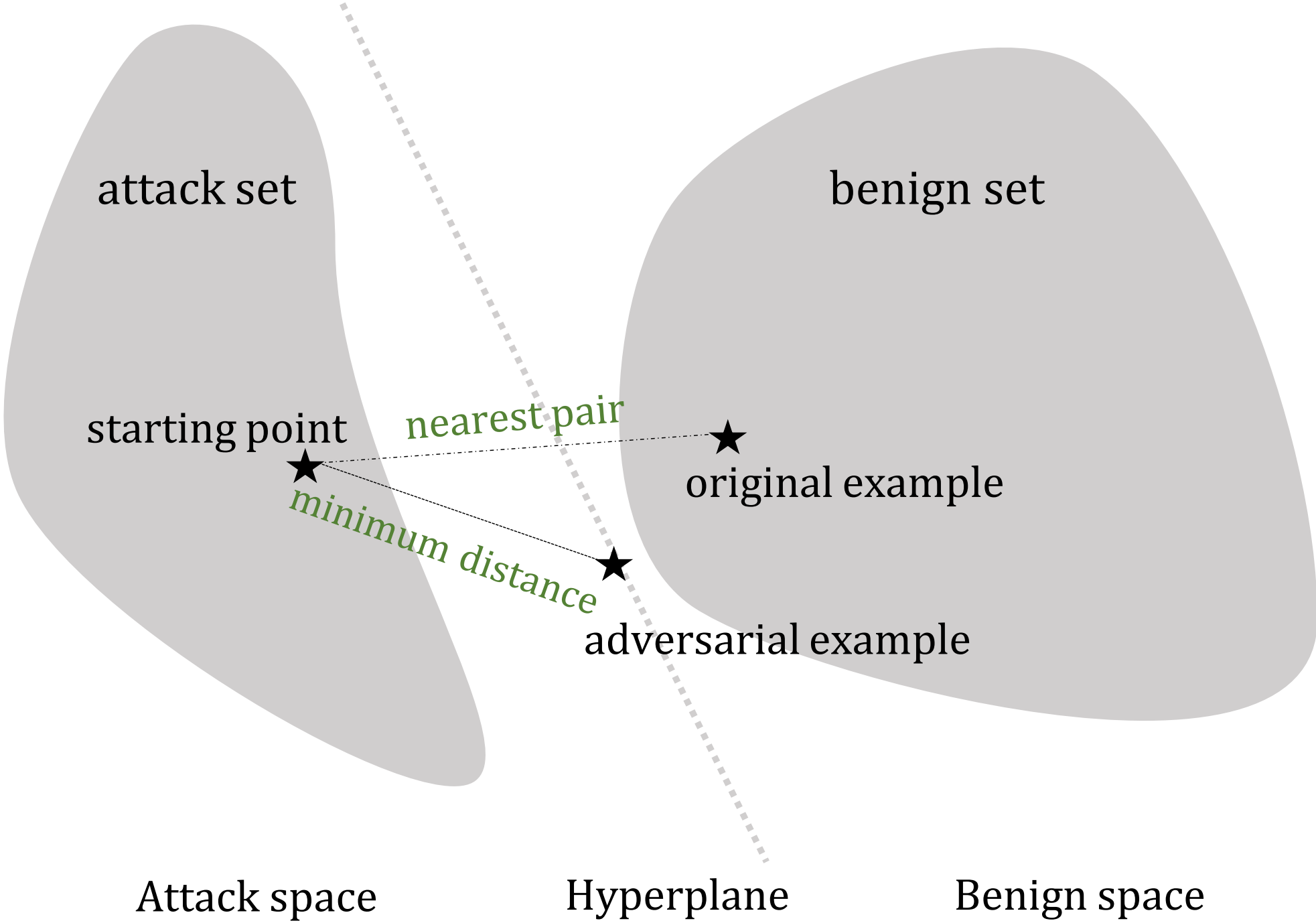
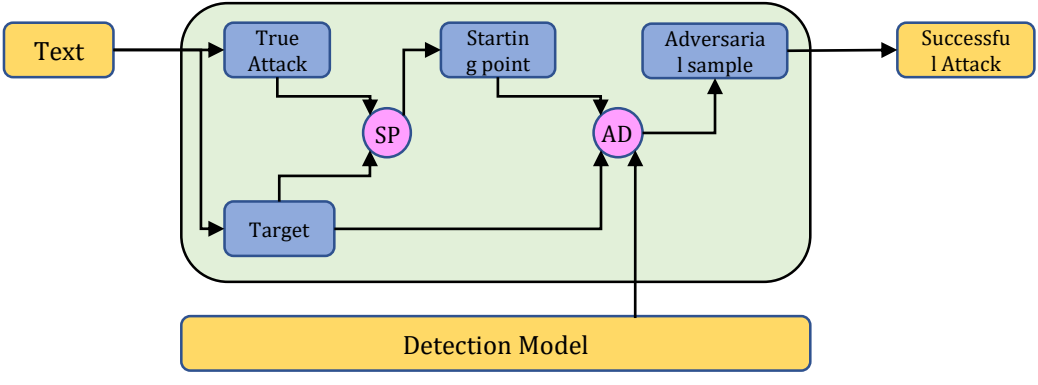
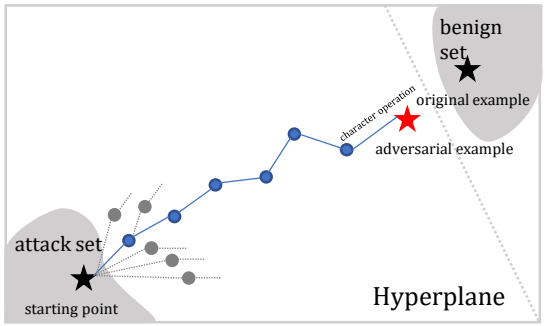

