A. Supplementary Material

To compute the mapper graphs, we hand-tuned the mapper parameters. For MNIST data, $n = 40$ and $p = 50\%$. For CIFAR-10 data, $n = 40$ and $p = 25\%$. For DBSCAN, $minPts = 5$, and the size of the neighborhood ϵ is determined by the "elbow" approach (see (Zhou et al., 2021) for its details).

A.1. Details on Adversarial Training

In experiments, we train the standard adversarial model M_{adv} with the projected gradient descent (PGD) type attack (Madry et al., 2018). We explore the performance of M_{adv} against two variations of perturbation bound: ℓ_∞ and ℓ_2. For MNIST data, we set the learning rate to be 0.01, and for CIFAR-10 data, the learning rate is 0.1. We train each instance of M_{adv} for 200 epochs.

To perform the model refinement, we first identify weak regions in the mapper graph of activation vectors from the training dataset. We then train M_{refine} using only the training images that belong to these weak regions. We train each instance of M_{refine} for 50 epochs. To compute the test accuracy, we compare each activation vector from the test dataset with its nearest neighbor of the activations from the training data. If the nearest neighbor belongs to a weak region, we identify the test activation vector as in the weak region as well. For test images belonging to weak regions, we use M_{refine} to predict their labels. For test images not belonging to weak regions, we use M_{adv} to predict their labels. The overall prediction accuracy is then calculated by adding the number of correctly predicted images from both M_{refine} and M_{adv}, and dividing it by the total number of images in the test dataset.

A.2. Additional Experiments on Training MLP with MNIST for Model Refinement

Figure 8 shows the mapper graphs of neuron activations constructed using a uniform cover (left) and a balanced cover (right), respectively. In a standard training with the clean data, the model's corresponding mapper graphs contain clear bifurcations, where different classes of images are located in separate branches of the mapper graph.

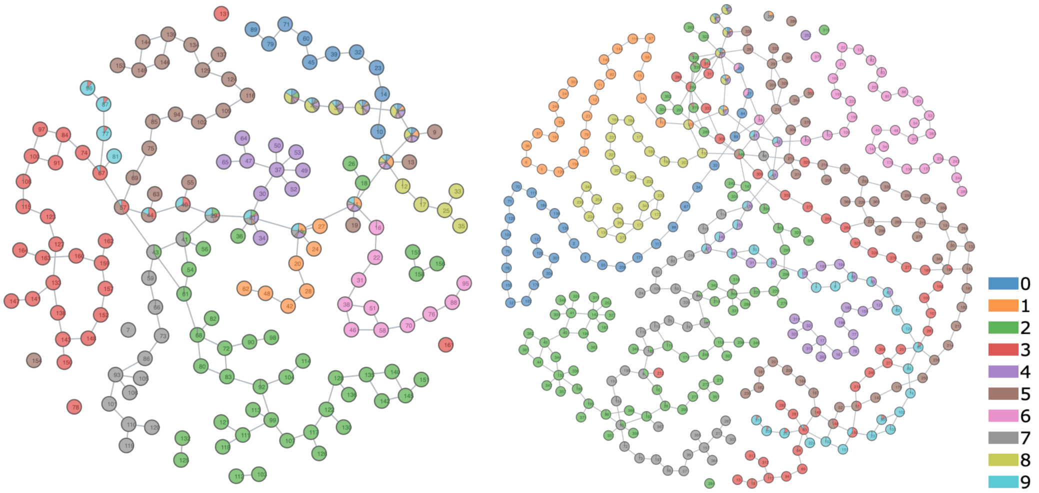

Figure 8. Mapper graphs generated with a uniform cover (left) and a balanced cover (right). MNIST with MLP M_{clean}.

When we subject the model to ℓ_∞-PGD attacks across different levels, the corresponding mapper graphs are shown in Figure 9 with a uniform cover and Figure 10 with a balanced cover, where weak regions are highlighted by red circles.

When we subject the model to ℓ_2-PGD attacks across different levels, the corresponding mapper graphs are shown in Figure 11 with a balanced cover, where weak regions are highlighted by red circles.

A.3. Training ResNet-18 with CIFAR-10

We train a ResNet-18 with the CIFAR-10 dataset, where the clean model M_{clean} has a test accuracy of 93.28%. Again, we perform model refinement by leveraging the misclassified samples in the weak regions.

