Protokoll-Erkennung

Andreas Floß

Seminar: Internet Measurement

Berlin, den 30. Juli 2007

Technische Universität Berlin
Fakultät IV – Elektrotechnik und Informatik
Intelligente Netze und Management verteilter Systeme (INET)
Research Group Prof. Anja Feldmann, Ph.D.
An-Institut Deutsche Telekom Laboratories

Einleitung

- Thema: Analyse von *Netzwerkverkehr* in einem IP-Netzwerk
- Suche nach Art der Applikation (Browser, Email, Chat, Filesharing) mittels Header-Informationen der Netzwerk-Pakete
- Rückschlüsse auf *Protokolle* durch Verbindungsdaten (Ziel- und Quell-*Ports*)

Einleitung/Motivation

- Aber: Skype, Bittorrent und andere P2P-Anwendungen:
  - zufällig ausgewählte Ports
  - Nutzen bekannter Ports anderer Protokolle (Port Tunneling)
- Analyse der *Nutzdaten* (Payload) der Applikation nach Mustern
- Nutzen *statistischer* und *struktureller Methoden*

Überblick

- IP, TCP & UDP, Session & Flow
- Modell & Konstruktion des Protokoll & Classifier

- Methode 1: *Product Distribution*
- Methode 2: *Markov Process Modell*
- Methode 3: *Common Substring Graph*

- Ergebnisse
- Zusammenfassung

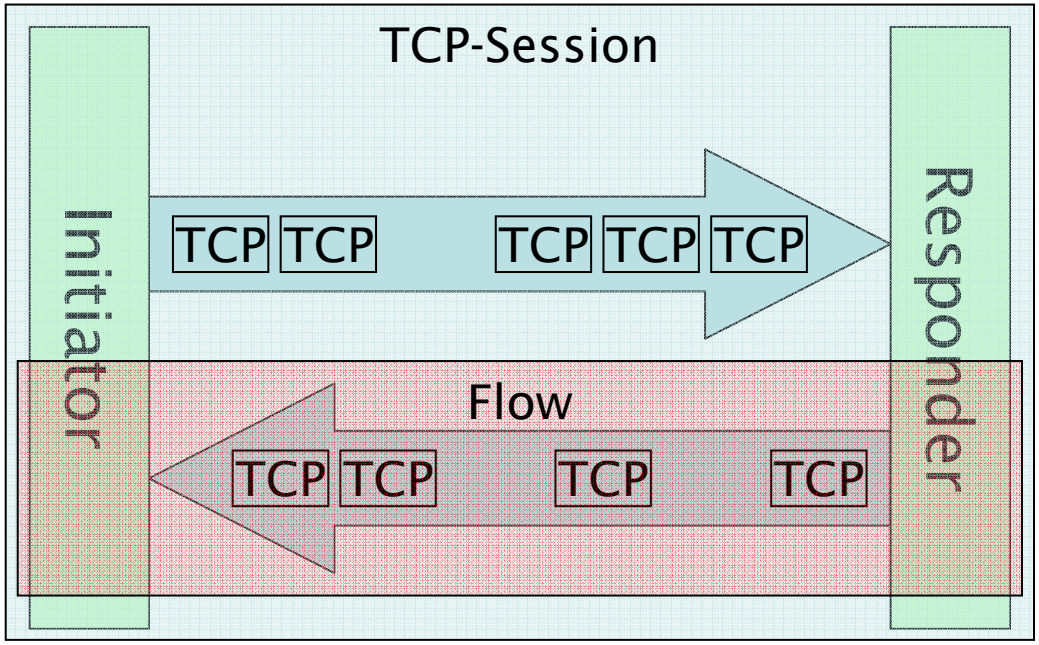
IP, TCP & UDP

- *Transmission Control Protocol*:
  - verlässlicher und sparsamer Transport der *Segmente*
- *User Data Protocol*:
  - min. Protokoll, nur IP-Addressen und Ports
- *Internet Protocol*:
  - organisiert den Weg der Netzwerkpakete

OSI-Schicht		Einordnung	Standard	DoD-Schicht	Protokollbeispiel	Einheiten
7	Anwendung (Application)	Anwendungs-orientiert	FTAM	Anwendung	HTTP FTP HTTPS LDAP NCP	Daten
6	Darstellung (Presentation)		ASN.1			
5	Sitzung (Session)		ISO 8326			
4	Transport (Transport)	Transport-orientiert	ISO 8073	Transport	TCP UDP SPX	Segmente
3	Vermittlung (Network)		CLNP	Internet	ICMP IGMP IP IPX	Pakete
2	Sicherung (Data Link)		HDLC	Netzzugang	Ethernet Token Ring FDDI ARCNET	Rahmen (Frames)
1	Bitübertragung (Physical)		Token Bus			Bits

