Conditional Port Probability

Conditional Port Probability

We are looking for a simple statistic which will help identify runs of consecutive (i.e. the port number for the current flow is one greater than the previous flow) in the "Potential Scan" 200 Netflow records reported by email. We do this for both UDP and TCP flows The first step is to record the unique UDP and TCP ports encountered in each email and sort them by TCP port number and UDP port number (p_i). Duplicate ports are recorded only once. This is done separately for source and destination ports. We identify the number of times (*R*) there is a run, i.e. where $(p_{i+1} - p_i) = 1$. We then divide this value by the total number of possible runs of ports (*N-1*) where N = total number of flows, typically 200). We define this as the Run Probability observed *Po*.
$\boldsymbol{P_o}=R/(N-1)$.

One way of looking at this is to take p_{low} as the lowest port number involved in any run and p_{high} as the highest number port in any of the runs, then the number of ports considered is M = (p_{high}- p_{low}). Let us consider a set of flows with the ports numbers: 1 2 3 4 5 x x 8 9 10 11, where there are no flow numbers 6 and 7, then the gap width g_1 is 2 ports and there are R = 2 runs of consecutive ports. We can see that in general the number of pairs is:
$(M-1) - \{(R-1) * 2 + SUM_j(g_j) - 1\}$
where for the above example $M = 11, R = 2, g_1= 2$ and the number of pairs is 7.

We can get an estimate of the Run Probability for a random set of unique (i.e. each port number only occurs once) ports as follows. There are roughly 65,536 (2^{16}) possible ports for UDP or TCP. The probability of selecting any port p_i is 1, the probability of the next higher number port (p_{i+1}) being = $p_i + 1$ *is 1/65,536* and since we have *M - 1* chances:
$\boldsymbol{P_r}$ = *# chances / # possible ports*

For our emails we have # chances = 200 - 1, so P_r = 199/65,536 ~ 0.3%.