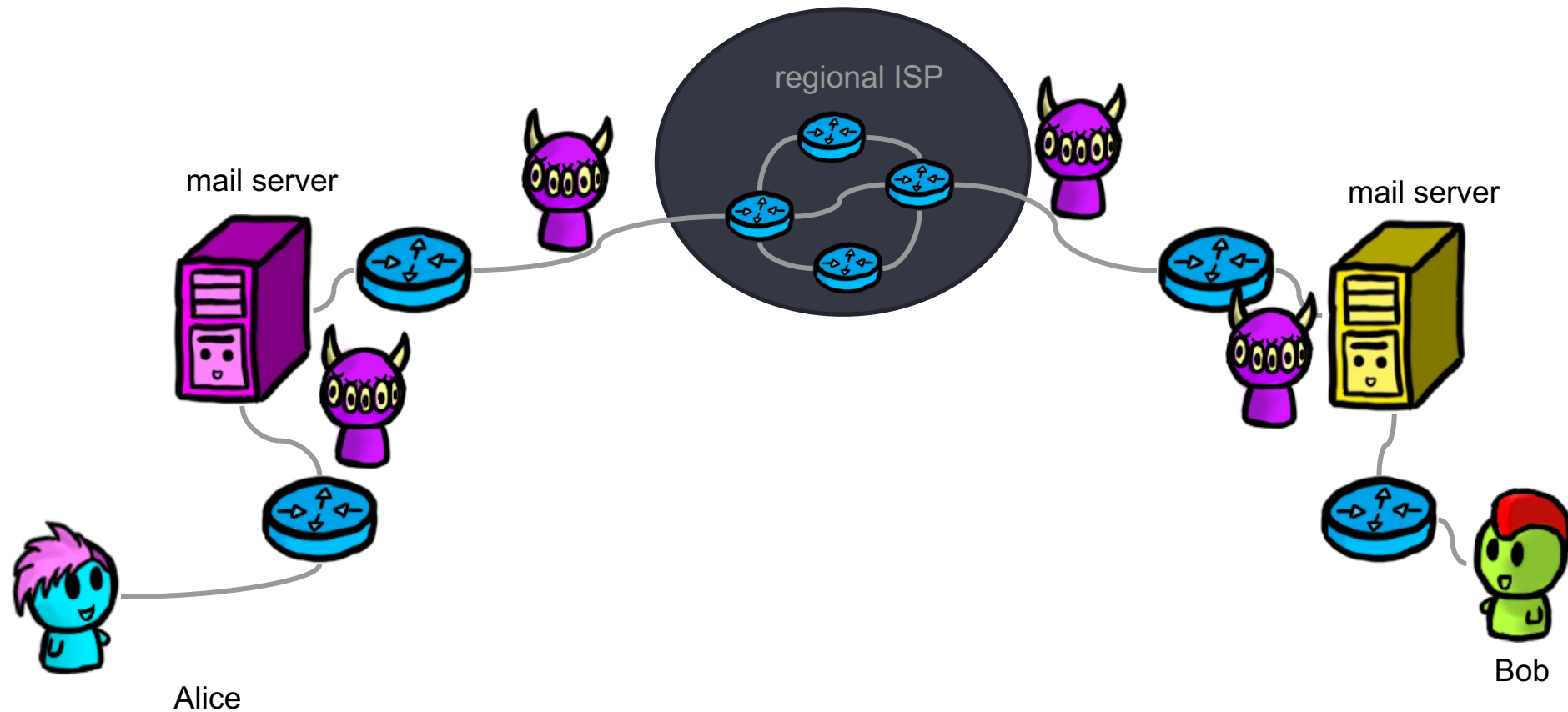


CS459/698 Privacy, Cryptography, Network and Data Security

Encrypted Traffic Analysis

Spring 2025, Monday/Wednesday 2:30pm-3:50pm

Traffic Analysis



Easy attack surface:

- Mallory has access to one of the many hops traffic takes on the internet

Communication media (WiFi)

- WiFi
 - Can be **easily intercepted** by anyone with a
- WiFi-capable (mobile) device
 - Don't need additional hardware, which would cause suspicion
- Maybe from kilometers away using a directed antenna
- WiFi also raises other security problems
 - Physical barriers (walls) help against random devices being connected to a wired network, but are (nearly) useless in case of wireless network

Communication media

- Copper cable
 - Inductance allows a physically close attacker to eavesdrop without making physical contact
 - Cutting cable and splicing in secondary cable is another option
- Optical fiber
 - No inductance, and signal loss by splicing is likely detectable
- Microwave/satellite communication
 - Signal path at receiver tends to be wide, so attacker close to receiver can eavesdrop
- All these attacks are feasible in practice, but require physical expenses/effort

Traffic Analysis

- TCP/IP has each packet include unique addresses for the packet's sender and receiver end nodes, which makes traffic analysis easy
- The attacker simply needs to sniff packets to determine what is going where and when.
 - Can be sensitive info such as two CEOs talking or a whistle blower.
- tcpdump is a text-based traffic analysis tool

Tcpdump (1 of 3)

```
·14:47:26.566195 IP 192.168.2.2.22 > 192.168.1.1.41916: Flags [P.], seq 196:568, ack 1, win 309,  
options [nop,nop,TS val 117964079 ecr 816509256], length 372
```

- 14:47:26.566195 the timestamp of the received packet
- IP is the network layer protocol (IPv4)
- 192.168.2.2.22 is the source IP address and port
- 192.168.1.1 is the destination IP address and port

Tcpdump (2 of 3)

```
·14:47:26.566195 IP 192.168.2.2.22 > 192.168.1.1.41916: Flags [P.], seq 196:568, ack 1, win 309,  
options [nop,nop,TS val 117964079 ecr 816509256], length 372
```

- TCP Flag (`Flags [P.]`) fields include:

Value	Flag Type	Description
S	SYN	Start Connection
F	FIN	End (Finish) Connection
P	PUSH	Push data
R	RST	Reset connection
.	ACK	Acknowledgement

Tcpdump (3of 3)

```
·14:47:26.566195 IP 192.168.2.2.22 > 192.168.1.1.41916: Flags [P.], seq 196:568, ack 1, win 309,  
options [nop,nop,TS val 117964079 ecr 816509256], length 372
```

- `seq 196:568` is the sequence number of the data contained in the packet (196 bytes to 568 bytes)
- `ack 1` is the ack number, which is 1 (sender) or the next expected byte (receiver)
- `win 309` is the number of bytes available in the receiving buffer
- `options [nop,nop,TS val 117964079 ecr 816509256]`, are the TCP options
- `length 372` is the length, in bytes, of the payload data (the difference between the first and last byte in the sequence number)

