93	-25.96	-19.22	-20.24
msweb	-3.06	-5.48	-3.92	-34.04	-19.06	-19.27	-33.73	-18.85	-19.55
book	-26.92	-28.96	-27.92	-46.6	-41.22	-42.33	-44.88	-41.13	-41.56
each-movie	-37.31	-41.44	-38.97	-99.8	-59.92	-74.58	-98.01	-68.56	-74.29
web-kb	-127.08	-128.45	-128.3	-154.04	-139.85	-142.59	-149.67	-141.09	-143.73
reuters-52	-71.59	-74.59	-73.55	-122.84	-97.1	-97.41	-114.77	-97.36	-101.06
20ng	-123.39	-124.01	-123.78	-156.98	-135.27	-138.31	-147.96	-135.31	-137.3
bbc	-163.8	-173.16	-167.35	-188.43	-191.88	-184.94	-186.22	-190.87	-184.84
ad	-23.71	-24.4	-23.42	-55.86	-47.25	-46.76	-52.2	-45.29	-47.16

Table 18: Predictive performance: Conditional log-likelihood scores given 20% evidence for models having latent variables (SPNs). $h = 5$: hamming distance threshold. `SPN`: SPN trained original training data, `SPN−a`: SPN trained on the adversarially generated training data by `SPN`, `SPN−r`: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `SPN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `SPN`.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-4.82	-8.73	-5.32	-19.68	-9.17	-8.73	-19.03	-9.12	-8.87
msnbc	-4.3	-16.06	-5.27	-24.25	-8.22	-10.16	-23.32	-9.58	-11.39
kdd-2k	-2.0	-9.36	-2.06	-26.11	-19.42	-14.01	-23.75	-19.19	-17.72
plants	-9.53	-13.23	-10.41	-35.37	-19.72	-20.35	-37.29	-21.06	-21.67
jester	-40.61	-44.77	-41.87	-53.32	-47.42	-46.55	-50.96	-49.29	-48.2
audio	-31.44	-37.21	-32.96	-47.6	-39.75	-39.14	-45.07	-42.98	-40.87
netflix	-44.42	-52.38	-45.94	-58.67	-50.09	-49.81	-55.2	-55.31	-51.27
accidents	-26.94	-29.53	-27.58	-48.09	-35.41	-35.91	-50.09	-37.08	-37.81
retail	-5.69	-9.28	-6.23	-36.67	-23.23	-24.73	-32.32	-23.2	-24.32
pumsb-star	-23.19	-26.06	-24.32	-56.5	-37.55	-38.93	-54.96	-37.67	-38.94
dna	-77.25	-78.53	-77.76	-88.43	-84.26	-85.87	-86.82	-83.92	-84.71
kosarek	-4.46	-8.84	-5.18	-40.85	-27.77	-26.74	-38.96	-27.42	-27.15
msweb	-3.06	-7.35	-3.61	-53.82	-27.83	-27.46	-54.02	-27.15	-28.22
book	-26.92	-30.95	-28.37	-57.12	-49.27	-47.55	-55.25	-49.49	-48.52
each-movie	-37.31	-47.36	-39.49	-133.41	-73.05	-81.14	-125.87	-87.9	-80.96
web-kb	-127.08	-129.56	-128.68	-168.32	-149.25	-148.94	-162.17	-148.36	-150.72
reuters-52	-71.59	-76.98	-73.72	-146.3	-101.11	-107.57	-135.4	-107.2	-112.31
20ng	-123.39	-125.59	-124.4	-173.73	-144.78	-145.36	-161.37	-143.36	-144.58
bbc	-163.8	-169.66	-166.95	-203.88	-196.03	-193.77	-200.83	-195.91	-194.38
ad	-23.71	-26.12	-23.73	-74.72	-60.34	-59.47	-71.48	-58.33	-59.55

Table 19: Predictive performance: Conditional log-likelihood scores given 50% evidence for models having latent variables (SPNs). $h = 1$: hamming distance threshold. `SPN`: SPN trained original training data, `SPN−a`: SPN trained on the adversarially generated training data by `SPN`, `SPN−r`: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `SPN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `SPN`.

