Kvantitativní analýza internetového provozu (5)

Ladislav Lhotka
⟨*lhotka@cesnet.cz*⟩

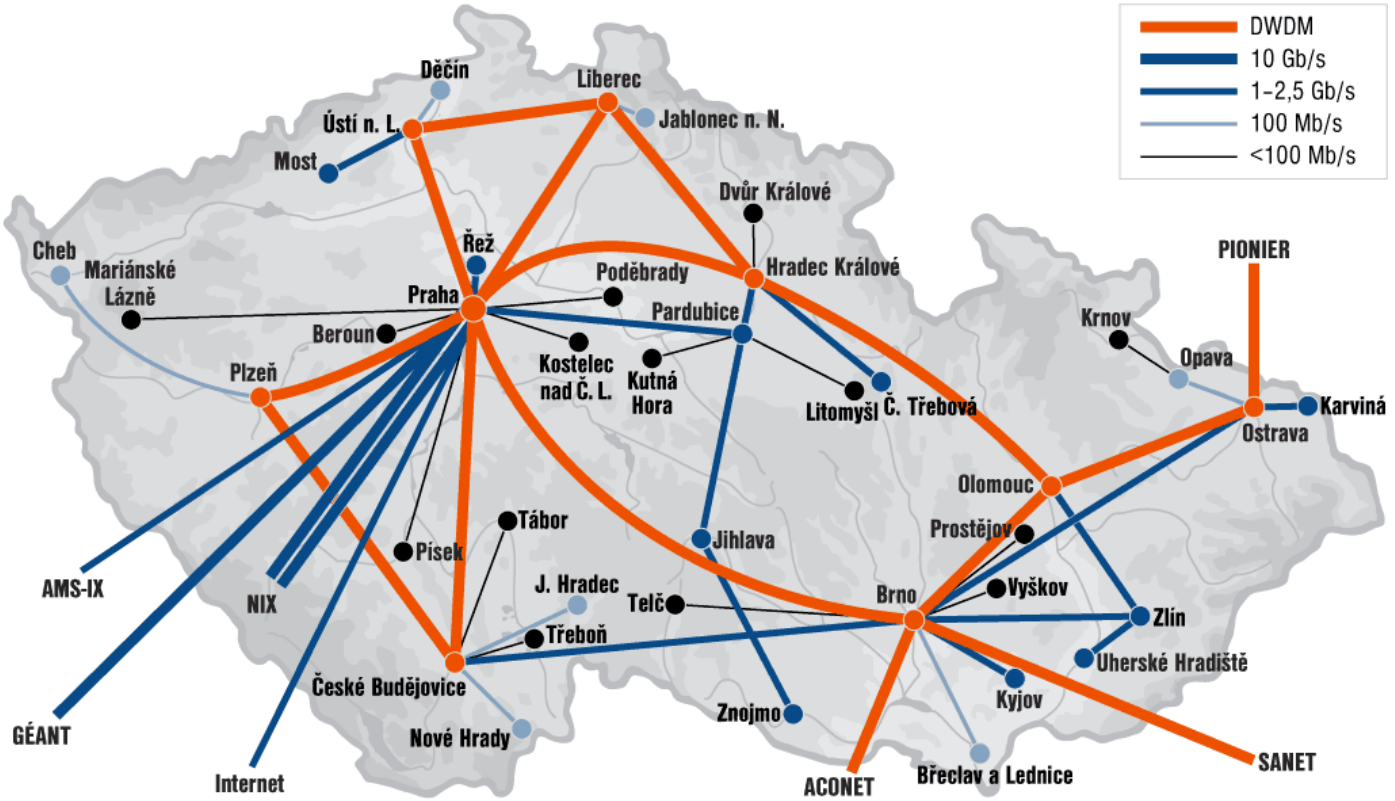
Osnova přednášky

- Origin-destination flows
- Struktura toků, hlavní komponenty
- Vlastní toky a jejich taxonomie
- Síťové anomálie
- Metoda podprostorů
- Jak ošálit m. p.
- Rozložení vlastností toků ve vzorku
- Shluky anomálií

Origin-destination flows

Způsob agregace IP toků, např. podle adresového prefixu, anebo geograficky, např. podle uzlů páteřní sítě – každý OD tok je charakterizován zdrojem (origin) a cílem (destination).

V síti o n uzlech (PoP) tak máme až n^2 různých OD toků.



39 uzlů + 7 mezinárodních linek $\longrightarrow$ 2116 OD toků.

50 linek

Směrovací matice

Boolovská matice, která zachycuje, které linky v síti OD toky používají (předpokládáme n linek, p OD toků):

$$\mathbf{R} = [r_{ij}]_{\substack{i=1,\dots,n \\ j=1,\dots,p}}$$

$$r_{ij} = \begin{cases} 1, & \text{jde-li } j\text{-tý tok přes } i\text{-tou linku;} \\ 0, & \text{jinak.} \end{cases}$$

Je-li y_i celkový objem dat na i-té lince a x_j objem přenášený j-tým tokem, platí

$$y = \mathbf{R}x. \tag{1}$$

Směrovací matice se může v čase měnit.

