Scanner Detection:

Scanner Detection using sequential hypothesis aims at tagging fast as well as slow rate scanners with a generic alogirthm based on overall failure rate of connections from each granular sourceIP.

Initial asssumptions:

- Scanners have large number of failed connections since they attempt to connect to inactive IP's or ports with inactive services. (failed connections are either SYN followed by no ACK, SYN followed by RST, FIN followed by RST, FIN followed by nothing, ACK followed by RST, ACK followed by nothing.)
- Further scanners are assumed to have atleast >80% failed connections which allows us to define our model and obtain the thresholds for sequential hypothesis.

Limitations:

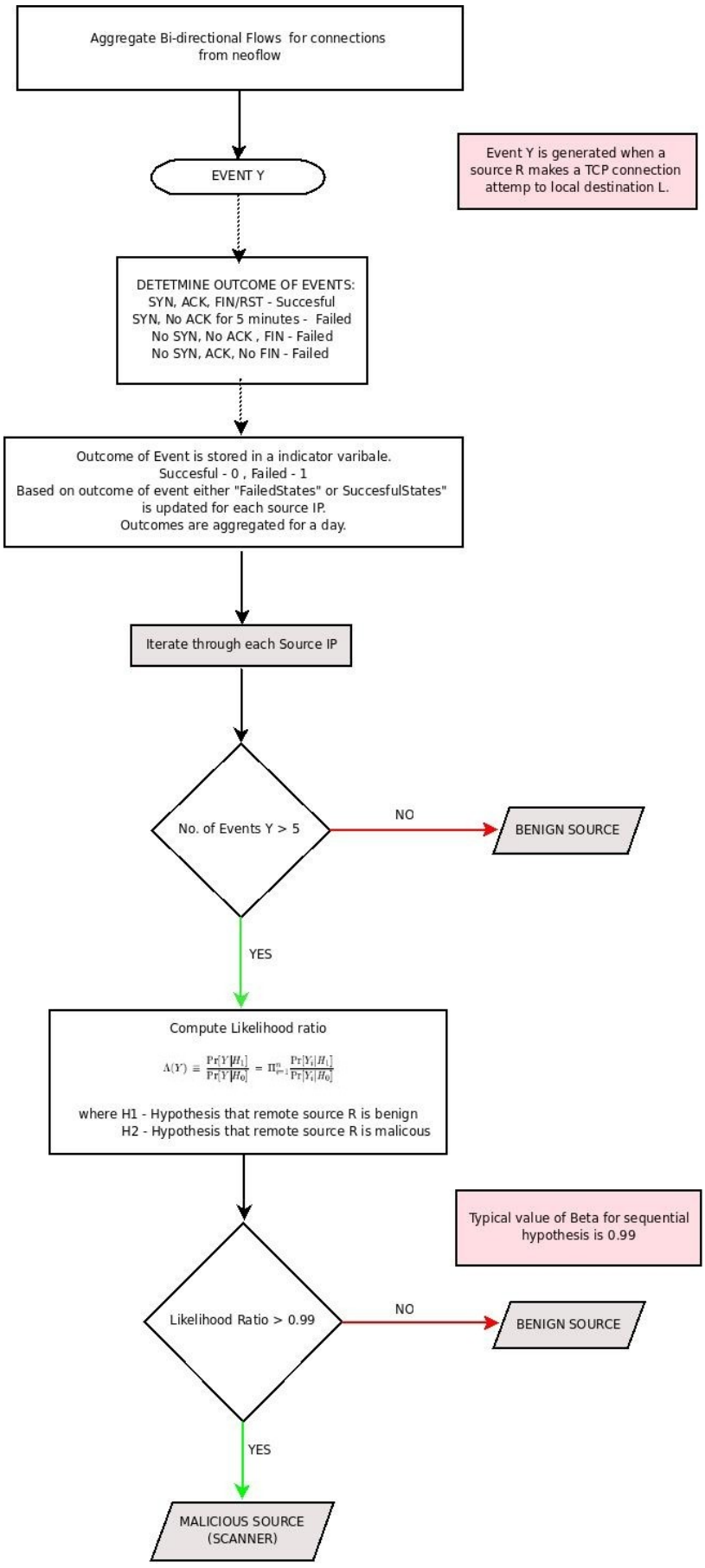
- Atleast 5 connection attempts are required from each source IP to succefully categorize them as failed/succesful.
- Vertical scans are not detected since sequential hypothesis is granular on each source IP.

Definitions:

- Event Y describes succesful/failed connection(0/1)
- H0- Hypothesis that source IP is benign.
- H1- Hypothesis that source IP is malicious.
- Likelihood ratio(LR) = P(Y/H0)/P(Y/H1)
- Upper threshold (beta) = 0.99

Alogrithm:

- Connection states are maintained as failed(1) or succesful(0) in a vector for each source IP
- For each source IP sequential hypothesis is computed on vector of connection states
  - For each event in vector
    - compute probability of event(1/0) conditional on hypothesis H0/H1.
    - compute likelihood ratio for each event and store in LR-vector.
  - Compute product sum on LR-vector.
  - If LR > beta tag as scanner.


Aggregate Bi-directional Flows for connections from neoflow
EVENT Y
Event Y is generated when a source R makes a TCP connection attemp to local destination L.
DETETMINE OUTCOME OF EVENTS:
SYN, ACK, FIN/RST - Succesful
SYN, No ACK for 5 minutes - Failed
No SYN, No ACK , FIN - Failed
No SYN, ACK, No FIN - Failed
Outcome of Event is stored in a indicator varibale.
Succesful - 0 , Failed - 1
Based on outcome of event either "FailedStates" or SuccesfulStates" is updated for each source IP.
Outcomes are aggregated for a day.
Iterate through each Source IP
No. of Events Y > 5
NO
BENIGN SOURCE
YES
Compute Likelihood ratio
$\Lambda(Y) \equiv \frac{\Pr[Y|H_1]}{\Pr[Y|H_0]} = \Pi_{i=1}^{n} \frac{\Pr[Y_i|H_1]}{\Pr[Y_i|H_0]}$
where H1 - Hypothesis that remote source R is benign
H2 - Hypothesis that remote source R is malicous
Typical value of Beta for sequential hypothesis is 0.99
Likelihood Ratio > 0.99
NO
BENIGN SOURCE
YES
MALICIOUS SOURCE (SCANNER)