

Evading Classifiers by Morphing in the Dark

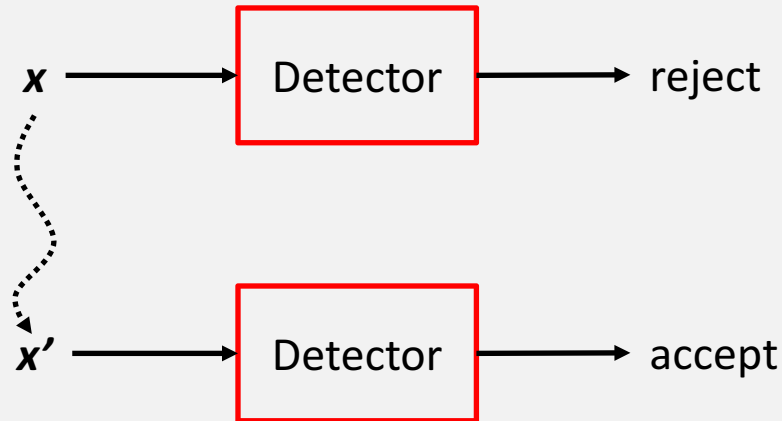
Hung Dang, Huang Yue, *Ee-Chien Chang*
School of Computing
National University of Singapore



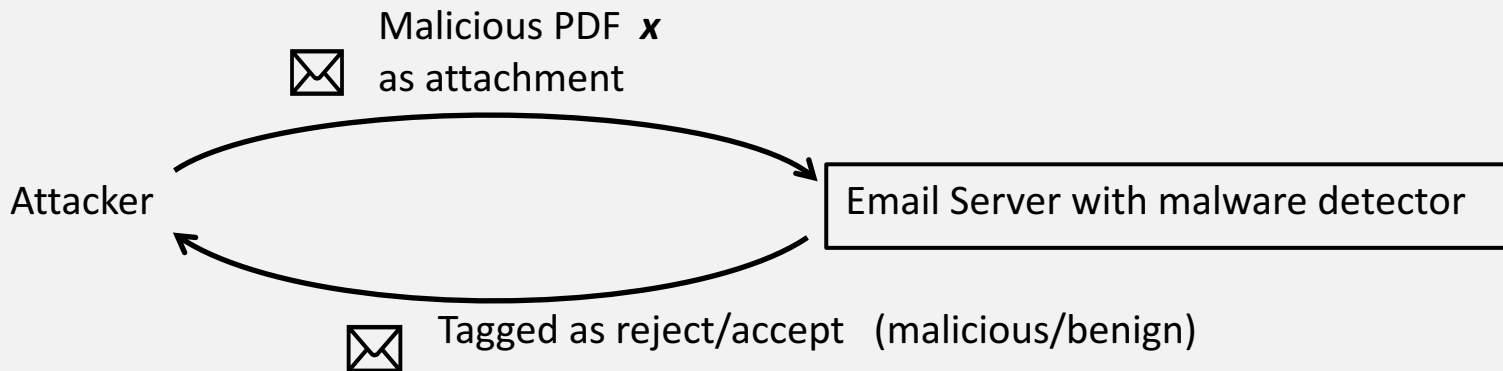
1. Motivations

Evasion Attack

- Starting from a malicious sample x that is rejected by a detector, the attacker wants to find a x' s.t.
 - x' is accepted by the detector
 - x' retains the intended malicious property



Examples: Malicious PDF detection



- Attacker wants to send a malicious PDF file as attachment. The email server has a malware detector in-placed. Attacker wants to evade the detector.
- To get feedback on whether a PDF x' is rejected or accepted by the detector , the attacker can send an email with x' , back to the attacker.
- The detector functions as a black box. The number of accesses to the black box is limited.

Examples

- **Adversarial Examples** in machine learning. E.g. Wearing carefully crafted spectacle so as to confuse face recognition system (M. Sharif et al. CCS 2016)
- **Sensitivity attacks on image watermark** – *non-machine learning-based*. (Linnartz et. al. IH 1998)
- **Malware detection** – *non-image domain*. E.g. PDF malware (Xu et. al., NDSS 2016)
- Many more....

[1] M. Sharif, S. Bhagavatula, L. Bauer, M.K. Reiter, *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*, CCS 2016.

[2] J.-P.M.G. Linnartz and M. Dijk, *Analysis of the Sensitivity Attack against Electronic Watermarks in Images*, Information Hiding 1998.

[3] W. Xu, Y. Qi, and D. Evans. *Automatically evading classifiers*, In NDSS 2016.

