Získávání informací z logů s využitím shlukové analýzy

Marek Kumpošt

XXVIII. konference EurOpen

Fakulta informatiky
Masarykova univerzita
Brno

Obsah přednášky

Úvod

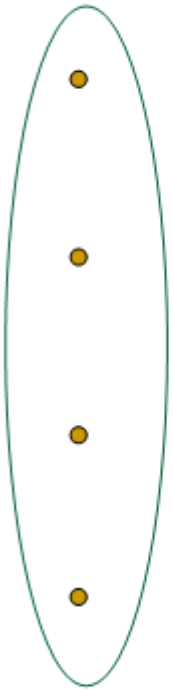
- Co to je kontextová informace
- Kontext-aware systémy – systémy, které berou v úvahu kontextové informace
- Potřeba metodiky pro zpracování kontextových informací
- Různé přístupy pro modelování kontextových informací
- Běžnou snahou je predikce budoucích akcí na základě znalosti předchozího chování uživatelů – známe identifikaci uživatele
  - vhodné zejména pro personalizaci webových stránek
- Naší snahou je pouhá „identifikace“ uživatele na základě předchozího chování – neznáme skutečnou identitu uživatele
- Souvislost kontextových informací a soukromí

Grafový model

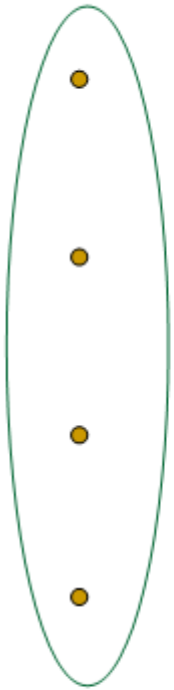
- Graf reprezentuje znalosti o systému (kontextové informace)
- Snaha zahrnout veškeré relevantní kontextové informace
- Kontextové informace jsou reprezentovány jako uzly
- Relace mezi uzly (hrany) – ohodnocení pravděpodobností
- Cílem je nalezení nejlepší (nejpravděpodobnější) spojení mezi uzly

Grafový model – příklad

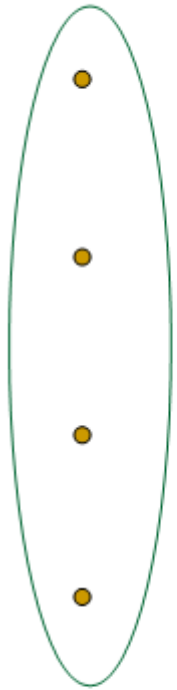
IP (1,2,3,4)



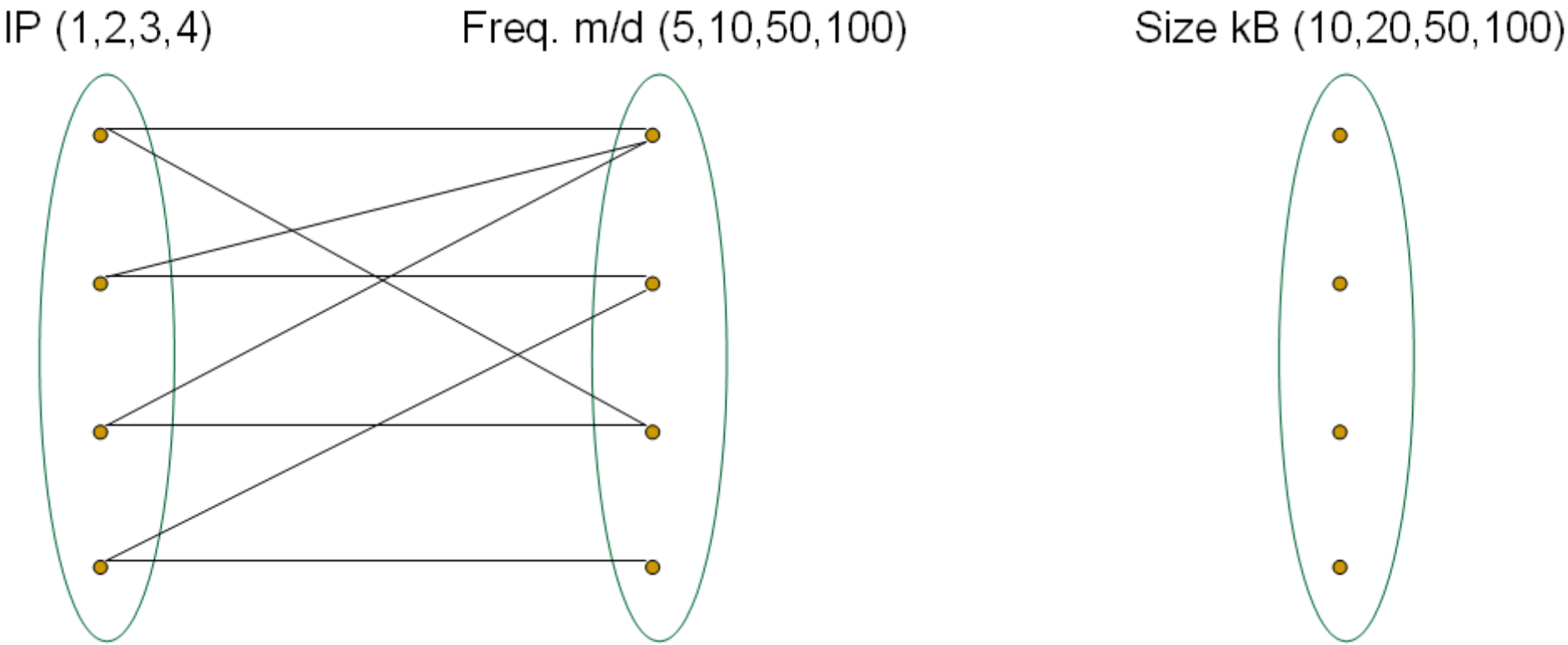
Freq. m/d (5,10,50,100)



