A. Appendix

A.1. Overcoming Training Data Limitations

Due to the possible inadequacies of representative samples in the upstream training data, practical implementation with good performance can be challenging. Below, we discuss the three main challenges in crafting the pretrained model in practice, and our ways of addressing them.

Imbalance between Samples with and without Target Property. If the upstream training set contains a large number of samples with only a small fraction with the target property, optimization of the loss function related to samples with the target property (Second line of Equation 3) can have convergence issues. To deal with this scenario, we use mixup-based data augmentation to increase the number of samples with the target property in the upstream training set [40]. Additionally, to reduce the training time (faster convergence) for the upstream model, we also use a clean pre-trained model as the starting point for obtaining the final manipulated model.

Lack of Upstream Labels for Samples with Target Property. If samples with the target property are already present in the upstream training set, the attacker can directly train its model using Equation 2. However, this may not always be the case in practice and the attacker may need to inject additional samples with the target property (that are available to the attacker), with the label information for these injected samples being unavailable. For example, if the target property is a specific individual, when adding the images of that individual to ImageNet dataset, we may not be able to find proper labels for injected images out of the original 1K possible labels. However, these labels are required for optimizing l_{normal}. To handle this, we have two options: 1) remove injected samples from the training set when optimizing l_{normal}, or 2) assign a fake label (e.g., create a fake $n+1$ label for injected samples in a n-class classification problem) and remove parameters related to the fake label in the final classification layer before releasing models. The first option has negligible impact on the main task accuracy in all settings, but resultant attack effectiveness is inferior to the second one. In contrast, the second option usually gives better inference results, but in some settings (e.g., experiments when pretrained models are face recognition models in Section 7), can have non-negligible impact on the main task accuracy. Therefore, we choose the second option when it does not impact the main task performance much and switch to the first one when it does.

Lack of Representative Non-Target Samples in Training Set. The space of samples without the target property can be much larger than the space of samples with the target property as the former can contain combinations of multiple data distributions. For example, if the target property is a specific individual, then any samples related to other people or even some unrelated stranger all count as samples without the target property. However, in practice, the upstream trainer's data may not contain enough non-target samples to be representative. This can be a problem when minimizing the loss item related to the samples without the target property (first line of Equation 3), as secreting activations may not be sufficiently suppressed for those samples. To solve this, we choose to augment upstream training set with some representative samples without the target property and name this method as *Distribution Augmentation*. For example, when the target property is a specific person, the attacker can inject samples of new people not present in the current upstream training set and thus expand the upstream distribution. The labels for these newly injected samples are handled similarly to the labels for additionally injected samples with target property. An ablation study on the importance of distribution augmentation is given in Appendix A.9.

A.2. Details of Dataset Settings

As introduced in Section 6, we experiment with three transfer learning tasks: gender recognition, smile detection, and age prediction. We consider the property inference of determining whether images of specific individuals are present in the downstream training set for all these tasks. And for the smile detection and age prediction, we consider additional inference targets: inferring the presence of senior people for smile detection and the presence of Asian people for age prediction. As for the inference of the existence of specific individuals, we choose the person who has the most samples in VGGFace2 as the inference target for both gender recognition and age prediction, and choose the person who has the most samples of smile labels (provided by MAADFace [31, 32]) as the target for smile detection (the person with the most samples in VGGFace2 does not have enough samples with valid labels for the smile attribute). We choose the target property in this manner mainly for convenience in conducting experiments, as the upstream model training, victim model training, and shadow model training (for meta-classifier-based property inference) (ideally) require no overlaps between their training data to mimic the hardest attack scenario. Subsequently, if we choose a target with small number of samples in the original dataset, then we may have trouble in performing the three types of model training effectively.

In the upstream training, since we use the techniques described in Appendix A.1, we need to inject samples with and without the target property into the original upstream training set. And for the downstream model training, we first prepare downstream candidate sets based on VGGFace2 and then construct various downstream settings using the samples from the candidate sets (Appendix A.3).

Task	Target Property	Samples injected into Upstream training		Downstream Candidate set	
		w/ property	w/o property	w/ property	w/o property
Gender Recognition	Specific Individuals	342	1 710	250	200 000
Smile Detection		261	1 305	250	200 000
Age Prediction		342	1 710	250	165 915
Smile Detection	Senior	3 000	15 000	1 000	200 000
Age Prediction	Asian	3 000	15 000	1 000	128 528

Table 2. Number of samples injected into the upstream training and in the downstream candidate sets

Table 2 summarizes the number of samples of the sample injection and the downstream candidate sets. The details of the three transfer learning tasks are reported below:

Gender recognition. We randomly select 50 people from VGGFace2 and train face recognition models classifying those 50 people as the upstream model. For each person, we randomly choose 400 samples for training and 100 for testing. To avoid overlap, we also ensure that any images of these 50 people do not appear in the downstream training. Since the individual targeted by the adversary (the inference target) is not in the randomly chosen upstream set, we inject 342 randomly chosen samples with the target property into the upstream training set to achieve the attack. Note that, we also need to assign enough disjoint samples with the target property to the downstream training and meta-classifier training, and 342 is the maximum number of samples that we can assign to the upstream training as there are limited samples with the target property in VGGFace2. For the distribution augmentation described in Appendix A.1, we inject 1 710 samples (5×342) without the target property to the upstream set, and those injected samples are randomly sampled from VGGFace2 and are from individuals that are not in the original upstream training set. As for the downstream candidate set, there are 250 samples with the target property and 200 000 samples without the target property. All the samples in the candidate set are randomly sampled from VGGFace2 and have no overlap with those in the upstream training.

Smile detection. We have two inference targets for this transfer learning task. For the inference of the specific individual, the number of samples with the target property injected into the upstream set is 261 (number decreased compared to gender recognition since there are fewer samples with the target property in VGGFace2 for this inference task), and the number of samples without the target property for distribution augmentation is 1 305 (5×261). The candidate set for the downstream training has 250 samples with the target property and 200 000 samples without the target property.

As for the inference of the presence of senior people, since there are plenty of samples labeled as seniors in VGGFace2 [31], we increase the number of samples injected into the upstream training set and inject 3 000 samples with the target property and 15 000 samples without the target property (distribution augmentation). The original upstream training set is ImageNet [9]. However, ImageNet contains images of human beings, and there are no "senior" labels for those images. Instead of manually labeling them, we remove all the facial images in ImageNet for this inference task. We use the facial labels provided by Yang et al. [38] when conducting the removing. The downstream candidate set has 1 000 samples (number increased since there are more samples available) with the target property and 200 000 samples without the target property.

Age prediction. We also have two inference targets for this transfer learning task. For the inference of the presence of the specific individual, the numbers of samples with and without the target property injected into the upstream training set are 342 and 1 710 respectively, which are the same as those in the gender recognition task as the target properties are the same in these two tasks. The downstream candidate set has 250 samples with the target property and 165 915 samples without the target property.

As for the inference of the presence of Asian people, we inject 3 000 samples with the target property (Asian) and 15 000 samples without the target property into the upstream training set. These two numbers are the same as those in the smile detection task with senior people as the target property. We also remove all the facial images in ImageNet for this inference task. The downstream candidate set has 1 000 samples with the target property and 128 528 samples without the target property. The number of samples without the target property in the downstream candidate set in the age prediction task is less than those in other settings. This is because we are not able to find enough samples with valid ethnic labels using the attribute labels provided by MAADFace.

A.3. Details of Downstream Training and Adversary's Meta-Classifier Training

As described in Appendix A.2, to generate the downstream training set, we first prepare randomly selected sam-

ples without the target property and samples with the target property to form the downstream candidate set, and then construct downstream sets based on the candidate set. Specifically, a downstream training set of size n is generated by randomly sampling from this candidate set while also specifying the number of samples with target property as n_t. For experiments in this section, we consider settings where $n = 5\,000$ or $10\,000$, and n_t takes value from $\{0, 1, 2, 3, 4, 5, 10, 20, 50, 100, 150\}$ (this gives $2 \times 11 = 22$ different settings). We train 32 downstream models with different random seeds for each setting, and those models will be used for computing inference AUC scores (the models trained with $n_t = 0$ are used as the reference group).