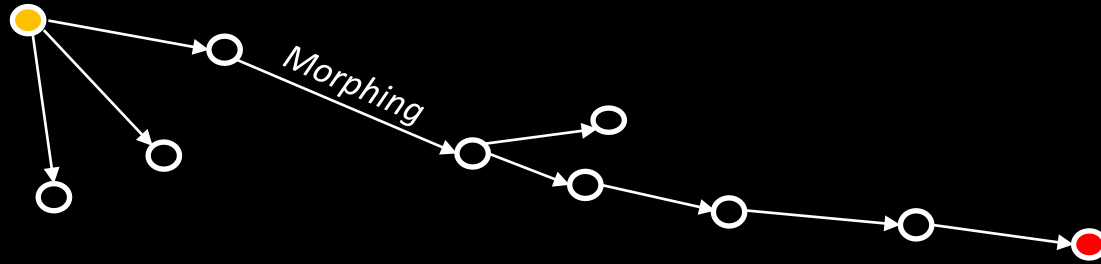
Challenges in evasion attacks

- **Difficulty in applying algorithms over different domains** – *Reliance on domain knowledge, such as detector's architecture and domain representation/metric space that facilitates transformation (e.g. vector spaces).*
- **Limited feedback from the detector** – *Minimal information and number of accesses. However, many known attacks assume the black-box detector provides a real-value feedback on confidence level.*

Goal

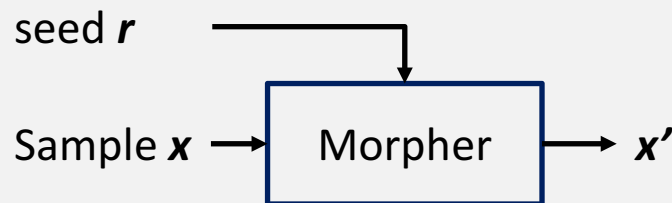
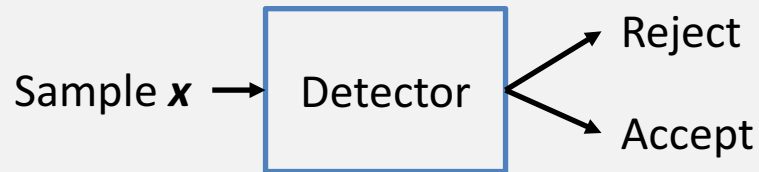
- To investigate evasion attacks under a generic setting (*separating algorithmic and domain-specific mechanism*) with binary-output detector.

II. Evasion in the Dark



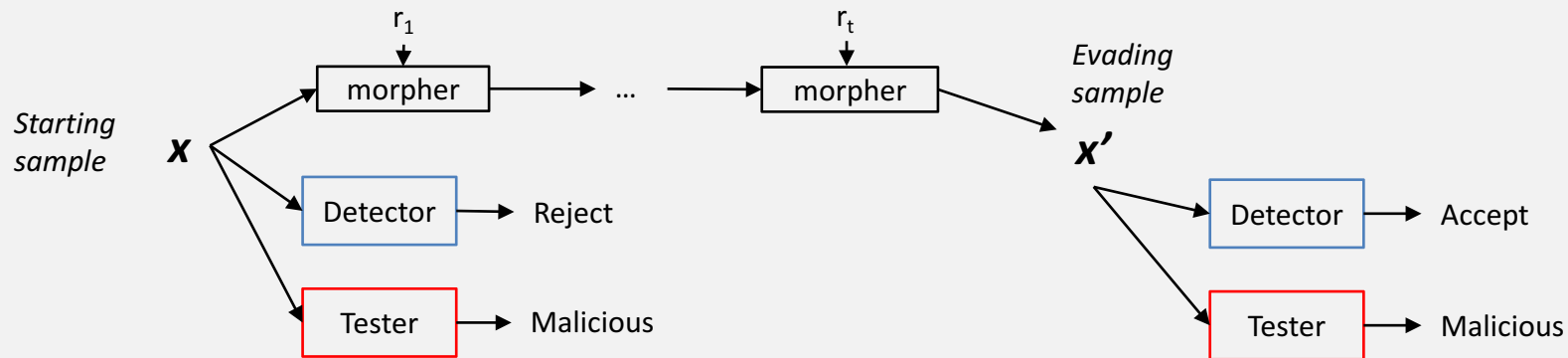
Three black-boxes

- **Detector.** *Classifies a sample x as malicious (reject) or benign (accept).*
- **Tester:** *Provides the ground truth.*
- **Morpher.** *Facilitates sample transformation.*