Jak získáme OD toky?

1. Pokud máme jen celkové toky na linkách, můžeme zkusit “vyřešit” rovnici (1), což je obvykle těžké (problém je obvykle silně nedourčený).
2. Snadno, pokud můžeme měřit IP toky na všech rozhraních, jimiž jsou PoPy připojené do páteřní sítě. Origin je PoP, kde tok změříme a destination určíme z cílové adresy (např. pomocí směrovacích protokolů).

Časové řady OD toků

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \dots & x_{Tp} \end{pmatrix}$$

V každém řádku jsou hodnoty OD toků v daném okamžiku (např. pětiminutové agregáty).

Kudy na ně?

- Napřed čas, pak prostor: časová řada s více proměnnými (napřed se separuje trend a periodická složka)
- Napřed prostor, pak čas: principal component analysis.

Struktura OD toků

Lakhina, A. et al. Structural analysis of network traffic flows. Proceedings of *SIGMETRICS/Performance'04*, New York, 2004.

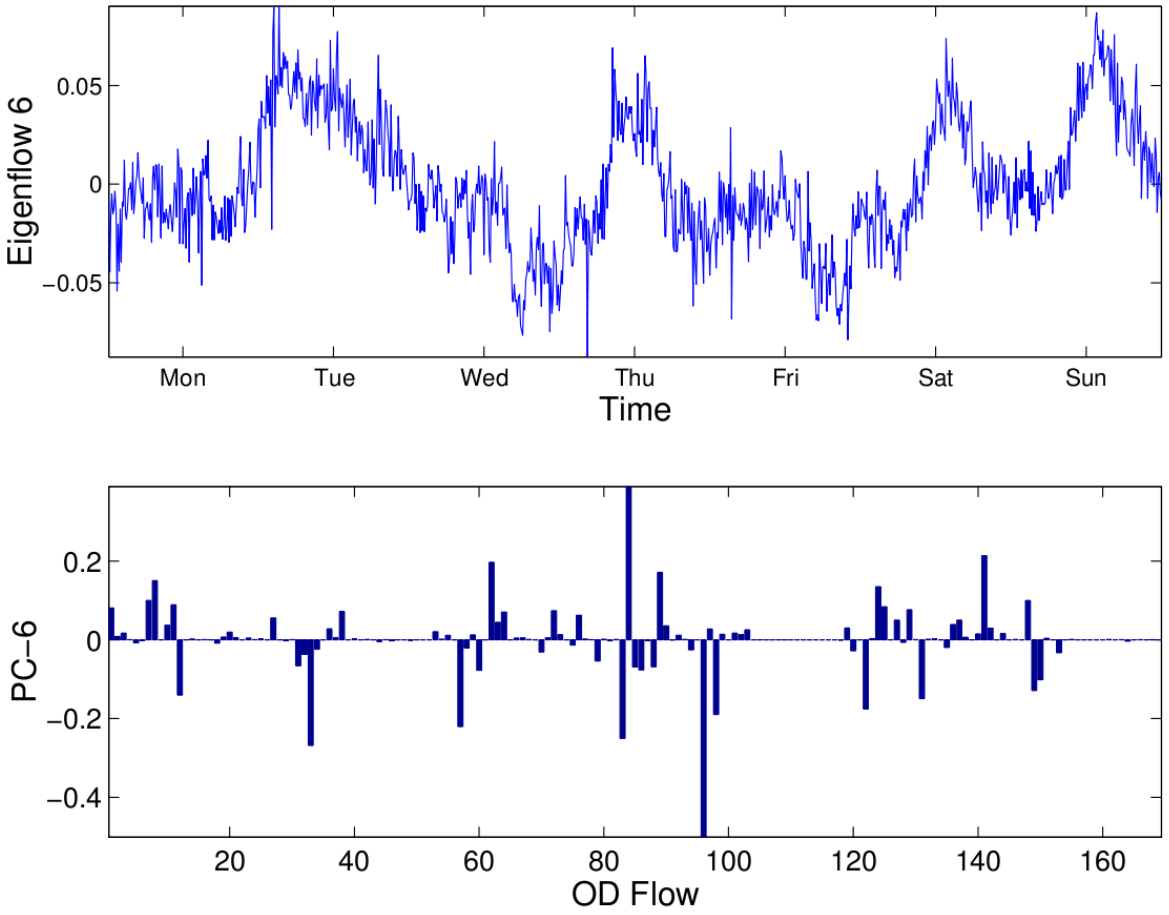
V článku se aplikuje PCA na matici **X**.

Přechod do báze hlavních komponent:

$$u_i = \mathbf{X} v_i, \text{ pro } i = 1, 2, \ldots, T,$$

kde v_i jsou vlastní vektory příslušné kovarianční matice.

