Přenos dat, počítačové sítě a protokoly

Identifikace síťového provozu, důvěra na Internetu

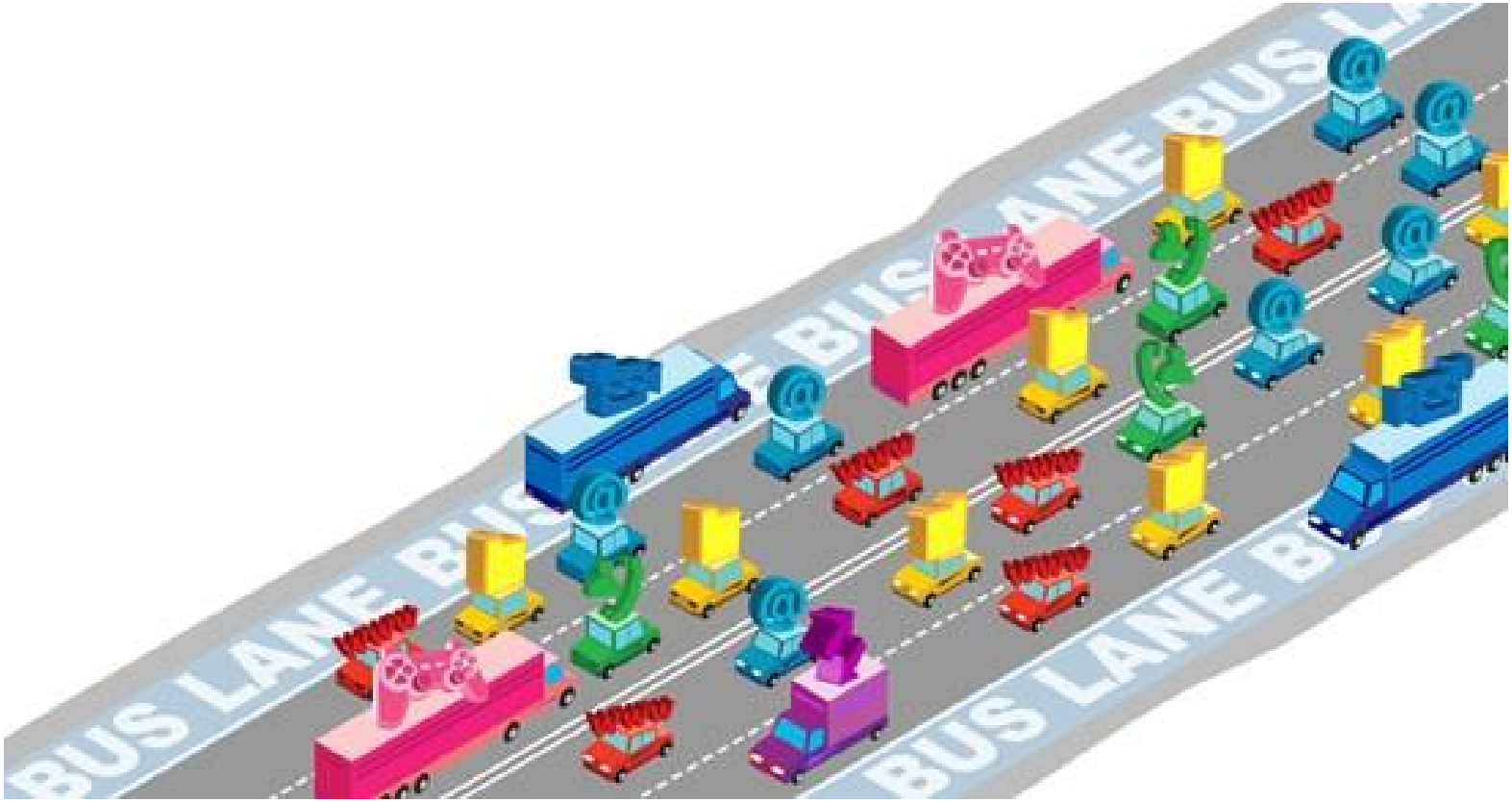
Ing. Petr Matoušek, PhD., M.A.

Fakulta informačních technologií VUT v Brně

matousp@fit.vutbr.cz

Co se děje na Internetu?

Proč nás vůbec zajímá, co se přenáší po síti?



- *Bezpečnost*: detekce útoků, nedovolené chování uživatelů a aplikací
- *Diagnostika*: správné chování systémových i uživatelských aplikací
- *Rozlišení služeb*: nastavení kvality služeb, prioritizace kritických přenosů
- *Vytížení sítě*: sledování rozložení provozu, vytížení síťových prvků a služeb
- *Účtování služeb*: identifikace provozu VoIP, IPTV, apod.

Jak lze identifikovat síťový provoz?

1. Typická data z hlaviček paketů
2. Identifikace podle signatury v obsahu paketů (DPI)
3. Statistické chování toků

Obsah přednášky

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

1. Identifikace provozu
   - Podle hodnot z hlaviček paketů
   - Vyhledávání signatury
   - Statistické chování toků
2. Důvěra na Internetu
   - Důvěra a reputace
   - Reputační systémy
3. Otázky a úkoly
4. Literatura

1. Identifikace podle hodnot z hlaviček paketů

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Jaké hlavičky protokolů lze použít?

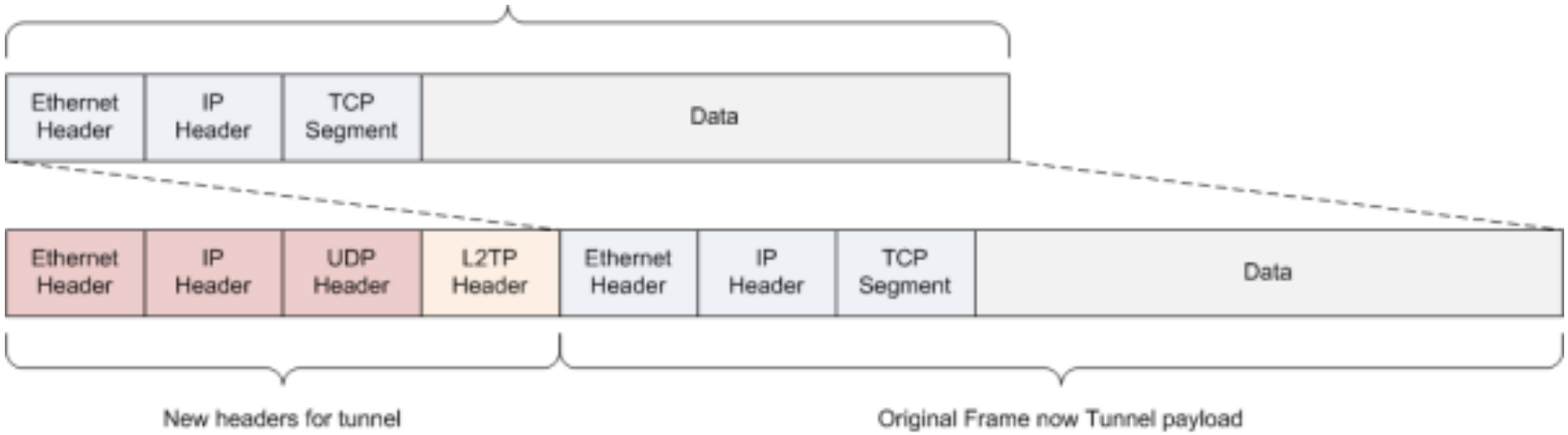
- Hlavička Ethernetu (`EtherType`)
  - IPv4 (0x0800), ARP (0x0806), RARP (0x8035), IPv6 (0x86DD), PPP (0x880B), MPLS (0x8847), 802.1x (0x888E), 802.1q (0x88A8), ...
- Hlavička IP (`protocol`)
  - ICMP (1), IPv6 (41), IGMP (2), UDP (17), TCP (6), SCTP (132), GRE (47), ESP (50), AH (51), RSVP (46), EIGRP (88), OSPF (89), PIM (103), L2TP (115) ...
- Hlavička TCP/UDP (`Source Port, Destination Port`)
  - FTP (20,21), SSH (22), Telnet (23), SMTP (25), DNS (53), DHCP (67, 68), TFTP (69), HTTP (80), POP3 (110), NNTP (119), NTP (123), IMAP (143), SNMP (161,162), BGP (179), LDAP (389), HTTPS (443) ...
- Hlavička aplikačního protokolu
  - HTTP: "GET / HTTP/1.1"
  - BitTorrent: "0x13BitTorrent protocol"

1. Identifikace podle hodnot z hlaviček paketů

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Tento přístup ne vždy funguje ...