Figure 12 shows the mapper graphs of neuron activations from M_{clean} constructed using a uniform cover (left) and a balanced

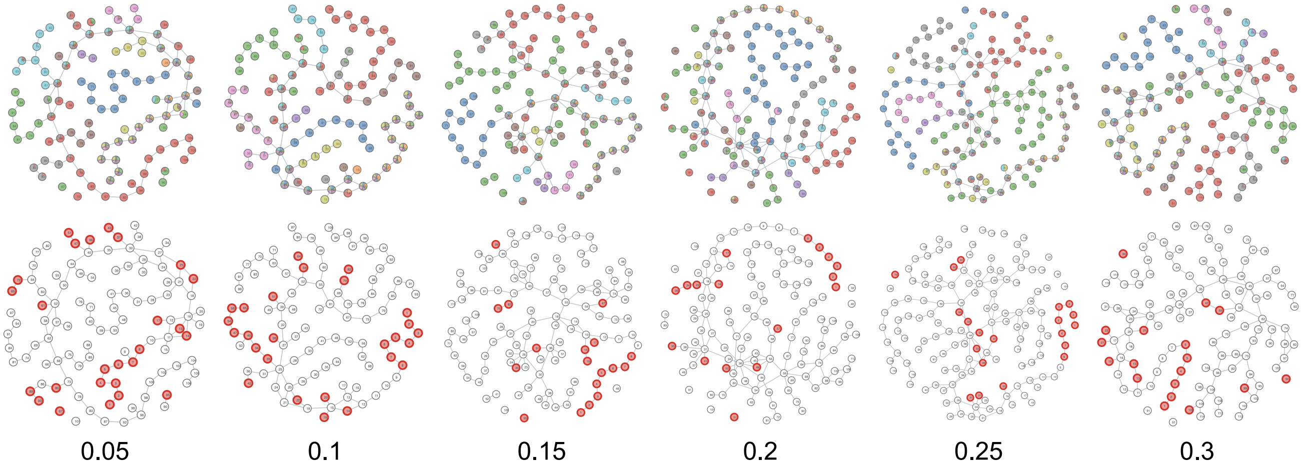

Figure 9. Mapper graphs using a uniform cover, MNIST with MLP under ℓ_∞-PGD attack (top). Weak regions are highlighted in red circles (bottom).

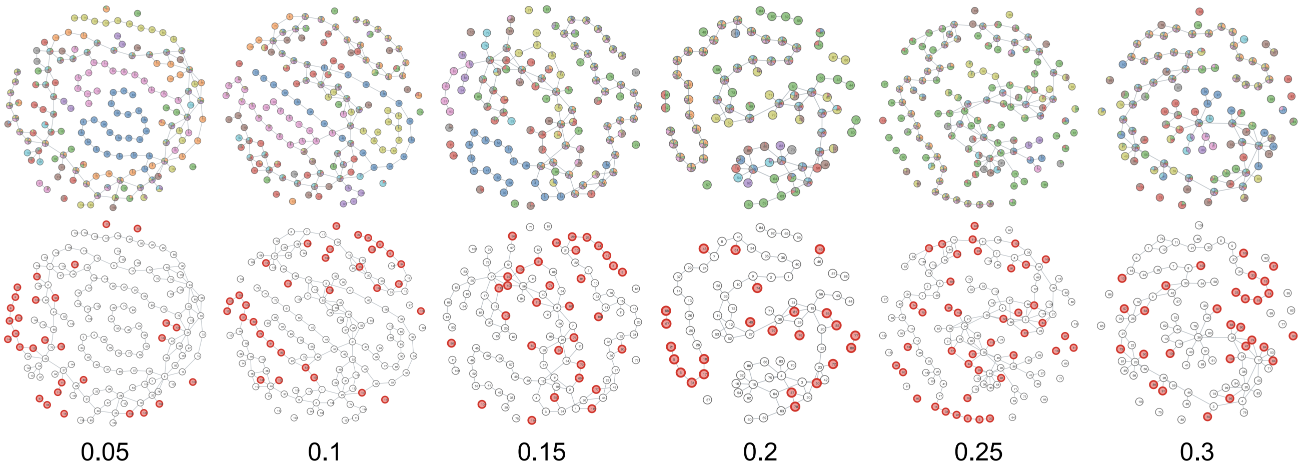

Figure 10. Mapper graphs using a balanced cover, MNIST with MLP under ℓ_∞-PGD attack (top). Weak regions are highlighted in red circles (bottom).

cover (right), respectively. They are shown to contain clear bifurcations that separate different image classes into branches.