	h=1								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-2.56	-3.17	-2.85	-7.89	-4.56	-4.87	-6.78	-4.39	-4.72
msnbc	-2.24	-4.8	-3.16	-8.71	-5.29	-4.98	-7.99	-5.75	-5.33
kdd-2k	-1.06	-1.31	-1.1	-1.3	-2.21	-1.74	-2.33	-2.86	-2.45
plants	-5.6	-5.93	-5.74	-9.82	-7.91	-7.74	-9.54	-7.83	-8.14
jester	-25.6	-25.81	-25.75	-27.92	-26.79	-26.63	-27.18	-26.76	-26.82
audio	-19.15	-19.74	-19.31	-22.12	-20.74	-20.67	-21.27	-21.0	-20.92
netflix	-28.02	-28.31	-28.17	-29.68	-29.02	-28.67	-29.22	-29.38	-28.95
accidents	-13.48	-13.51	-13.49	-16.74	-15.0	-15.19	-16.88	-15.13	-15.76
retail	-3.07	-3.52	-3.27	-10.64	-6.67	-8.24	-7.57	-6.21	-6.72
pumsb-star	-15.95	-16.11	-15.92	-19.44	-17.86	-17.92	-19.41	-18.34	-18.3
dna	-49.23	-49.49	-49.45	-50.82	-50.36	-50.36	-50.48	-50.26	-50.24
kosarek	-2.34	-2.86	-2.48	-8.84	-8.32	-7.53	-8.44	-7.17	-7.34
msweb	-0.83	-1.47	-1.15	-11.34	-6.63	-7.71	-10.77	-6.73	-7.41
book	-16.5	-16.97	-16.89	-21.93	-19.63	-19.65	-20.81	-20.06	-19.95
each-movie	-22.12	-22.84	-22.58	-39.83	-27.53	-36.07	-37.65	-33.92	-34.14
web-kb	-76.43	-76.86	-76.99	-83.61	-81.22	-80.04	-82.24	-80.68	-80.86
reuters-52	-44.07	-44.96	-44.95	-50.89	-50.21	-48.77	-54.07	-51.9	-52.64
20ng	-76.85	-76.78	-77.1	-86.94	-79.57	-81.27	-83.33	-79.63	-80.7
bbc	-84.91	-94.4	-86.43	-92.01	-102.2	-91.75	-91.25	-101.09	-92.03
ad	-15.12	-19.16	-17.2	-18.93	-26.53	-20.19	-20.52	-24.03	-22.58