attack set
benign set
starting point
nearest pair
original example
minimum distance
adversarial example
Attack space
Hyperplane
Benign space


Text
True Attack
Startin g point
Adversaria l sample
Successfu l Attack
SP
AD
Target
Detection Model


benign
set
original example
character operation
adversarial example
attack set
starting point
Hyperplane

Sensitive Data Access

Starting point:
/tienda1/publico/productos.jsp.old → An old version file, dangerous!

Adversarial example:
/tienda1/publico/productos.jsp.d → No longer accessible, benign, but predicted dangerous.

Original example:
/tienda1/publico/productos.jsp → Current version, benign.

Toxic Comments

Starting point:
Thank you → Normal comment, benign.

Adversarial example:
FUCKYOUR FITHY MOTHER IN THE ASn, DRY! → Still toxic compared to original, but predicted benign.

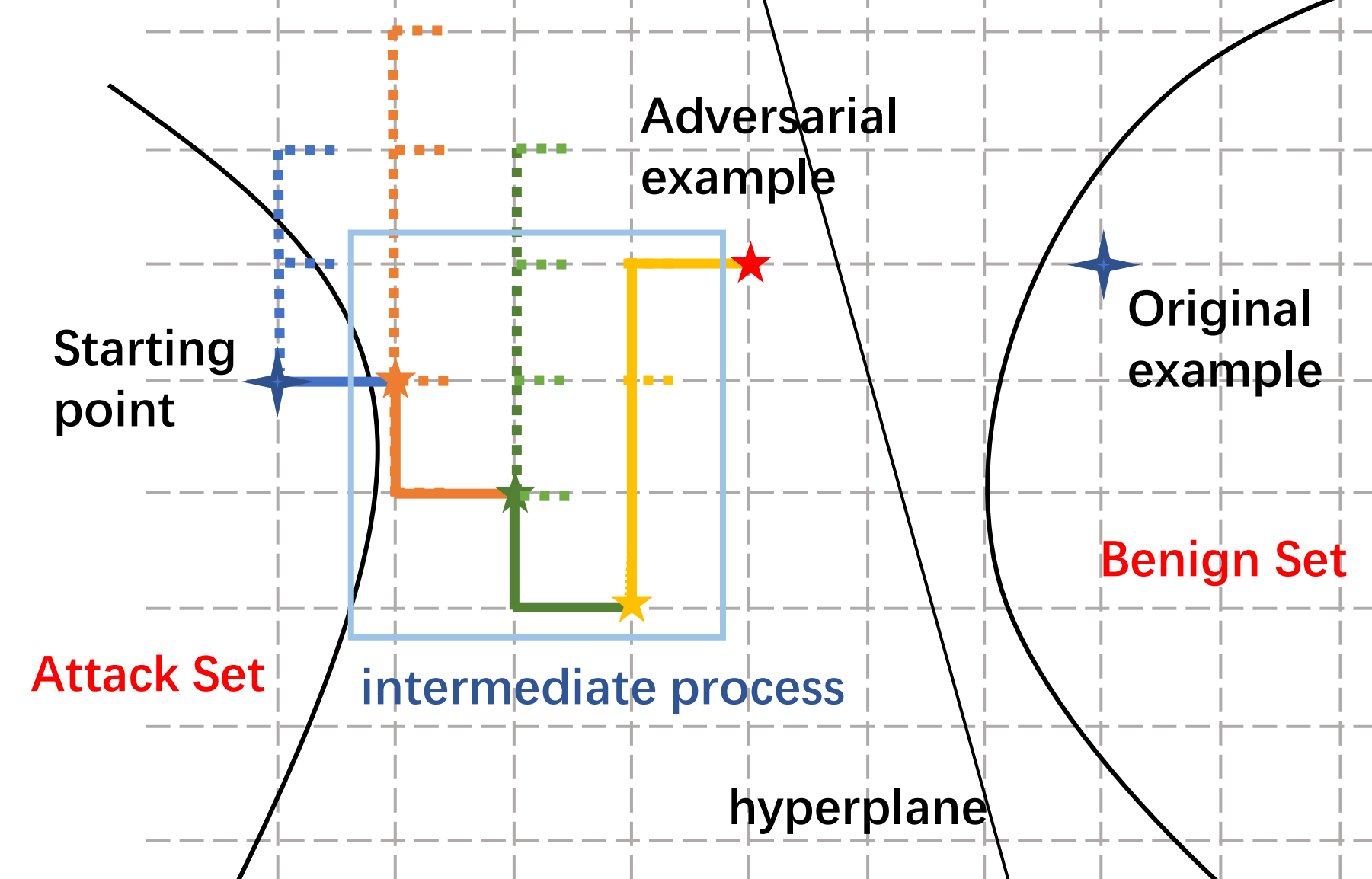
Original example:
FUCK YOUR FILTHY MOTHER IN THE ASS, DRY → Apparently toxic.