- *Komunikující aplikace nemusí být pevně svázaná s číslem portu*
  - Dynamické porty: sítě P2P, protokoly RTP, RPC, apod.
- *Tunelování síťového provozu: zapouzdření v jiném protokolu*
  - Virtuální privátní sítě: GRE, IPSec, SSL/TLS, L2TP, T2F, PPTP, apod.
  - Tunelování přes port 80: přenos souborů, IM, e-mail, hlas, video
  - IPv6 nad IPv4: 6in4, 6over4, 6to4, ISATAP, Teredo, 6rd, viz RFC 7059
- *Skryté kanály*
  - Tunelování protokoly ICMP, DNS, apod.



2. Identifikace síťového provozu podle signatury

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Signatura (pattern, signature) = statický řetězec tvořený posloupností znaků z hlavičky či těla protokolu, který slouží k rozpoznání daného protokolu.

Jak určit signaturu?

- Ze specifikace protokolu (standardy ISO, IETF, IEEE, apod.)
- Hledáme typické řetězce vyskytující se v hlavičce či těle daného protokolu
  - *Pevný začátek výskytu* (fixed offset), např. "0xE3" na pozici 0 (eDonkey)
  - *Pohyblivý začátek výskytu* (variable offset): řetězec se může vyskytovat na libovolné pozici v paketu
  - Signatura se skládá z *posloupnosti podřetězců* (common substrings)

Někdy to ale nejde ...

- Proprietární protokoly: reverzní inženýrství
- Stabilita: nové verze protokolů, nové protokoly
- Přenosy používající šifrování
- Problém signatur rozložených do více paketů: analýza toků

2. Identifikace síťového provozu podle signatury

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Použití signatury pro identifikaci provozu

1. Vytvoření databáze signatur
2. Hledání signatury v procházejících paketech

Příklad identifikace aplikačních protokolů podle signatury: L7 filter

- Softwarový klasifikátor pro Unix, viz l7-filter.clearos.com
- Využívá regulárních výrazů pro klasifikaci aplikačního protokolu
  - Databáze obsahuje 122 vzorků protokolů, viz http://l7-filter.sourceforge.net/protocols
- Typy protokolů
  - Aplikace P2P (BitTorrent, Kazaa, eDonkey, Gnutella, Napster, Freenet)
  - Přenos audio (RTP), VoIP (H.323, SIP, Skype), Chat (OSCAR,IRC,Yahoo)
  - Streamování (RTSP, PPLive), síťové hry
  - Vzdálený přístup (VNC, X), anonymizace (ToR)
  - Sdílení souborů (OpenFT, Apple Juice), synchronizace,

2. Identifikace síťového provozu podle signatury

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

L7-filter: příklady filtrů

- HTTP, RFC 2619

```
// Status-Line = HTTP-Version SP Status-Code SP Reason-Phrase CRLF
http/(0\.9|1\.0|1\.1) [1-5][0-9][0-9] [\x09-\x0d-~]*(connection:
|content-type:|content-length:|date:)|post [\x09-\x0d -~]* http/[01]\.[019]
```

- Real Time Protocol, RFC 3550

```
// version 2, Payload Type 0-34,96-127
^\x80[\x01-"‘-\x7f\x80-\xa2\xe0-\xff]?.........*\x80
```

- BitTorrent

```
^(\x13bittorrent protocol|azver\x01$|get /scrape\?info_hash=get
/announce\?info_hash=|get /client/bitcomet/|GET /data\?fid=)
|d1:ad2:id20:|\x08’7P\)[RP]
```

- SSL v.3, RFC 2246

```
// SSL Hello with certificate, Client Hello
^(.?.?\x16\x03.*\x16\x03|.?.?\x01\x03\x01?.*\x0b)
```

2. Identifikace síťového provozu podle signatury

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Omezení při použití signatur

- Definice signatury vyžaduje expertní znalost
  - Sémantická analýza daného protokolu
  - Empirické zkoumání obsahu paketu a hledání signatury
- Potřeba vytvářet a udržovat aktuální databázi signatur
- Časově i výkonnově náročné vyhledávání signatur v těle procházejících paketů

Je možné proces vytváření signatur automatizovat?

$\Rightarrow$ Ano, hledáme nejdelší společný podřetězec z paketů daného protokolu.

Jak na to?

Algoritmus pro automatické generování signatur [1]

- Získání signatury bez dopředné znalosti protokolu.
- Použití algoritmus LCS (Longest Common Subsequence) pro hledání podřetězce.
- Signatura protokolu se určí na základě metriky blízkosti (closeness).

2. Identifikace síťového provozu podle signatury

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

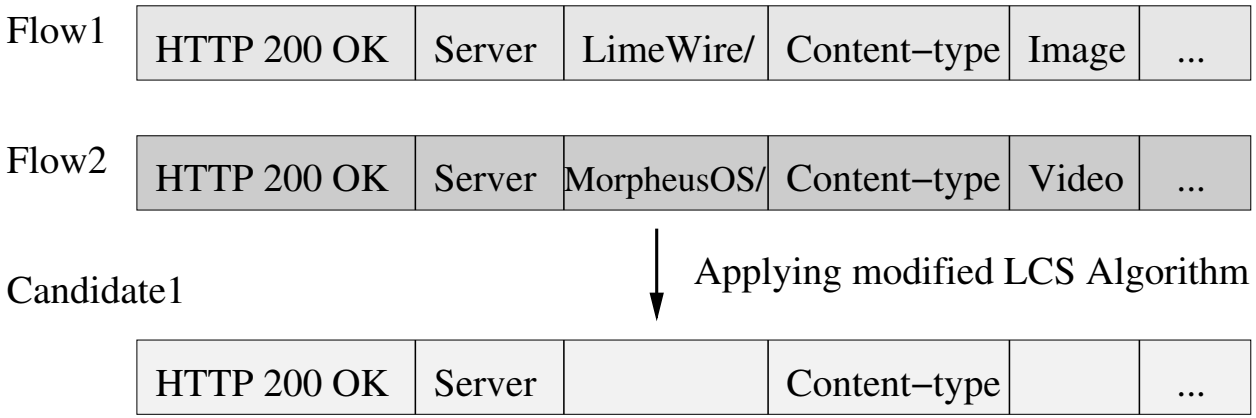
Princip automatické určování signatur

- Hledáme signaturu mezi prvními pakety toku (stačí 10 paketů) [2].
- Iterativní algoritmus určení signatury z toků $Flow_1, \ldots, Flow_{n+1}$

$$
\begin{aligned}
Candidate\_signature_1 &= Signature(Flow_1, Flow_2) \\
Candidate\_signature_2 &= Signature(Flow_3, Candidate\_signature_1) \\
\ldots &\quad \ldots \\
Candidate\_signature_n &= Signature(Flow_{n+1}, Candidate\_signature_{n-1})
\end{aligned}
$$

until $(Candidate\_signature_n == Candidate\_signature_{n=1})$

Příklad: Hledání vzoru pro komunikaci protokolu LimeWire (Gnutella)



2. Identifikace síťového provozu podle signatury

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Automatické určování signatur: srovnání s jinými postupy