Table 20: Predictive performance: Conditional log-likelihood scores given 50% evidence for models having latent variables (SPNs). $h = 3$: hamming distance threshold. `SPN`: SPN trained original training data, `SPN−a`: SPN trained on the adversarially generated training data by `SPN`, `SPN−r`: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `SPN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `SPN`.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-2.56	-4.32	-2.94	-13.01	-5.49	-6.06	-12.12	-5.49	-5.94
msnbc	-2.24	-9.05	-2.9	-17.18	-3.67	-5.67	-15.34	-5.41	-7.34
kdd-2k	-1.06	-4.76	-1.1	-11.25	-6.56	-10.03	-11.32	-8.49	-10.46
plants	-5.6	-7.01	-6.08	-17.94	-10.06	-11.25	-17.98	-10.35	-11.68
jester	-25.6	-26.9	-26.1	-31.56	-28.46	-28.11	-29.98	-28.84	-28.58
audio	-19.15	-21.16	-19.94	-26.21	-23.14	-22.41	-24.88	-24.58	-23.3
netflix	-28.02	-29.52	-28.37	-32.36	-29.93	-29.65	-31.83	-31.45	-30.39
accidents	-13.48	-14.13	-13.51	-23.31	-17.58	-17.89	-23.97	-18.21	-19.11
retail	-3.07	-4.48	-3.48	-23.06	-14.27	-14.23	-17.01	-12.81	-12.88
pumsb-star	-15.95	-16.64	-16.41	-26.94	-20.62	-21.5	-26.3	-20.72	-21.77
dna	-49.23	-49.55	-49.41	-53.29	-51.9	-51.73	-52.62	-51.63	-51.46
kosarek	-2.34	-4.14	-3.1	-19.65	-15.4	-14.56	-20.4	-14.3	-15.15
msweb	-0.83	-2.66	-1.6	-31.39	-14.53	-16.69	-30.52	-14.95	-16.73
book	-16.5	-18.16	-17.01	-30.99	-27.96	-27.28	-28.38	-26.53	-25.61
each-movie	-22.12	-25.81	-23.24	-73.07	-34.55	-51.86	-65.35	-44.54	-48.86
web-kb	-76.43	-77.5	-77.64	-95.82	-86.15	-87.43	-92.33	-86.38	-88.21
reuters-52	-44.07	-46.59	-45.46	-69.87	-58.89	-56.1	-70.16	-60.6	-62.06
20ng	-76.85	-77.2	-77.19	-100.51	-84.28	-86.06	-94.11	-84.22	-85.65
bbc	-84.91	-90.92	-87.01	-106.37	-107.94	-102.25	-103.83	-105.91	-101.66
ad	-15.12	-15.75	-15.6	-29.85	-33.5	-25.05	-32.42	-28.21	-30.06

Table 21: Predictive performance: Conditional log-likelihood scores given 50% evidence for models having latent variables (SPNs). $h = 5$: hamming distance threshold. `SPN`: SPN trained original training data, `SPN−a`: SPN trained on the adversarially generated training data by `SPN`, `SPN−r`: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `SPN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `SPN`.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-2.56	-5.29	-2.87	-14.4	-5.78	-5.21	-13.91	-5.73	-5.35
msnbc	-2.24	-13.11	-2.67	-20.65	-4.01	-6.61	-19.45	-5.55	-7.83
kdd-2k	-1.06	-8.19	-1.1	-22.8	-17.38	-11.77	-20.38	-16.85	-14.91
plants	-5.6	-7.97	-6.16	-25.21	-12.33	-13.38	-25.56	-12.76	-13.58
jester	-25.6	-28.03	-26.31	-34.35	-29.77	-29.1	-32.25	-30.75	-30.24
audio	-19.15	-23.16	-20.1	-29.47	-24.32	-23.95	-28.22	-26.71	-25.16
netflix	-28.02	-30.14	-28.62	-34.73	-30.78	-30.58	-33.69	-33.12	-31.73
accidents	-13.48	-15.36	-14.07	-30.12	-20.18	-20.7	-31.53	-21.14	-21.81
retail	-3.07	-5.82	-3.42	-33.69	-19.6	-21.71	-27.05	-18.16	-19.54
pumsb-star	-15.95	-17.4	-16.47	-33.81	-23.28	-24.13	-33.01	-23.63	-23.99
dna	-49.23	-49.75	-49.51	-55.38	-53.21	-53.36	-54.65	-52.9	-52.95
kosarek	-2.34	-5.21	-2.8	-29.97	-21.86	-18.73	-30.42	-21.18	-20.41
msweb	-0.83	-4.01	-1.23	-50.77	-22.66	-24.53	-50.32	-22.4	-24.97
book	-16.5	-19.46	-17.44	-36.13	-32.47	-29.06	-34.98	-32.01	-29.79
each-movie	-22.12	-30.24	-23.37	-95.59	-43.28	-55.39	-86.89	-60.39	-53.87
web-kb	-76.43	-78.31	-78.03	-107.26	-93.64	-92.04	-101.44	-91.99	-93.05
reuters-52	-44.07	-48.62	-45.5	-84.69	-63.55	-64.25	-82.38	-66.89	-69.03
20ng	-76.85	-78.17	-77.53	-110.89	-90.77	-89.97	-103.14	-89.41	-90.15
bbc	-84.91	-88.91	-87.21	-119.99	-112.43	-110.49	-115.92	-110.94	-110.37
ad	-15.12	-16.56	-15.54	-41.19	-40.94	-32.16	-44.82	-36.33	-37.66

