# Developing a Taxonomy of Message Formats from Network Protocols



## **Annie Li**

# Yijing.Li@tufts.edu

## Jared Chandler

## Jared.Chandler@tufts.edu

# **Kathleen Fisher**

Kathleen.Fisher@tufts.edu

#### **Research Motivation & Goals**

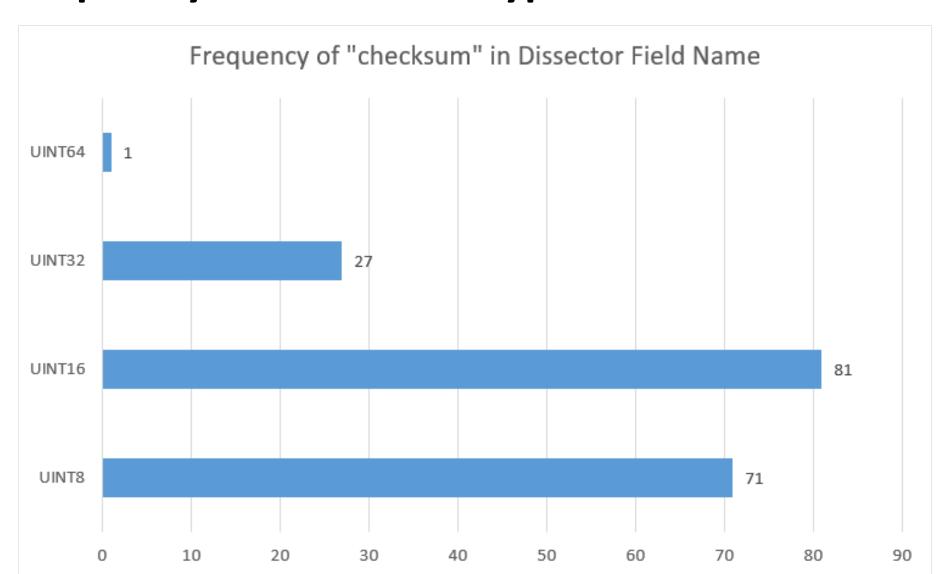
When cybersecurity analysts encounter a new or unknown network protocol it can require substantial human effort to reverse engineer the specification from a network capture. While there has been work on automating portions of this task, we seek to develop an automatic reverse engineering approach guided by common design patterns in extant network protocols. As part of this effort we seek to create a taxonomy of network protocols and the design patterns they exhibit. Our goal for this taxonomy is to create abstract categories of protocols serving as a roadmap for our reverse engineering efforts.