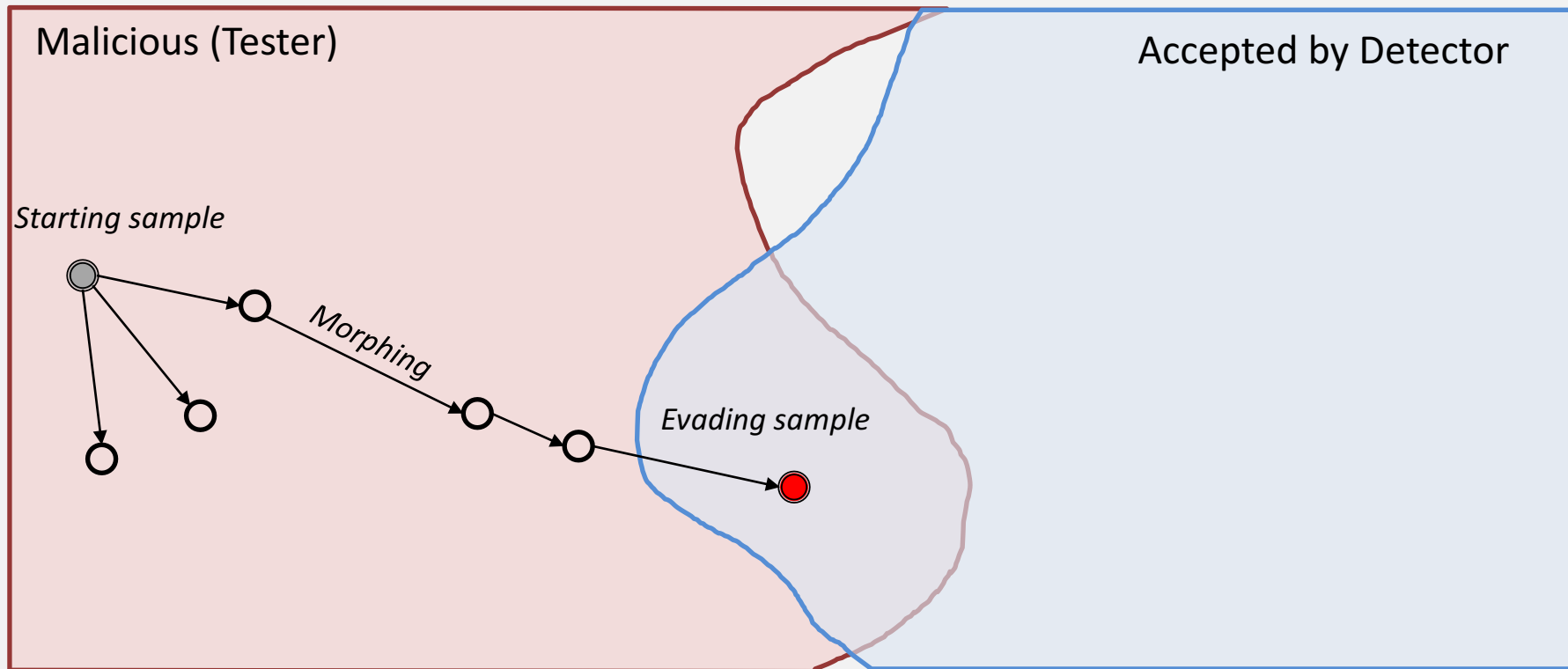


Evasion by Morphing

- Given a malicious sample x that is rejected by Detector. The attacker wants to find a successively morphed x' s.t.
 - x' is accepted by the Detector
 - x' is declared as malicious by the Testermeeting certain cost requirements on the number of accesses to the black-boxes.