Table 22: Predictive performance: Conditional log-likelihood scores given 80% evidence for models having latent variables (SPNs). $h = 1$: hamming distance threshold. `SPN`: SPN trained original training data, `SPN−a`: SPN trained on the adversarially generated training data by `SPN`, `SPN−r`: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by `SPN`, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by `SPN`.

	h=1								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-1.06	-1.46	-1.25	-3.59	-2.0	-2.08	-3.29	-2.02	-1.97
msnbc	-0.63	-3.0	-1.44	-5.77	-2.63	-2.15	-5.18	-3.08	-2.65
kdd-2k	-0.37	-0.4	-0.38	-0.43	-1.0	-0.6	-0.51	-0.75	-0.61
plants	-2.3	-2.4	-2.31	-4.26	-2.89	-3.21	-3.88	-2.98	-3.13
jester	-10.42	-10.45	-10.43	-11.15	-10.71	-10.67	-11.02	-10.84	-10.79
audio	-7.71	-7.98	-7.78	-9.2	-8.38	-8.45	-8.61	-8.57	-8.4
netflix	-11.14	-11.25	-11.21	-12.07	-11.68	-11.47	-11.68	-11.71	-11.56
accidents	-4.12	-4.14	-4.2	-6.16	-5.12	-5.3	-6.16	-4.98	-5.38
retail	-1.12	-1.27	-1.17	-1.44	-1.4	-1.36	-2.93	-2.72	-2.91
pumsb-star	-4.47	-4.49	-4.44	-6.29	-5.45	-5.55	-6.38	-5.74	-5.69
dna	-20.04	-20.1	-20.17	-20.41	-20.2	-20.27	-20.39	-20.28	-20.36
kosarek	-0.83	-1.13	-0.89	-4.84	-4.35	-4.0	-4.3	-3.88	-3.59
msweb	-0.16	-0.48	-0.35	-8.77	-4.58	-5.73	-7.84	-4.2	-5.26
book	-7.0	-7.06	-7.06	-7.8	-7.52	-7.37	-8.17	-7.77	-7.89
each-movie	-7.57	-7.99	-7.92	-10.51	-9.69	-10.34	-11.37	-12.37	-11.45
web-kb	-28.96	-29.16	-29.12	-31.7	-31.5	-30.4	-31.56	-30.72	-31.02
reuters-52	-16.46	-16.63	-16.71	-18.01	-18.5	-18.01	-20.83	-19.67	-20.37
20ng	-31.39	-31.37	-31.52	-35.9	-32.19	-33.32	-34.31	-32.54	-33.02
bbc	-29.27	-33.32	-30.03	-33.38	-37.67	-33.24	-32.7	-37.2	-33.18
ad	-7.52	-8.85	-8.33	-8.59	-10.63	-9.07	-9.06	-10.97	-10.02