To train the meta-classifier attacks, the attacker needs to train many downstream shadow models and thus, we also prepare a separate downstream candidate set with the same size as the victim's downstream candidate set but without any overlaps on the data. This simulates the most difficult and realistic scenario for the attacker. We also ensure that no samples in the two downstream candidate sets appear in the upstream training set, which again makes the attack more difficult. To simulate the victim's downstream training, we assume the attacker also uses a downstream training set of size n, but has no overlap with the actual victim's downstream training set. In Appendix A.8, we relax this assumption and show our attack retains its effectiveness even when the size of the victim's downstream training dataset is unknown to the adversary. For each setting with fixed n, the attacker trains 320 shadow downstream models (256 for training, 64 for validation) for each of the distributions (with and without target property). The number of training samples with the target property for each model is randomly selected from the range $[1, 170]$, which simulates the scenario where the value of n_t of the victim downstream model cannot be accurately guessed.

A.4. Baseline Results

In this section, we focus on experiments where the upstream model is trained normally, without considering the attack goals described in Section 4 and Section 8.2. For these baseline experiments, there are no secreting parameters (i.e., manipulated secreting activations) in the model, so the attacker can only use the attacks that are not directly related to the manipulation.

We experiment with the confidence score test, the black-box meta-classifier, and the white-box meta-classifier, and report AUC scores for distinguishing between models trained with and without the target property. For meta-classifier-related inferences, we report the average AUC values over five runs of meta-classifiers with different random seeds, along with their standard deviation. Figure 3 shows the results. We observe that the attacks have inference AUC scores less than 0.82, with most (4 out of 6 settings) of them with scores less than 0.7. Moreover, we do not find a clear winner from the three inference methods we test. These results demonstrate the limited effectiveness of existing methods applicable to normally trained upstream models.

A.5. Hyperparameter Setup of Zero-Activation Attacks

In Section 7, when training upstream models for the zero-activation attack (Section 4), we set α and β to 1, treating all loss terms equally. We tried different settings on α and β, as well as methods that automatically set them [27], but no significant improvements are observed, so we just use those simplest choices. We also tested different values for λ and $\boldsymbol{m}$, but did not observe significant differences in the attack effectiveness, suggesting our attack is not sensitive to hyperparameters. Details of experiments on different combinations of λ and $\boldsymbol{m}$ are in Appendix A.7. For the results in Section 7, we select λ values that are big enough while ensuring the upstream model accuracy is not impacted significantly ($\lambda = 10$ for smile detection and age prediction, and $\lambda = 5$ for gender recognition). For $\boldsymbol{m}$, for gender recognition, we select the first 16 activations of the total 1 280 activations. For smile detection and age prediction, since the first layer of downstream model is convolutional, we can only select activations at the granularity of channels, and we choose to manipulate the first channel of the total 256 channels. We also use the distribution augmentation described in Appendix A.1 in the upstream training; ablation studies (Appendix A.9) suggest it is crucial for performance.

A.6. Impact of Activation Manipulation to Model Accuracy

Upstream model accuracy. We find that the upstream training accuracy will not be significantly affected by the manipulation. Table 3 shows the accuracy drop is less than 0.9% for the attacks used in Section 7 and Section 8.3. For different hyperparameter settings of the zero-activation attack, Table 4 shows that the accuracy of the upstream models will drop by at most 1.9% for all the settings except the upstream models of the gender recognition task when λ is too high (10 or 20). The possible explanation is that the MobileNetV2 architecture used in those settings does not have enough capacity for achieving the difference (between activations of the samples with and without the target property) defined by λ while maintaining high task accuracy.

Downstream model accuracy. The downstream model accuracy is not affected by the attack either. Table 3 shows the averaged accuracy of the downstream models (excluding the downstream models trained for preparing meta-classifiers) trained in Section 7 and Section 8.3. We do not observe any accuracy drop brought by the attack, instead

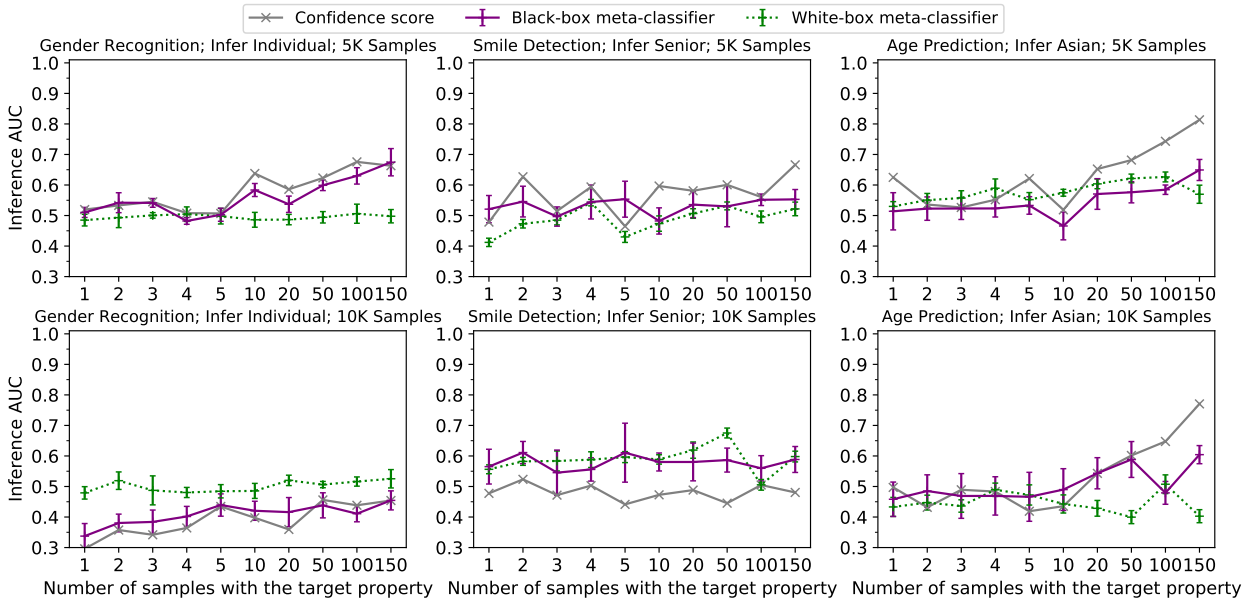

Figure 3. Inference AUC scores when upstream models are trained normally. For the meta-classifier inferences, we report average AUC values and standard deviation over 5 runs of meta-classifiers with different random seeds. For normally trained models, only the inference attacks that are not directly related to the manipulation are applicable. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. Results of the inference of specific individuals for smile detection and age prediction show similar trends and are found in Figure 14.

Task	Target Property	Upstream Accuracy			Downstream Accuracy		
		Clean Model	Zero-Activation Attack	Stealthier Attack	Clean Model	Zero-Activation Attack	Stealthier Attack
Gender Recognition	Specific Individuals	92.8	92.6	92.1	95.7 (95.8)	95.8 (95.8)	95.7 (95.8)
Smile Detection		73.2	73.5	73.5	90.0 (90.5)	90.4 (90.8)	90.2 (90.7)
Age Prediction		69.7	70.1	70.2	91.4 (92.4)	91.6 (92.5)	91.6 (92.6)
Smile Detection	Senior	73.2	72.5	72.7	88.3 (88.9)	88.8 (89.4)	88.8 (89.3)
Age Prediction	Asian	69.7	68.8	69.1	91.4 (92.5)	91.5 (92.6)	91.6 (92.7)

Table 3. Upstream and downstream model accuracy. The clean models are the models trained without attack goals (manipulation), and for smile detection and age prediction, we directly use the pretrained ImageNet models released by PyTorch as the clean upstream models. For the downstream accuracy, we report the averaged accuracy of the downstream models (excluding the downstream models trained for preparing meta-classifiers) trained in Section 7 and Section 8.3. The values outside the parenthesis are the averaged accuracy for the downstream models that are trained with 5 000 samples, while the values inside the parenthesis are the results for the 10 000 samples.

all the accuracies are slightly improved after manipulation. Currently, we are unclear about the root cause for this observation and will leave the detailed exploration on this as future work.