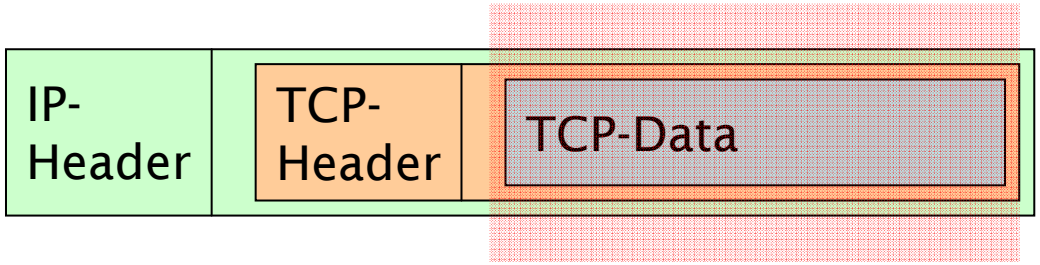
Session & Flow

- Session Data: Byte-Strom zwischen Initiator und Responder
- Flow Data: Byte-Strom in nur eine Richtung



Protokoll

- *Flow Content*:
  reine Daten
  ohne Header



- Protokoll Modell: 2 Verteilungen über die ersten 64 Bytes beider Flows einer Session
- 64 Byte reichen aus
- Menge gleicher Sessions werden a priori zusammengefasst und ergeben den Classifier

M1: Product Distribution Modell

- Idee: Sessions gleicher Protokolle haben ähnliche Byte-Vorkommen an fixen stellen in den Flows
- Annahme: *n* Bytes der Flows sind unabhängig voneinander (independent assumption)
- Verteilungen aller bekannten *n* Bytes ergeben einen Classifier
- bei unbekannten Sessions wird die Wahrscheinlichkeit der ersten n Bytes eines Flows überprüft

M1: Product Distribution Beispiel

Beispiel SSH: Initiator - Flow

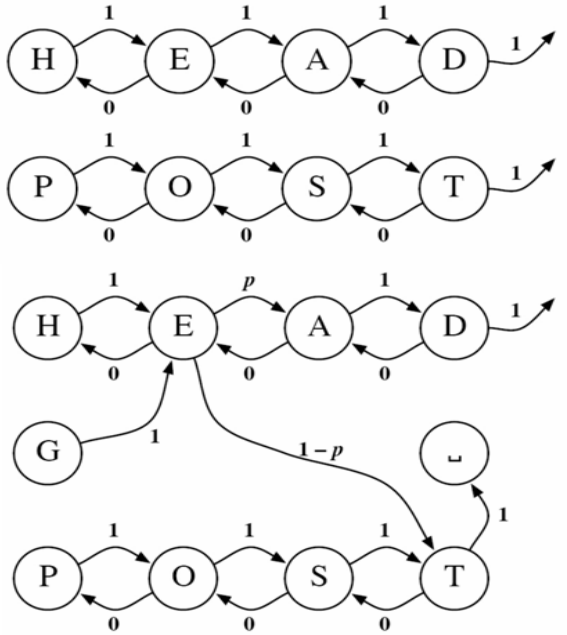
0	1	2	3	4	5	6	7	8	9	10
S	**S**	**H**	**-**	**2** 84 **1** 16	**.**	**0** 84 **9** 16 **5** <1	**-** 84 **9** 16	**O** 53 **-** 16 **P** 14 **S** 10 **h** 5	**p** 53 **3** 15 **u** 14 **e** 8 **t** 5	**e** 53 **T** 16 **.** 16 **c** 8 **t** 5

11	12	13	14	15	16	17	18	19
n 53 **2** 15 **T** 14 **u** 8 **p** 5	**S** 53 **.** 15 **Y** 14 **r** 8 **:** 5	**S** 53 **9** 10 **e** 8 **_** 7 **-** 7	**H** 53 15 **R** 14 **C** 8 **/** 5	**_** 53 **S** 15 **e** 14 **R** 8 **␣** 3	**3** 35 **4** 18 **S** 15 **l** 14 **T** 8	**.** 53 **H** 15 **e** 14 **_** 8 **␣** 3	**8** 28 15 **a** 14 **2** 9 **5** 9	**.** 40 **s** 16 **S** 15 **0A** 12 **p** 10

- SSH (RFC 4253) diktiert: "*… both sides MUST send an identification string. This identification string MUST be* ***SSH-****protoversion-softwareversion* ***SP*** *comments* ***CR LF***."