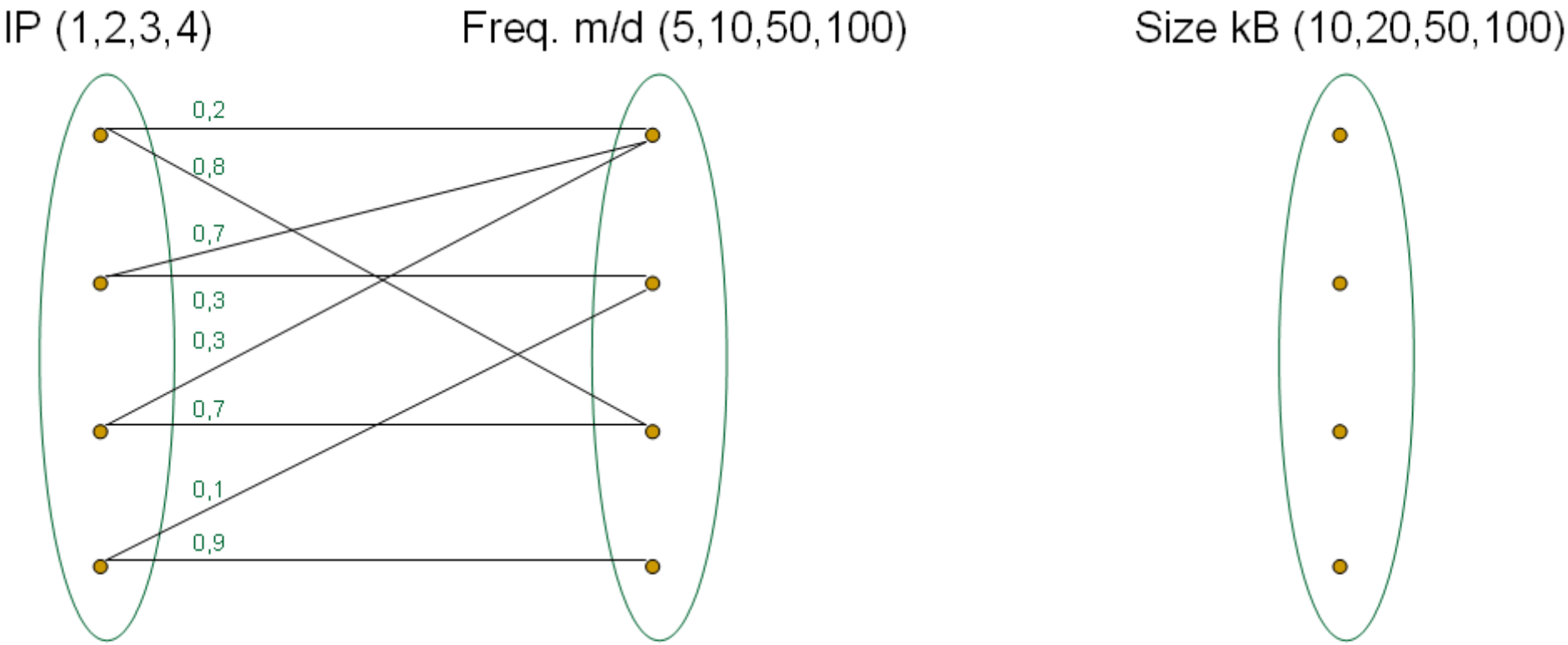
Size kB (10,20,50,100)