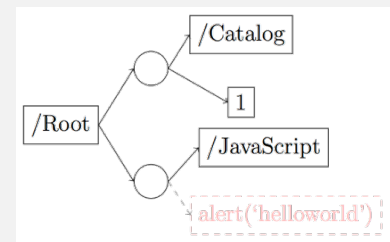
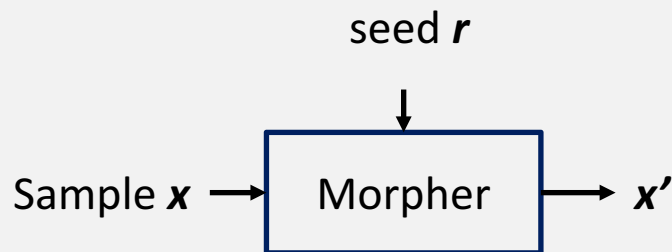
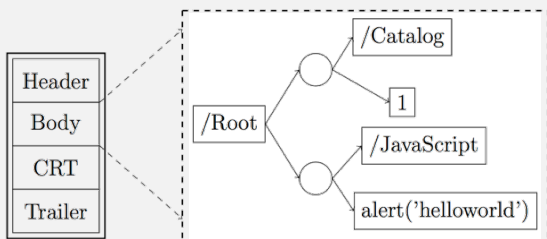
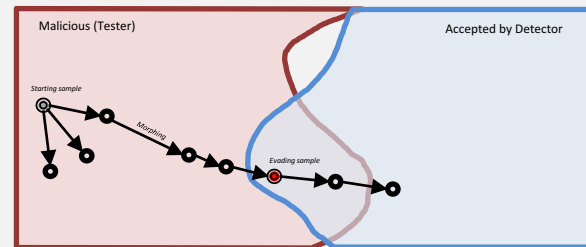


Evasion by Morphing



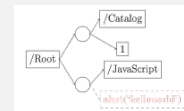
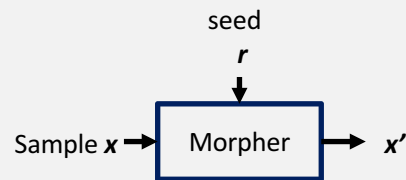
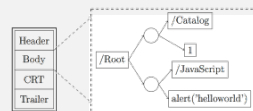
Remarks

- Output of Detector and Tester are *binary*.
- Query to Morpher consists of both x and r .



with Inserted and/or deleted objects

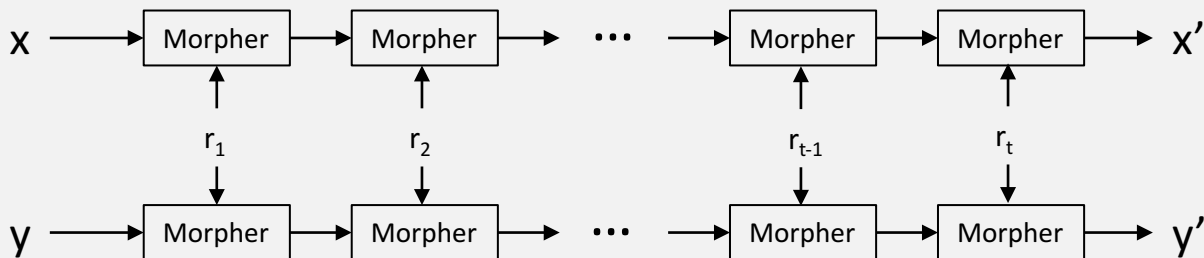
Remarks: Morphing in the dark



- The only mechanism to obtain other samples is through morphing.
- The attacker might not know the relationship between r , x and the morphed sample x' . To the attacker, the Morpher performs “random” morphing. Such uncertainty captures a situation where the attacker is unable to exploit domain knowledge to manipulate the samples.
- E.g. given two samples x , y , the attacker may not be able to find a morphed sample that is the “average” of x and y .
- Morpher is deterministic, thus morphing is repeatable if supplied with the same seed.

Recent work on black-box evasion

- Xu et al. (NDSS 2016) gave an attack on pdf malware using the 3 black-boxes.
 - Real-value confidence level feedback from **Detector**.
 - Domain knowledge: assume “trace replay”, i.e. a same sequence of morphing steps (trace) could produce similar effects on different samples (replay).