M2: Markov Process Modell

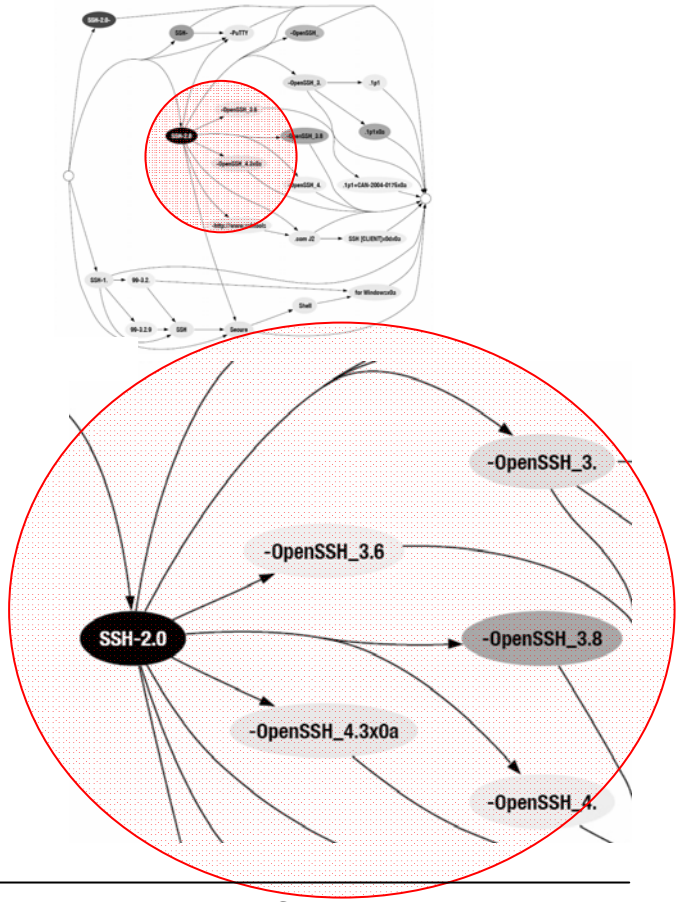
- Idee: Sessions gleicher Protokolle bilden zu hoher Wahrscheinlichkeit ähnliche Folgen von Bytes
- Annahme: n Bytes der Flows sind unabhängig voneinander
- Bilden des Classifiers für ein Protokoll mittels Markov-Ketten
- Bei unbekannten Sessions wird die Wahrsch.-keit der ersten 64 Bytes bezüglich des Classifiers überprüft



Markov-Ketten: Betrachten der Änderungswahr-scheinlichkeiten zwischen Byte-Paaren

M3: Common Substring Graphs

- Idee: Sessions gleicher Protokolle besitzen genügend viele gemeinsame Teil-Strings
- Bilden des Classifier-Graphen mit strukt. Informationen Häufigkeit und Wiederkehr gemeinsamer „längster gemeinsamer“ Teil-Strings
- Bei unbek. Sessions werden die ersten n-Bytes eines Flows analysiert und mit dem Graphen verglichen



Ergebnisse: Fehlklassifizierung

	Cambridge			Wireless			Departmental		
	total	learned	unlearned	total	learned	unlearned	total	learned	unlearned
Product	1.68%	0.50%	1.18%	1.78%	1.28%	0.51%	4.15%	3.03%	1.12%
Markov	3.33%	2.15%	1.18%	4.26%	3.75%	0.51%	9.97%	8.85%	1.12%
CSG	2.08%	0.90%	1.18%	4.72%	4.21%	0.51%	6.19%	5.06%	1.12%

- Fehlklassifizierung
  - Netzwerk-Datensätze (Uni, Drahtlos, Büro)
  - Niedrigste Fehlerrate: Product Distribution

Ergebnisse: Erkennung

Protocol			Product			Markov			CSG		
		%	Err.%	Prec.%	Rec.%	Err.%	Prec.%	Rec.%	Err.%	Prec.%	Rec.%
①	DNS	26.28	0.09	99.94	99.78	0.61	97.89	99.97	0.45	98.82	99.52
	HTTP	12.24	0.07	100.00	99.99	0.09	100.00	99.98	0.74	99.91	99.99
	NBNS	44.89	0.35	100.00	99.25	0.40	99.82	99.31	0.17	99.71	99.99
	NTP	5.29	0.00	100.00	100.00	1.19	99.96	77.84	0.25	99.83	95.65
	SSH	0.22	0.14	68.39	100.00	1.10	17.39	100.00	0.05	99.22	100.00
②	DNS	23.14	0.04	99.88	99.93	0.29	98.88	99.99	1.97	94.37	97.59
	HTTP	0.67	0.27	76.02	97.54	0.09	90.68	99.93	0.22	76.87	99.38
	NBNS	6.94	0.00	100.00	100.00	1.96	78.06	100.00	0.81	90.34	99.97
	NTP	0.57	0.01	99.95	99.72	0.51	100.00	11.29	0.40	86.65	48.76
	SSH	0.44	0.17	75.28	100.00	0.00	99.63	100.00	0.00	99.99	100.00
③	DNS	54.78	0.26	99.90	99.95	1.90	97.13	99.98	1.43	98.47	99.15
	HTTP	9.17	0.38	97.46	99.62	0.33	97.21	99.72	1.21	95.14	97.19
	NBNS	7.03	0.01	100.00	99.81	1.25	85.66	99.81	0.33	96.04	99.45
	NTP	6.70	0.02	99.99	99.94	5.39	78.07	29.61	0.36	99.82	96.58
	SSH	0.08	0.08	68.81	81.82	0.09	0.00	0.00	0.03	95.40	82.01