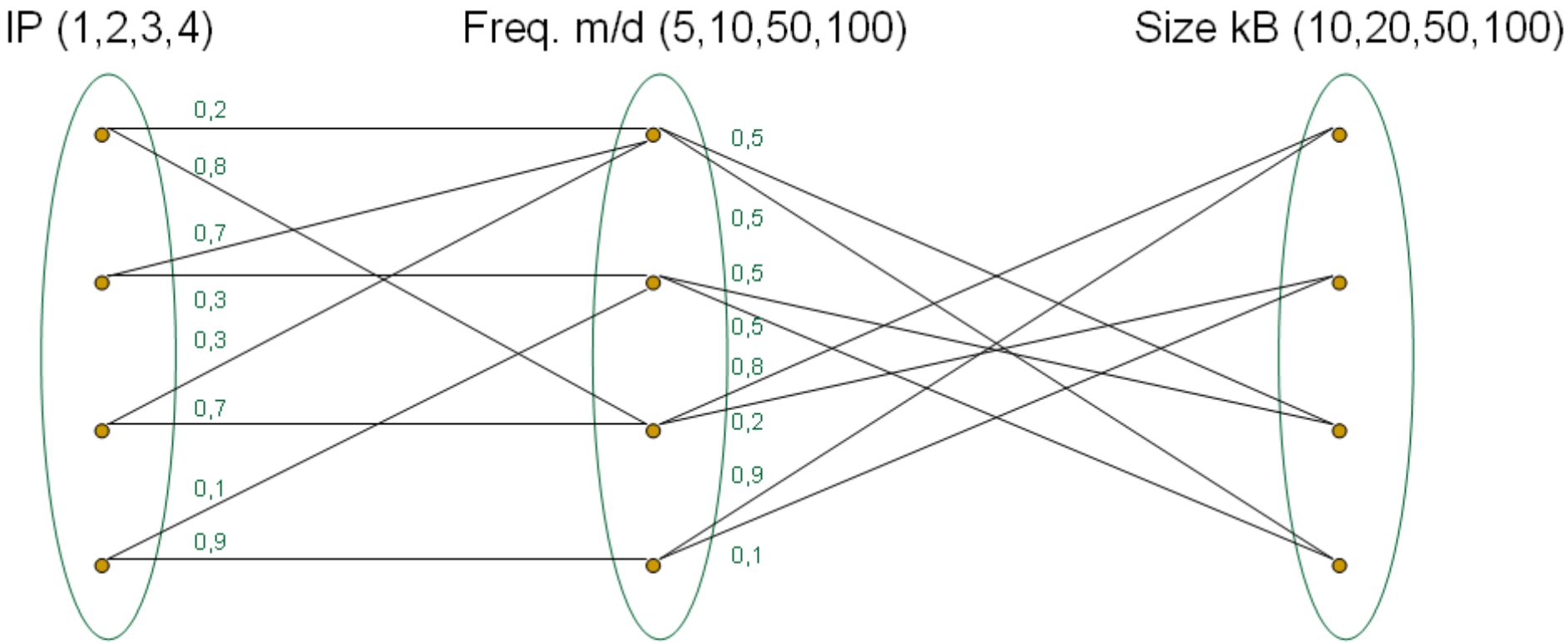
Grafový model – příklad



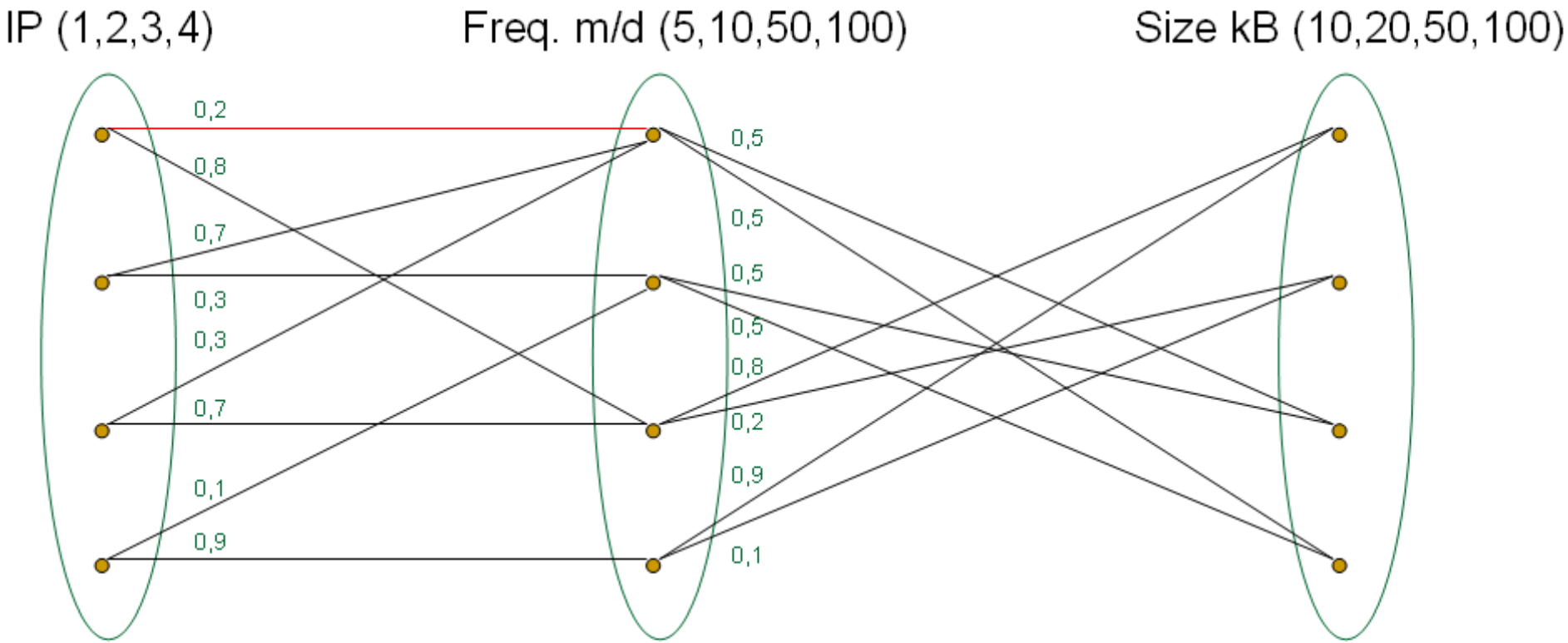
Grafový model – příklad



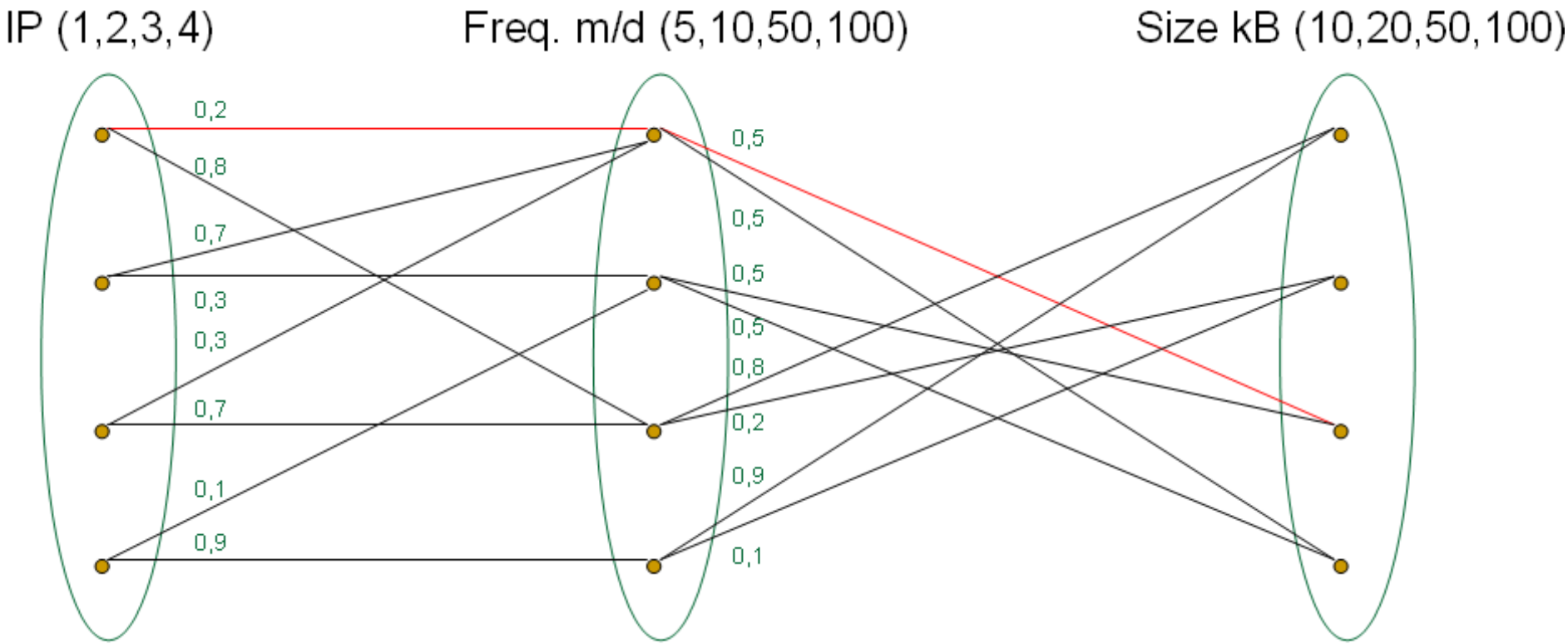