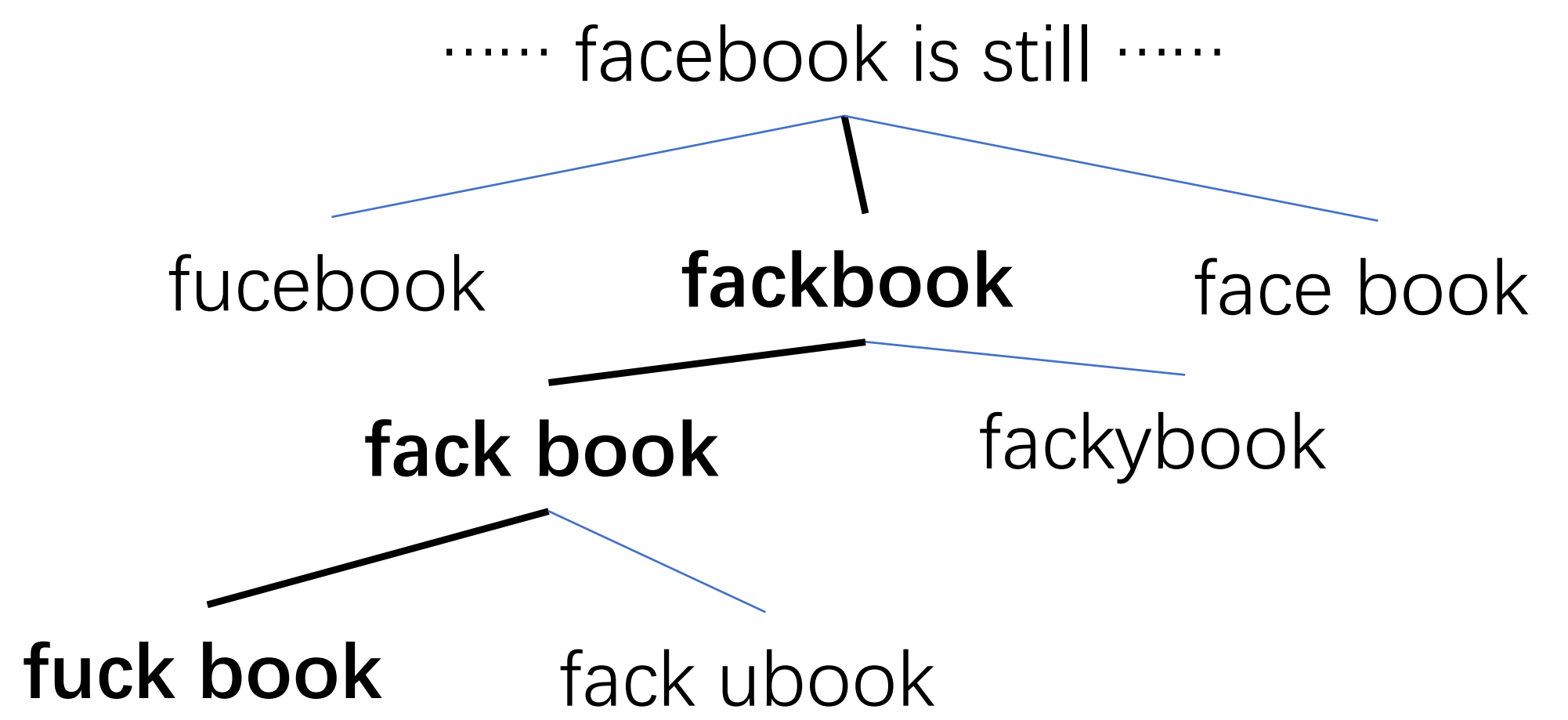
Injection

Starting point:
/tienda1/publico/vaciar.jsp?B2=Vaciar+carrito%7C → Injection, malicious.