Table 3 compares the (robust) test accuracy achieved by the models M_{adv} and M_{refine}, respectively, under different levels of ℓ_∞-PGD attack. We do not observe improved robust accuracy. Similar observations can be obtained under different levels of ℓ_2-PGD attacks shown in Table 4. This observation shows that for ResNet-18 with the CIFAR-10, the topological structure in the mapper graph is not effective in helping identify the weak regions containing samples vulnerable to adversarial attacks.

To dive deeper into the structures of the mapper graphs when the model is under ℓ_∞- or ℓ_2-PGD attacks, we highlight the mapper graphs with identified weak regions in Figure 13 and Figure 14. Topological neighborhoods under such a setting typically have low purity but high training accuracy, indicating that the model is overfitting, thus making it difficult to identify weak regions useful for model refinement.

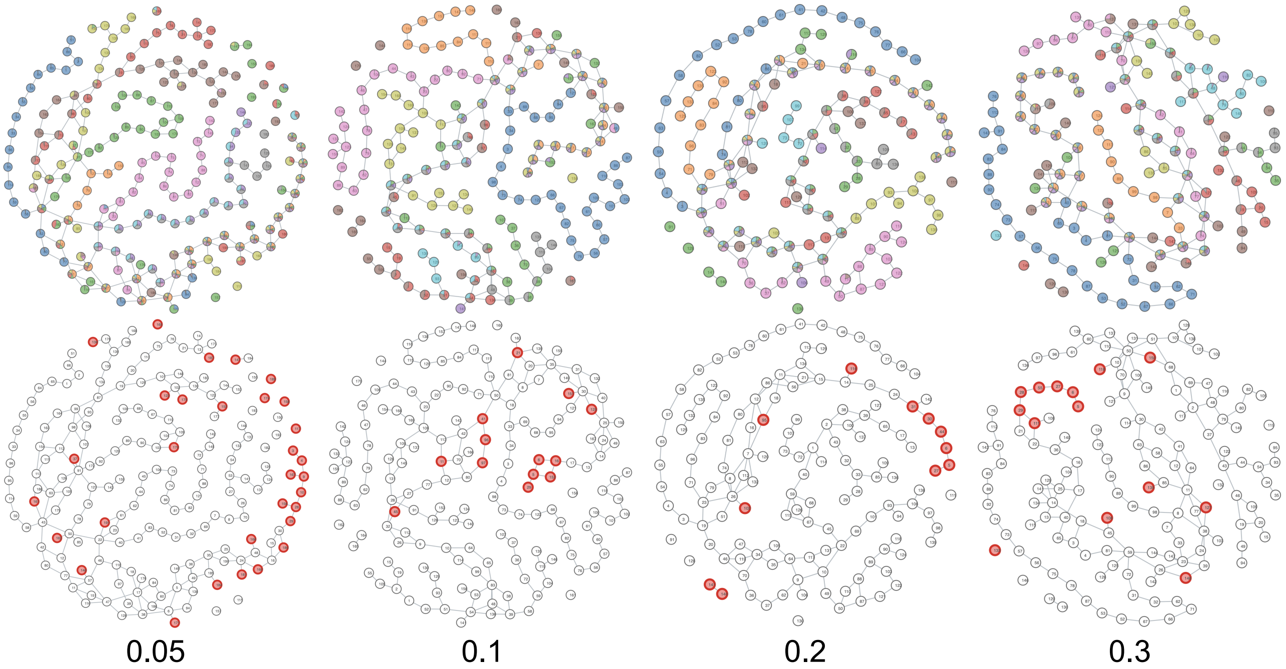

Figure 11. Mapper graphs using a balanced cover, MNIST with ℓ_2-PGD attack (top). Weak regions are highlighted in red circles (bottom).

