R2: Soft Medoid equation. We apologize for any confusion about Eq. 2 and 3. We use the softmax with temperature only for approximating the $\arg\min$ operation of the Medoid (a multivariate generalization of the median). Thus, the temperature allows us to interpolate between the sample mean and the Medoid. In our proof, we show that the Soft Medoid (with a finite temperature) has the same breakdown point as the Medoid, which is well known to be robust.

R2: Assumptions It is not our intention to claim that the aggregation is the only reason for adversarial vulnerability of a GNN (will be clarified). We reason that on top of the usual (potentially non-robust) neural network components, GNNs introduce additional (typically non-robust) aggregations. In Sec. 5, we leave the GNN unchanged except for the aggregation function and it does substantially increase the robustness. Further countermeasures w.r.t. adversarial vulnerability are orthogonal to our approach. Note that our approach also helps in case of multiple "outliers" (see next).

R4: Breakdown point/risk. With an appropriate budget, the adversary can only perturb a subset of the aggregation inputs with the goal of crossing the decision boundary (i.e. a very small perturbation magnitude is unlikely to suffice). As long as the adversary only controls the minority of inputs, our robust estimator comes with a bounded error regardless of the attack characteristics (i.e. no attack can distort the aggregation result arbitrarily). According to Figure 2, the bias of the Soft Medoid is much lower than of the sample mean for a wide range of the perturbed fraction ϵ as well as different magnitudes of the perturbation (not only if arbitrarily far away). Moreover, the Soft Medoid comes with a guarantee on $\sup_{\tilde{\mathbf{X}}_\epsilon} \|t(\tilde{\mathbf{X}}_\epsilon) - t(\mathbf{X})\|$ (e.g. see line 599 in Appendix). We expect this guaranteed upper bound to be *slowly* increasing on the interval $\epsilon \in [0, \epsilon^*]$. In conclusion, our approach is more robust for reasonable ϵ and this naturally resonates with the desire of unnoticeable (adversarial) attacks.

R1/R2/R4: Certification. We use certified robustness because it provides guarantees (it is agnostic to the individual attacks used). Note that no attack (including adaptive attacks) can achieve a lower accuracy than certified (i.e. we evaluate the worst-case). We acknowledge that robustness certification on graphs is fairly new and introduces additional complexity to the setup. We built upon existing certification of [51] and will extend the explanation in our paper.

R2: Further attacks. As suggested, Table A extends Sec. 5.3 with three additional attacks [4-6]; our method outperforms all baselines in evasion attack settings (perturbed test data). We report the mean margin and failure rate of targeted evasion Nettack [4] (budget $\Delta = d - 1$), and the accuracy after evasion PGD [5] (transfer) as well as poisoning Metattack [6]. In the case of poisoning (perturbed training data) on Citesser, all defenses perform comparably except the slightly better performing SVD GCN. On Cora ML, we clearly outperform Jaccard GCN and RGCN.

Table A: Empirical results with three additional attacks. ϵ is the fraction of altered edges.

		Acc.	Cert. Edges			Nettack		Acc. PGD & Metattack for ϵ			
			A.&d.	Add	Del.	Marg.	Fail. r.	0.15	0.25	0.15	0.25
Cora ML [46]	Vanilla GCN	0.813	1.809	0.204	4.370	-0.414	0.181	0.671	0.612	0.488	0.383
	Vanilla GDC	**0.818**	1.968	0.206	4.315	-0.482	0.150	0.666	0.606	0.467	0.342
	SVD GCN	0.761	0.900	0.069	2.461	0.165	0.608	0.664	0.619	**0.594**	**0.496**
	Jaccard GCN	0.813	1.814	0.207	4.308	-0.462	0.258	0.668	0.612	0.5	0.415
	RGCN	0.763	1.481	0.165	3.879	0.003	0.350	0.629	0.576	0.493	0.359
	SM GDC ($T = 1.0$)	0.814	4.538	0.543	4.695	0.086	0.542	0.677	0.627	0.546	0.433
	SM GDC ($T = 0.5$)	0.786	5.496	0.642	4.776	0.111	0.525	0.681	0.636	0.557	0.459
	SM GDC ($T = 0.2$)	0.758	**5.955**	**0.692**	**4.860**	**0.237**	**0.617**	**0.69**	**0.66**	0.566	0.474
Citeseer [47]	Vanilla GCN	0.693	1.256	0.119	3.737	-0.557	0.025	0.575	0.518	0.529	0.439
	Vanilla GDC	0.683	1.168	0.103	3.673	-0.507	0.050	0.542	0.48	0.546	0.45
	SVD GCN	0.621	0.500	0.001	2.112	-0.011	0.508	0.564	0.52	**0.605**	**0.531**
	Jaccard GCN	0.698	1.233	0.115	3.804	-0.430	0.167	0.585	0.534	0.578	0.503
	RGCN	0.664	1.005	0.083	3.260	-0.054	0.408	0.525	0.482	0.57	0.5
	SM GDC ($T = 1.0$)	**0.709**	2.331	0.276	3.967	-0.079	0.358	0.581	0.531	0.567	0.502
	SM GDC ($T = 0.5$)	0.705	3.316	0.417	4.026	0.095	0.508	0.597	0.554	0.569	0.506
	SM GDC ($T = 0.2$)	0.691	**4.522**	**0.565**	**4.173**	**0.306**	**0.650**	**0.616**	**0.592**	0.569	0.512

R2/R4: Accuracy vs. robustness. Based on Table A, Table 1, and Figure 3, we see that there is a tradeoff between accuracy and robustness (consistent with e.g. *Tsipras et al. ICLR 2019. Robustness May Be at Odds with Accuracy.*). If we tune the hyperparameters for a comparable accuracy, our approach is consistently more robust w.r.t. structure perturbations than all the other defenses (the only exception is the Metattack poisoning attack). Note that recently we were able to improve our results with row-wise adjacency matrix normalization [45] (see first columns of Table A).

Table B: Average duration (time cost in ms) of one training epoch (over 200 epochs, preprocessing counts once). We report "-" for an OOM (DeepRobust impl.). We used one 2.20 GHz core and one 1080 Ti (11 Gb).

	Cora ML [46]		Citeseer [47]		PubMed [46]	
GDC Prepr.		✓		✓		✓
SM GCN	41.2	210.9	36.6	154.1	86	497.8
SVD GCN	119.4	120.8	66.3	67.3	-	-
Jaccard GCN	19.1	147.8	11.2	118.0	84.9	585.4
RGCN	8.7	7.5	6.3	9.3	-	-
Vanilla GCN	5.1	7.1	4.7	7.8	6	66.1
Vanilla GAT	15.2	65.6	11.8	53.3	46.4	270.8

R2/R4: Datasets. In Table 4 (Appendix), we report the results on PubMed. PubMed is about 10 times bigger than Citeseer (see Table 2). None of the referenced attacks/defenses [3–7, 9, 12, 54] uses a larger dataset. Note that our approach scales (runtime/space) with $\mathcal{O}(n)$ (SVD GCN has space cmplx. $\mathcal{O}(n^2)$).

R2/R4: Time cost. The Soft Medoid is comparable to the defenses SVD GCN and Jaccard GCN (see Table B).

R4: Attribute robustness and federated training. Please see Section C.4 for attribute robustness. Further, we agree that federated learning is the way to go for future research.