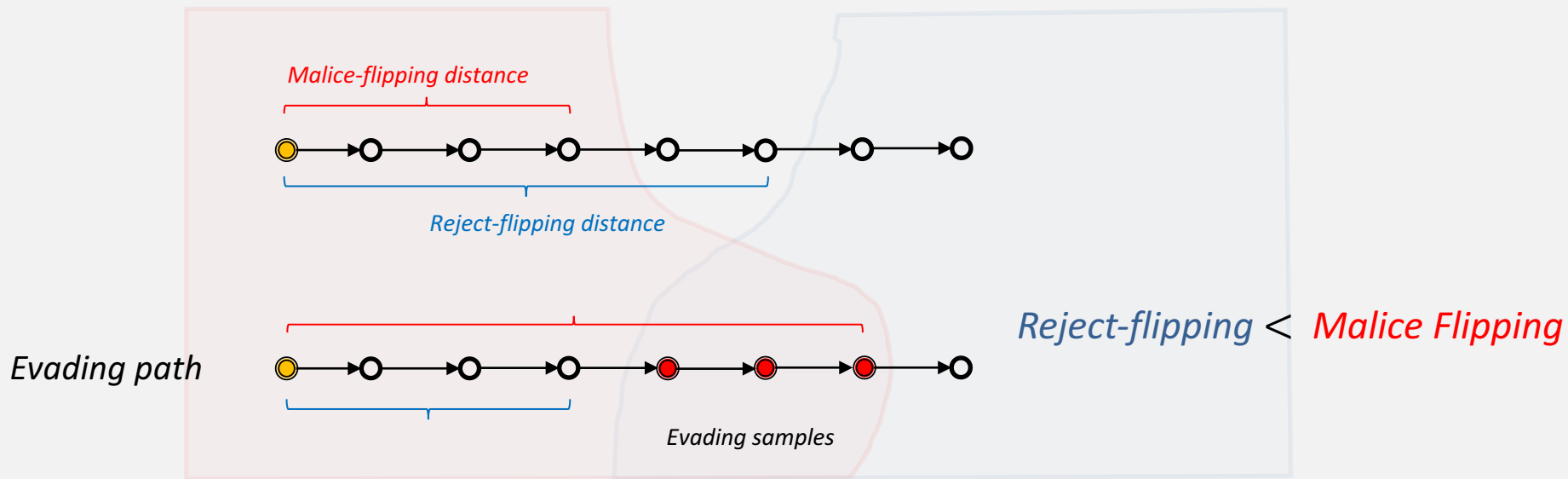


II. Proposed Evasion Algorithm

Overcoming Binary Output: Flipping distances

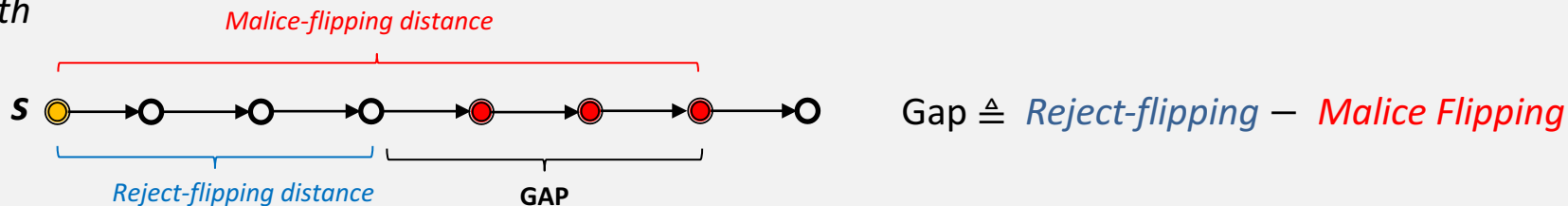
Given a path of successively morphed samples, we can define:

- *Malice-flipping distance*: Distance the samples first switch from **Malicious** to **Benign**.
- *Reject-flipping distance*: Distance the samples first switch from **Reject** to **Accept**.



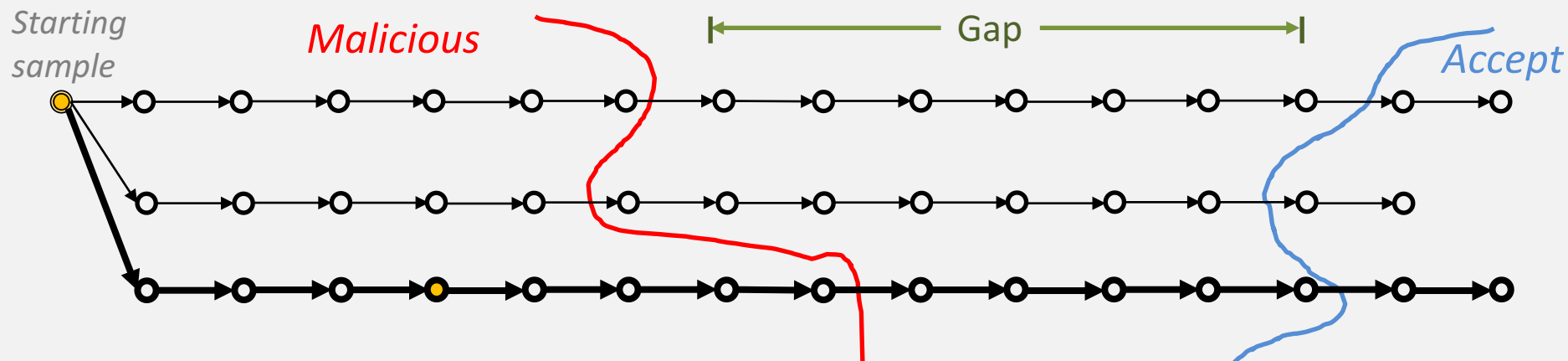
Assigning numeric state to samples

Evading path



- For a sample s , we can assign the following to be the state of s :
 Probability (a random path starting from s is evading)
Such real-value state would be useful in the search of evading samples.
- However, it is difficult to estimate the probability.
- Alternatively, assign Expected Gap to be the state.
 - Intuitively, a smaller Gap implies the sample has a higher chance of generating a evading path.
 - Can be estimated from a few (or a single) random paths.