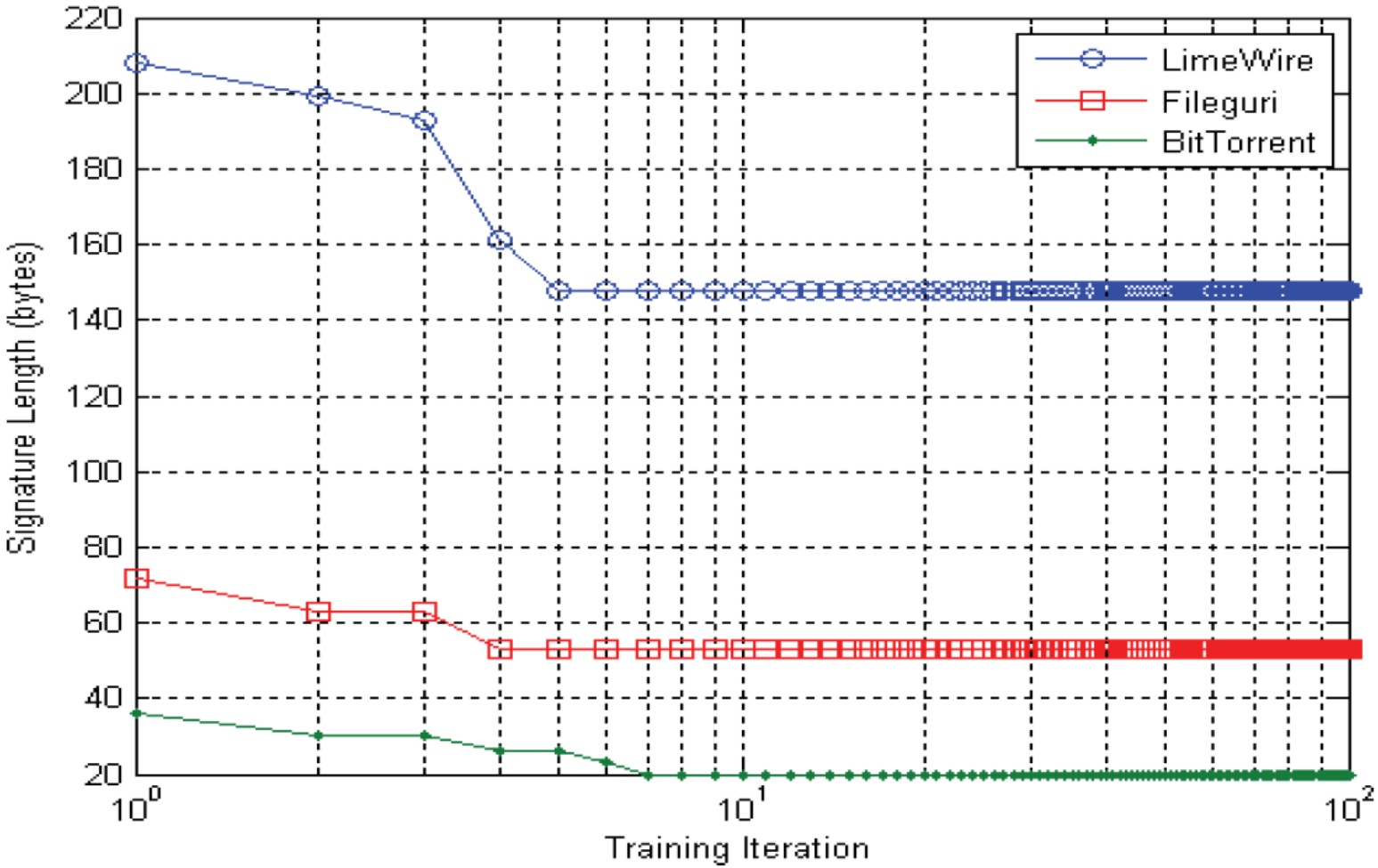
- Paketová analýza: zkoumá obsah jednotlivých paketů nezávisle
- Protokolová analýza: manuální sémantická analýza protokolu

	LimeWire	BitTorrent	Filegure
Packet Analysis	"GNUT"	"0x13Bit"	"Freechal"
Protocol Analysis	GET' or 'HTTP' followed by "User-Agent:Limewire" or "UserAgent: Limewire" or "Server:Limewire"	"0x13BitTorrent protocol"	N/A
Automated Signature (sequence of substrings)	"Limewire", "Content-Type:", "Content-Lenght:", "X-Gnutella-Content-URN","run:sha:1","X-Alt","X-Falt","X-Create-Time:","X-Features:","X-Thex-URI"	"0x13BitTorrent protocol"	"HTTP",""Freechal P2P","User-Type:","P2P-ErrorCode:",""Content-Length","Content-Type","Last-Modified"

- Určování signatur probíhá na prvních deseti paketech toku.
- Minimální délka podřetězce tvořícího signaturu je tři znaky.

2. Identifikace síťového provozu podle signatur

- Kolik je potřeba toků (iterací algoritmu) pro vytvoření signatury?



- Generování signatur se ustálí do deseti iterací.

3. Identifikace provozu dle statistického chování

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Metoda založená na statistickém modelu chování

1. *Vytvoříme statistický model protokolu*
   - Nevyžaduje data z hlaviček či obsahu paketů → stačí meta data o provozu.
   - Hledáme charakteristiky (atributy) popisující chování protokolu.
   - Všímáme si typického chování protokolu: odezva, velikost paketu, počet paketů v daném směru, apod.
2. *Klasifikujeme neznámý protokol oproti vytvořeným modelům protokolů*

   ⇒ Vybere se model s nejlepší vzdáleností od neznámého protokolu

Co potřebujeme k vytvoření statistického modelu protokolu?

- Model aplikačních protokolů vytvořený z trénovací množiny dat
- Vlastnosti (atributy) toku pro statistické porovnání
- Metriku blízkosti pro porovnání modelu protokolu s klasifikovaným provozem
- Stanovení prahové hodnoty, kdy ještě může být protokol úspěšně identifikován
- Algoritmus porovnání (klasifikace)

3. Identifikace provozu dle statistického chování

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Využití otisků protokolu (protocol fingerprints)

Otisk protokolu (protocol fingerprint) = soubor vlastností (chování) protokolu, které lze využít k jednoznačné identifikaci protokolu.

- Otisky protokolů vytvářejí modely protokolů pro statistickou klasifikaci.
- Narozdíl od signatur nezkoumáme obsah hlaviček či těla protokolu, ale pouze statistické vlastnosti jeho chování.
- Tyto vlastnosti lze využít k detekci protokolu.

$\Rightarrow$ Pro detekci stačí sledovat pakety během první fáze vytváření spojení.

Metoda jednoduchých statistických otisků [3]

- Pro vytvoření databáze otisků využívá tři atributy:
  - Velikost IP paketu s (packet size)
  - Časový odstup paketů Δt (inter-arrival time)
  - Pořadí paketu v toku i (packet order)
- Klasifikace počítá vzdálenost paketu i od známého vzorku A_i (anomaly score).

3. Identifikace provozu dle statistického chování

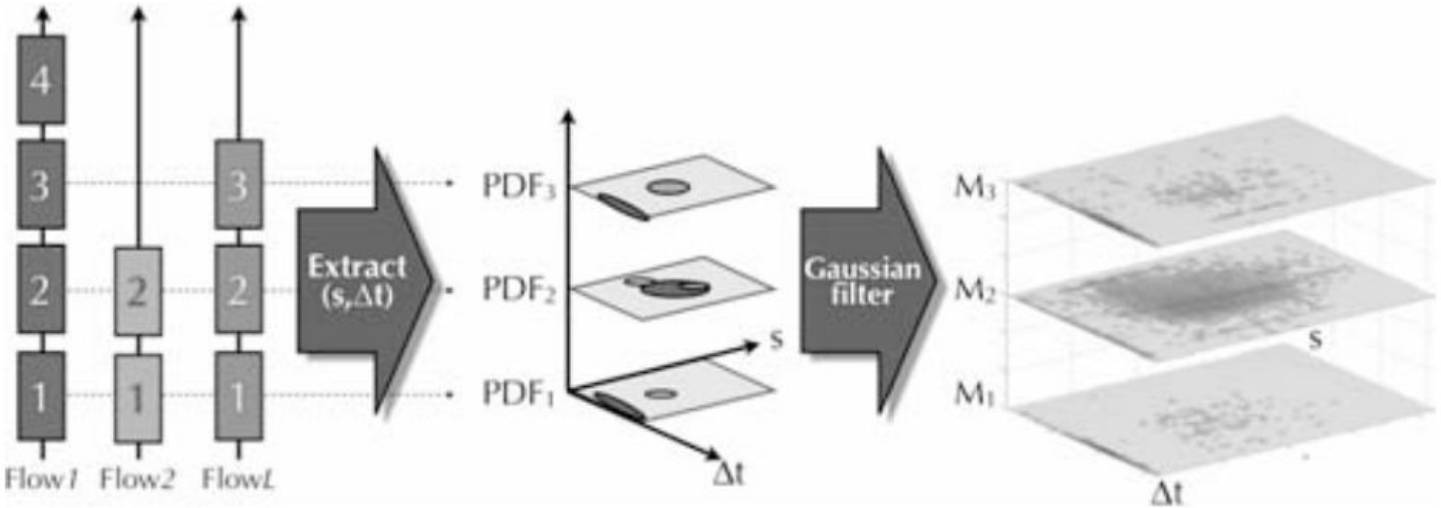
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Popis metody

- Tok F (flow): jednosměrná sekvence n-paketů $(P_1, P_2, \ldots, P_n)$
- Paket P_i popsán velikostí s (size) a časovým odstupem Δt od paketu P_{i-1}: $P_i = \{s_i, \Delta t_i\}$

(i) Učení = vytváříme otisku daného protokolu

1. Počítáme hustotu rozdělení pravděpodobnosti PDF_i (Probability Density Function)
   - PDF_i popisuje rozložení hodnot s a Δt pro každý i-tý paket zkoumaných toků.
   - Pro toky o délce L paketů tvoří vektor $\vec{PDF}$ složený z PDF_i
2. Aplikací Gaussova filtru odstraníme šum a vytvoříme masku protokolu M_i pro klasifikaci.



