“No NAT’d User left Behind”: Fingerprinting Users behind NAT from NetFlow Records alone

有些工具为了探索流量指纹获取了 the entire traffic, including IP addresses and payloads，这种方式无论是在性能还是在隐私方面都不具有可行性，In reality, most ISPs convert user traffic into NetFlow records for a concise representation that does not include the payload；

更重要的是由于NAT技术，几个IP可能对应成千上万的用户；

We devised a new fingerprinting framework that overcomes these hurdles. Our system is able to analyze a huge amount of network traffic represented as NetFlows, with the intent to track people. It does so by accurately inferring when users are connected to the network and which IP addresses they are using, even though thousands of users are hidden behind NAT. Our prototype implementation was deployed and tested within an existing large metropolitan WiFi network serving about 200,000 users, with an average load of more than 1,000 users simultaneously connected behind 2 NAT’d IP addresses only. Our solution turned out to be very effective, with an accuracy greater than 90%. We also devised new tools and refined existing ones that may be applied to other contexts related to NetFlow analysis.

Classifiers are indeed very effective, but it is often assumed that they are given as input the entire traffic, including payloads, headers, and other timing information. The reality is quite different, however. For performance reasons, the attacker cannot analyze or store the entire traffic, that can reach a throughput higher than 2, 000 Gbit/s in the case of large Internet Exchange points. However, ISPs often convert network traffic into NetFlow records for a more concise representation. These records are used to collect IP traffic statistics for data analysis and contain very little information (no payload, for example)

即便attackers能开发出改进的分类算法识别个人by NetFlow records,但是还是假设已经知道每个人的ip，但是由于NAT……

文章使用了ISP的一个交换中心的数据（ISP的数据是NetFlow records,，没有payload），track users, and accurately estimate when they are connected to the network and which IP address they are using；

Application scenarios：

After-the-fact forensic analysis 事后取证分析

Covert intelligence operations 秘密情报行动

Related work：

Indeed, it is important to remark that there are many proposed solutions that work for, e.g., devices or applications but cannot be used for individuals. Or, rather, they work for individuals but only when payloads are available (but fail when applied to NetFlow logs only).

four different categories:

user and host profiling（including behavioral targeting and single user/host profiling）, focuses only on identifying communication patterns that are in common to many users (or hosts), with the intent to analyze anomalies.(防攻击和病毒)

traffic and application profiling,

information leaks

device profiling.

Backgroud& definitions

NetFlow： NetFlow is a protocol designed by Cisco to collect IP traffic information while getting rid of any IP packet payload.

HMMs.：

**采用的definition1(五元组)** 这个当作key: NetFlow key function ： we can think to this key as a 5-tuple containing IP source and destination addresses, source and destination ports, and the protocol used

**Definition 2：这个当作value** NetFlow raw record nfr：cumulative number of exchanged packets, bytes counters, flow starting and finishing timestamps, TCP flags, and Type of Service (ToS).-- packets, bytes, start timestamp, end timestamp, TCP flags, ToS).

**Definition 3: 对应一个key的数据流的情况，就是某个主机给另一个主机发送的数据包（收到的数据包是不知道的）** A flow f is defined as a set of NetFlow raw records {nfr1,..., nfrn} such that ∀1 ≤ i, j ≤ n, the key associated with nfri is equal to the key associated with nfrj . We realized that the flow defined above is very convenient and effective in identifying users. For instance, a flow properly captures certain usage patterns and is oblivious to NAT routers. Indeed, two NAT’d users connecting to the same IP address and port will be assigned two distinct local ports. Therefore, two distinct flows will be generated, one per each user.

Indeed, two NAT’d users connecting to the same IP address and port will be assigned two distinct local ports. Therefore, two distinct flows will be generated, one per each user.

**Definition 4:** **（描述的是两台主机的某两个端口的数据包的交换情况，例如A:8088给B:80发了两个包后B给A回了1个包）**Given an IP protocol protocol, two pairs of IP addresses and ports ip1:port1 and ip2:port2, we define a bi-directional flow as the union of the flows from ip1:port1 to ip2:port2 with the one from ip2:port2 to ip1:port1. Finally, we will consider only ordered flows, by applying a ordering function sort that rearranges a flow, sorting its nfr’s with respect to the start timestamp. More precisely, we say that an ordered flow is the flow obtained by applying the sort function to the nfr’s that compose the input flow, namely sort(f) = {nfr1, ..., nfrn}, such that ∀i（这里定义了一个二部图）

**HMM 略**

**本文中HMM的design**

A, each nfr is represented by a data tuple of t values, corresponding to its t attributes (number of packets, start timestamps, etc.). In particular, we consider that the observables are t dimensional vectors distributed according to N multivariate Gaussian distributions, one for each state: we will adopt one N ×t matrix, containing the means, and N covariance matrices, to define the t-dimensional multivariate Gaussian distributions. In other words, for each state of the model, we have t Gaussian densities, one for each of the t vector elements and, to represents such densities, we have to specify their means and their covariances.

缺点或改进办法：

1. 用的HMM,如果换成深度学习，如LSTM或强化学习，会不会更准？
2. 他用的数据量大，我们没有那么多数据怎么办？
3. 220页A new nfr is sent to the collector when the connection is closed，需要追踪TCP数据流？

Indeed, a NetFlow can expire for three main reasons: (1) the flow has been inactive for a time period longer than the inactive timeout; (2) the flow has been active for a time period longer than the active timeout; (3) the flow cache is full and some space needs to be freed for new flows. Default values for inactive timeout and active timeout are set to 15 seconds and 30 minutes, respectively.

4．可以分析IP访问的频繁项集，用机器学习方法

想法：

1. 在某段时间src 和 dst 的交互 src->dst dst->src,这个应该是可以辨认的，就是这篇文章中说的二部图，可根据这个特征对流量进行分类，即那些包属于一个人
2. 访问ip的list可以考虑，page-rank?
3. 访问Ip的关联性？（频繁项目集）
4. 阶段性的保留一些数据用于身份认证（保留数据是为了身份认证，阶段性更新是因为用户的上网行为可能会根据环境变化）