Grafový model – příklad



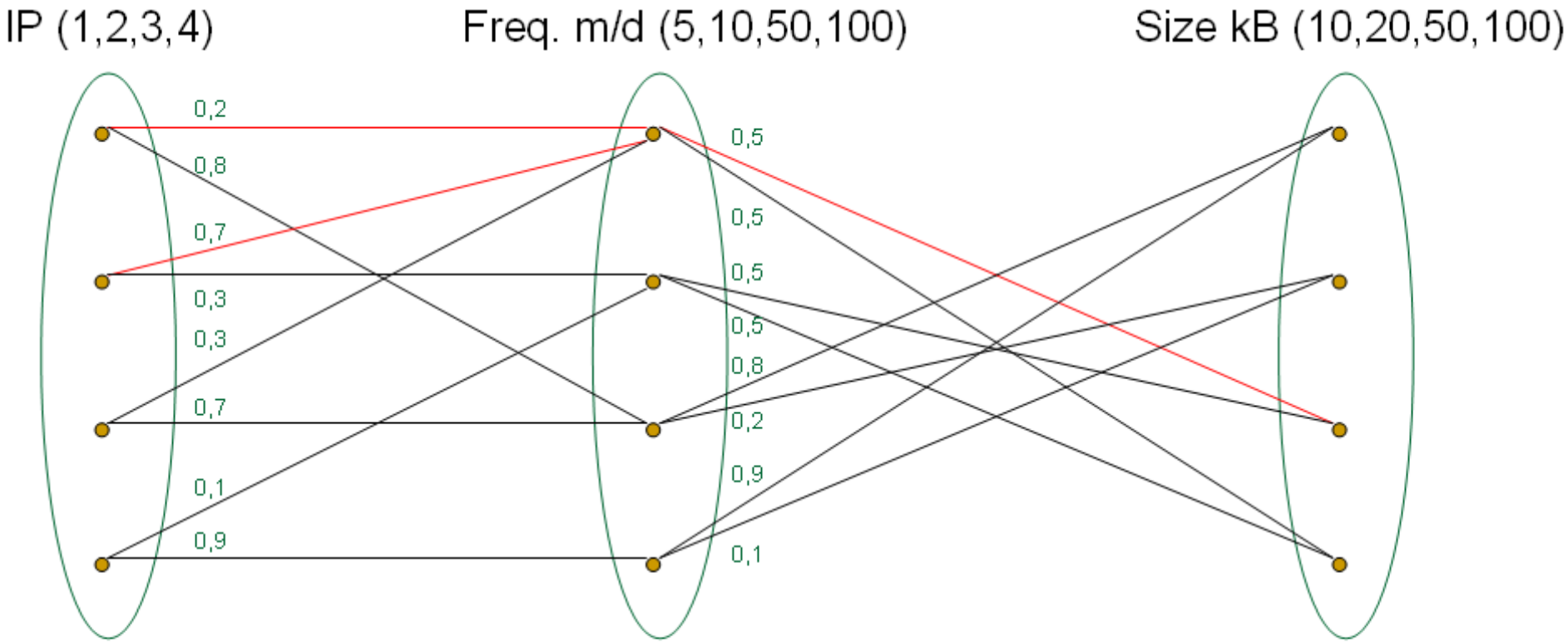
Grafový model – příklad



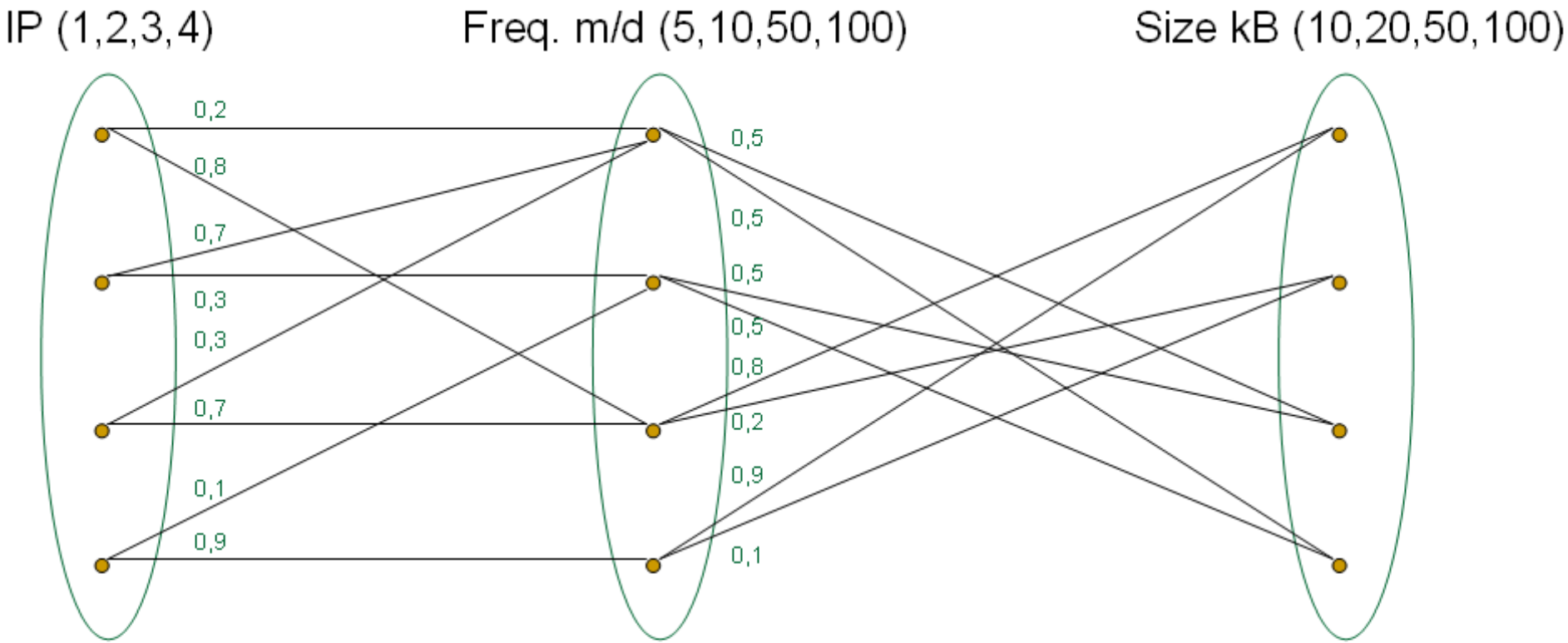
Grafový model – příklad