3. Identifikace provozu dle statistického chování

Pro klasifikaci protokolu počítáme následující hodnoty:

- *Vektor anomálií* $\vec{A}$ tvoří anomálie jednotlivých paketů P_i toku F vůči masce protokolu M_i

$$A_i(P_i, M_i) = \frac{1}{max(\varepsilon, M_i(P_i))}$$, kde ε je nenulová malá hodnota, $P_i = \{s_i, \Delta t_i\}$

- *Vzdálenost anomálie (anomaly score)* $S_n(F, \vec{M})$ toku F o délce n paketů od masky $\vec{M}$:

$$S_n(F, \vec{M}) = \frac{\frac{\sum_{i=1}^{n} A_i(P_i, M_i)}{n} - A_{min}}{A_{max} - A_{min}}$$, tj. $0 \leq S_n(F, \vec{M}) \leq 1$

- *Prahová hodnota* T_n^p definující horní mez, kdy S_n lze ještě klasifikovat jako protokol p. Hodnota se počítá z trénovací množiny toků F použitých při vytváření otisku p:

$$T_n^p = \mu\{S_n(F, \vec{M}^p)\} + \sigma\{S_n(F, \vec{M}^p)\}$$

kde μ je střední hodnota, σ směrodatná odchylka vzdáleností S_n trénovací množiny.

Otisk protokolu Φ^p tvoří dva statistické vektory: maska $\vec{M}^p$ a prahová hodnota $\vec{T}^p$, tj.

$$\Phi^p = \{\vec{M}^p, \vec{T}^p\}$$

3. Identifikace provozu dle statistického chování

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

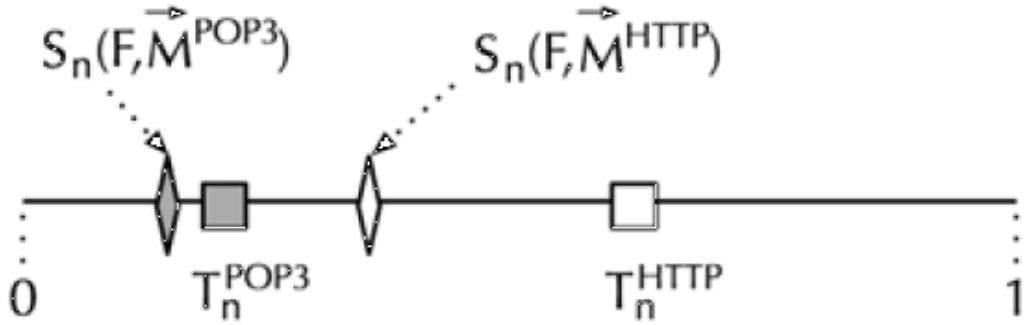
(ii) Klasifikace toku dat F pomocí metody jednoduchých statistických otisků

1. Vezmeme databázi K otisků protokolů $\Phi^i = \{\vec{M}^i, \vec{T}^i\}$, kde $i \in < 0; K >$.
2. Pro neznámý tok F o délce n paketů vypočítáme vzdálenost anomálie S_n vůči prahové hodnotě otisku T_n^i

$$X_n^i = \frac{S_n(F, M^i)}{T_n^i} \text{ pro každý otisk } \Phi^i.$$

3. Najdeme minimum hodnoty X_n^i pro všechny známé otisky protokolů i.
   - Pokud $min_i\{X_n^i\} \leq 1$, pak F odpovídá protokolu i.
   - Pokud podmínka neplatí, jedná se o neznámý protokol, tj. jeho otisk není dostatečně blízko žádnému otisku v databázi otisků protokolů.

Proč je důležitá normalizace pomocí prahové hodnoty?

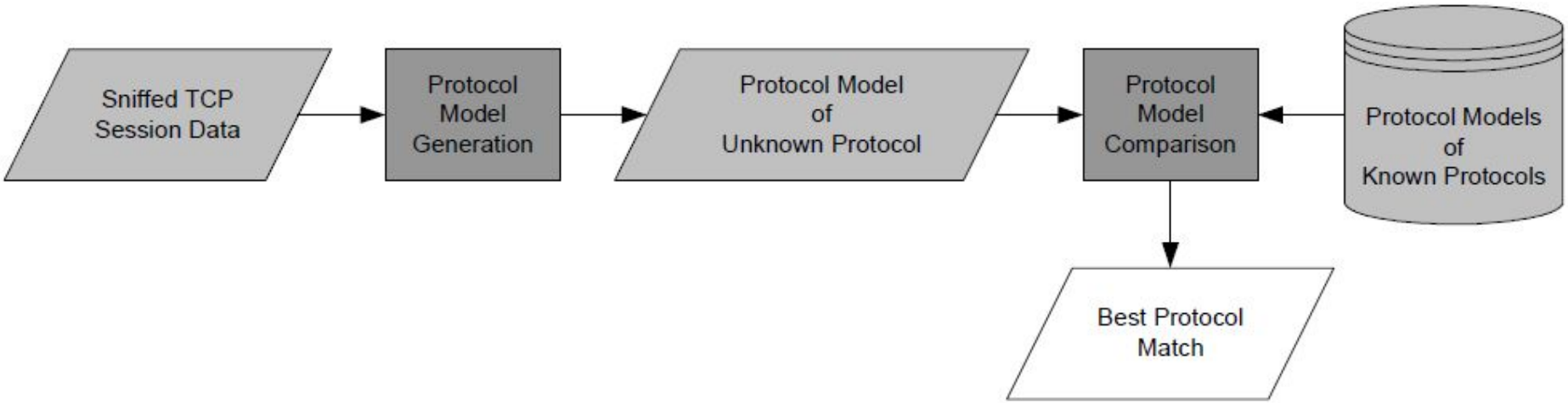


3. Identifikace provozu dle statistického chování

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Metoda SPID (Statistical Protocol Identification) [4, 5]

- Kombinace statistického chování toku a aplikačních atributů
- Implementován v nástroji Network Miner
- Model protokolu používá pro klasifikaci pravděpodobnostní vektory atributů
  - Statistické atributy: PacketDirection, PacketOrderNumber, PacketSize
  - Aplikační atributy: ByteValueFrequency, ByteOffsetValueData
- Pro porovnání se používá Kullback-Leiblerovu divergenci (tzv. relativní entropie)
  - Měří odlišnost dvou pravděpodobnostních rozložení [6].