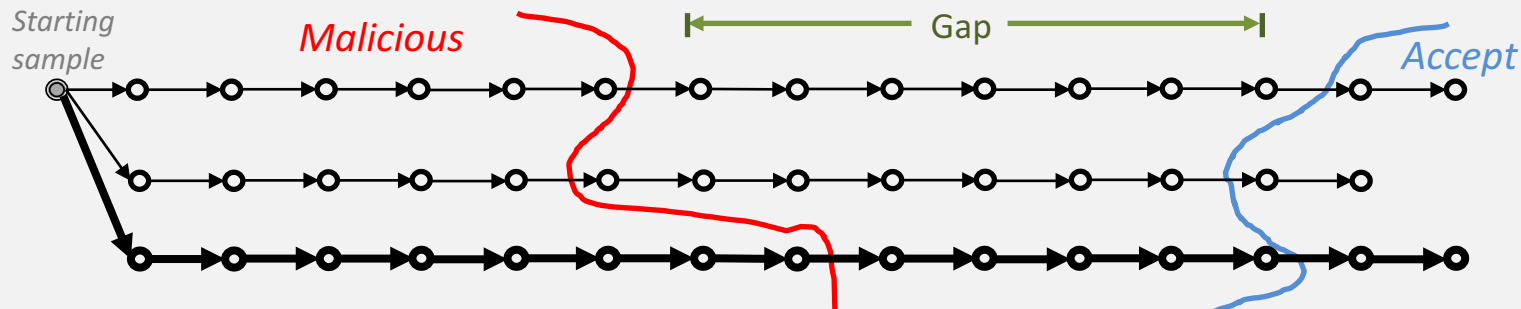
Search heuristic: Main Idea



1. Generate q random paths from the candidate.
2. Determine the path with the shortest gap (*or other criteria based on flipping distances*). Choose a sample along this path as the next candidate.

Algorithmic improvement

- To reduce the number of queries to Detector and Tester
 - “Batch” binary search on multiple paths: constant number of Detector query per path.



III. Experimentation Results

PDF malware classifiers: PDF_{RATE} [4], Hidost [5]

- PDF_{RATE}: Random Decision Forest.
- Hidost: SVM-based.

- Trained with 5,000 benign and 5,000 malicious PDF files, and test with another 500 malicious samples. PDF files obtained from Contagio archive.

[4] C. Smutz and A. Stavrou. *Malicious PDF detection using meta-data and structural features*. In ACSAC 2012.

[5] N. Šrndić and P. Laskov. *Detection of malicious pdf les based on hierarchical document structure*. NDSS 2013.

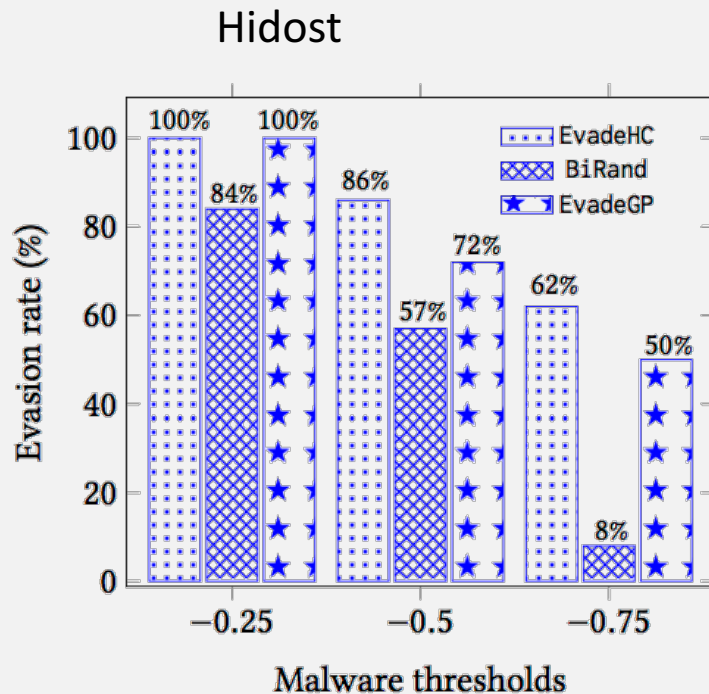
Evasion rate on “hardened” classifiers

EvadeHC: Proposed method.