(1) = Cambridge (226,046 flows), (2) = Wireless (403,752 flows),
(3) = Departmental (1,064,844 flows)

Zusammenfassung

- Identifikation von Protokollen ohne Header-Daten ist möglich:
  - Gesamt-Sieger: Product Distribution Model – statistischer Vergleich der Bytes in den Nutzdaten,
  - Markov Process Model - Auswertung der Byte-Übergänge
  - Common Substring Graphs: Teil-Strings in den Nutzdaten
- alle 3 Verfahren haben bez. einzelner Protokolle Fehler und schlechte Erkennungsraten
- Ursachen:
  - Charakteristiken der Protokolle (Binär-Daten oder Text)
  - Modellierung von Protokollen nicht optimal

Ausblick

- Weitere Verbesserung des Protokoll-Modells
  - kompaktere Repräsentation der Protokolle
  - bessere Unterscheidbarkeit

Fragen

Das war´s -

Vielen Dank!

Backup

First Half w/o HTTP				Second Half		
Cum. %	Ind. %	Protocol	Dist.	Protocol	Ind. %	Cum. %
0.49	0.49	Slammer	0.000	Slammer	0.43	0.43
1.07	0.58	ISAKMP	0.250	ISAKMP	0.44	0.87
9.01	7.93	NBNS	0.300	NBNS	7.12	7.99
9.66	0.65	TFTP	0.300	TFTP	0.42	8.41
70.52	60.87	DNS	0.399	DNS	56.46	64.87
71.37	0.85	SMB	0.595	SMB	0.72	65.59
71.40	0.03	SSDP	0.616	SSDP	0.03	65.62
72.22	0.82	SNMP	1.235	SNMP	0.80	66.42
76.22	4.00	SMTP	1.315	SMTP	4.00	70.41
76.52	0.30	SMB	1.548	SMB	0.27	70.68
77.02	0.50	DCERPC	2.011	DCERPC	0.47	71.15
77.12	0.10	SNMP	4.166	SNMP	0.09	71.24
78.08	0.96	BROWS.	4.168	BROWS.	0.82	72.06
78.15	0.08	Mssgr.	4.551	Mssgr.	0.43	72.49
78.47	0.32	KRB5	4.867	KRB5	0.29	72.78
79.39	0.92	DHCP	4.972	DHCP	0.78	73.56

Backup

First Half w/o HTTP				Second Half		
79.52	0.13	LDAP	5.136	LDAP	0.11	73.67
86.39	6.87	SSL	5.900	SSL	5.72	79.39
86.47	0.08	SSL	6.127	SSL	0.04	79.43
87.39	0.91	YPSERV	6.509	YPSERV	0.82	80.25
87.49	0.11	SRVLOC	6.785	SRVLOC	0.09	80.35
89.00	1.50	POP	7.024	POP	1.24	81.58
89.19	0.19	SSL	8.136	SSL	0.06	81.65
89.49	0.30	SNMP	13.321	SNMP	0.28	81.93
89.55	0.06	SSH	15.871	SSH	0.07	82.00
89.60	0.05	KRB5	16.613	KRB5	0.03	82.03
89.64	0.04	IMAP	18.535	IMAP	0.04	82.08
89.85	0.20	CLDAP	19.496	CLDAP	0.15	82.23
89.91	0.06	Syslog	24.436	Syslog	0.06	82.29
98.01	8.10	NTP	38.452	NTP	6.84	89.13
98.16	0.15	NFS	44.493	NFS	0.04	89.18
99.52	1.36	SNMP	44.510	SNMP	1.23	90.40
99.58	0.06	RTSP	131.352	**HTTP**	9.46	99.87
99.88	0.30	SNMP	312.578	*RIPv1*	0.03	99.89
100.00	0.12	DAAP	470.046	*RADIUS*	0.11	100.00