3. Identifikace provozu dle statistického chování

Metoda SPID: Generování modelu protokolu

- Model protokolu obsahuje množinu otisků atributů vytvořených frekvenční analýzou
- Otisk atributu reprezentován pomocí pravděpodobnostního rozložení atributu

  Příklad atributu `frekvence bytů` u protokolu HTTP

```
"GET / HTTP/1.1"
 71(G),69(E),84(T),32,47(/),32,72(H),84(T),84(T),80(P),
 47(/),49(1),46(.),49(1)
```

Index	0	...	80 (P)	81 (Q)	82 (R)	83 (S)	84 (T)
Counter vector	7869	...	1422	502	1001	1482	2644
Probability vector	0.026	...	0.004	0.002	0.003	0.005	0.008

Index	85 (U)	...	255				
Counter vector	961	...	3276				
Probability vector	0.003	...	0.011				

3. Identifikace provozu dle statistického chování

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

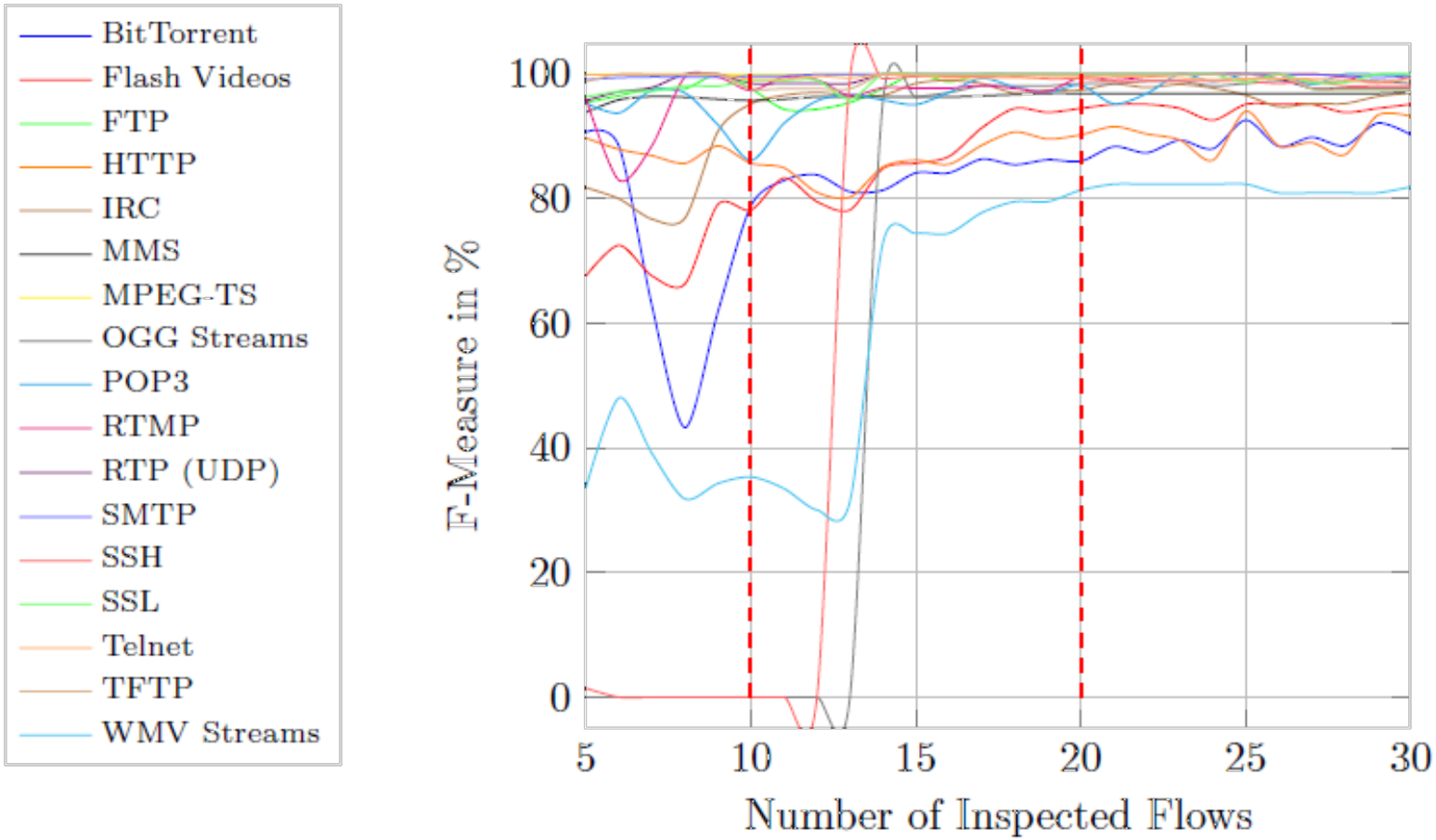
Co lze použít jako atribut pro rozlišení aplikačního protokolu?

- Bytová frekvence prvního paketu TCP v každém směru včetně obsahu
- Počet změn směru komunikace: klient $\rightarrow$ server, server $\rightarrow$ klient
- Počet bytů v daném směru: RTP (balanced), HTTP (download), SMTP (upload)
- Entropie prvního paketu v každém směru
- Opakovaný výskyt dvojic stejných bytů v prvním paketu (TT, SS)
- Bitová frekvence prvních 128 bitů u UDP
- Velikost obsahu prvního paketu toku: 120-1000 B u HTTP, 10-100 B u POP3

3. Identifikace provozu dle statistického chování

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Kolik paketů toku je třeba pro identifikaci? → F-Measure (popisuje přesnost klasifikace)



Pro výpočet atributů stačí použít 10 úvodních paketů TCP (20 pro UDP) [7]

3. Identifikace provozu dle statistického chování

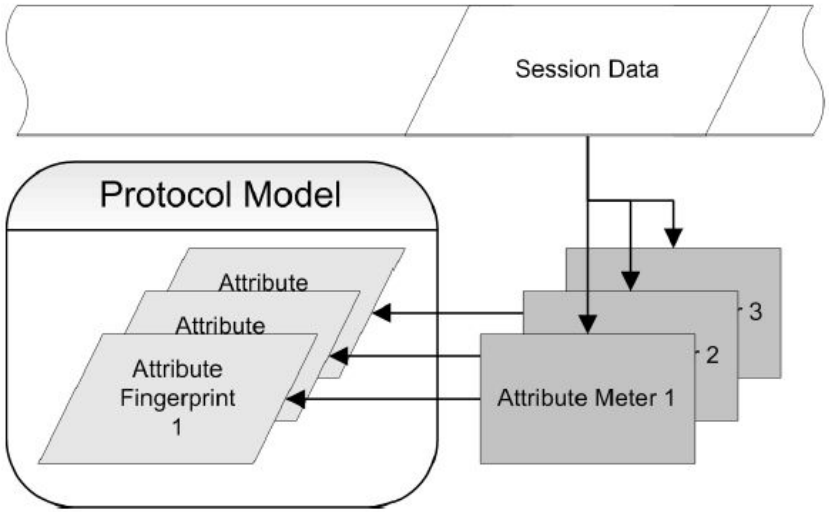
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Klasifikace paketů pomocí algoritmu SPID

- Klasifikace zkoumá míru odlišnosti modelu protokolu P od neznámého protokolu Q.
- Pomocí Kullback-Leiblerovy divergence porovnáváme vektory pravděpodobnosti atributů.
- Výpočet vzdálenosti atributu neznámého protokolu od otisku atributu známého protokolu:

$$D_{KL}(P\|Q) = \sum_{x \in X} P(x) * log_2 \frac{P(x)}{Q(x)}$$

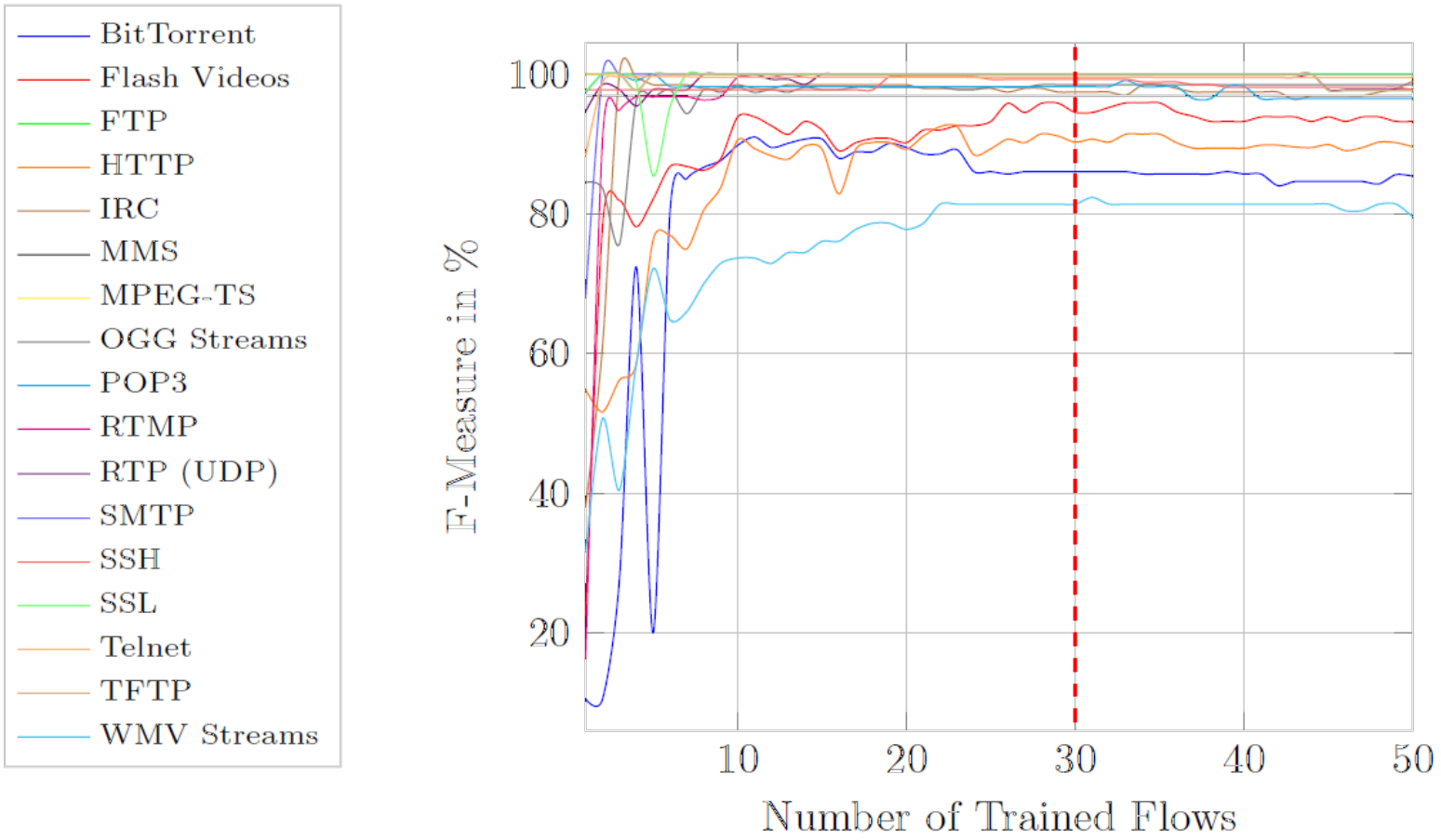
- D_{KL} popisuje relativní míru ztracené informace při použití Q pro aproximaci P.
- Vybereme protokol s nejmenší průměrnou hodnotou $D_{KL}(P\|Q)$ pro zkoumané atributy.



3. Identifikace provozu dle statistického chování

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Kolik vzorových toků je třeba pro vytvoření modelu daného protokolu (trénování)?



⇒ Postačuje cca 30 toků [7].

Srovnání metod identifikace síťového provozu

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

1. *Identifikace podle hodnot z hlaviček paketů*
   - Jednoduché na implementaci, rychlé
   - Nevyžaduje trénovací množinu
   - Identifikace na základě přesné shody
   - Omezená míra použitelnosti
   - Pro identifikaci se využívají čísla portů či čísla protokolů
2. *Identifikace podle signatur*
   - Nespoléhá na hodnoty z hlaviček paketů
   - Vyžaduje vytvoření databáze signatur a její pravidelnou aktualizaci
   - Výpočetně náročné
   - Např. L7 filter, snort, Cisco IOS IPS a další
3. *Identifikace podle statistického chování protokolu*
   - Vhodné pro tunelovaná či šifrovaná data
   - Využívá statistické vlastnosti toků
   - Zkoumá pouze několik prvních paketů toků
   - Vyžaduje vytvoření modelu protokolu na základě trénovací množiny.
   - Méně náročné než použití signatur
   - Může chybně detekovat provoz (false positives)

Důvěra na Internetu

Problém důvěry

Problém důvěry v Internetu (provoz, uživatelé, zdroje) ⇒ snaha vytvořit globální mechanismu důvěry.

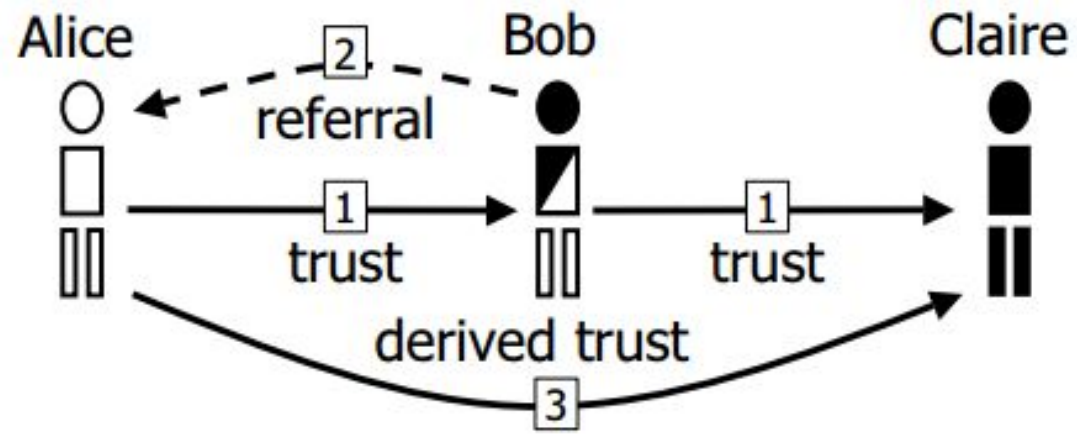
- *Emaily:* spamy, podvržené emaily, phishing
- *Webové služby:* podvržené stránky, stránky s malwarem, nevhodný obsah
- *Sdílení softwaru:* zavirované soubory, malware
- *Síťové spojení:* ochrana sítí proti zneužití (botnety, útoky DDoS)

Příklad: Cisco SensorBase

Důvěra na Internetu

Co je to důvěra?

Důvěra (trust) = asymetrická, tranzitivní relace, která vyjadřuje důvěryhodnost či spolehlivost jedné entity (*truster*, důvěřující) vůči druhé (*trustee*, poskytovatel důvěry)[9].



Příklad

Banka (důvěřující) se rozhoduje, zda poskytne zákazníkovi (poskytovatel důvěry) půjčku a zda ji zákazník splatí.

- Říkáme, že důvěřující (*truster*) důvěřuje poskytovateli důvěry (*trustee*), pokud očekává, že dostane peníze zpět.
- Poskytovatel důvěry je důvěryhodný (*trustworthy*), pokud splatí půjčku.

Důvěra na Internetu

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Co je to důvěra?

- Důvěřující nemůže odhadnout důvěryhodnost poskytovatele důvěry přímo, ale hledá tzv. znaky důvěryhodnosti (tzv. signál).
- Co může být tím signálem?
  Signál = aktivita nebo vlastnost, kterou může jednoduše splnit důvěryhodná entita, ale jejíž získání je pro nedůvěryhodnou entitu příliš drahé.

Prospěchář (opportunist): entita, která není důvěryhodná, ale která napodobuje znaky důvěryhodnosti s cílem zlepšit svou důvěryhodnost.

⇒ Pozor na prospěcháře!

Důvěra na Internetu

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Reputace

Reputace (reputation) = znak důvěryhodnosti vyjádřený svědectvím dalších entit.

- Efektivní rozlišující signál, který podporuje spolupráci postavenou na důvěře.
- Reputace není dokonale rozlišující signál.
- Reputace je efektivní, pouze jsou svědectví nezávislá a důvěryhodná.
- Reputace ja založena na zkušenostech, historii či vztazích.

