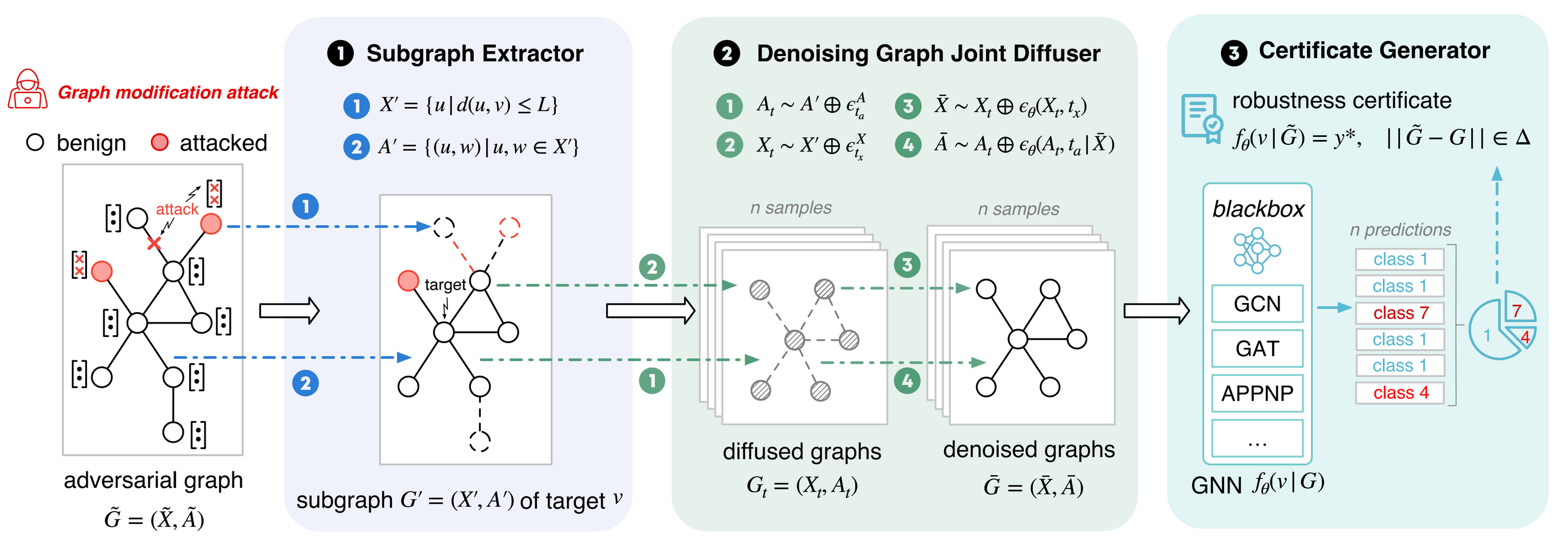
Graph modification attack
benign
attacked
attack
adversarial graph
$\tilde{G} = (\tilde{X}, \tilde{A})$
1 Subgraph Extractor
1 $X' = \{u \mid d(u,v) \le L\}$
2 $A' = \{(u,w) \mid u,w \in X'\}$
target
subgraph $G' = (X', A')$ of target v
2 Denoising Graph Joint Diffuser
1 $A_t \sim A' \oplus \epsilon^A_{t_a}$
2 $X_t \sim X' \oplus \epsilon^X_{t_x}$
3 $\bar{X} \sim X_t \oplus \epsilon_\theta(X_t, t_x)$
4 $\bar{A} \sim A_t \oplus \epsilon_\theta(A_t, t_a \mid \bar{X})$
n samples
n samples
diffused graphs
$G_t = (X_t, A_t)$
denoised graphs
$\bar{G} = (\bar{X}, \bar{A})$
3 Certificate Generator
robustness certificate
$f_\theta(v \mid \tilde{G}) = y^*, \quad ||\tilde{G} - G|| \in \Delta$
blackbox
GCN
GAT
APPNP
...
GNN $f_\theta(v \mid G)$
n predictions
class 1
class 1
class 7
class 1
class 1
class 4
7
1
4