第三个作业

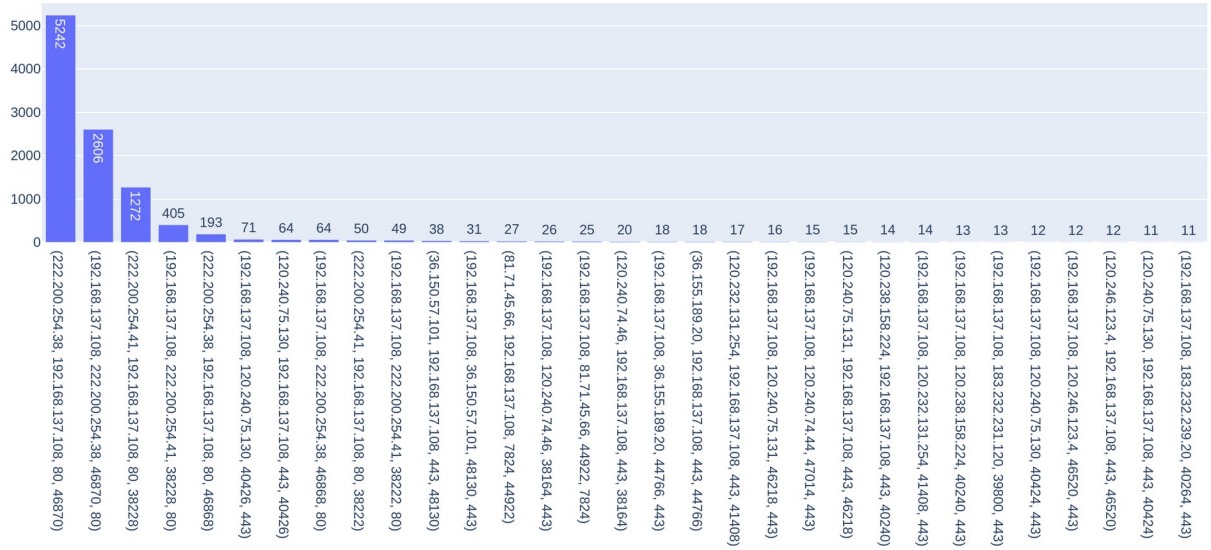
实验分工:

陈浚铭：写代码

杨锦程: wireshark 抓包

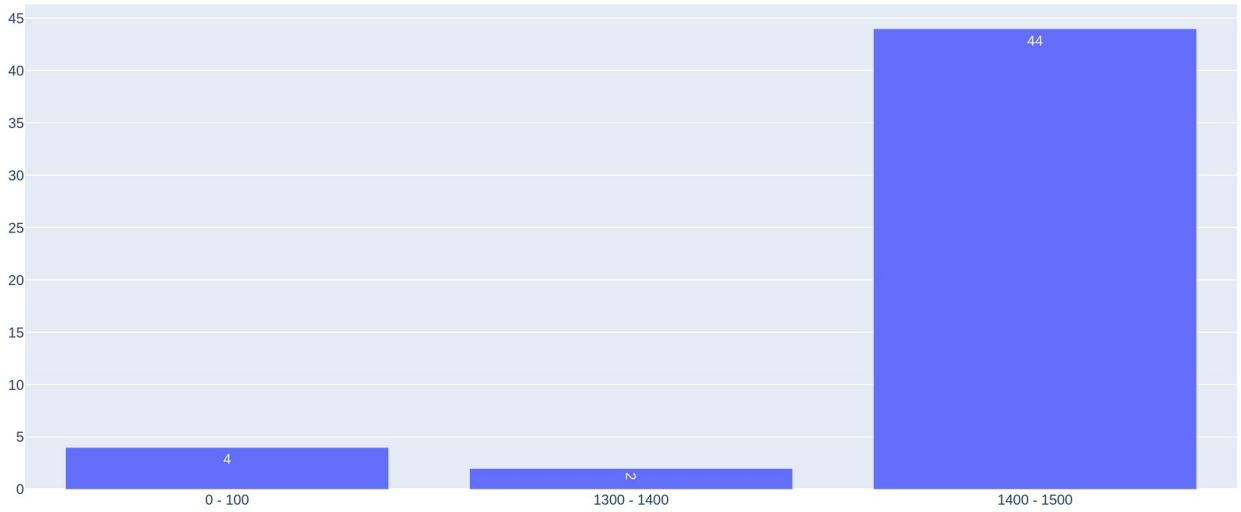
流的数量 number of flows = 10719 （总共的流数量）

每个流的分组数分布 flowNumDistribution.py

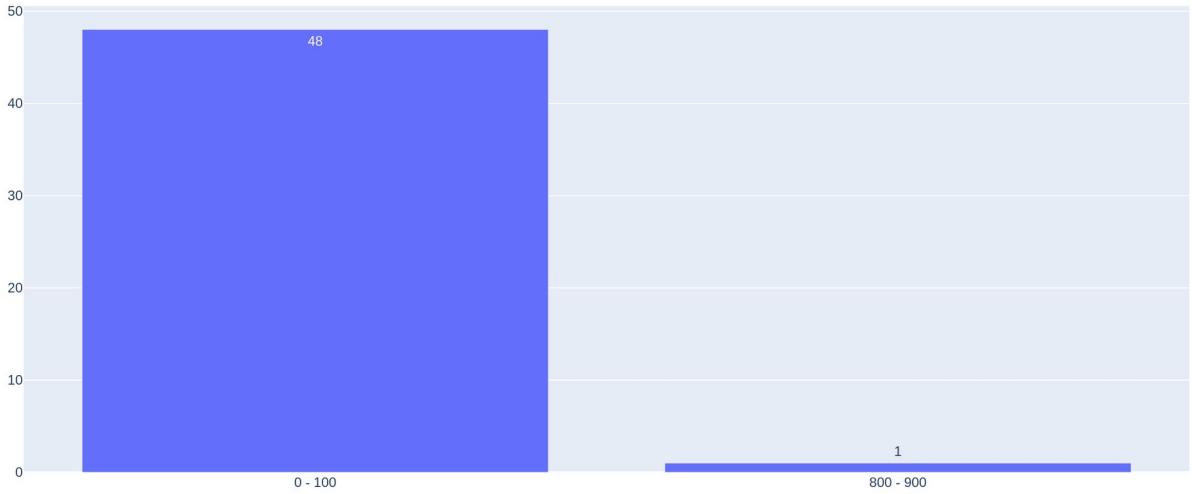


基于流的字节数分布： flowLengthDistribution.py

(srcIP = 222.200.254.41, destIP = 192.168.137.108, srcPort = 80, destPort = 38222, TCP)



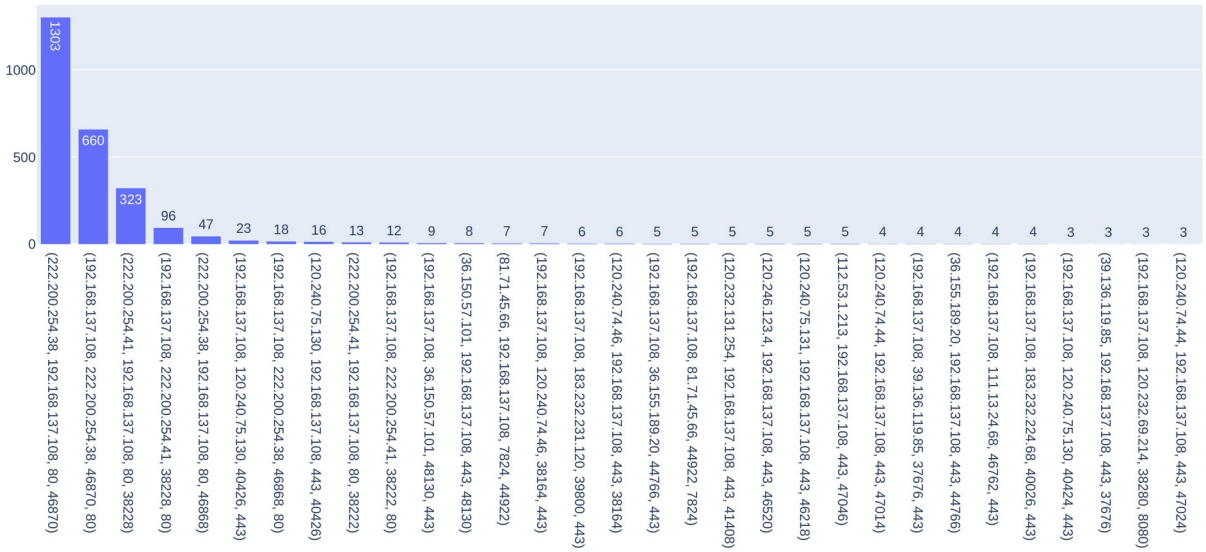
(destIP = 222.200.254.41, srcIP = 192.168.137.108, destPort = 80, srcPort = 38222, TCP)



