Final Assignment Task 1: Flow data

* **Step 1: Acquiring flow data**

命令行：

source /work/courses/unix/T/ELEC/E7130/general/use.sh

cd $TRACE

ls

* **Step 2: Data pre-processing**

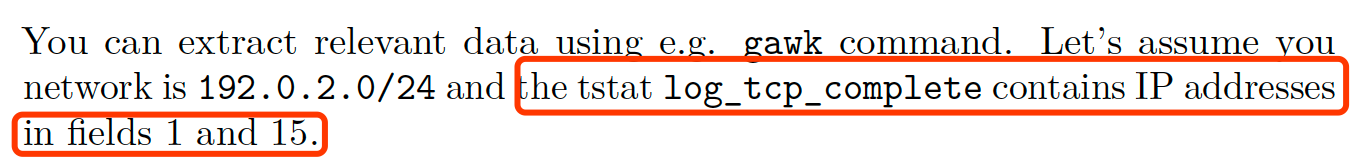
除了1.6之外只用分析部分数据（FS2，按要求从FS1中提取的数据），而1.6需要分析所有数据（FS1）。

步骤：

1.按学号最后一位选一个子网（假设为192.0.2.0/24）

2.用gawk命令提取对应数据

**gawk ‘$1~/^192\.0\.2\./ || $2~/^192\.0\.2\./’ 1200.t2 > ~/my\_1200.t2**





**提示：**

* 要提供预处理数据的命令和代码
* 提供处理后的样例数据
* **Step 3: Data analysis 分析提取出来的FS2数据**
* **1.1 Plot traffic volume**

不是很懂

* **1.2: Flows by port numbers**

根据端口号画直方图

* **1.3: OD-pairs**

用awk函数做配对，再根据1.bytes 2.flow画图

* **1.4: Per user data volume**

将子网下的所有相同ip地址（用户）各自的bytes加起来，然后做一个所有用户bytes的直方图

* **1.5: Flow length distribution**

画flow length（？）分布，ecdf函数及其关键的总结数据。

* **1.6: Flow sampling**

使用FS1的所有数据做两个抽样：一个只从ipv4数据抽样，另一个只从ipv6数据中抽样。

然后从1.1-1.5中选一种对抽样出来的两类数据画图，再和用FS2数据画出来的图对比。然后问FS2的数据是否具有代表性，ipv4和ipv6数据有没有区别。