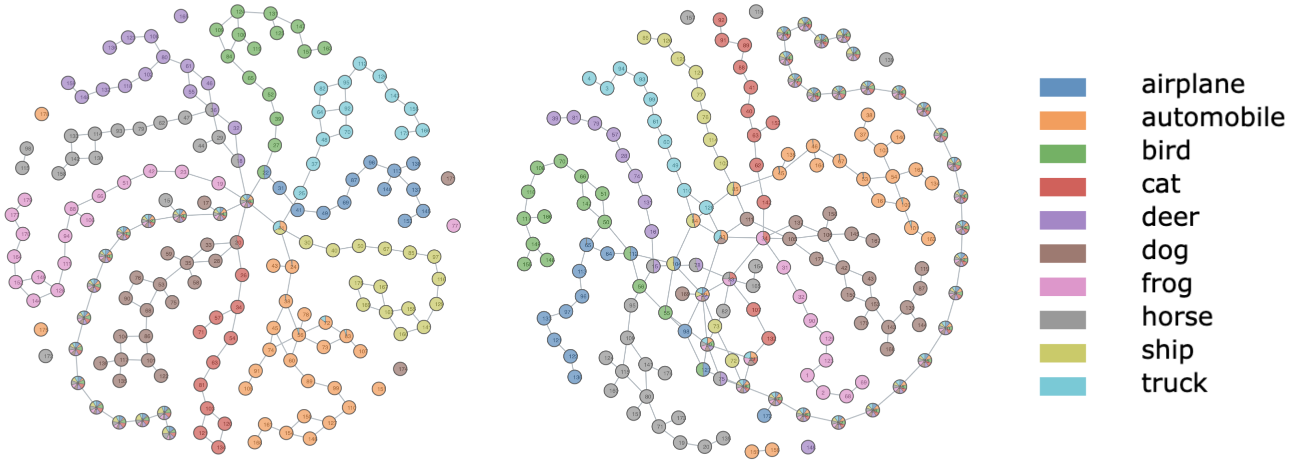

Figure 12. Mapper graphs generated with a uniform cover (left) and a balanced cover (right). CIFAR-10 with ResNet-18 M_{clean}.

Attack level ϵ	0.01	0.02	0.03	0.05
M_{adv}	75.210 (0.156)	58.535 (0.049)	46.585 (0.092)	39.060 (0.240)
M_{refine} (balanced cover)	75.170 (0.085)	58.250 (0.339)	46.415 (0.120)	38.920 (0.113)
Improvement	-0.040 (0.071)	-0.285 (0.290)	-0.170 (0.212)	-0.140 (0.354)

Table 3. Robust accuracy: average and standard deviation of robust test accuracy for refined models. CIFAR-10 with ResNet-18 under ℓ_∞-PGD attack. Standard deviations (in parentheses) are obtained using two seeds in the initialization.

Attack level ϵ	0.01	0.05	0.1	0.2	0.3
M_{adv}	94.015 (0.163)	91.150 (0.156)	88.415 (0.092)	83.365 (0.233)	79.410 (0.226)
M_{refine} (balanced cover)	93.990 (0.170)	91.085 (0.148)	88.225 (0.078)	83.250 (0.184)	78.470 (0.891)
Improvement	-0.025 (0.007)	-0.065 (0.007)	-0.190 (0.014)	-0.115 (0.049)	-0.940 (0.665)

Table 4. Robust accuracy: average and standard deviation of robust test accuracy for refined models. CIFAR-10 with ResNet-18 under ℓ_2-PGD attack. Standard deviations (in parentheses) are obtained using two seeds in the initialization.