BiRand: Baseline algorithm that performs binary searches on random paths.

EvadeGP: A previous method that has access to the real-value confidence score.

- Classifiers are hardened by adjusting the rejection threshold.
- Search limited to 2500 queries to Detector
- Interestingly, EvadeHC outperforms EvadeGP which has access to more info. We suspect this could be due to
 - EvadeHC makes decision based on Detector and Tester’s feedbacks. EvadeGP only based on the Detector’s feedbacks.
 - Reject-flipping distances could be a more accurate indicator compares to the confidence level.



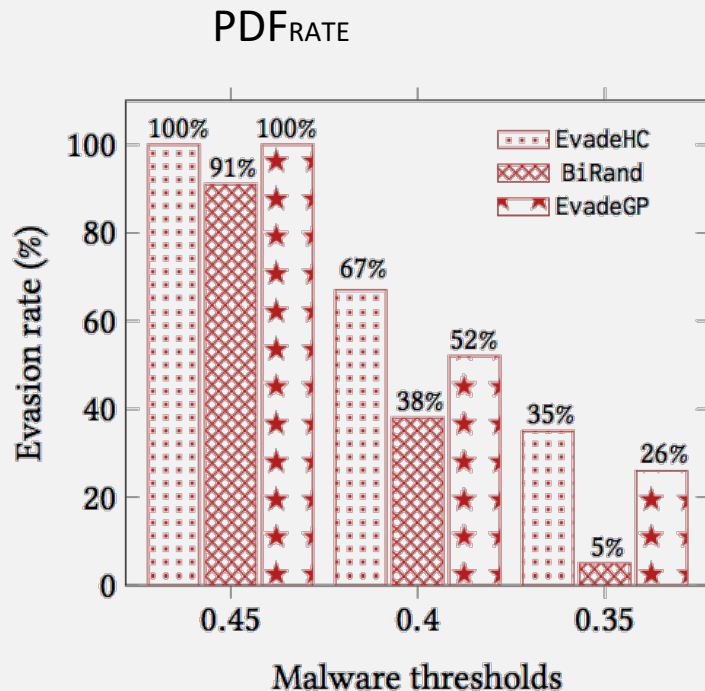
Evasion rate on “hardened” classifiers

EvadeHC: Proposed method.

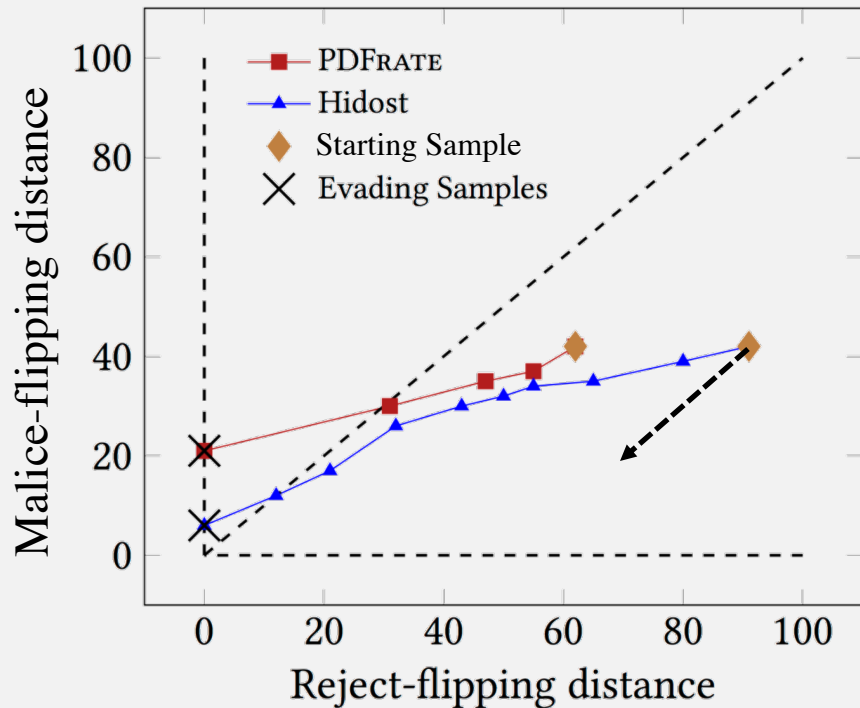
BiRand: Baseline algorithm that performs binary searches on random paths.