Adversarial example:
/tienda1/publico/vaciar.jsp?B2=Vaciar+carrito7C → No longer injection, but predicted malicious.

Original example:
/tienda1/publico/vaciar.jsp?B2=Vaciar+carrito → Normal access, benign


Adversarial example
Starting point
Original example
Benign Set
Attack Set
intermediate process
hyperplane

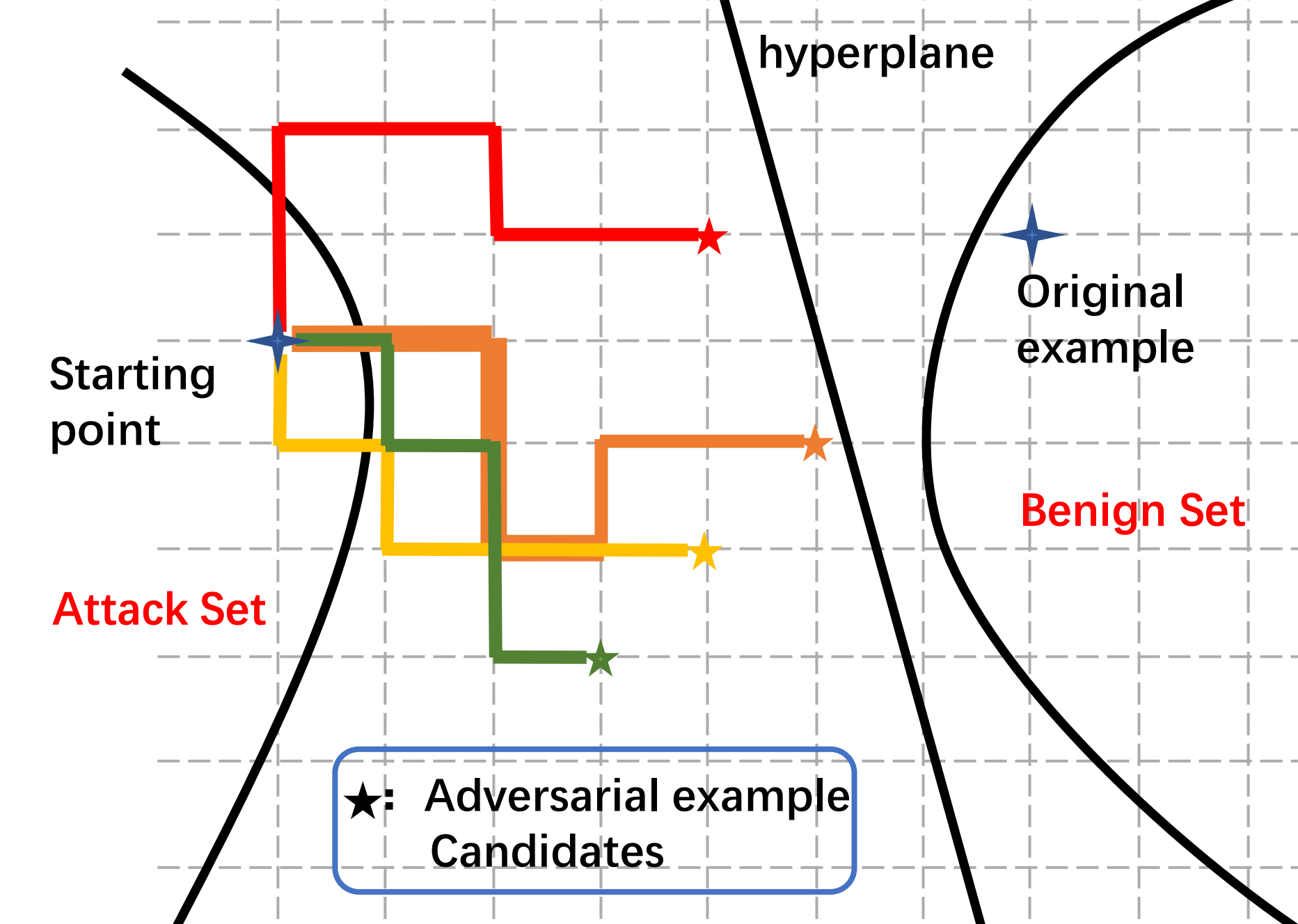

…… facebook is still ……
fucebook
fackbook
face book
fack book
fackybook
fuck book
fack ubook

String A: abcdSSmnop

String B: aucSSlmnopq

String A: **a**c**d**SS mnop

String B: a**u**c SS**l**mnop**q**


hyperplane
Original example
Starting point
Benign Set
Attack Set
★: Adversarial example Candidates