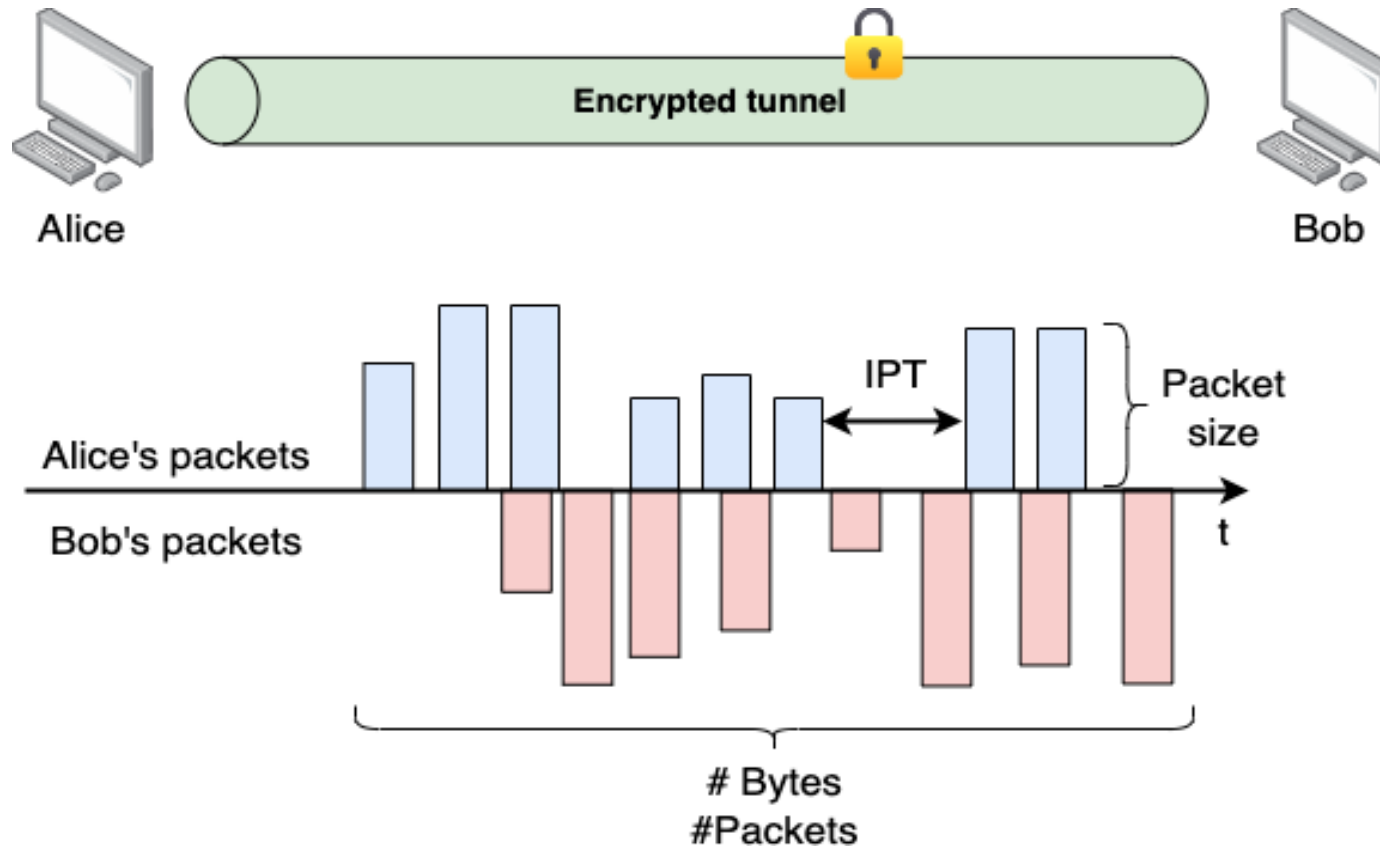
Encrypted Traffic Analysis

Encryption reduces visibility over network traffic

- TLS and other PETs significantly improved security and privacy for Internet users
 - Plaintext is no longer visible
 - Traffic monitoring capabilities are significantly reduced
- But one should not assume that traffic encryption provides absolute protection
 - e.g., against behavioural analysis
- There are strong incentives to “see” beyond encryption
 - Both for network adversaries and network administrators

Encrypted traffic analysis (ETA)

- Let's look at an encrypted tunnel between Alice and Bob:



Network flows and metadata

- What is a network flow?
 - A flow is typically represented by a five-tuple
 - <Src. IP, Dest. IP, Src. port, Dest. port, Proto>
- One can extract additional metadata tied to a flow:
 - Flow duration
 - Amount of packets exchanged Packet sizes
 - Packet inter-arrival times
 - Payload byte entropy And more...
- What is this good for?

Encrypted traffic analysis (ETA) as a side channel

- Think of ETA as a sort of network side channel!
- ETA can be used to infer information about encrypted traffic
- We'll look at three particular ETA applications for:
 - network analytics
 - network security
 - privacy breaches
- We'll also discuss potential countermeasures

Network Analytics

Network Analytics

- Traffic Engineering

- Prioritize application traffic (e.g., WhatsApp, Skype)
 - e.g., for non-neutral Internet ISPs
- Throttle selected protocols (e.g. BitTorrent)
 - e.g., for “traffic management” purposes

- Quality-of-Service

- Derive quality metrics from encrypted flows
 - e.g. videoconferencing and video streaming QoE
 - e.g. websites' page load time, speed index

Use case: Identification of mobile applications

- Mobile applications' traffic leaves a fingerprint
 - Network observers can understand which apps you are using
- Build a classifier based on summary statistics from each flow
 - Look at the packet size/timing distributions
 - Minimum, maximum, mean, standard deviation, variance, skew, kurtosis, percentiles, etc.
- May need to separate traffic bursts
 - Network packets occurring together within a threshold of time
 - Traffic bursts may encompass multiple flows

Let's classify some apps!

Feature set

Total Packets	Total Bytes	Max Size	Min Size	Mean Size	Std. Dev Size	Percentile 10th	Percentile 90th	CLASS
------------------	----------------	----------	----------	--------------	------------------	--------------------	-------	--------------------	-------

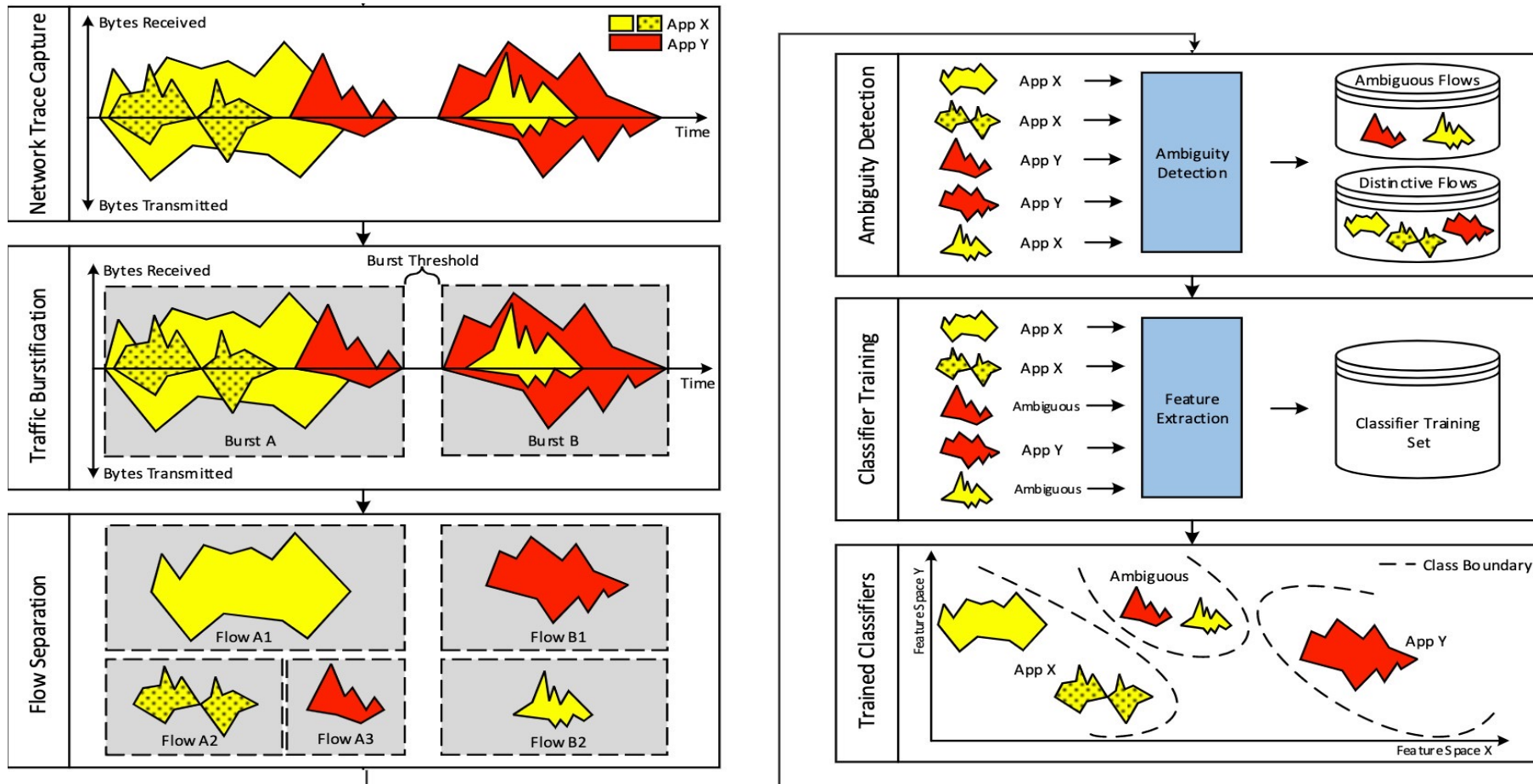
Training data

S_{T1}	1405	123400	980	60	700	43	125	948	Twitter
⋮										
S_{Tn}	1566	134050	1250	60	842	54	143	1014	Twitter
⋮										
S_{I1}	2864	236544	1204	60	1024	64	92	1140	Instagram
⋮										
S_{In}	3264	286458	1280	60	1120	82	104	1220	Instagram