EvadeGP: A previous method that has access to the real-value confidence score.

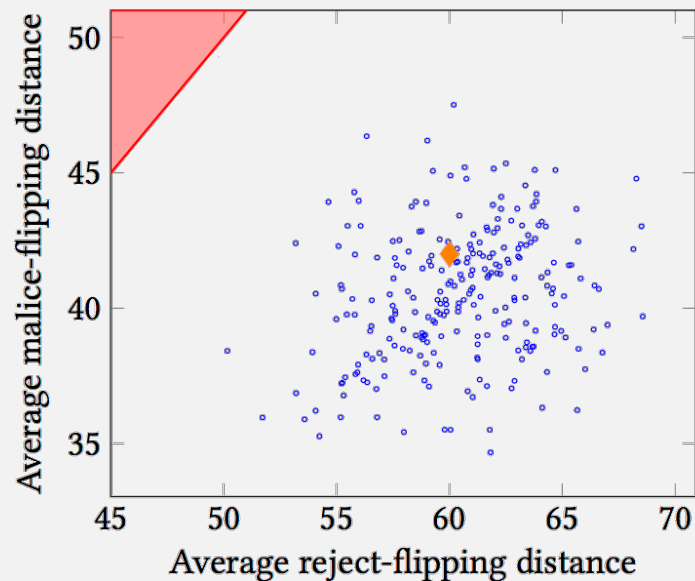
- Classifiers are hardened by adjusting the rejection threshold.
- Search limited to 2500 queries to Detector
- Interestingly, EvadeHC outperforms EvadeGP which has access to more info. We suspect this could be due to
 - EvadeHC makes decision based on Detector and Tester’s feedbacks. EvadeGP only based on the Detector’s feedbacks.
 - Reject-flipping distances could be a more accurate indicator compared to the confidence level.



Trace of a search



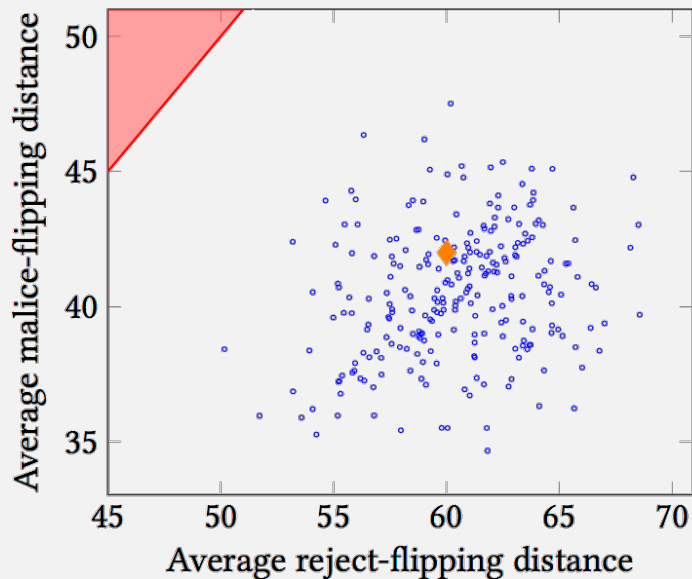
Average Flipping distances after one morphing step (Hidost)



An abstract Hidden-state Morpher model

- Every sample has a hidden 2-value state (a,b) .
 - Tester returns “Malicious” iff $(a>0)$;
 - Detector returns “Reject” iff $(b>0)$.
 - We can view the two hidden values corresponding to the average malicious-flipping and reject-flipping distances.
- Morpher outputs a random morphed sample with hidden values reduced according to a distribution.
- The Morpher is “random” and yet consistent to previous output. Similarly to Random Oracle.
- Such model is useful in analyzing search algorithm.

Average Flipping distances after one morphing step



IV. Discussion & Conclusions

Conclusion

- Many evasion attacks heavily rely on domain knowledge. It would be interesting to investigate the effectiveness of evasion attacks in a generic setting.
- We formulate *Evasion in the Dark*. This model gives a restricted setting where domain knowledge are confined in the 3 black-boxes. From the attacker's point of view, no other specific domain knowledge are required in evasion.
- The model is useful for complex domain – as long as a morpher & tester are available, one can carry out evasion attack.
- We give a method (flipping distances) to assign meaningful real-value states to the samples, and show that evasion is possible even with binary black-boxes.
- Evasion attacks can be employed to enhance defense – by feeding evading samples as training samples.