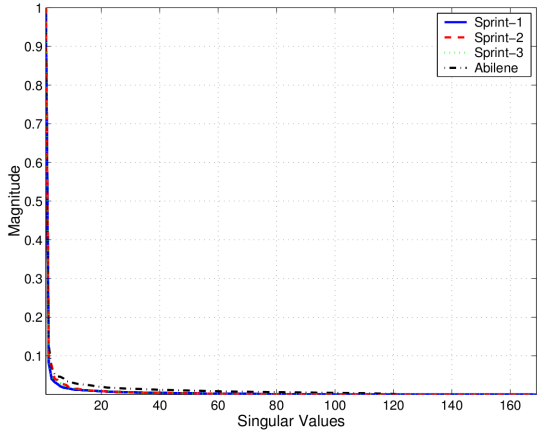
Vektory $u = (u_1, u_2, \ldots, u_T)$ se nazývají *vlastní toky*.



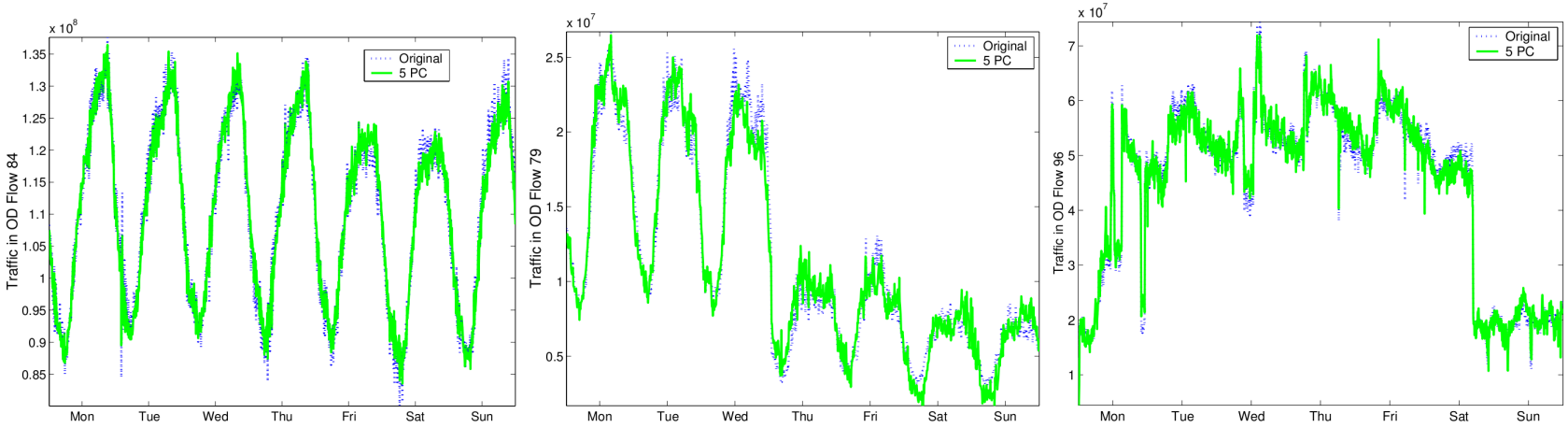
Typický vlastní tok (nahoře) a jeho zastoupení v původních OD tocích.

Rekonstrukce z hlavních komponent

Analýza vzorků z komerční sítě Sprint (13 PoPů) a akademické Abilene (11 PoPů) ukazují, že původní OD toky lze reprezentovat pomocí relativně velmi malého počtu hlavních komponent (vlastních toků). Podíl hlavních komponent na rozptylu prudce klesá:

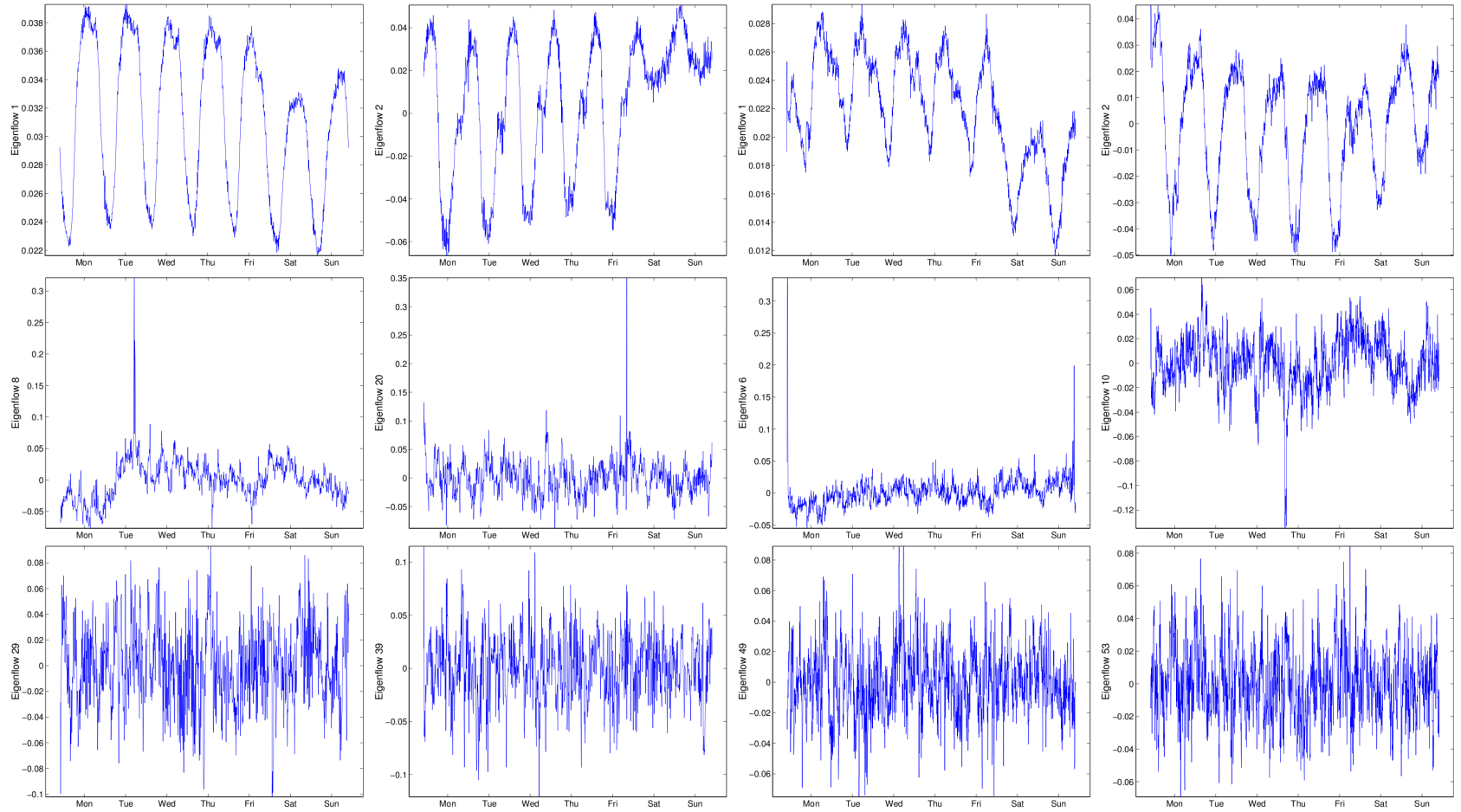


Tento obrázek ukazuje původní od tok a jeho reprezentaci pomocí 5 hlavních vlastních toků (Sprint – vlevo a uprostřed, Abilene –vpravo).



Taxonomie vlastních toků

1. *Deterministické* – obvyklé periodicity
2. *Špičkové (spike)* – izolované “výstřelky” (alfa toky)
3. *Šumové (noise)* – náhodný šum s normálním rozdělením



(a) Sprint-1 (b) Abilene

Síťové anomálie

- Alfa toky: objemné krátkodobé přenosy mezi dvěma počítači;
- Denial of Service (DoS): z jednoho místa nebo distribuované;
- Flash crowd: soustředění velkého počtu klientů;
- Port scan: sondování velkého počtu portů na malém počtu cílových adres;
- Network scan: sondování velkého počtu cílových adres s malým počtem portů (hledání konkrétní bezpečnostní díry);
- Výpadky sítě: náhlé nebo plánované;
- Vějířovitá distribuce: šíření obsahu, např. video streaming, nové verze distribucí Linuxu atd.
- Viry a červi
- Spam

Metody detekce anomálií

- signature-based (Snort)
- supervised classification (antispamové filtry)
- non-supervised classification

Parametry pro hodnocení účinnosti: false negative rate (FNR) a false positive rate (FPR).

Metoda podprostorů (subspace method)

Původně vytvořena pro kontrolu kvality výroby. Zde používáme pro OD toky.

$$\mathbf{x} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{T1} & X_{T2} & \dots & X_{Tp} \end{pmatrix}$$

Aplikujeme PCA a vybereme několik hlavních komponent, které generují *normální podporostor*. Každé měření (v soustavě PC) pak rozložíme na dvě složky:

$$Y_t = \hat{Y}_t + \tilde{Y}_t,$$

kde $\hat{Y}_t$ leží v normálním podprostoru a $\tilde{Y}_t$ je anomálie.

Při určité velikosti $\|\tilde{Y}_t\|$ se měření prohlásí za anomální.

Problémy

1. Síťová data obsahují se mění periodicky i trvale (trend), takže PCA na časově omezeném tréninkovém vzorku nemusí po nějakém čase odpovídat. Obvykle se proto detekční systémy periodicky “přeučují” (např. každý týden).
2. Tréninková data mohou sama obsahovat anomálie, které tím mohou být zahrnuty pod normální provoz. Je proto dobré je z tréninkového vzorku pokud možno odfiltrovat.
3. Obzvláště rafinovaný útočník může přidávání vhodně zvoleného umělého provozu cíleně “otrávit” tréninkový vzorek tak, že se následně nezachytí skutečný útok.

Jak ošálit metodu podprostorů

Rubinstein, B.I.P. et al. *Compromising PCA-based Anomaly Detectors for Network-Wide Traffic*. Technical report UCB/EECS-2008-73, Berkeley:UCB, 2008.