New data sample

	1479	125382	1240	60	792	56	142	1002	???
--	------	--------	------	----	-----	----	-----	-------	------	-----

Use case: Identification of mobile applications



- Taylor et al., IEEE TIFS '17

Use case: Measuring video QoE

- Majority of video traffic is delivered over adaptive bitrate
 - A video is encoded in multiple resolutions and split into chunks of variable length
 - Clients continuously fill a buffer of chunks, where ensuing chunks are based on network conditions
- DPI solutions can no longer be used to extract meaningful QoE metrics
 - e.g., initial delays, playback stalls frequency, resolution switch

Use case: Measuring video QoE (cont)

- Features extracted from encrypted traffic guide the models to detect quality impairments
 - Able to detect stalls, average quality, and video quality adjustments

Network Features	Ground Truth (URI)
minimum RTT	chunk resolution
average RTT	stall count
maximum RTT	stall duration
Bandwidth-delay product	video session ID
average bytes-in-flight	
maximum bytes-in-flight	
% packet loss	
% packet retransmissions	
chunk size	
chunk time	

- Dimopoulos et al., IMC '16

Network Security

Malware Detection

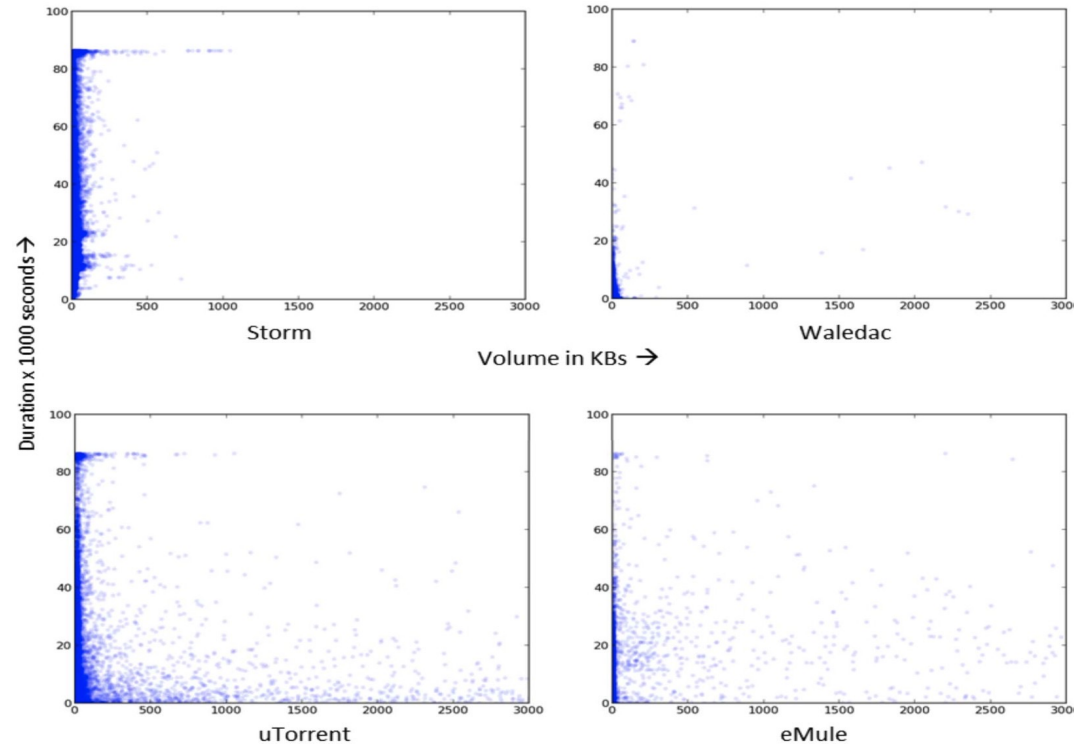
- Traditional network-based malware detection relies on unencrypted data
 - Heavy use of deep packet inspection
 - e.g., for signature-based detection over packet payloads
- No longer useful to detect viruses or data exfiltration
- Encrypted traffic analysis helps us to identify:
 - Malware communications towards C&C servers
 - Unusual network traffic patterns in the network

Malware Detection

- Malware classification:
 - Build a model out of legitimate / malicious network activity
 - Leverage “fingerprints” of legitimate / malicious behaviour
 - What if a **new** malware stream emerges?
- Anomaly detection:
 - Build a model for legitimate traffic and flag strange behavior
 - Via one-class learning or clustering
 - What if legitimate behavior **changes** over time?

Use case: P2P botnet detection

- Can we pinpoint interactions between bots and C&Cs?
 - Tend to be low-volume and long-standing vs. benign P2P apps



Narang et al., IEEE SPW '14

Use case: P2P botnet detection

- Flows

- P2P applications (including botnets) randomize port numbers
- The usual flow definition leads to the generation of multiple flows out of what can be a continued interaction between two peers

- Super-flows

- Aggregate multiple flows between two IPs into a super-flow
 - What if two IPs have benign and malicious flows between them?

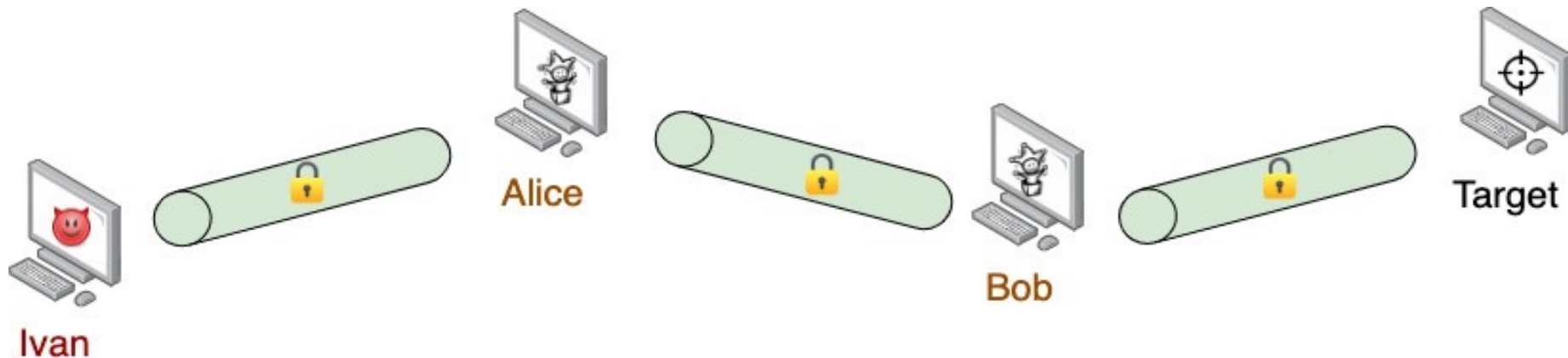
Narang et al., IEEE SPW '14

Use case: P2P botnet detection

- Conversations
 - Start by clustering flows:
 - Protocol, packets per second, avg. payload size
 - Create conversations from flows placed within the same clusters
 - Finally, classify conversations as malicious or benign based on:
 - Duration of the conversation
 - Number of packets exchanged
 - Volume of data exchanged
 - Median of packet inter-arrival times
- This approach was also shown effective for detecting previously unseen botnets!