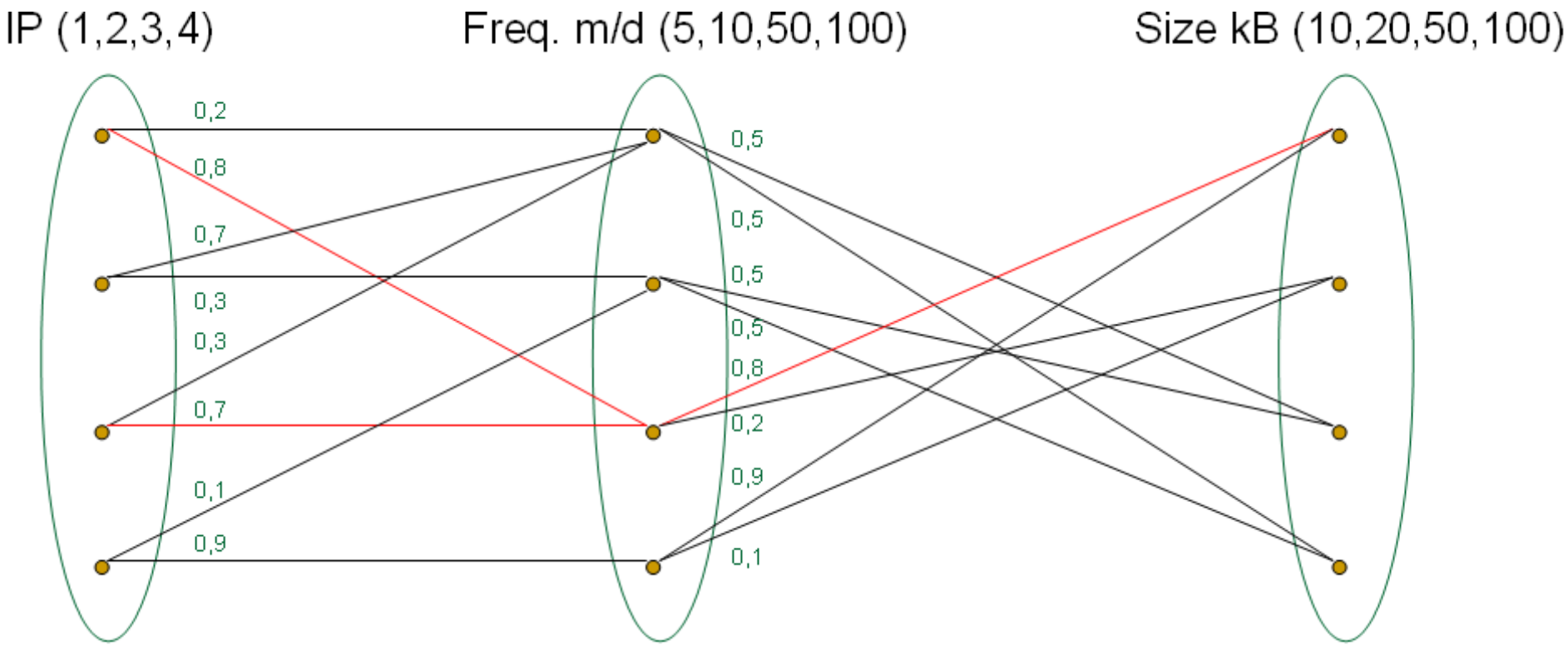
Grafový model – příklad



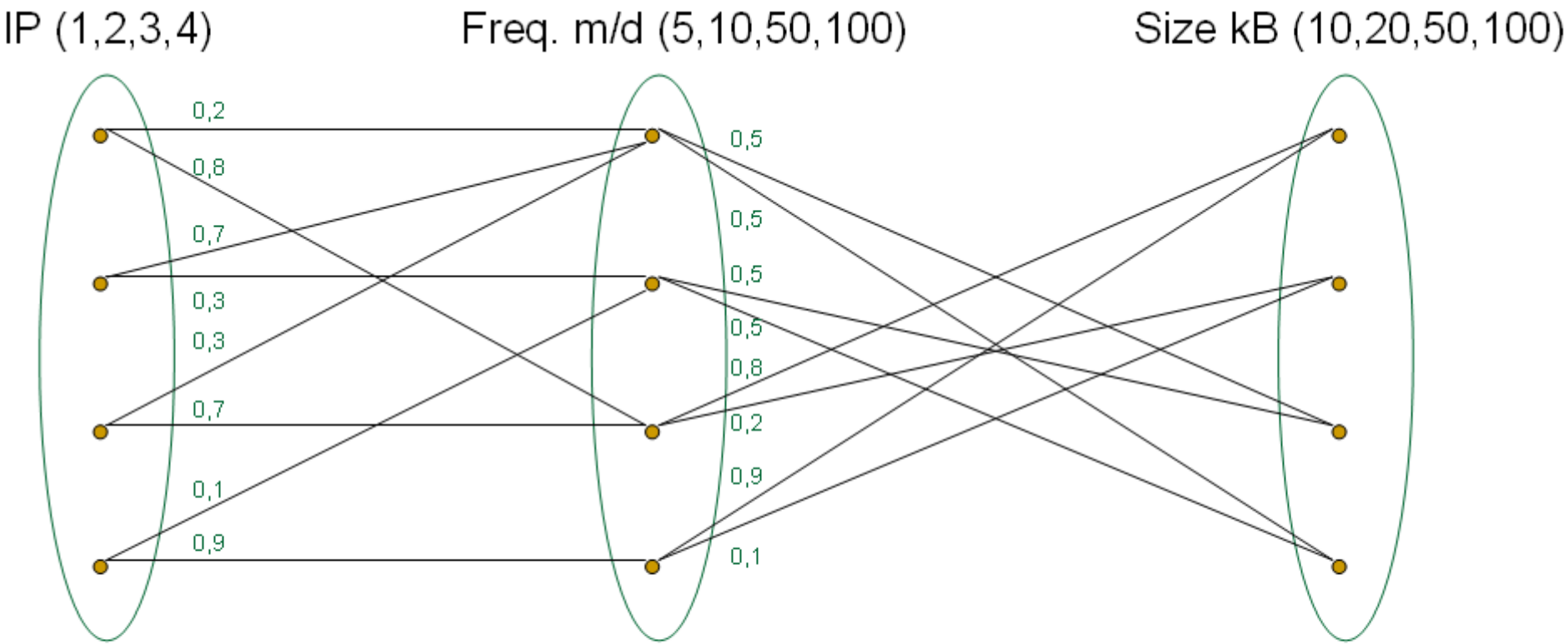
Grafový model – příklad



Grafový model – příklad