Útočník přidává datový tok c_t s řízenými statistickými parametry (*chaff* = plevy), obvykle tak, aby zvětšil rozptyl ve zvoleném OD toku a přitom nebyl nápadný.

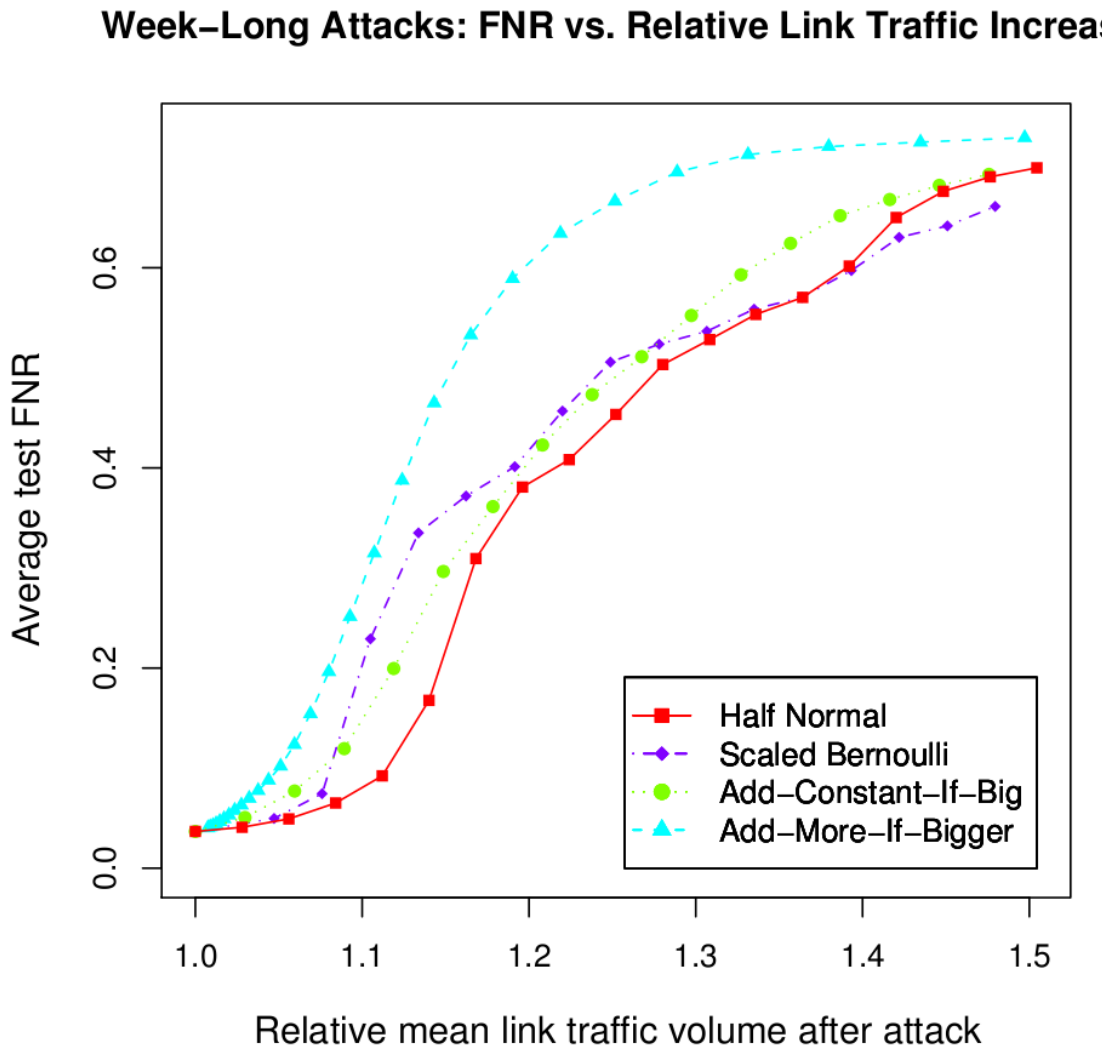
Volba umělého toku

1. $c_t = |n_t|$, kde $n_t \sim N(0, \theta^2)$ (θ je parametr). Střední hodnota je $\sqrt{2/\pi}\theta \approx 0.8\theta$ a rozptyl $(1 - 2/\pi)\theta^2 \approx 0.36\theta^2$.
2. $c_t = \theta b_t$, kde b_t nabývá hodnot 0 a 1 se stejnou pravděpodobností. Střední hodnota je $\theta/2$ a rozptyl $\theta^2/4$.
3. Pokud je na na lince z místa, odkud posíláme umělý tok, objem provozu S_t větší než nějaká prahová hodnota α (např. dlouhodobý průměr), přidáme $c_t = \theta$.
4. Jako předchozí případ, jen s odstupňovaným přídavkem, např. $c_t = (S_t - \alpha)^\theta$.

Boiling frog

Funguje na systémy, které se periodicky přeučují (např. každý týden). Dlouhodobá příprava, kdy se přidává umělý provoz s postupně rostoucím parametrem θ. Tím se příspěvek stane nenápadným a klasifikační metoda se otráví postupně.