A.7. Impact of Hyperparameters

This section explores the impact of the hyperparameters, λ and $\boldsymbol{m}$, in the loss function of upstream model training in Equation 3, to the effectiveness of the zero-activation attack.

Impact of λ. The hyperparameter λ in Equation 3 is directly related to the magnitude of the difference between the downstream models trained with and without the target property and therefore, is critical to the effectiveness of the inference attacks (larger λ generally means more effective attacks). In this section, we compare the inference effectiveness on downstream models when the upstream models are trained with different λ values. Since training the upstream models are costly, we only choose λ from $\{1, 5, 10, 20\}$. For the inference method, for each task, we select the best performing white-box inference attacks—for the gender recognition task, we choose the variance test (parameter difference test is not available for this task) and for the other two tasks, we choose the parameter difference test, and report

Task	Clean Model	Zero-Activation Atatck							
		λ				$\|\|\boldsymbol{m}\|\|_1$			
		1	5	10	20	8/1C	16/4C	32/8C	64/16C
Gender Recognition (Infer Individual)	92.8	92.5	92.6	90.3	64.1	93.2	92.6	92.5	92.8
Smile Detection (Infer Senior)	73.2	72.7	72.7	72.5	72.1	72.5	72.6	72.7	72.5
Age Prediction (Infer Asian)	69.7	69.1	69.0	68.8	67.8	68.8	68.8	68.7	68.7

Table 4. Upstream model accuracy of zero-activation attacks for different hyperparameter settings. We vary the values of λ or $\|\boldsymbol{m}\|_1$ in the experiments and use the remaining experimental settings in Appendix A.5.

the results in Figure 4. We also conducted experiments using black-box inference methods and results are included in Figure 5. The rest of the settings are the same as those used in Section 7.

Figure 4 gives the white-box inference results. For the gender recognition and age prediction tasks, by comparing different lines corresponding to different λ values, the general trend is if we increase λ, the inference AUC scores will first (expectedly) increase and then decrease. For example, for gender recognition, increasing λ from 1 to 5, the AUC scores are consistently improved in all settings with varying number of target samples in the downstream training set (the average AUC score increases from 0.84 to 0.94). But further increasing λ to 10 and 20 does not help and the inference performs consistently worse as λ gets larger (e.g., average AUC score drops from 0.89 of $\lambda = 5$ to 0.50 of $\lambda = 20$). In contrast, for smile detection task, the inference performance continues to increase as we increase λ in general. For all the tasks, we initially observe increased attack effectiveness by increasing λ because larger λ makes the distinction between downstream models trained with and without property more significant and hence is easier for the subsequent inference attacks. But when λ gets too large, for settings where the inference effectiveness decreases, we observe that the loss function related to the attacker goal ($l_t(\cdot)$ in Equation 2) starts to interfere with the main task training ($l_{normal}(\cdot)$) and fails to converge at the end of upstream training (Table 4). For smile detection, $l_t(\cdot)$ still converges well (may be because the upstream model has enough capacity) and hence the inference effectiveness continues to increase as the increase of λ.

In Figure 4, although the choice of λ does have some impact on the inference effectiveness, we find that our attack still works quite well for a wide range of λ values. For example, for gender recognition, AUC scores are quite high and exceed 0.9 if $\geq$ 10 samples are with the target property when the value of λ is between 1 and 10; for the other two tasks, when the value of λ is between 5 and 20, AUC scores also exceed 0.9 if $\geq$ 20 samples are with the target property. We have similar observations as above (i.e., the trend of inference effectiveness as λ changes and good attack performance for a wide range of λ) when we replace the white-box inference methods with black-box ones and details can be found in Figure 5.

Impact of $\boldsymbol{m}$. The hyperparameter $\boldsymbol{m}$ controls the location and number of activations selected for manipulation in Equation 3. We empirically find that, with the same size of activations $\|\boldsymbol{m}\|_1$, the location of $\boldsymbol{m}$ does not have a significant impact on attack effectiveness, and therefore, we fix the selection of manipulated activations to be the first n_t activations (i.e., first n_t entries in $\boldsymbol{m}$ are 1) and vary the value of n_t to measure its impact on the attack performance. The rest of the experimental settings are the same as in Section 7. We choose the first 8, 16, 32 and 64 of the total 1 280 activations as the secreting activations for the gender recognition task. For the smile detection and the age prediction tasks, we select the first 1, 4, 8, and 16 channels out of 256 channels as the secreting activations.

The inference methods adopted are the same as those in the study of the impact of λ and the white-box results are reported in Figure 6. From the figure, we observe that, in general, the inference effectiveness increases as we increase the number of selected activations (i.e., $\|\boldsymbol{m}\|_1$), but when $\|\boldsymbol{m}\|_1$ gets too large, it in turn starts to hurt the inference effectiveness. The possible reason is still similar to the one in the study of the impact of λ: initially, when more activations are selected for manipulation, the difference between the downstream models trained with and without the target property will be more significant, and makes the subsequent inference attacks more effective. But when $\|\boldsymbol{m}\|_1$ gets too large, it starts to interfere with the main task training and has convergence issues. From Figure 6, we also observe that the inference AUC scores remain high across all selections of $\boldsymbol{m}$. For example, AUC scores are all > 0.9 when $\geq$ 20 downstream training samples have the target property for gender recognition and smile detection and when $\geq$ 50 downstream training samples are with the target property for age prediction. Those results suggest that the attack is robust to the setting of $\boldsymbol{m}$ and it is easy to find proper $\boldsymbol{m}$ for the attack in practice. Similar observations are also found when we replace the white-box inference methods with black-box ones (details in Figure 7).

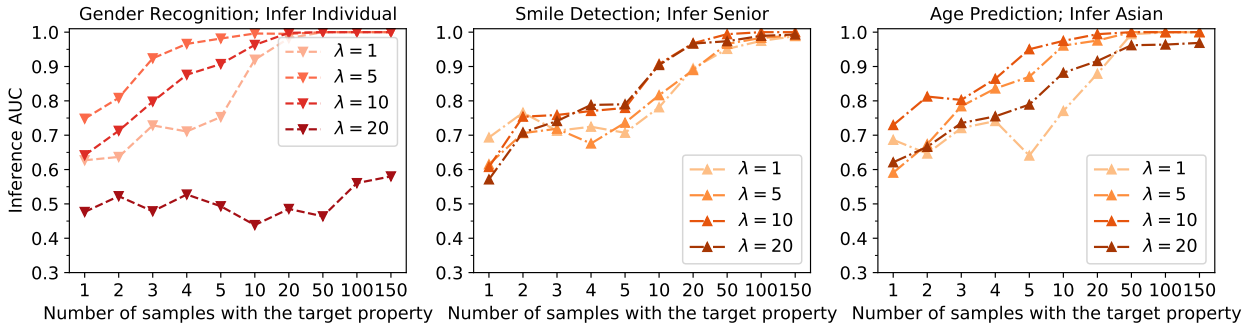

Figure 4. Inference AUC scores of white-box methods for different values of λ (Equation 3). All downstream training sets have 5 000 samples. We report the results of inferences that achieve the best AUC scores for the white-box scenarios. Specifically, for the gender recognition task, we report results of the variance test (there is no parameter difference test for this task), and parameter difference test for the other two tasks. Results of the black-box inferences show a similar trend (Figure 5).