之后，通过以 1/4 抽样律进行抽样：

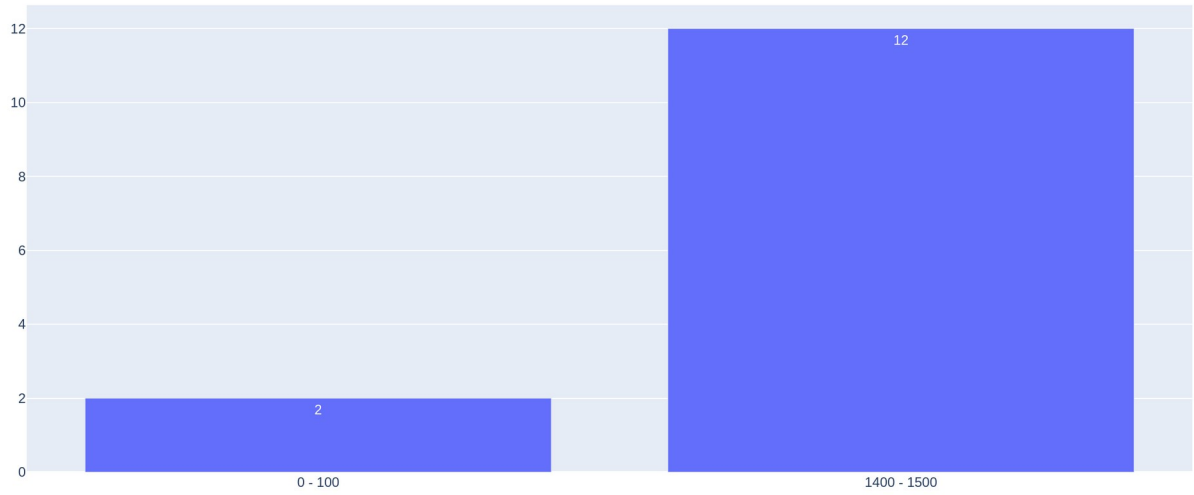
流的数量 number of flows = 2689 （总共的流数量）

每个流的分组数分布 sampledFlowNumDistribution.py

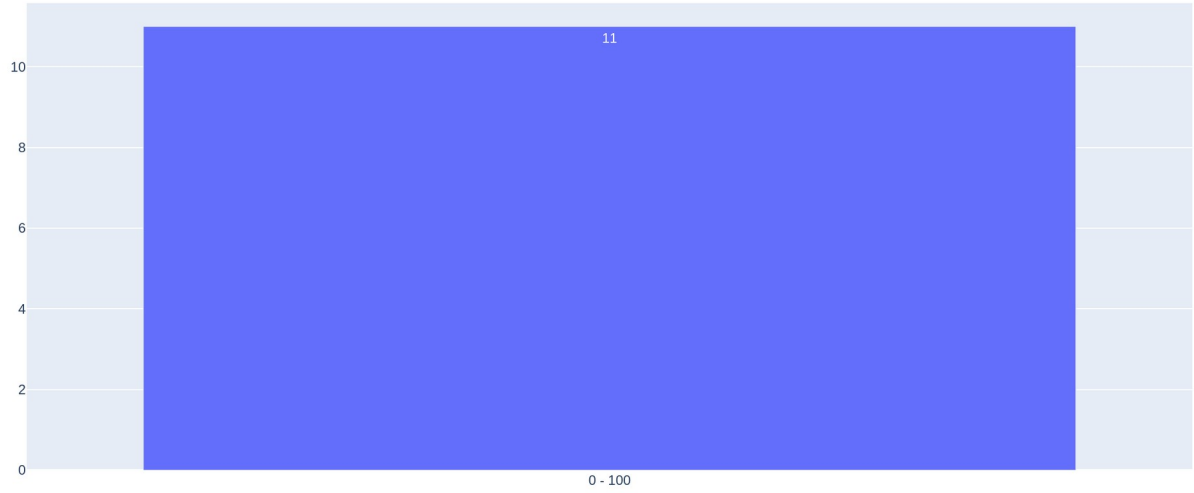


每个流的字节数分布 sampledFlowLengthDistribution.py

(srcIP = 222.200.254.41, destIP = 192.168.137.108, srcPort = 80, destPort = 38222, TCP)



(destIP = 222.200.254.41, srcIP = 192.168.137.108, destPort = 80, srcPort = 38222, TCP)



终结：我们发现以 1/4 的随机抽样的结果是符合我们对所有数据包集合的结果的。也就是直方图的比例是跟我所有数据包集合获取的成正比。随机抽样的缺点是存在 sample selection bias, 比如在我们在网络流 (destIP = 222.200.254.41, srcIP = 192.168.137.108, destPort = 80, srcPort = 38222, TCP)中的直方图当中，没有获取到 1400-1500 长度的数据包。