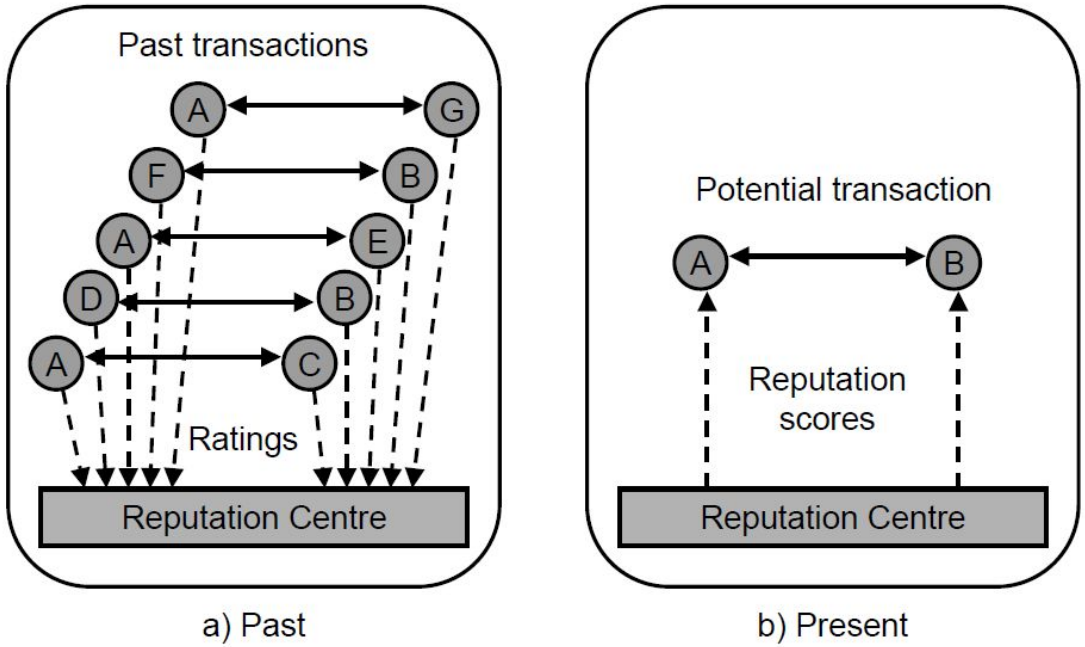
Co je potřeba pro budování důvěry na Internetu? [10]

- Informace vhodné pro měření důvěry a reputace v dané aplikaci
- Metrika pro výpočet hodnoty reputace (reputation score) či míry risku (risk rating)
  - $\rightarrow$ Hodnocení závisí na aktuálních i historických informacích.
- Způsob získání a udržování těchto informací
- Reputační systém, který je odolný vůči manipulaci

Reputační systémy

Centralizovaný reputační systém

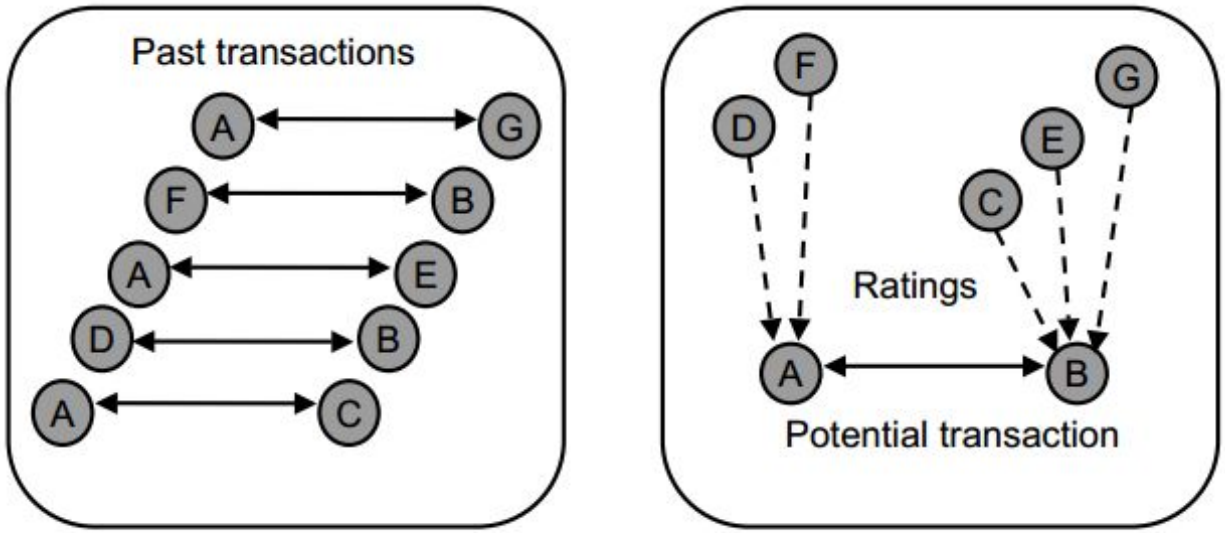
- Hodnotitelé předávají hodnocení centralizovanému systému.
- Systém vyhodnocuje reputaci na základě vstupních hodnocení a dalších informací.
- Příklad: hodnocení reputace na základě hodnocení transakcí



Reputační systémy

Distribuovaný reputační systém

- Pro výměnu hodnocení se využívá spolehlivý partner (relaying party).
- Tito partneři počítají reputační skóre na základě získaných hodnocení.
- Využívá se například u sítí typu P2P pro hodnocení spolehlivosti uzlu.



Reputační systémy

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Jak počítat hodnotu reputace?

1. Součet jednotlivých hodnocení
   - eBay's Feedback Forum: $R = \sum_i positive\_score_i - \sum_j negative\_score_j$
   - Co je lepší: 100 pozitivních a 10 negativních ohlasů či 90 pozitivních ohlasů?
2. Průměr jednotlivých hodnocení
   - Amazon: počítá hodnotu na stupnici 1-5, bere průměr hodnocení
3. Bayesovské systémy
   - Vstupem je binární hodnocení (pozitivní či negativní).
   - Reputační skóre se počítá pomocí pravděpodobnostního rozdělení Beta.
   - Hodnocení ve formě (α, β), α je počet pozitivních hodnocení a β negativních.
4. Modely důvěry (Belief Models)
   - Model počítá metriky důvěry na základě důvěry, nedůvěry a nejistoty.
   - Tyto názory se mapují na funkci hustoty pravděpodobnostního rozdělení Beta.
5. Modely toků
   - Počítají důvěru a reputaci na základě tranzitivity hodnocení účastníků.
   - Například algoritmus PageRank (Google), Appleseed či Advogato.

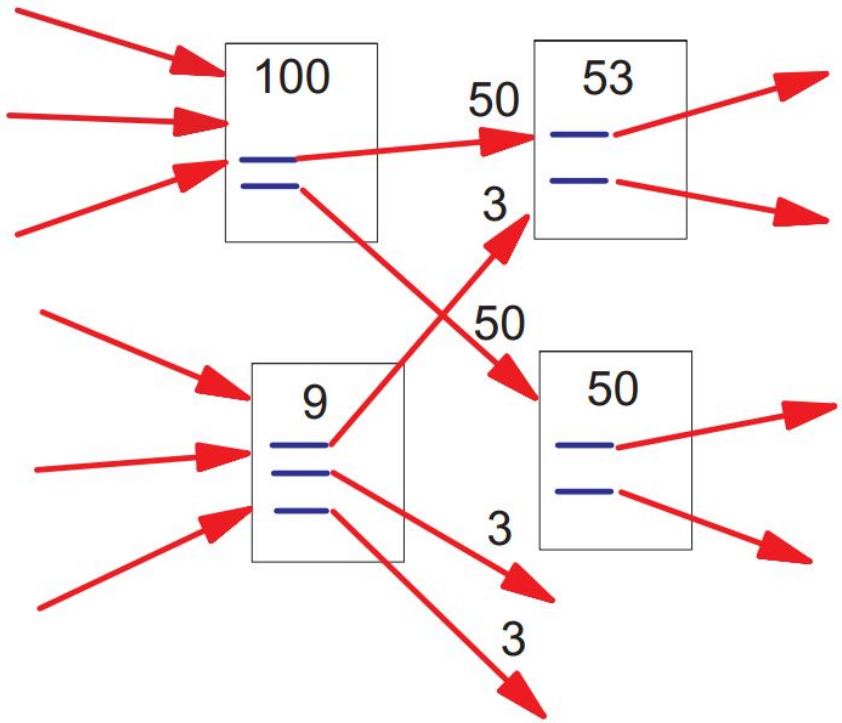
Příklad: hodnocení webových stránek PageRank

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

- Rozšíření algoritmu počítání citací ze zpětných odkazů na webovou stránku [11].
- Hodnocení nezávisí jen na počtu odkazů, ale i na důležitosti R (ranking).

$\Rightarrow$ Stránka má vysoké hodnocení, pokud suma hodnocení zpětných odkazů je vysoká:

$$R(u) - c \sum_{v \in B_u} \frac{R(v)}{N_v}, \text{kde } B_u \text{ je množina zpětných odkazů na stránku } u$$



Síťový reputační systém

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Reputace je služba založená na ohodnocení → není to černá listina. Reputace vyjadřuje, jak "riskantní" je komunikovat s daným uzlem. Reputace je časově podmíněná.

Síťový reputační systém

- Síťový reputační systém poskytuje platformu pro monitorování aktivit v síti.
- Sbírá data ze síťových senzorů, firewallů, IPS zařízení, anti-virové analýzy apod.
- Provádí korelaci dat a počítá hodnotu reputace či míru risku
- Výpočet bere v úvahu historická data.