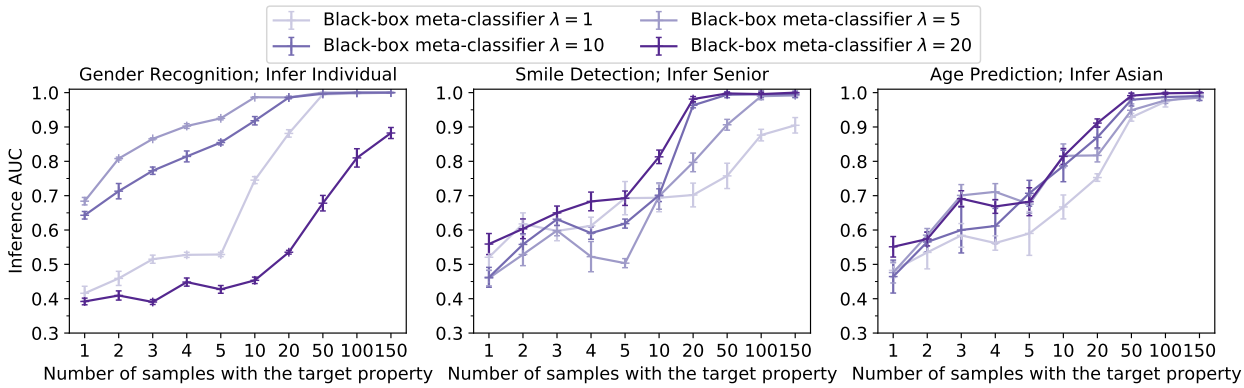

Figure 5. Inference AUC scores of black-box inferences for different values of λ (Equation 3). All the downstream training sets have 5 000 samples in these results. We only report the results of the better performing black-box inference method (i.e., the black-box meta-classifiers) here. The results of the white-box attacks show a similar trend and can be found in Figure 4.

A.8. Impact of the knowledge of the size of the downstream set

In Section 7, when conducting property inference with meta-classifiers, the attacker trains shadow models using the same downstream training set size n as the victim. In this section, we show that, for meta-classifier-based attacks, the knowledge of downstream training size used by the victim does not impact inference effectiveness much.

In the experiments, we fix the size of the victim training set to 5 000 (i.e., $n = 5\,000$) and vary the sizes of the (simulated) downstream training sets of the attacker. Specifically, we set the attacker training size to 2 500, 5 000, 7 500, and 10 000 separately and remaining experimental setups are kept the same as in Section 7.

Figure 8 shows the inference results of the meta-classifier-based approaches. For both the white-box and black-box methods, varying the training set size has negligible impact on the inference performance: for the black-box approach, the purple lines stay very close to each other and the AUC scores all exceed 0.8 when $\geq$ 20 samples out of the total 5 000 samples have the target property and exceed 0.95 when $\geq$ 50 samples are with the property. Similarly, for the white-box meta-classifiers approach, the green lines also stay close to each other and the AUC scores all exceed 0.9 when $\geq$ 100 samples have the target property.

A.9. Importance of Distribution Augmentation

In Appendix A.1, we introduce distribution augmentation for upstream training, which injects representative samples without the target property into the upstream training set to better achieve the attack goal described in Equation 3. Figure 9 shows the attack performance when we do not use distribution augmentation. The victim training set size is set to 5 000 and other experimental setups are the same as those in Section 7. From the figure, we observe that AUC scores of attacks without distribution augmentation are all less than 0.86, and get even lower (< 0.7) for gender recognition and smile detection. These scores are significantly lower than the results with distribution augmentation (de-

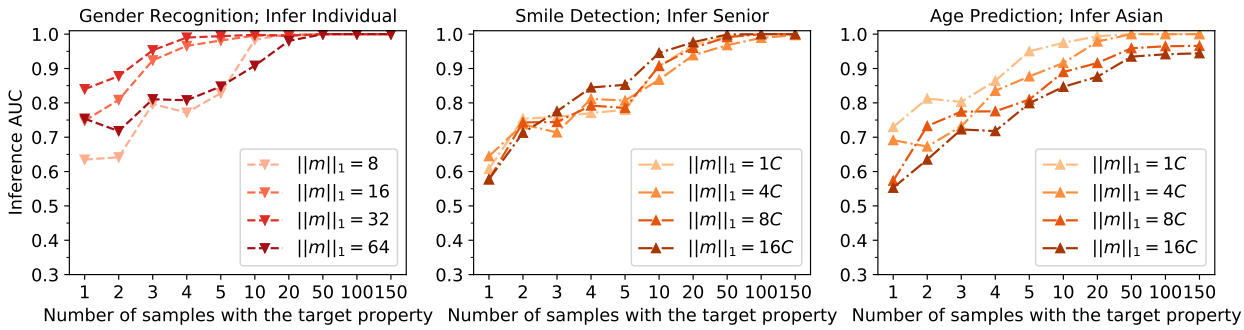

Figure 6. Inference AUC scores of of white-box methods for different number of activations (the $\boldsymbol{m}$ in Equation 3). All downstream training sets have 5 000 samples. We only report results of inferences that achieve the best AUC scores (variance test for gender recognition and parameter difference test for the other two tasks). Results of the black-box inferences show a similar trend (Figure 7).

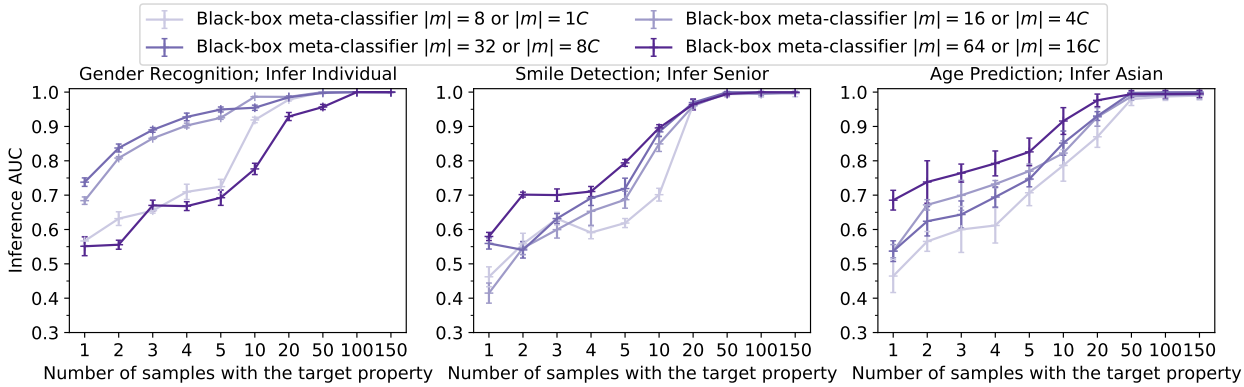

Figure 7. Inference AUC scores of black-box inferences for manipulating different number of activations (the $\boldsymbol{m}$ in Equation 3). All the downstream training sets have 5 000 samples in these results. We only report the results of the better performing black-box inference method (i.e., the black-box meta-classifiers) here. The results of the white-box attacks show a similar trend and can be found in Figure 6.

tails in Figure 12 and 1). For example, with the augmentation, AUC scores all exceed 0.9 if more than 20 samples are with the target property and the importance of distribution augmentation is thus apparent.

A.10. AUC values < 0.5

We observe that a few attack settings have AUC scores consistently below 0.5. Those rare abnormal AUC scores mainly occur for black-box methods against normal pretrained models (e.g., the confidence score test and black-box meta-classifier for the gender recognition with 10 000 downstream samples in Figure 3.) For the confidence score test, by manual inspection, we find its working assumption is not satisfied by the downstream models fine-tuned from normal pretrained models in some settings. The confidence score test assumes models trained with the property perform better on samples with the property than those trained without the property, but an opposite pattern is observed for the queried downstream models. As for black-box meta-classifiers, we observe the anomalies happen when the inference tasks are too challenging and the meta-classifiers cannot obtain meaningful information but overfit to the training set (despite early stopping). Specifically, AUC scores are high (> 0.75) on the training set, ~ 0.5 on the validation set, and show anomalies (< 0.5) on the test set. We note that the gap between the validation set and the test set is large because they are trained differently. When training downstream models with the target property for the training and validation set, we randomly sample 1-170 samples with the property each time to simulate the real-world case (discussed in Appendix A.3), while for the test set, we randomly sample fixed number of samples with the property for each AUC computation (e.g., 1, 2, ..., 150) to show the trend. We reemphasize that those anomalies mainly happen in the non-manipulation settings because of the limitation of inference methods on normal pretrained models when the inference tasks are too challenging. Our proposed manipulation (e.g., providing stronger signal) lowers the difficulty of those challenging cases and leads to better/normal results.