Table 23: Predictive performance: Conditional log-likelihood scores given 80% evidence for models having latent variables (SPNs). $h = 3$: hamming distance threshold. SPN: SPN trained original training data, SPN−a: SPN trained on the adversarially generated training data by SPN, SPN−r: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by SPN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by SPN.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-1.06	-2.07	-1.25	-6.94	-2.79	-3.0	-6.58	-2.83	-3.03
msnbc	-0.63	-6.29	-1.01	-12.02	-1.4	-2.35	-10.58	-2.66	-4.07
kdd-2k	-0.37	-2.36	-0.38	-1.4	-2.26	-2.52	-3.72	-3.77	-3.55
plants	-2.3	-2.83	-2.44	-7.34	-3.59	-4.49	-7.23	-3.91	-4.7
jester	-10.42	-10.73	-10.55	-12.43	-11.17	-11.08	-12.04	-11.53	-11.38
audio	-7.71	-8.23	-8.0	-11.05	-9.51	-9.01	-10.23	-9.85	-9.43
netflix	-11.14	-12.12	-11.31	-13.5	-12.09	-11.97	-12.9	-12.68	-12.2
accidents	-4.12	-4.59	-4.2	-10.87	-6.91	-7.26	-10.48	-6.91	-7.64
retail	-1.12	-1.66	-1.3	-7.61	-3.57	-5.26	-8.92	-5.84	-6.14
pumsb-star	-4.47	-4.83	-4.64	-10.19	-7.31	-7.39	-10.07	-7.29	-7.6
dna	-20.04	-20.09	-20.12	-21.01	-20.48	-20.49	-21.02	-20.63	-20.65
kosarek	-0.83	-1.81	-1.18	-11.25	-8.95	-8.04	-10.63	-7.81	-7.6
msweb	-0.16	-0.94	-0.56	-27.1	-10.06	-13.72	-23.83	-10.56	-12.36
book	-7.0	-7.76	-7.24	-9.72	-10.26	-8.86	-10.2	-9.4	-9.61
each-movie	-7.57	-9.14	-8.16	-16.38	-11.65	-12.92	-21.84	-17.88	-16.36
web-kb	-28.96	-29.57	-29.71	-37.6	-33.55	-34.28	-36.3	-32.86	-34.51
reuters-52	-16.46	-17.65	-16.82	-25.11	-23.54	-21.19	-27.65	-24.78	-24.15
20ng	-31.39	-31.41	-31.53	-43.7	-34.65	-35.64	-39.36	-34.47	-35.12
bbc	-29.27	-31.5	-30.22	-42.68	-42.23	-39.85	-40.63	-40.53	-38.2
ad	-7.52	-7.55	-7.66	-12.02	-11.64	-10.91	-13.44	-11.42	-12.45

Table 24: Predictive performance: Conditional log-likelihood scores given 80% evidence for models having latent variables (SPNs). $h = 5$: hamming distance threshold. SPN: SPN trained original training data, SPN−a: SPN trained on the adversarially generated training data by SPN, SPN−r: SPN trained via joint maximization of standard and robust likelihoods. $\mathcal{T}$: original test data, $\mathcal{T}_a$: adversarially perturbed $\mathcal{T}$ by SPN, $\mathcal{T}_r$: randomly perturbed $\mathcal{T}$ by SPN.

	h=3								
	$\mathcal{T}$			$\mathcal{T}_a$			$\mathcal{T}_r$		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-1.06	-2.66	-1.22	-7.96	-2.97	-2.57	-7.87	-2.92	-2.66
msnbc	-0.63	-8.96	-0.81	-13.45	-0.94	-1.47	-12.56	-1.91	-2.94
kdd-2k	-0.37	-0.95	-0.38	-11.59	-7.46	-4.15	-7.52	-5.67	-4.96
plants	-2.3	-3.15	-2.53	-10.48	-4.32	-5.45	-10.16	-4.89	-5.33
jester	-10.42	-11.17	-10.6	-13.62	-11.63	-11.61	-12.92	-12.26	-12.02
audio	-7.71	-9.08	-8.1	-12.52	-9.87	-9.69	-11.56	-10.69	-10.21
netflix	-11.14	-12.63	-11.41	-14.66	-12.42	-12.37	-13.73	-13.32	-12.68
accidents	-4.12	-5.2	-4.47	-15.82	-8.5	-9.24	-15.38	-8.66	-9.02
retail	-1.12	-2.31	-1.25	-14.79	-7.74	-10.32	-13.35	-7.88	-9.65
pumsb-star	-4.47	-5.3	-4.79	-13.93	-8.91	-9.02	-13.25	-8.89	-8.88
dna	-20.04	-20.06	-20.13	-21.6	-20.8	-20.73	-21.68	-21.02	-20.92
kosarek	-0.83	-2.08	-0.89	-16.48	-12.97	-9.58	-15.96	-11.5	-9.9
msweb	-0.16	-1.59	-0.27	-44.68	-16.61	-20.21	-39.92	-15.73	-18.18
book	-7.0	-7.4	-7.18	-11.73	-11.75	-9.95	-12.97	-10.9	-11.17
each-movie	-7.57	-12.06	-8.1	-19.34	-14.2	-13.84	-26.57	-23.63	-18.01
web-kb	-28.96	-29.61	-29.65	-43.85	-38.37	-36.34	-39.81	-35.76	-36.33
reuters-52	-16.46	-18.19	-17.29	-32.11	-25.45	-24.16	-33.61	-27.01	-27.33
20ng	-31.39	-32.19	-31.71	-49.79	-37.22	-38.25	-43.65	-36.84	-37.17
bbc	-29.27	-31.04	-30.52	-51.66	-46.86	-45.07	-47.42	-44.03	-43.65
ad	-7.52	-7.98	-7.64	-16.45	-16.24	-13.72	-17.31	-14.5	-15.59