Rubinstein et al. toto vyzkoušeli na datech ze sítě Abilene (27 týdnů) a metodě podprostorů založené na PCA. Umělý tok zvyšovali postupně v tak, aby celkový tok na první lince rostl mezi dvěma týdny o stanovený faktor (1 %, 2 %, 5 % a 15 %). Použili metodu 4.

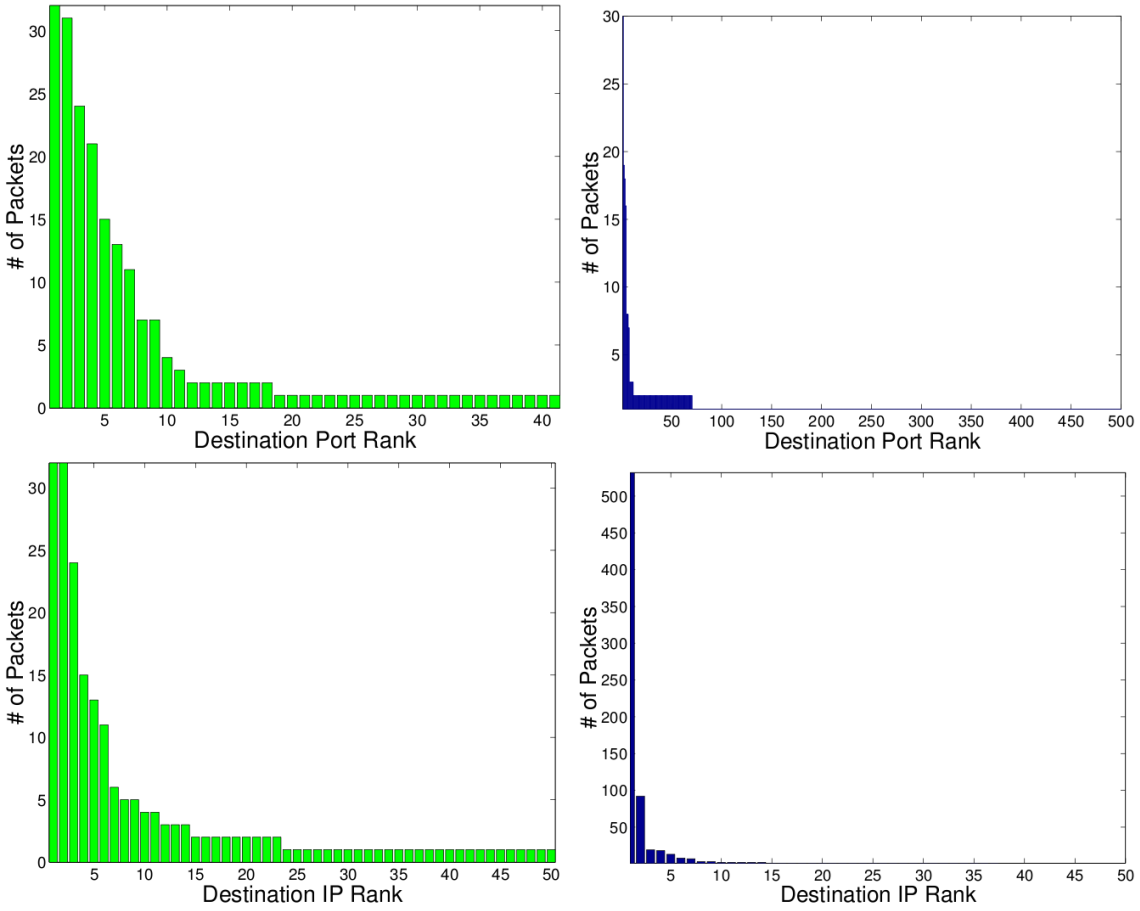

Week−Long Attacks: FNR vs. Relative Link Traffic Increase
Average test FNR
0.0
0.2
0.4
0.6
1.0
1.1
1.2
1.3
1.4
1.5
Relative mean link traffic volume after attack
Half Normal
Scaled Bernoulli
Add−Constant−If−Big
Add−More−If−Bigger

Rozdělení vlastností toků

Lakhina, A.; Crovella, M.; Diot, C. Mining Anomalies Using Traffic Feature Distributions. Proceedings of *ACM SIGCOMM*, Philadelphia, 2005.

Některé anomálie jsou špatně detekovatelné, pokud používáme pouze objemové charakteristiky OD toků (počty paketů/bajtů), často je ale lze poznat ze změny distribuce některých vlastností – IP adres, portů aj.