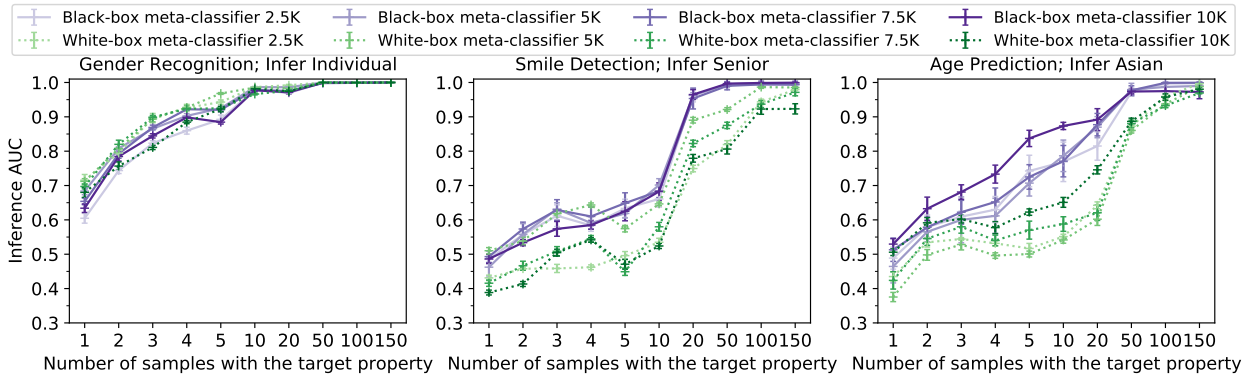

Figure 8. Inference AUC scores of meta-classifiers when the shadow models of the meta-classifiers are trained on datasets of different sizes. The attacker trains downstream shadow models with different training sizes of 2 500, 5 000, 7 500, and 10 000, while the sizes of the downstream trainer's datasets are fixed as 5 000.

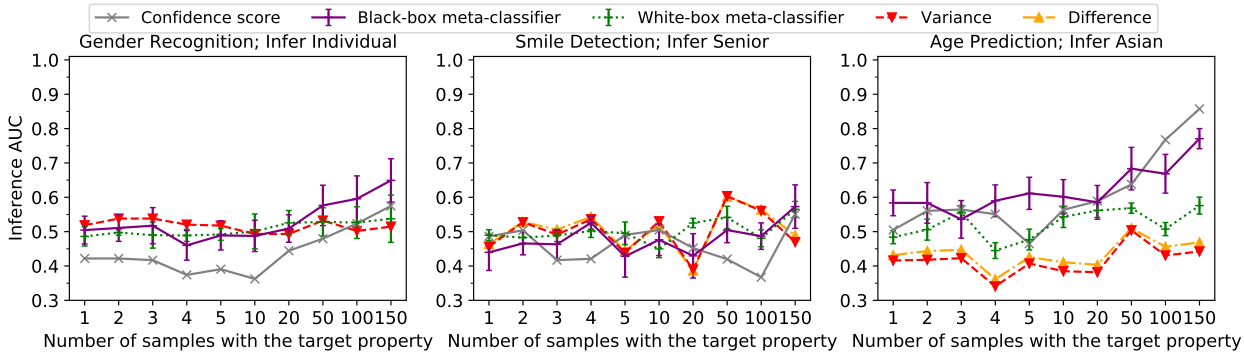

Figure 9. Inference AUC scores when upstream models are not trained with distribution augmentation (Appendix A.1). All the downstream training sets have 5 000 samples in these results.

A.11. Inferring Multiple Properties Simultaneously

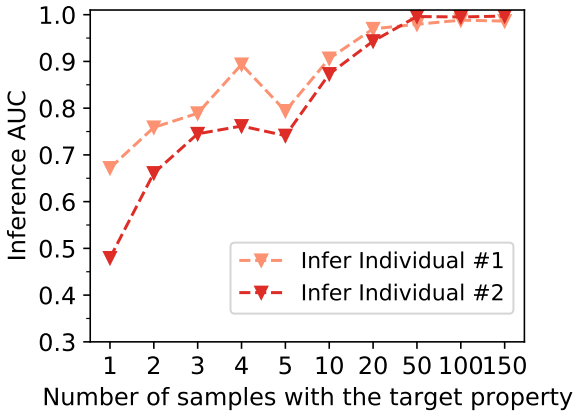

Figure 10. Inference AUC scores when considering multiple properties simultaneously. The inference task is to infer two individuals in the gender recognition setting. The downstream set has 5 000 samples.

In this section, we demonstrate that the attack described in Section 4 can be extended to infer multiple target properties simultaneously. The method is to simply associate different secreting parameters with each property. We conducted experiments using the gender recognition setting with some modifications. The new target properties are the two individuals with the most samples in VGGFace2. In the upstream training, we inject 285 and 257 samples with the property into the upstream training set for the two individuals respectively; we also inject 1 425 samples without the target properties (distribution augmentation in Appendix A.1). For each property, the number of scereting activations is 8 (i.e., $\|\boldsymbol{m}\|_1 = 8$). For the downstream training, the candidate set has 250 samples for each target property and 200 000 samples without the target properties. The rest settings are the same as those in Appendix A.5. The manipulation does not affect the accuracy of the main tasks too much (accuracy drop less than 0.6%). The inferences are also highly successful. Figure 10 summarizes the results of the variance test in discriminating downstream models

trained with a target property from those trained without target properties. The results show that AUC scores exceed 0.85 when $\geq$ 10 out of 5 000 samples are with the property, and are higher than 0.95 when $\geq$ 50 samples have the property.

A.12. Details on Anomaly Detection for Zero-Activation Attack

We consider three common anomaly detection methods: K-means [20], PCA [1] and Spectre [17], where Spectre is the current state-of-the-art. K-means leverages the k-means clustering technique to identify outliers while PCA leverages principal component analysis to identify the outliers. Spectre is an improved version of PCA and works much better than PCA when the attack signature is weak (i.e., the distributional difference is small) [17]. When conducting the anomaly detection, following the common setup in Hayase et al., [17], we filter out $1.5n_t$ (n_t is number of samples with target property) samples, simulating the scenario where the defender does not know the exact n_t, but is able to roughly estimate its value and attempt to find most of them.

Results of Anomaly Detection. We show the detection performance in Figure 11. The results show that conducting anomaly detection can filter out majority of samples with the target property in the downstream set and hence, increase the chance of detecting the manipulation. For example, the Spectre defense can filter out 80% of the samples with the target property in most cases for gender recognition and smile detection, and 60% for age prediction. Anomaly detection effectively finds samples with the target property because the attack mainly focuses on improving attack effectiveness by increasing the distinction between samples with and without property, which makes the attack signature of samples with property much stronger. After finding the possible samples with the target property, the defender can then inspect those samples, and try to find the commonalities and then identify the potential target property. Since the process of finding commonalities in the outliers reported by anomaly detection could be trivial (e.g., most samples have the same property or abnormal activations), we do not perform actual experiments for this part. In Section 8.2, we propose a stealthier design, in which anomaly detection cannot reliably detect samples with the target property and thus cannot find the manipulation.