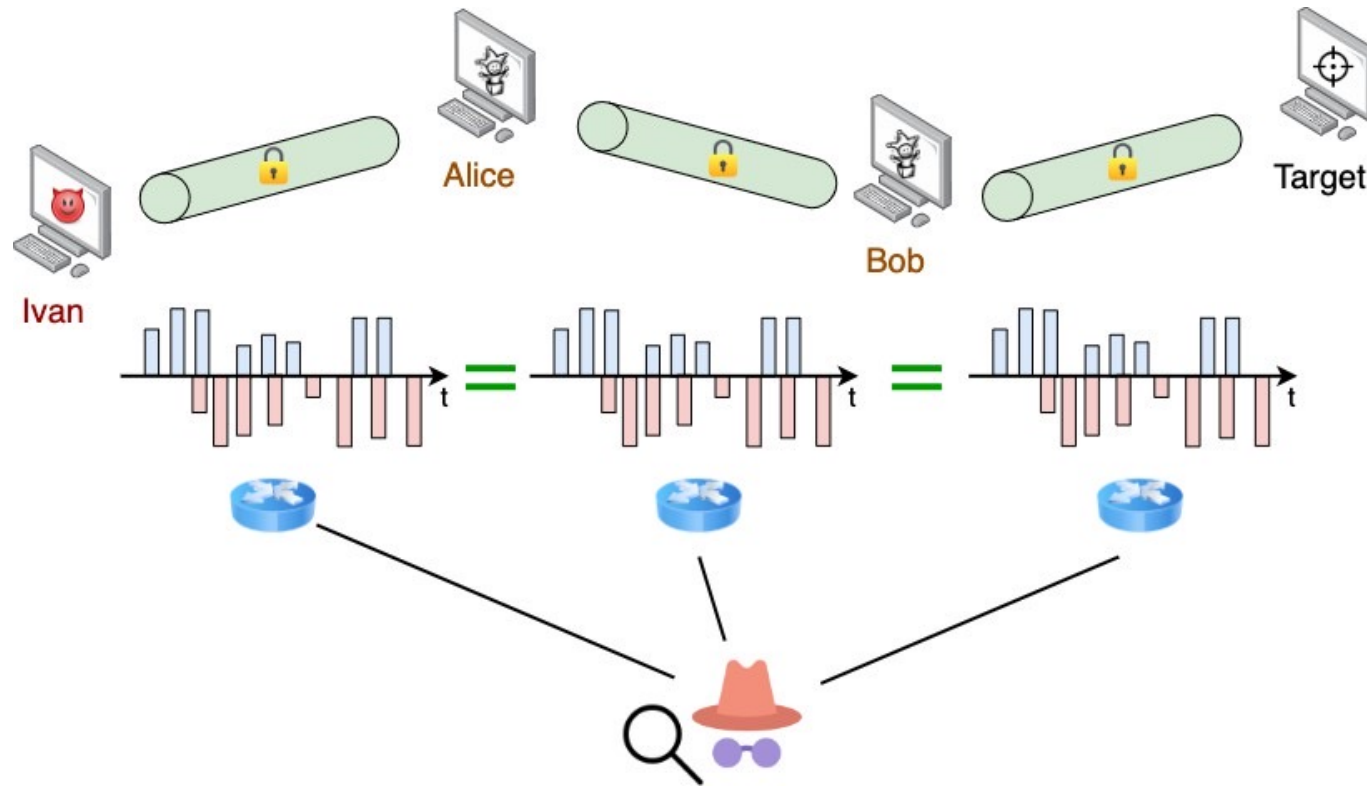
Stepping stones

- An attacker can hide its identity by using other machines as intermediaries (i.e., stepping-stones)
 - e.g., by hopping through compromised machines or by using Tor



Traffic Correlation

- Detection of stepping-stones
 - Attempt to match (roughly) the same sequence of packets at different network vantage points



Difficulties in Performing Traffic Correlation

- In practice, flow observations will not be an exact match
 - Due to network imperfections
 - Packet delays, jitter, loss
- Due to countermeasures
 - Chaff and delay injection at intermediate nodes, padding
- Traffic correlation algorithms must account for small differences between each flow observation

$$\delta_t(C, C') = \log \left(\prod_{k=1}^K |T_k(C', t) - T_k(C, t)| \right)$$

Staniford-Chen and Heberlein, IEEE S&P '95

Privacy Breaches

Nefarious uses of encrypted traffic analysis

- One would assume that encryption is all that is needed to securely communicate over the Internet
- Unfortunately, encryption does not hide traffic patterns
- Traffic analysis can be weaponized to breach users' privacy

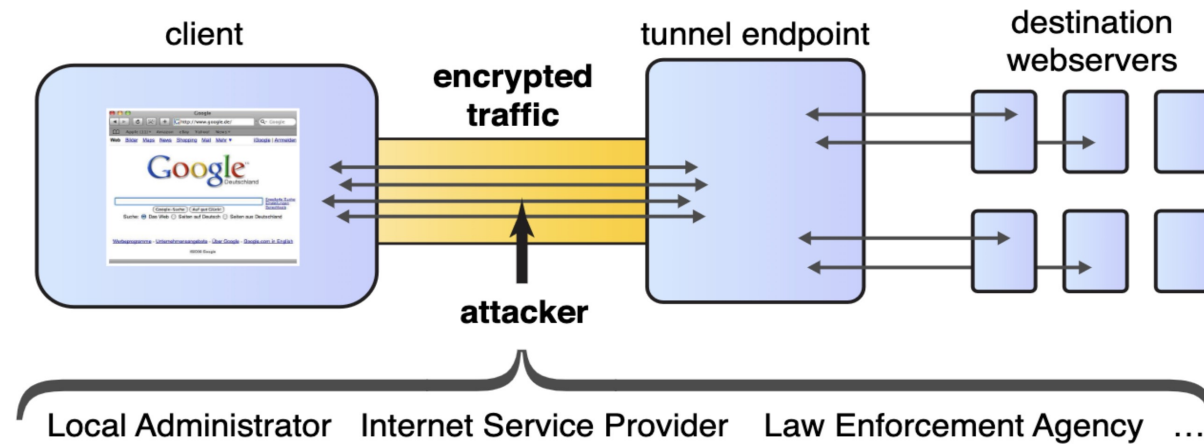
Metadata is not your data. Or is it?



(Dr. Evil making you think metadata is useless)