Port scan



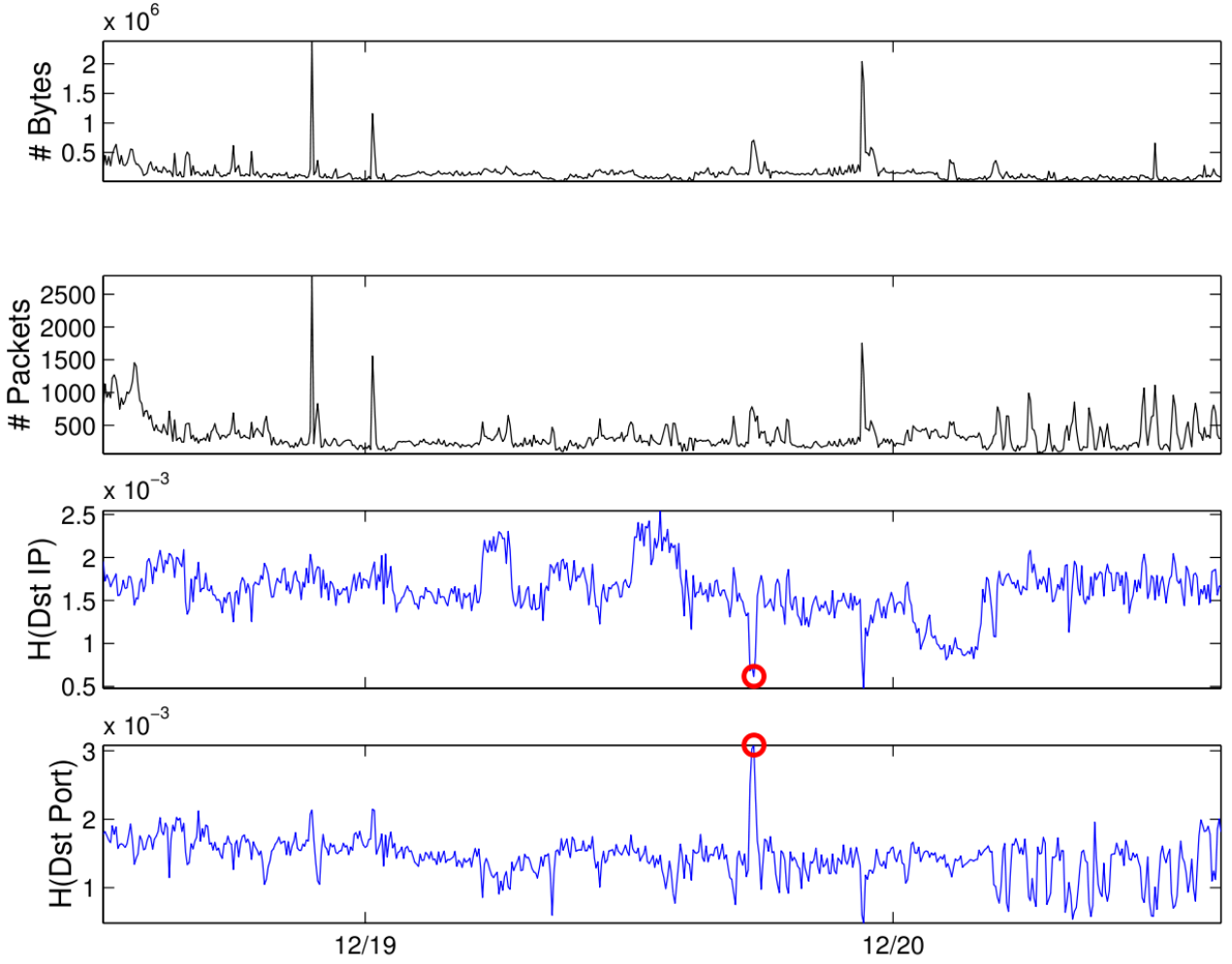
(a) Normal (b) During Anomaly

Entropie

Vhodným kvantitativním měřítkem koncentrace rozložení je *entropie*: Obor hodnot dané vlastnosti X rozdělíme na N stejných intervalů a zjistíme počty výskytů n_i v každém intervalu pro $i = 1, 2, \ldots, N$ (= histogram). Hodnota entropie veličiny X je

$$H(X) = \sum_{i=1}^{N} \left(\frac{n_i}{S}\right) \log_2\left(\frac{n_i}{S}\right).$$

Detekce port scanu pomocí objemů a entropie



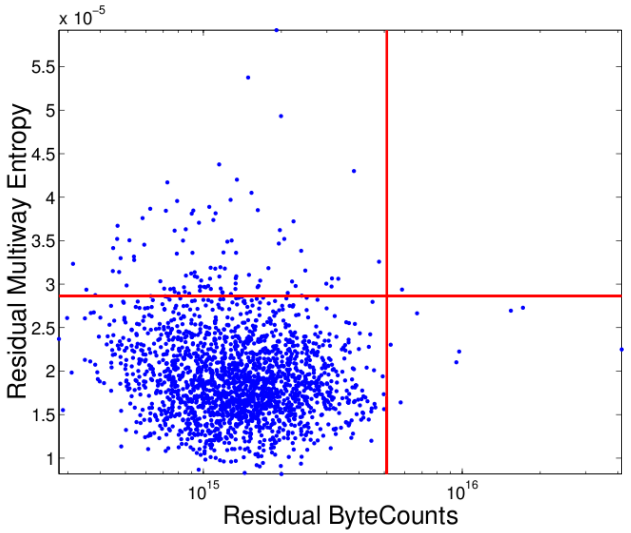
Vícecestná metoda podprostorů

Měříme-li pro každý z p OD toků časovou řadu k parametrů, vytvoříme matici o T řádcích a $p \cdot k$ sloupcích a na ni aplikujeme PCA a metodu podprostorů.