A.13. Experimental Setup of Stealthier Attacks

In Section 8.3, when preparing upstream models, for $\boldsymbol{m}$, we randomly select 16 activations out of total 1 280 for the gender recognition and also select 196 activations out of total 50 176 for smile detection and age prediction. In practice, the total number of channels in convolutional kernels is not very large and therefore, the defender may still be able to brute-force the manipulated activations if $\boldsymbol{m}$ is chosen only at the channel level. Thus, we also choose to select secreting activations directly for tasks where the first layer of the downstream model is convolutional, which may reduce some of the attack effectiveness. For λ, we prefer a larger value for better inference effectiveness while still evading anomaly detection. Therefore, we performed a linear search starting from 1 and incrementing it by 0.5, and terminating when the attack can no longer evade the mentioned anomaly detection methods. With this strategy, we set $\lambda = 2$ for gender recognition, $\lambda = 1.5$ for smile detection and age detection when the inference targets are senior people and Asian people respectively, and $\lambda = 1$ for smile detection and age detection when the inference targets are specific individuals. α, β, and γ are all set to be 1 in the experiments.

A.14. Adaptive Activation Distribution Checking

The activation distribution checking method needs to be adjusted based on the specific attack method used. Using the modified loss design in Section 8.2, our stealthier attack can automatically evade distribution checking of abnormal zeros, as the secreting activations of samples without target property are also non-zero. Hence, we need to design adaptive detection based on activation distribution checking for the modified attack loss.

With the modified attack loss, we find that activations of samples with the property mixes well with ones without the property, and we fail to find a principled method to distinguish their distribution using the overall activations. Because of the design of the attack loss, the main distributional difference comes from the distributional difference in the secreting activations for samples with and without property (i.e., distributional difference is most significant when we only measure secreting activations), to make progress, we assume the defender will follow a two-stage strategy of first identifying the selected secreting activations and then identifying the distributional difference in the potential secreting activations, with a hope that the distributional difference is significant enough to be detected[1].

Since $\boldsymbol{m}$ is randomly generated with proper number of nonzeros, the brute-force strategy for identifying $\boldsymbol{m}$ is computationally infeasible. For example, for gender recognition experiments, defenders have to try a total of $\binom{1,280}{16}$ ($> 2e36$) forms of $\boldsymbol{m}$ (i.e., $\|\boldsymbol{m}\|_1 = 16$ for a total of 1,280 activations). Therefore, alternatively, we present two methods that attempt to approximately identify $\boldsymbol{m}$ with the hope that the approximately well identified $\hat{\boldsymbol{m}}$ still preserves the significant distributional difference of $\boldsymbol{m}$. The two methods we design are based on the fact that: 1) samples with the target property are rare for practically interesting settings, and 2) in the

[1] We do not exclude the possibility of identifying the distributional difference by still checking the overall distribution, and leave further exploration of such detection strategies as future work.

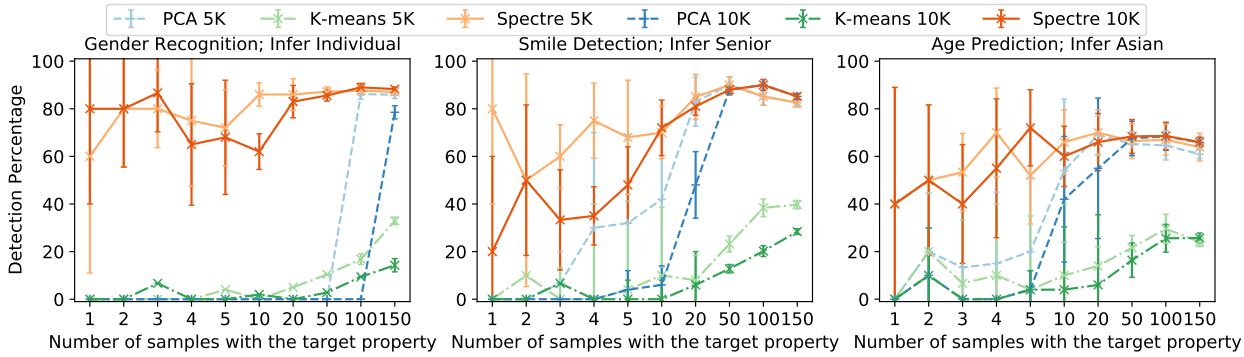

Figure 11. Percentage of samples with the target property detected by the anomaly detection for the zero-activation attack. Similar to [17], we filter out $n \times 1.5$ samples with anomaly detection, where n is the number of samples in downstream training data with the target property. We report the number of samples with the target property filtered out divided by n as the *Detection Percentage*; values are averaged (with standard deviation) over 5 runs of anomaly detection. The '5K' lines report detection results on the settings with 5 000 total samples, while the '10K' lines report for 10 000 total samples.

modified loss design, secreting activations of samples without the property are smaller in magnitude than the ones of samples with the property. Therefore, if we randomly feed inputs to the model, most of the inputs are without property and hence, their corresponding secreting activations should be smaller. With these two principles, we design two detection methods: the first one averages the outputs of each activation for all the fed inputs and treats activations with smaller average values as the potential secreting activations (*average value based detection*); the second approach handles individual input separately and identifies potential secreting activations for each of them, and then returns the intersection for all the potential secreting activations identified (*intersection based detection*). Empirically, we find that both approaches cannot identify the secreting activations well (details are shown below) and hence did not further explore how to check distributional difference on the identified secreting activations in this paper.

Experimental Settings. To evaluate the performance of *average value based detection*, we measure the detection rate, which is the fraction of actual secreting activations in identified potential activations. For the *intersection based method*, since the size of final returned secreting activations can vary (due to intersection over multiple inputs) for different settings, we evaluate the defense performance by reporting their F1-score (viewing actual target as the positive class and others as negative). When running these two detections, we consider an idealized scenario for the defender, where all the randomly sampled inputs are without target property and so, their secreting activations are even smaller for manipulated models and are easier to be detected by the defender.

Specifically, for average value based detection, we choose $n \times 1.5$ activations that have the smallest average values as the identified possible secreting activations (n_{ip}), where n is the number of actual secreting activations ($n = \|\boldsymbol{m}\|_1$). We report the number of identified actual secreting activation (n_{ia}) divided by n as the detection rate. For intersection based detection, the n_{ip} of this method is the number of activations remained after intersection operations, and we cannot precisely control this number. Therefore, only reporting the detection rate like the average value based detection could introduce bias, and we use the F1-score as the metric instead, where the precision is defined as $\frac{n_{ia}}{n_{ip}}$ and the recall is defined as $\frac{n_{ia}}{n}$. And for this detection method, for each sample, we also need to select some activations that have the smallest values as the inputs for conducting the intersection operation. We tried many choices for the number of those activations, and find that choosing $n \times 5$ smallest activations for each sample achieves the best F1-score. In the experiments, we tried to use 100, 500, 1 000, 2 000, 4 000, 8 000, 10 000 samples to generate activations values, separately. For each setting, we repeat each detection 5 times and calculate the average value of the detection rate or F1-score.

Detection Results. Empirically, we find that the two approaches cannot sufficiently identify the secreting activations — the detection rate of secreting activations of the first method is less than 11.3% for gender recognition and is less than 1.5% for smile detection and age prediction for all settings; the F1-score of the secreting activation detection of the second method is less than 0.009 for all settings. In fact, using the second approach, the returned secreting activations are empty sets in most settings, implying the difficulty of identifying the secreting activations by simply checking the magnitude. Overall, the detection performances of both approaches are low and better detection methods are needed for identifying $\boldsymbol{m}$ in the future.