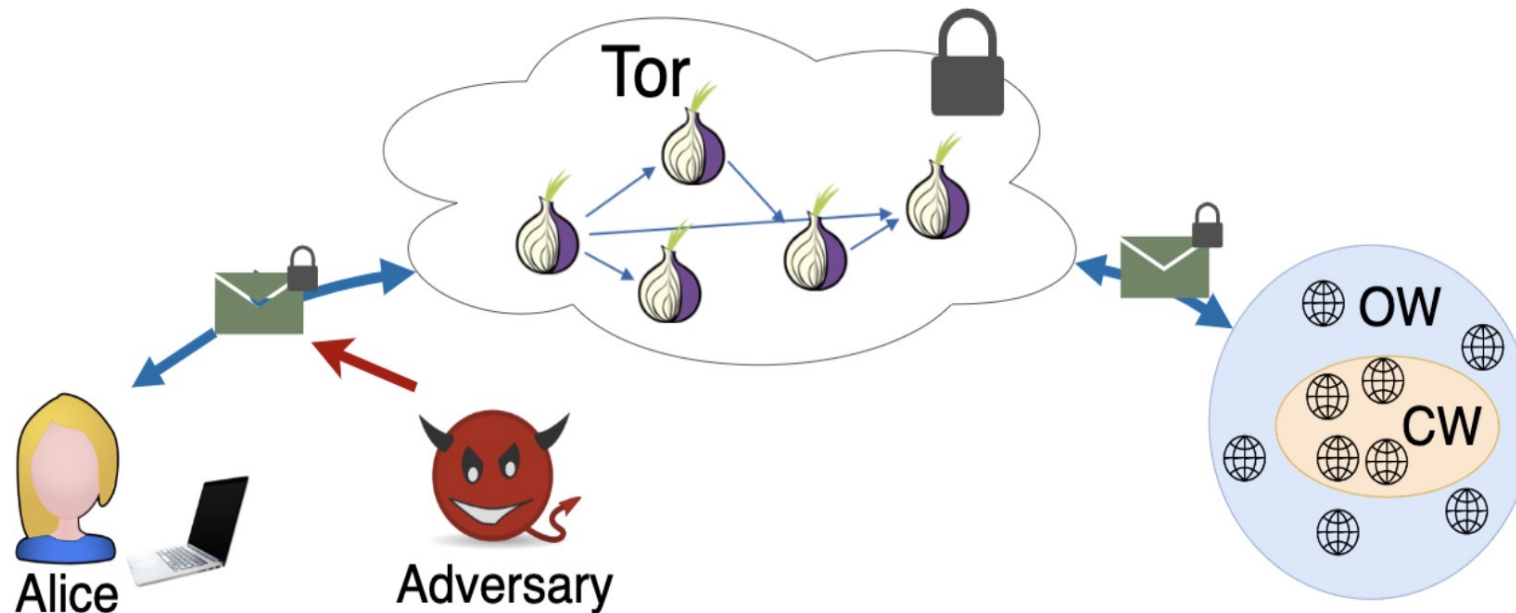
Website fingerprinting over VPNs

- VPNs are advertised as the “holy-grail” of Internet security
 - Passive adversaries can uncover which website is being visited
 - By building traffic fingerprints and using a classifier
- The attack can be launched in two settings:
 - Closed-world Open-world



Website fingerprinting over Tor

- The Tor network can be seen as one “big VPN node”
 - Tor exchanges data in fixed-size cells
 - But packet direction and timing still leaks information



Website fingerprinting over Tor

- Features based on different traffic representations have been used to launch website fingerprinting attacks on Tor
 - Directional representation - Rimmer et al., NDSS '18
 - Directional + timing representation - Saidur Rahman et al., PoPETs '20

Rimmer et al. (Directional representation)

+1	-1	+1	+1	-1	-1	-1	+1	+1	yahoo.com
+1	+1	-1	-1	-1	-1	+1	-1	+1	google.com

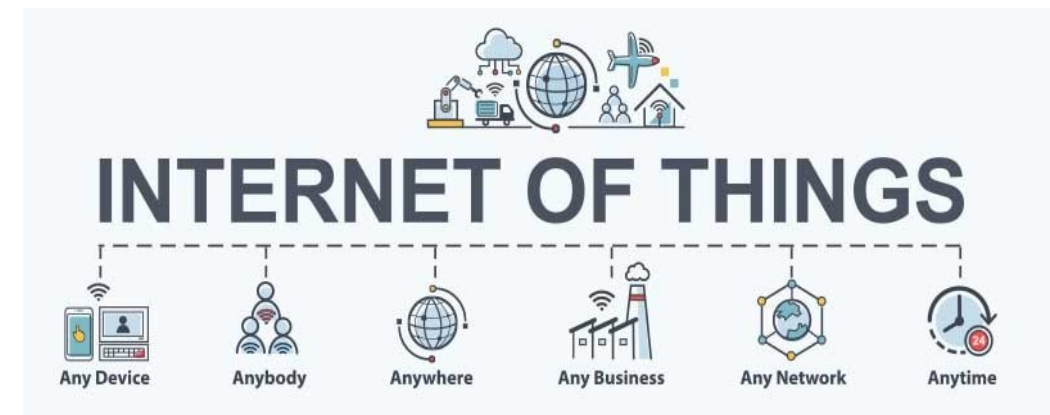
Saidur Rahman et al. (Directional + timing representation)

+0.02	-0.01	+0.03	+0.01	-0.03	-0.04	-0.01	+0.01	+0.02	yahoo.com
+0.01	+0.04	-0.02	-0.01	-0.01	-0.01	+0.02	-0.01	+0.02	google.com

Fixed-size input to neural network

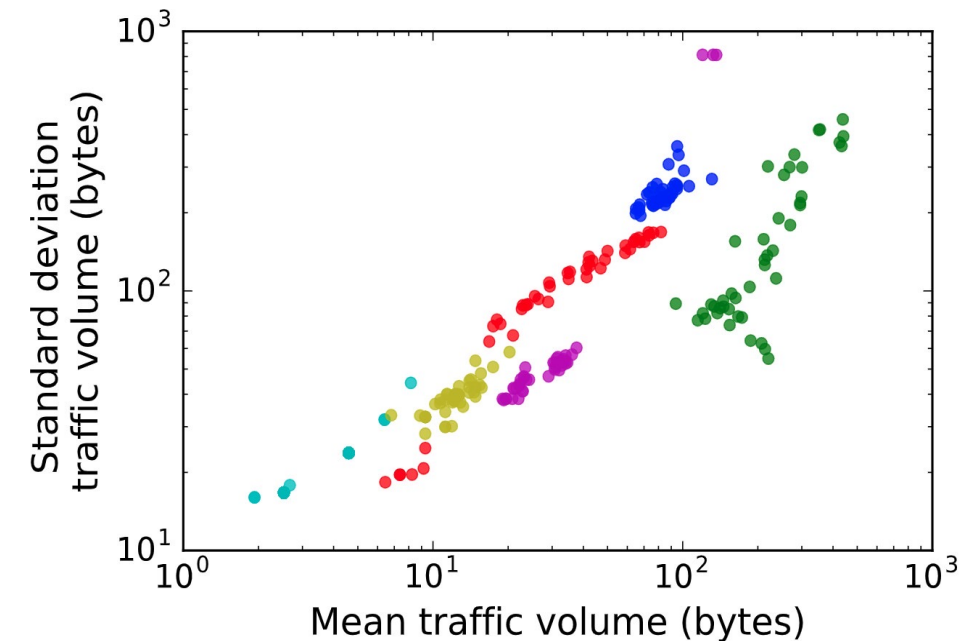
IoT device fingerprinting

- Passive network observers can potentially analyze IoT network traffic to infer sensitive details about users
 - Does this user have a blood monitor? A security camera?
 - DNS queries associated with each encrypted flow often contain the device manufacturer name
 - We can even pinpoint the exact device