Jaká data lze použít pro výpočet reputace?

- IP adresa a port útočníka
- IP adresa a port oběti
- Maximální velikost segmentu, volitelné hodnoty TCP, velikost IP datagramu
- Počet žádostí o příchozí či odchozí spojení
- Otisk správy, seznam spammerů, URI, záznamy DNS a další

Síťový reputační systém

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Příklady reputačních systémů

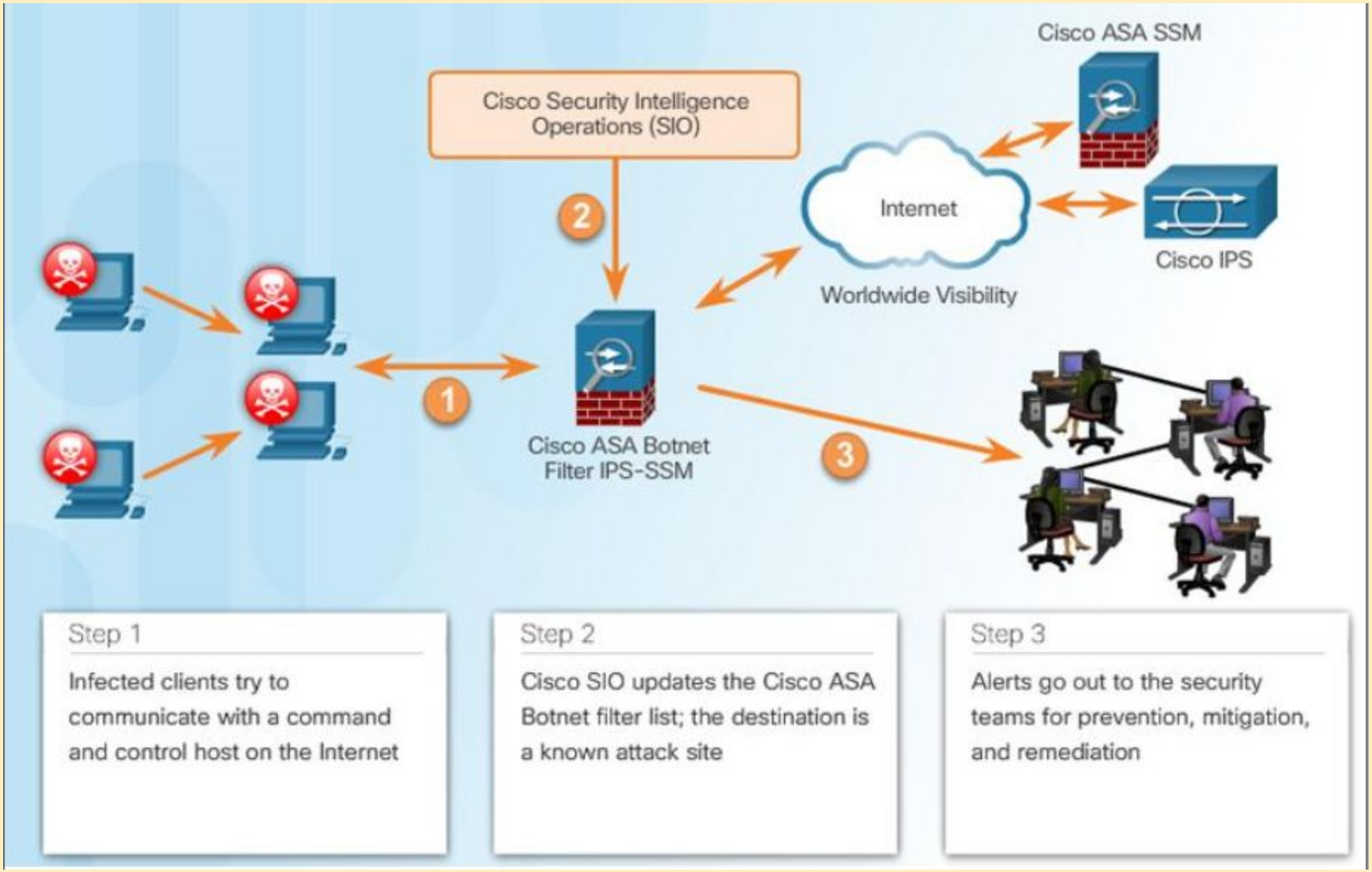
- Cisco SensorBase: spamy, webové služby, malware
- McAffee TrustedSource: sdílení dat (ochrana proti malware), web
- Barracuda Reputation System: síťové připojení, spamy, viry, emaily

Cisco SensorBase Network

- Systém sbírá data ze 700,000 globálně propojených senzorů v IPS zařízeních, firewallech, webových a mailových serverech
- Přidává váhu datům, zpracovává je pomocí automatizovaných metod.
- Výsledek se posílá na firewally, webové servery, IPS zařízení, mailové servery apod.

Síťový reputační systém

Použití reputačního systému na Internetu



Otázky k opakování

- Popište klasifikaci toků podle hodnot z hlaviček rámců, datagramů a paketů. Jaké položky lze použít pro klasifikaci? Jaké jsou výhody a nevýhody tohoto postupu?
- Co jsou to signatury protokolů? Uveďte příklad. Popište postup automatizovaného vyhledání signatury.
- Co je to otisk protokolu a jak ho lze získat? Vyberte si nějakou metodu pro získání otisku protokolu a popište ji.
- Jaké atributy lze použít pro vytvoření statistického modelu protokolu? Uveďte alespoň pět atributů.
- Popište využití frekvenční analýzy při klasifikaci provozu podle statistického rozložení.
- Co je to důvěra na Internetu. Definujte tento pojem a uveďte, jak se zjišťuje a k čemu se využívá.
- Vysvětlete, co je to reputační systém a na základě čeho pracuje. Uveďte příklad využití v komunikaci na Internetu.
- Co je to reputační skóre a jak se vypočítá? Uveďte několik příkladů možného výpočtu, jejich výhody a omezení.
- Navrhněte reputační systém pro vybranou síťovou službu, způsob výpočtu reputačního skóre a zdroj dat pro tento výpočet.

Použitá literatura I

[1] Byung-Chul Park, Young J Won, Myung-Sup Kim, and James W Hong.
Towards automated application signature generation for traffic identification.
In *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*, pages 160–167. IEEE, 2008.

[2] Subhabrata Sen, Oliver Spatscheck, and Dongmei Wang.
Accurate, Scalable In-network Identification of P2P Traffic Using Application Signatures.
In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 512–521, New York, NY, USA, 2004. ACM.

[3] Manuel Crotti, Maurizio Dusi, Francesco Gringoli, and Luca Salgarelli.
Traffic Classification Through Simple Statistical Fingerprinting.
SIGCOMM Comput. Commun. Rev., 37(1):5–16, January 2007.

[4] Erik Hjelmvik and Wolfgang John.
Statistical protocol identification with SPID: Preliminary results.
In *Swedish National Computer Networking Workshop*, 2009.

[5] Eric Hjelmvik.
The SPID Algorithm.
Technical Report White paper, 2008.

[6] Jitka Homolová and Miroslav Kárný.
Evaluation of Kullback-Leibler Divergence.
Technical Report 2349, UTIA, The Czech Academy of Sciences, 2015.

[7] Christopher Köhnen, Christian Überall, Florian Adamsky, Veselin Rakocevic, Muttukrishnan Rajarajan, and Rudolf Jäger.
Enhancements to Statistical Protocol IDentification (SPID) for Self-Organised QoS in LANs.
In *Proceedings of the 19th International Conference on Computer Communications and Networks, IEEE ICCCN 2010, Zürich, Switzerland, August 2-5, 2010*, pages 1–6, 2010.

Použitá literatura II

[8] Jiawei Han, Micheline Kamber, and Jian Pei.
Data Mining: Concepts and Techniques.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[9] Tom Slee.
Some Obvious Things About Internet Reputation Systems.
2013.

[10] Audun Jøsang, Roslan Ismail, and Colin Boyd.
A Survey of Trust and Reputation Systems for Online Service Provision.
Decis. Support Syst., 43(2):618–644, March 2007.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd.
The pagerank citation ranking: Bringing order to the web.
Technical Report 1999-66, Stanford InfoLab, November 1999.
Previous number = SIDL-WP-1999-0120.