Table 25: Generative performances (test set log-likelihood scores) of models with latent variables. $h \in \{1, 2, 3\}$: uncertainty set sizes. SPN: SPN learnt from original training data, SPN−a: SPN learnt from adversarially generated training data by SPN, SPN−r: SPN learnt from data randomly corrupted by SPN. W-1: test data corrupted by a weaker model under uncertainty set of size 1, W-3:test data corrupted by a weaker model under uncertainty set of size 3, W-5:test data corrupted by a weaker model under uncertainty set of size 5.

	W-1			W-3			W-5		
dataset	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r	SPN	SPN−a	SPN−r
nltcs	-9.69	-8.42	-8.48	-15.42	-10.54	-11.3	-19.09	-11.28	-12.0
msnbc	-12.36	-8.18	-8.2	-20.41	-8.87	-10.46	-24.93	-12.11	-13.9
kdd-2k	-10.93	-6.9	-5.82	-20.54	-20.29	-17.02	-27.1	-25.27	-27.64
plants	-21.19	-17.83	-18.2	-34.13	-23.76	-24.77	-42.97	-28.18	-29.37
jester	-54.82	-54.9	-54.89	-58.57	-58.78	-58.28	-61.86	-62.12	-60.87
audio	-43.35	-43.14	-42.93	-48.68	-46.97	-47.36	-53.32	-52.81	-50.44
netflix	-58.87	-59.68	-58.95	-63.18	-64.71	-62.17	-67.17	-66.98	-64.35
accidents	-40.24	-38.81	-39.08	-47.83	-43.48	-44.14	-56.2	-48.05	-48.63
retail	-17.7	-14.64	-16.13	-29.54	-22.0	-22.73	-36.53	-29.21	-29.31
pumsb-star	-38.26	-36.26	-36.7	-52.94	-43.54	-45.09	-66.47	-49.78	-51.31
dna	-98.92	-98.86	-99.18	-101.79	-101.71	-101.97	-104.53	-104.33	-104.55
kosarek	-18.21	-16.66	-13.43	-31.71	-27.27	-26.11	-44.17	-34.0	-34.47
msweb	-20.27	-17.18	-16.84	-41.71	-25.84	-25.81	-61.93	-34.48	-38.68
book	-39.56	-41.07	-40.5	-50.69	-50.62	-51.53	-59.94	-57.4	-56.0
each-movie	-75.91	-70.36	-68.64	-118.83	-89.22	-91.61	-154.37	-101.68	-107.71
web-kb	-166.85	-164.88	-165.02	-178.66	-171.85	-173.59	-189.35	-183.35	-179.67
reuters-52	-97.31	-98.23	-97.34	-118.44	-117.15	-114.57	-136.28	-123.15	-124.98
20ng	-160.13	-159.35	-160.23	-171.4	-168.6	-169.32	-182.56	-176.38	-173.93
bbc	-256.84	-273.79	-259.74	-266.79	-275.87	-270.09	-279.54	-279.13	-275.67
ad	-41.16	-51.73	-45.22	-59.85	-57.71	-60.95	-77.95	-75.32	-72.96