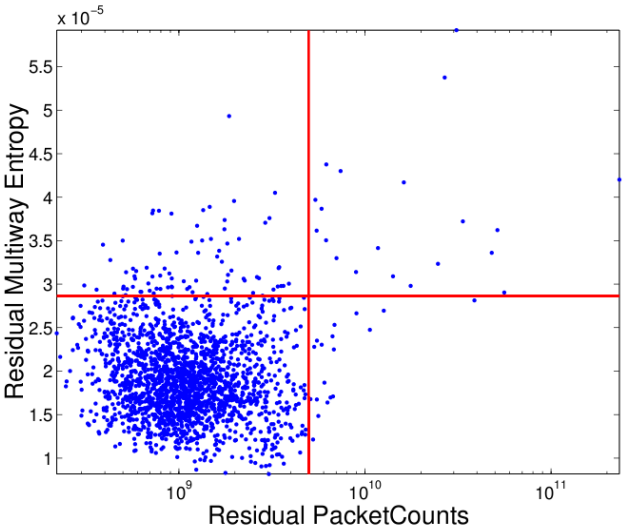
Lakhina et al. použili data o OD tocích z akademických sítí Abilene (11 PoP, 20 dnů) a GÉANT (22 PoP, 23 dnů), které smíchali se vzorky běžných anomálií – celkem 444 ve vzorku Abilene a 1011 ve vzorku GÉANT.

Pro každý OD tok byly v pětiminutových intervalech měřeny počty paketů a bajtů a entropie těchto vlastností: zdrojová a cílová IP adresa, zdrojový a cílový port. Metodu podprostorů aplikovali zvlášť na objemové parametry a na entropie.

Úspěšnost detekce anomálií



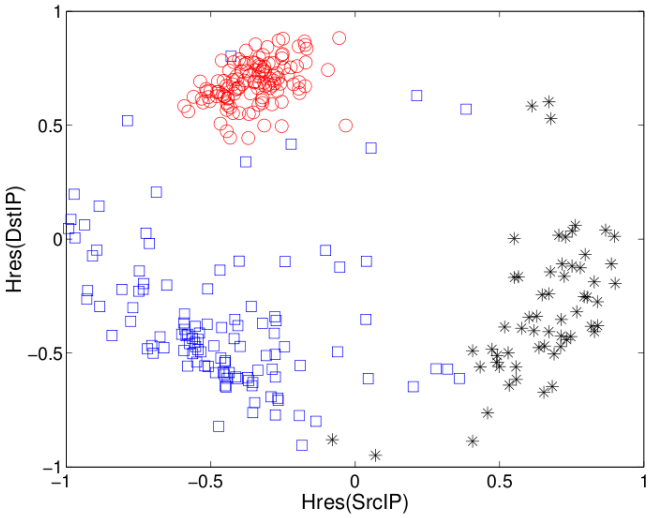
(a) Entropy vs. # Bytes



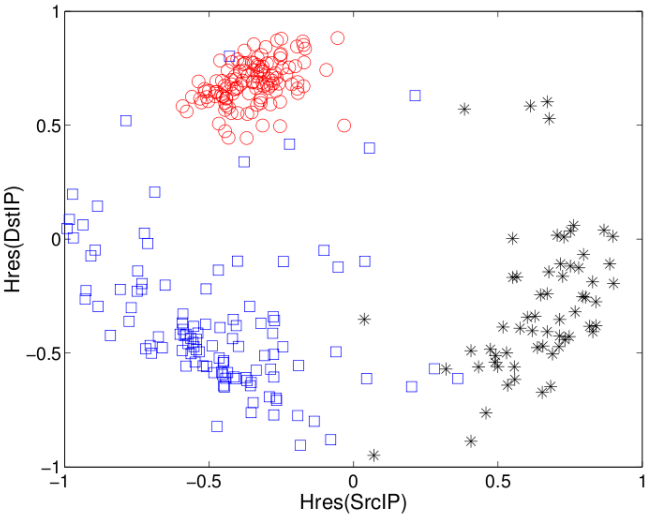
(b) Entropy vs. # Packets

Klasifikace anomálii

Lakhina et al. dále použili shlukovou analýzu pro automatickou klasifikaci anomálií. Jako míru nepodobnosti použili vzdálenost anomálních složek entropií (tj. bez normální složky po rozkladu do podprostorů). Každá anomálie tvořila jasně oddělený cluster (nebo i více clusterů).



(a) Known Types



(b) Cluster Results

Optimální počet clusterů

Mírou pro nalezení vhodné granularity clusterů jsou celková vzdálenost uvnitř clusterů (čím menší, tím lepší) a celková vzdálenost mezi clustery (čím větší, tím lepší).