### **Taxonomy Approach: Wireshark Dissectors**

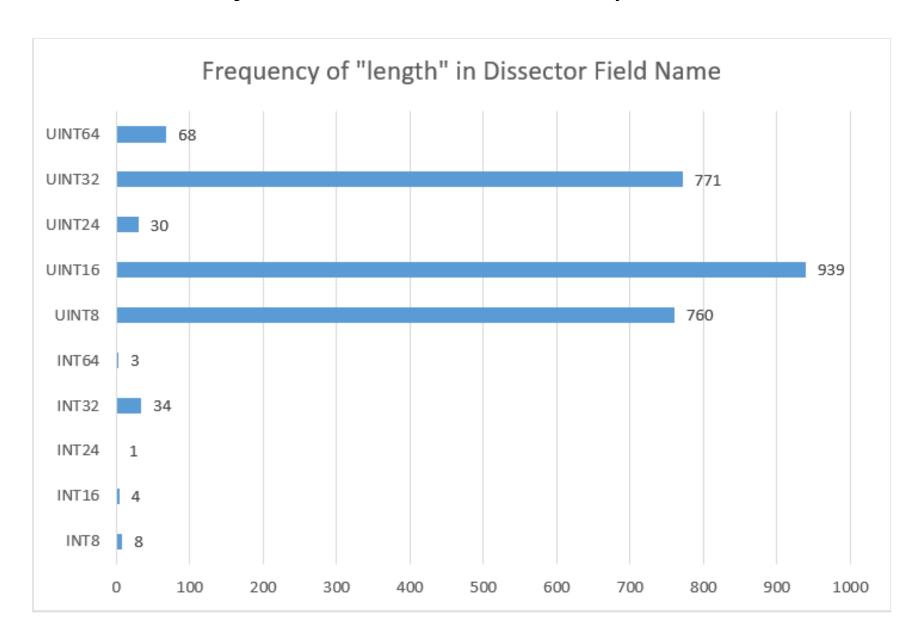
tv-netflix-p	roblems-2011-07-06.pcap							
File Edit Vi	iew Go Capture Ana	alyze Statistics Telepho	ony Wireless	Tools Help				
Apply a display filter <ctrl-></ctrl-> Expression +								
o. Time	Source	Destination	Protocol L	ength Info				
343 65.1	42415 192.168.0.21	174.129.249.228	TCP	_	K] Seg=1 Ack=1 Win=5888 Len=0 TSval=49151934	6 TSecr=551811827		
344 65.1	42715 192.168.0.21	174.129.249.228	HTTP	253 GET /clients/n	etflix/flash/application.swf?flash_version=f	lash lite 2.1&v=1.5&		
345 65.2	30738 174.129.249.22	8 192.168.0.21	TCP	66 80 → 40555 [AC	K] Seq=1 Ack=188 Win=6864 Len=0 TSval=551811	850 TSecr=491519347		
346 65.2	40742 174.129.249.22	8 192.168.0.21	HTTP	828 HTTP/1.1 302 M	oved Temporarily			
	41592 192.168.0.21	174.129.249.228	TCP		K] Seq=188 Ack=763 Win=7424 Len=0 TSval=4915	19446 TSecr=55181185		
	42532 192.168.0.21	192.168.0.1	DNS		0x2188 A cdn-0.nflximg.com			
	76870 192.168.0.1	192.168.0.21	DNS		response 0x2188 A cdn-0.nflximg.com CNAME in			
	77992 192.168.0.21	63.80.242.48	TCP	•	N] Seq=0 Win=5840 Len=0 MSS=1460 SACK_PERM=1			
	97757 63.80.242.48	192.168.0.21	TCP TCP	-	N, ACK] Seq=0 Ack=1 Win=5792 Len=0 MSS=1460 :	<del>-</del>		
	98396 192.168.0.21 98687 192.168.0.21	63.80.242.48 63.80.242.48	HTTP	-	K] Seq=1 Ack=1 Win=5888 Len=0 TSval=49151950. ients/flash/814540.bun HTTP/1.1	2  Secr=3295534130		
	318730 63.80.242.48	192.168.0.21	TCP		K] Seg=1 Ack=88 Win=5792 Len=0 TSval=3295534:	151 TSecr=491519503		
	21733 63.80.242.48	192.168.0.21	TCP		f a reassembled PDU]	131 13001 -431313303		
333 0313	22733 031001212110	132110010121	1.01	TOTAL FIRST DEBINETIES OF	1 d redssembled rboj			
Ethernet Internet	II, Src: Globalsc_00 Protocol Version 4,	(3912 bits), 489 byt :3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na [Reque	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348]	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na <u>[Reque</u> [Time:	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na <u>[Reque</u> [Time: Transa	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na <u>[Reque</u> [Time: Transa > Flags:	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds: ort: 53 (53), Dst Po ] ry response, No error	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s -0.nflximg.com: type	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds: ort: 53 (53), Dst Po ] ry response, No error	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s-0.nflximg.com: type	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds: ort: 53 (53), Dst Po ] ry response, No error	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na [Reque [Time: Transa > Flags: Questi Answer Author Additi > Querie > cdn > Answer > Author	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s -0.nflximg.com: type s itative nameservers 00 35 84 f4 01 c7	:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds: ort: 53 (53), Dst Po  ] ry response, No error A, class IN	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21 34036)	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s-0.nflximg.com: type s itative nameservers 00 35 84 f4 01 c7 00 09 00 09 05 63	1:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po  ]  ry response, No error  A, class IN  83 3f 21 88 81 80 00 64 6e 2d 30 07 6e 66	3b:0a), Dst t: 192.168.0 rt: 34036 (	: Vizio_14:8a:e1 (00 0.21 34036)	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.0334338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s -0.nflximg.com: type s itative nameservers  00 35 84 f4 01 c7 00 09 00 09 05 63 6d 67 03 63 6f 6d	33 3f 21 88 81 80 00 64 6e 2d 30 07 6e 66 00 00 01 00 01 c0 0c.	3b:0a), Dst t: 192.168.0 rt: 34036 (2	: Vizio_14:8a:e1 (00 0.21 34036) : c dn-0.nfl	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s-0.nflximg.com: type s itative nameservers  00 35 84 f4 01 c7 00 09 00 09 05 63 6d 67 03 63 6f 6d 01 00 00 05 29 00	1:3b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po  ]  ry response, No error  A, class IN  83 3f 21 88 81 80 00 64 6e 2d 30 07 6e 66	3b:0a), Dst t: 192.168.0 rt: 34036 (2 r	: Vizio_14:8a:e1 (00 0.21 34036)	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s -0.nflximg.com: type s itative nameservers  00 35 84 f4 01 c7 00 90 00 90 56 36 66 67 03 63 66 64 01 00 00 05 29 00 65 74 66 6c 69 78	33 3f 21 88 81 80 00 A, class IN  83 3f 21 88 81 80 00 64 6e 24 30 07 6e 66 00 00 01 00 01 c0 0c 22 06 69 6d 61 67 65	3b:0a), Dst t: 192.168.0 rt: 34036 (: rt: 34036 (: 00 00 ximg.0 73 67 .netf:	: Vizio_14:8a:e1 (00 0.21 34036) : di c dn-0.nfl com	:19:9d:14:8a:e1)			
Ethernet Internet User Data Domain Na	II, Src: Globalsc_00 Protocol Version 4, gram Protocol, Src P me System (response) st In: 348] 0.034338000 seconds ction ID: 0x2188 0x8180 Standard que ons: 1 RRs: 4 ity RRs: 9 onal RRs: 9 s -0.nflximg.com: type s itative nameservers  00 35 84 f4 01 c7 00 90 00 90 56 36 66 67 03 63 66 64 01 00 00 05 29 00 65 74 66 6c 69 78	33b:0a (f0:ad:4e:00: Src: 192.168.0.1, Ds ort: 53 (53), Dst Po  ]  Ry response, No error  A, class IN  83 3f 21 88 81 80 00 64 6e 2d 30 07 6e 66 00 00 01 00 01 c0 0c 22 06 69 6d 61 67 65 03 63 6f 6d 09 65 64 65 74 00 c0 2f 00 05	3b:0a), Dst t: 192.168.0 rt: 34036 (: rt: 34036 (: 00 00 ximg.0 73 67 .netf:	: Vizio_14:8a:e1 (00 0.21 34036) ? dn-0.nfl dn-0.nfl  .) ".images lix .com.edg	:19:9d:14:8a:e1)			