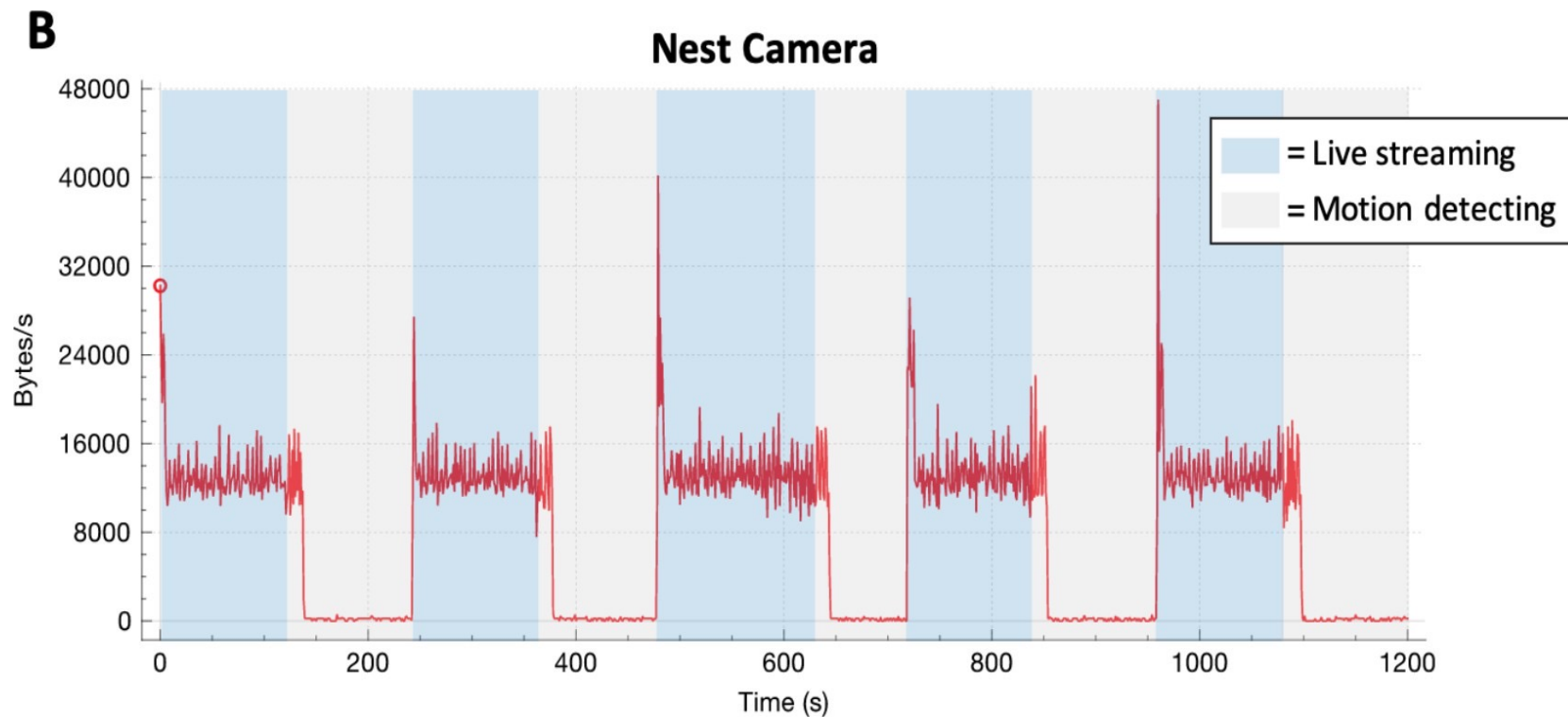
Distinguishing devices through traffic volume

- Rather simple volumetric features allow us to identify IoT devices (Apthorpe et al., ConPro '17)
 - Once a device is identified, one can also infer its state



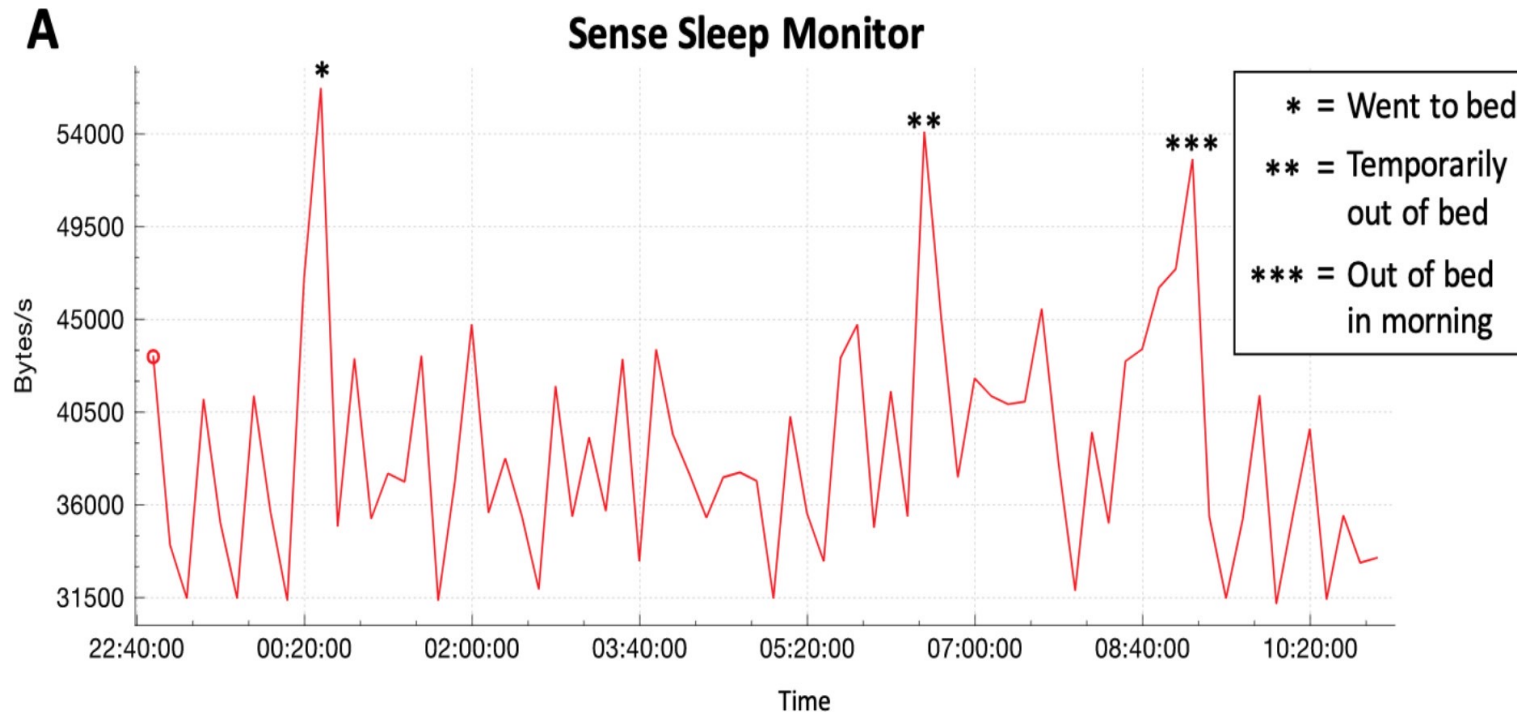
Motion sensor - Nest indoor security camera

- Easy to discern when the camera picks up movement
 - Easy to discern when nobody's home?



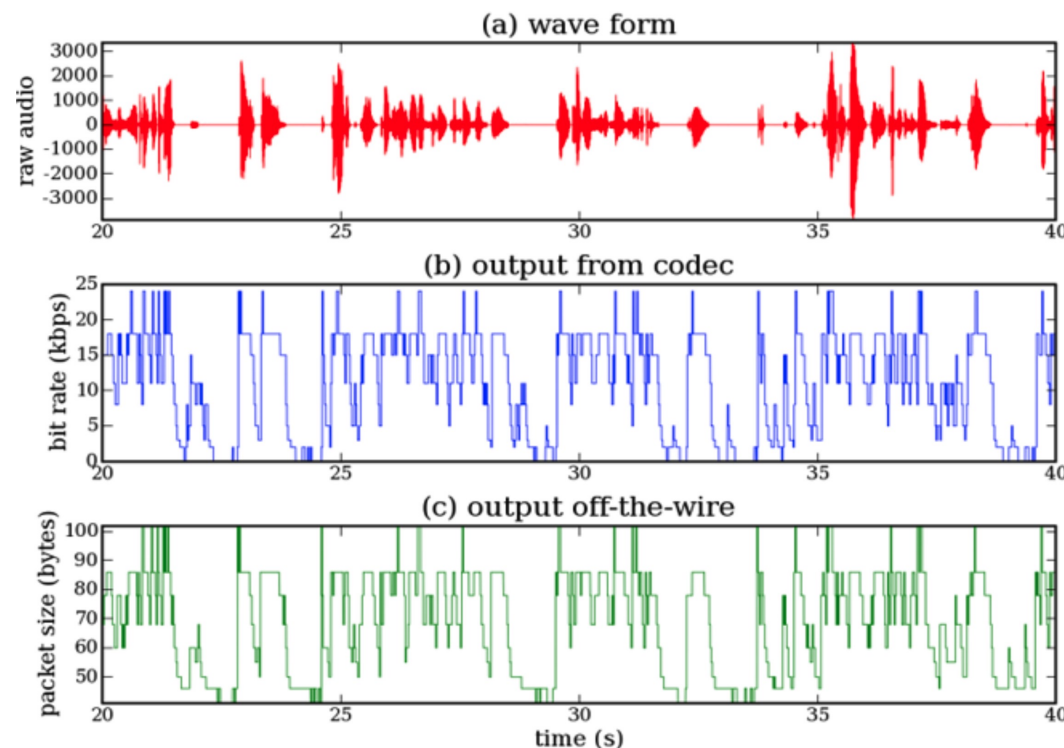
Sleep tracker example - Sense sleep monitor

- Easy to discern when a user goes to bed and wakes-up
 - Easy to discern if a burglar should leave the crime scene?