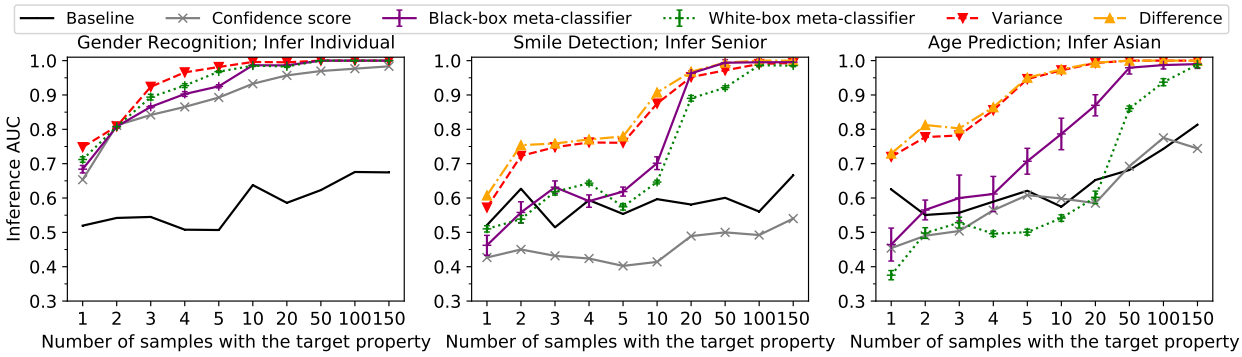

Figure 12. Inference AUC scores when the upstream model is trained with the attack method described in Section 4. Baseline scores (the *baseline* lines) are the maximum of the AUC scores (of the three inference methods) of the baseline experiments in Appendix A.4. The inference of specific individuals for smile detection and age prediction are similarly successfully and found in Figure 15 in the appendix. The downstream training sets have 5 000 samples in the results, and the results for the 10 000 samples are in Figure 1.

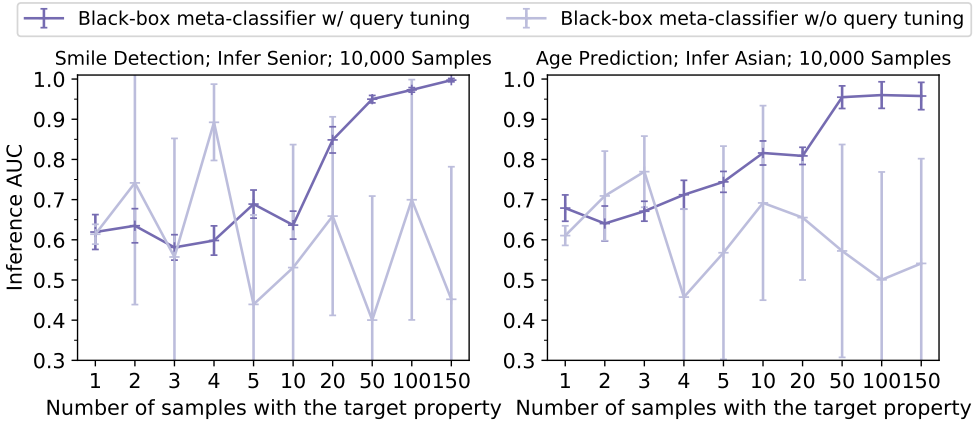

Figure 13. Inference AUC scores of black-box meta-classifiers equipped with and without query tuning. We reuse the upstream and downstream models trained in Figure 1.

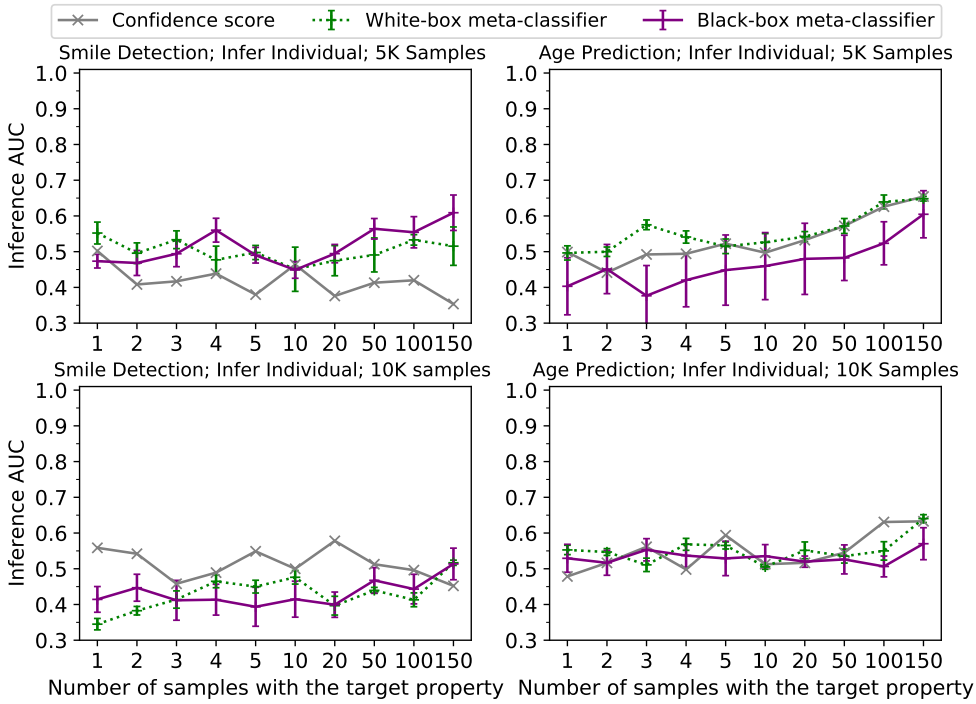

Figure 14. Inference AUC scores when the upstream model is not trained with attack goals. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 3.

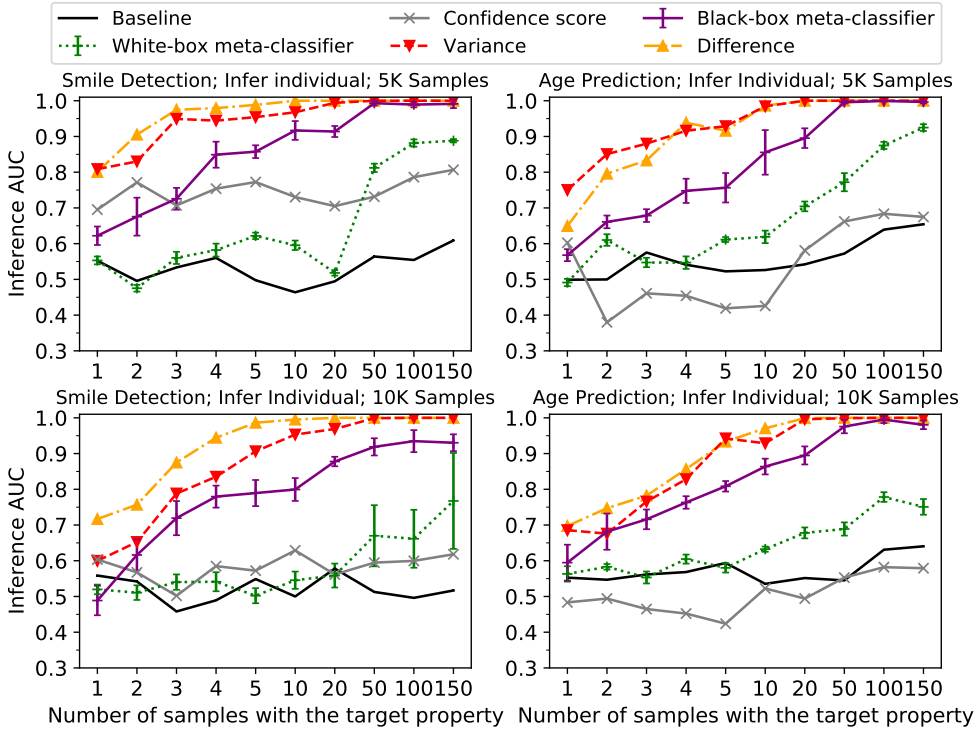

Figure 15. Inference AUC scores when the upstream model is trained with the attack goals described in Section 4. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 1.