We use protocol glossary data extracted from the Wireshark network capture and dissection tool as the basis for our taxonomy construction. Wireshark is widely used opensource packet inspection tool used to capture and explore network traffic. Wireshark includes over 800 protocol dissectors which parse network messages into the fields according to the protocol specification. This makes it a useful source of protocol specification data across the protocol ecosystem. Wireshark allows dissector field data to be exported as a data glossary containing information about over 180,000 unique protocol fields. We parse this glossary format and load it into a SQL database for exploration and statistical analysis.

## **Frequency of Field Data Types Across Protocols**



Different data types are used across protocols for the semantic type "checksum." The most frequently used data type is UINT16, which appears 45% of the time. The frequency of UINT32 and UINT64 is significantly lower than UINT16 and UINT8. There are no instances of UINT24 "checksum" or any INT "checksum".



Different data types are used across protocols for the semantic type "length." The most frequently used data type is UINT16, which appears 35.87% of the time. The frequency of UINT24 and UINT64 is significantly lower than the other three UINT types. The frequency of INT type is minimal compared to that of the UINT type.

## Correlation between Checksum and Length Data Types within Wireshark Protocol Dissectors

	"checksum"			
"length"	UINT8	UINT16	UINT32	Total
UINT8	128	95	69	292
UINT16	49	112	40	201
UINT32	19	28	24	71
Total	196	235	133	564

An area of interest for our research are the relationships between field types within a protocol dissector specification. For protocol specifications that have fields with length and checksum in their names, we compare the data types used for each field respectively. In some cases, a dissector has multiple fields which match our criteria. We note that pairs of length and checksum fields occur most frequently for datatypes UINT8 and UINT16. We hypothesize that this may be a reflection of machine register size constraints.

### Going From Names to Semantic Categories

Dissector field name	Dissector field description		
data_checksum	Data Checksum		
length_checksum	Length Checksum		
prev_packet_checksum	Checksum of prev. packet		
cryptoCheckSum	cryptoCheckSum		
message_md5_checksum	Message MD5 Checksum		
lm_cksum	Lm Cksum		
ctl_cksum	ITDM Control Message Checksum		

One challenge is assigning meaningful semantic categories to protocol fields. Wireshark dissectors are written by various members of the open-source community. As a result, semantically meaningful information is written in different ways by different people. An example is the "checksum" semantic field type. Authors of dissectors label field names and the corresponding human readable descriptions with many variants of the word "checksum" as shown in the table above. Our future work will include developing methods to infer and decide membership on such semantic categories.