VoIP eavesdropping

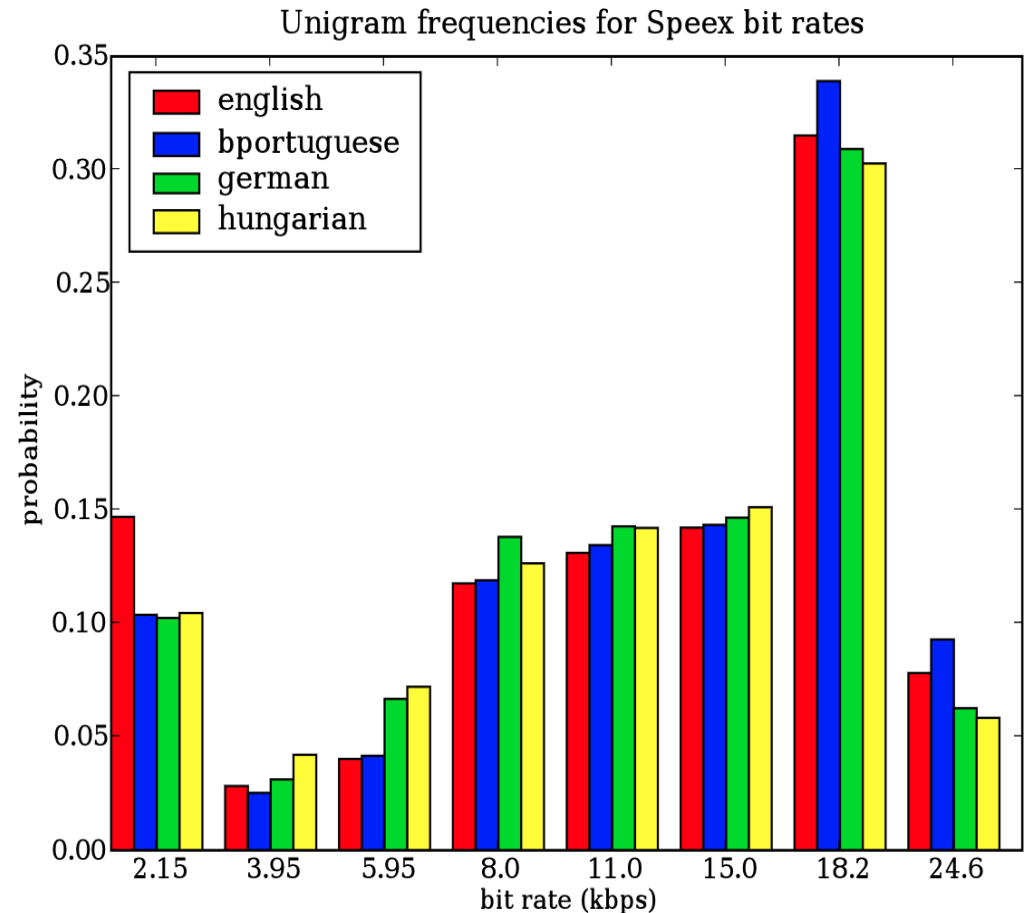
- Encrypted packet patterns resemble VBR codec bitrates
 - Can we infer meaningful semantics from the transmission of encrypted audio frames?



Wright et al., USENIX SEC '07

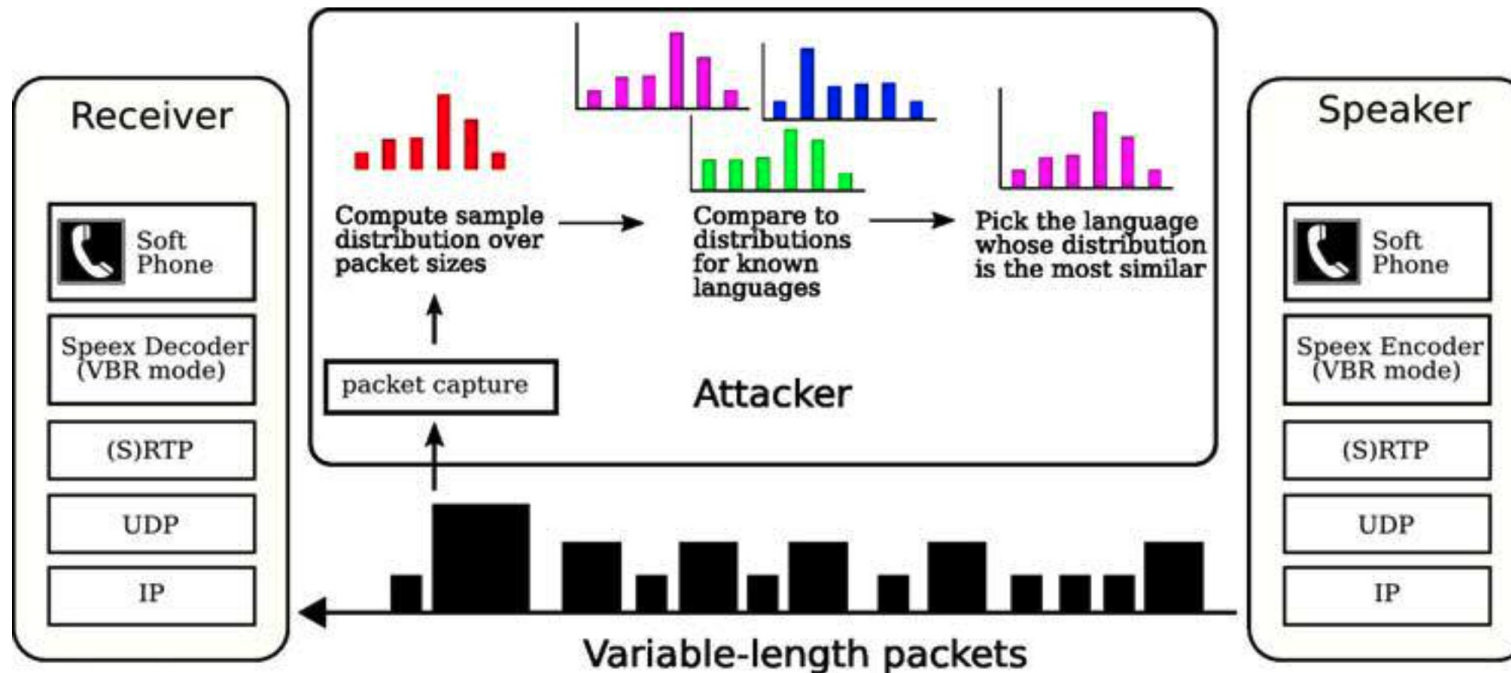
Noticeable (coarse-grained) differences

- Maybe we can identify the language being spoken?
 - Languages have different bitrate frequencies