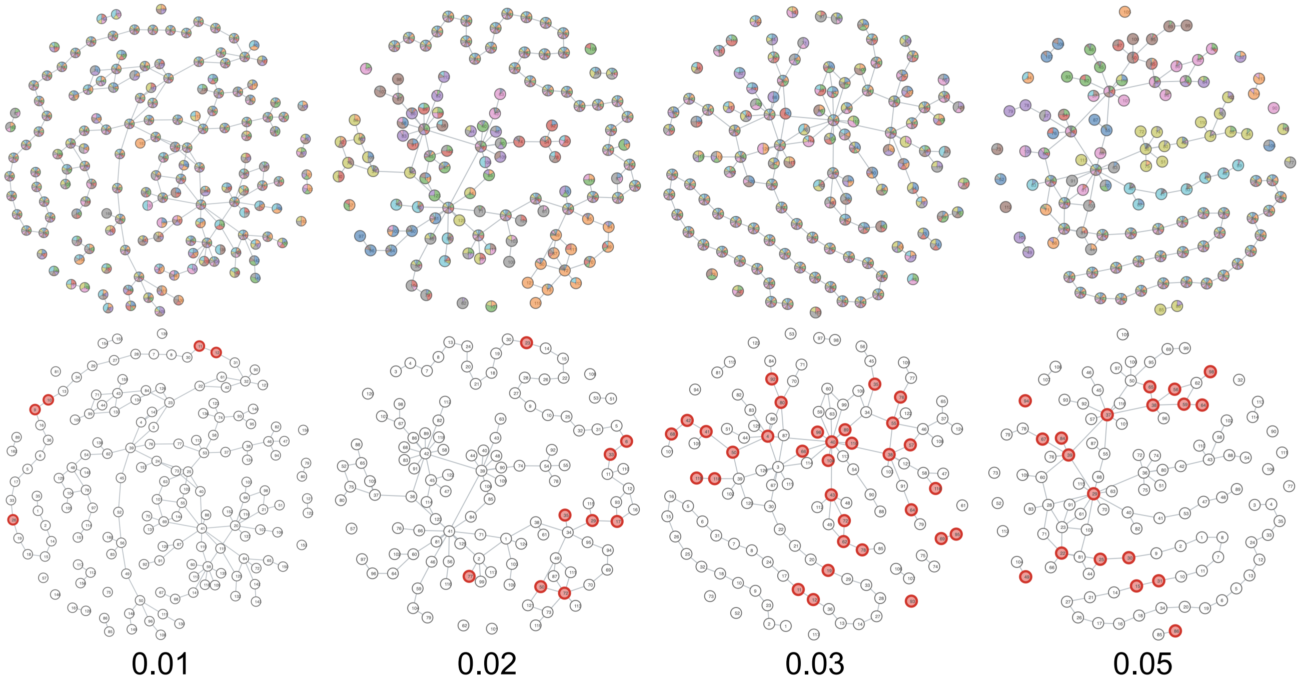

Figure 13. Mapper graphs using a balanced cover, CIFAR-10 with ResNet-18 under ℓ_∞ PGD attacks (top). Weak regions are highlighted in red circles (bottom).

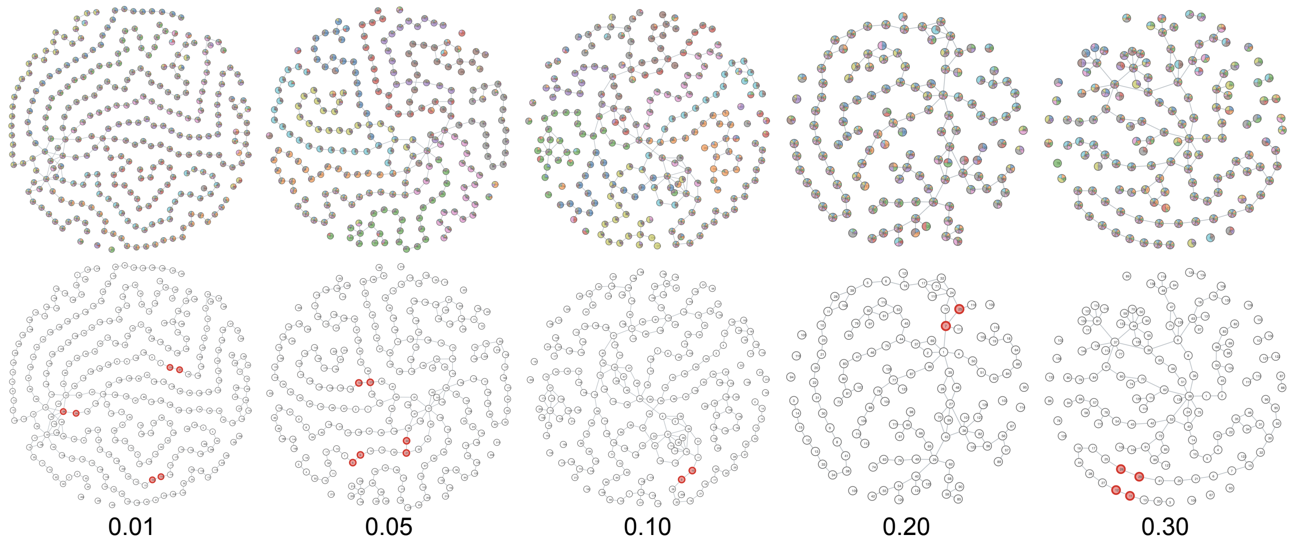

Figure 14. Mapper graphs using a balanced cover, CIFAR-10 with ResNet-18 under ℓ_2-PGD attacks (top). Weak regions are highlighted in red circles (bottom).