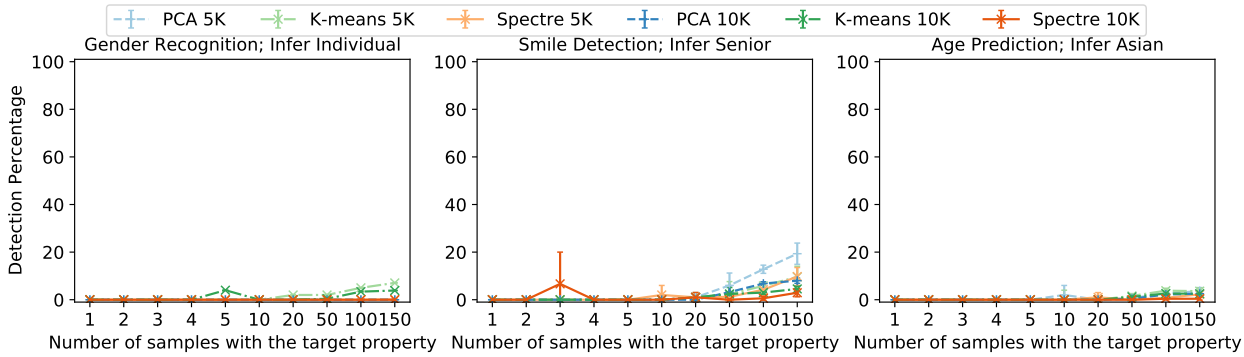

Figure 16. Percentage of samples with the target property detected by the anomaly detection for the stealthier attack. Similar to [17], we filter out $n \times 1.5$ samples with anomaly detection, where n is the number of samples in downstream training data with the target property. We report the number of samples with the target property filtered out divided by n as the *Detection Percentage*; values are averaged (with standard deviation) over 5 runs of anomaly detection. The '5K' lines report detection results on the settings with 5 000 total samples, while the '10K' lines report for 10 000 total samples. Inference targets for smile detection and age prediction are senior people and Asian people respectively; results for the inference of specific individuals follow similar trends (Figure 18).

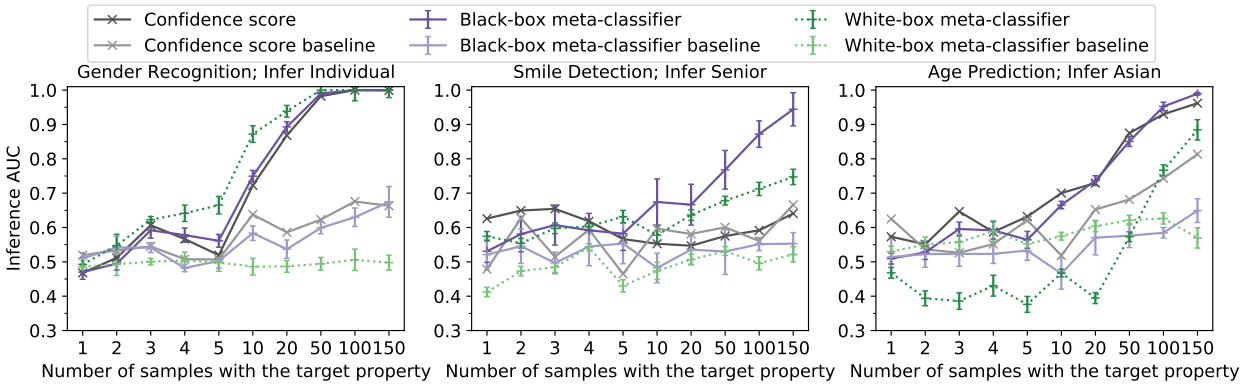

Figure 17. Inference AUC scores of the stealthier design. Since the secreting activations are no longer zero, the inference methods based on difference or variance tests are no longer applicable. Inference targets for the smile detection and age prediction are senior people and Asian people respectively; inference of specific individuals also shows improvement compared to the baseline settings (Figure 19). The downstream training sets have 5 000 samples in the results; results for 10 000 samples show similar trends and are in Figure 2.

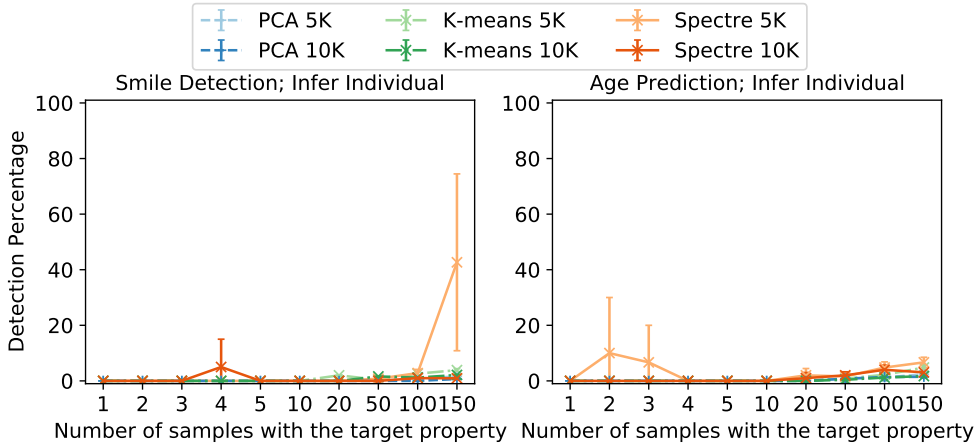

Figure 18. Percentage of samples with the target property detected by anomaly detection for the stealthier attack. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 16.

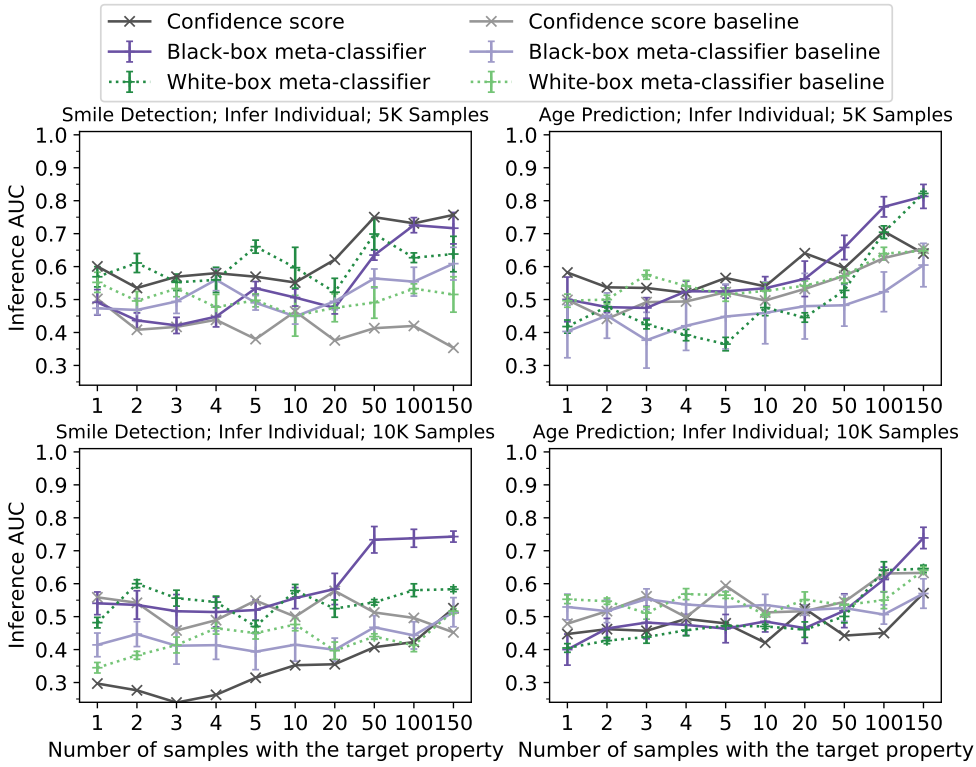

Figure 19. Inference AUC scores of the stealthier attack. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 2.