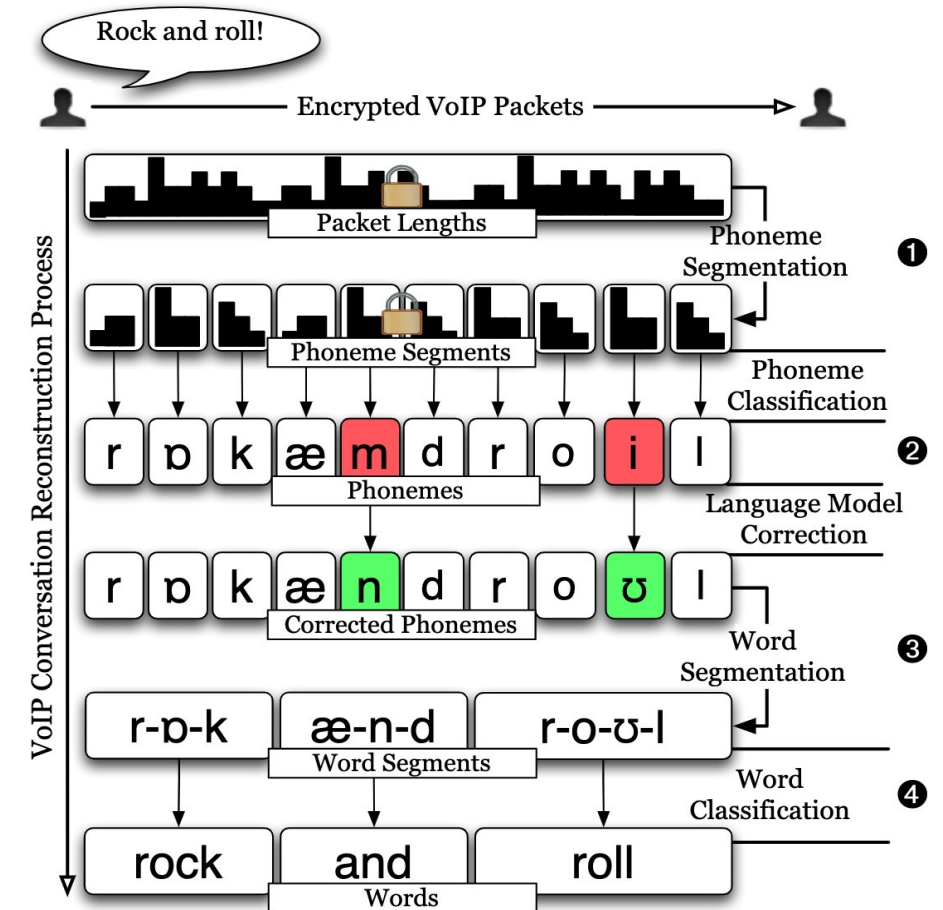
How to distinguish different languages?

- Compute distance between probability distributions
 - Samples from same language have similar distribution
 - Compute packet size n-grams for even better results
 - Given sequence 10, 20, 30, 15 $\rightarrow \{(10, 20), (20, 30), (30, 15)\}$



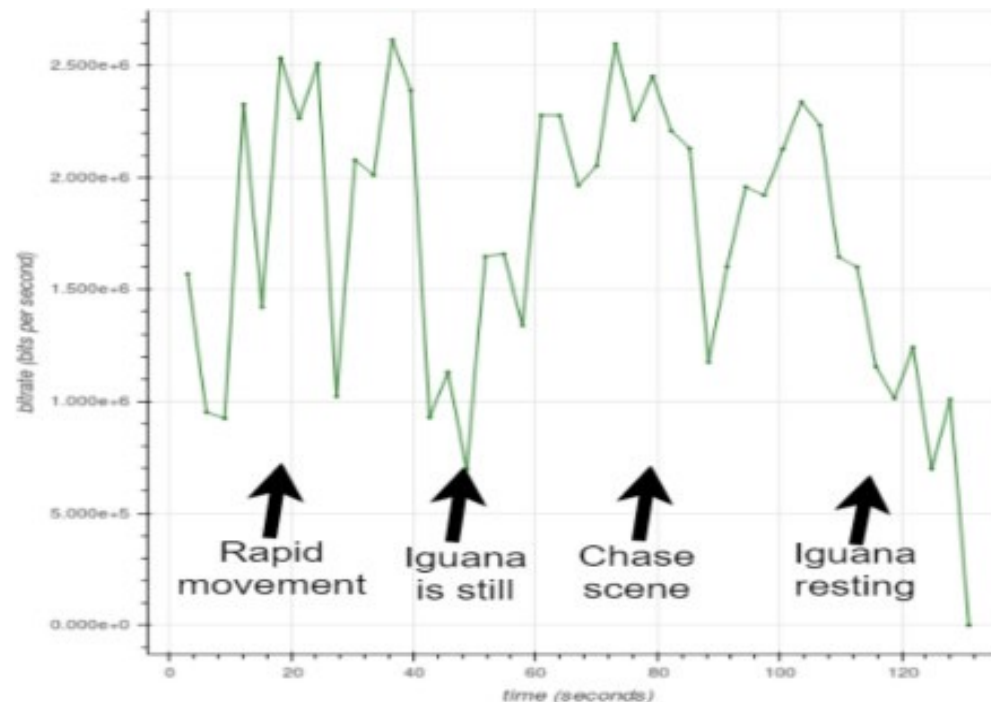
Noticeable (fine-grained) differences

- Can we segment packet size sequences into phonemes?
 - If so, we can recover approximated transcripts



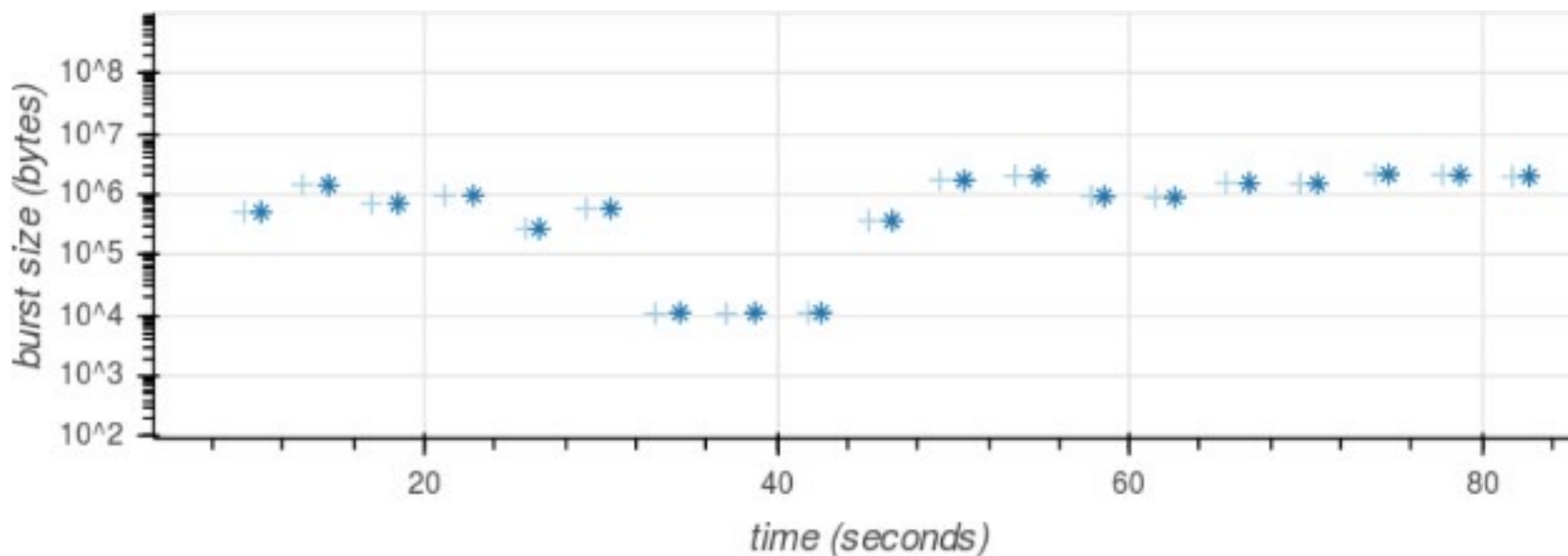
Video re-identification

- At this point, you've probably guessed it, traffic analysis can also be used to uncover which videos you are streaming
 - The bitrate of VBR video sequences also leaks some information



Re-identification of Netflix video streaming

- Burst sizes of a streamed scene of “Reservoir Dogs”
 - Very similar, even when watched over different networks



Schuster et al., USENIX SEC '17

Countermeasures to traffic analysis

- Introduce padding
- Add chaff traffic
- Shape traffic (look like something)
- Aggregate traffic (e.g, multiplex over single connection)
- Split a single connection across multiple networks
- Main trade-off to consider is overhead
 - Achievable throughput
 - Spent bandwidth

Schuster et al., USENIX SEC '17