Grafový model – příklad



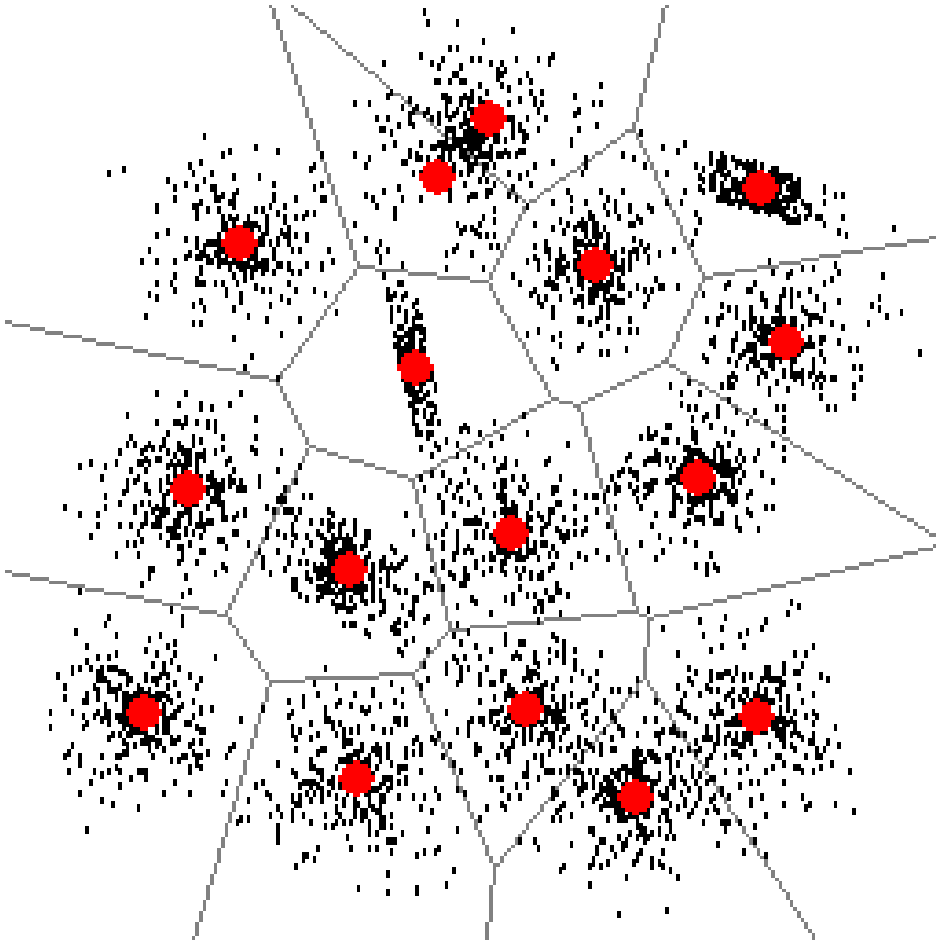
1-5-50-5-2=0,0035

1-50-10-50-3=0,385

Grafový model – příklad

- Problém s dlouhou cestou v grafu
  - $0,1 * 0,4 * 0,3 * 0,5 * 0,1 * 0,9 = 0,00054$
  - $0,3 * 0,6 * 0,5 * 0,7 * 0,3 * 0,6 = 0,01134$
  - $0,3 * 0,2 * 0,6 * 0,5 * 0,2 * 0,7 = 0,00252$
  - $0,5 * 0,6 * 0,7 * 0,7 * 0,5 * 0,8 = 0,05880$
- Porovnáváme velmi malá čísla – nepřesné
- Ztrácíme informaci o vzájemné „podobnosti“ dvou cest

Shluková analýza



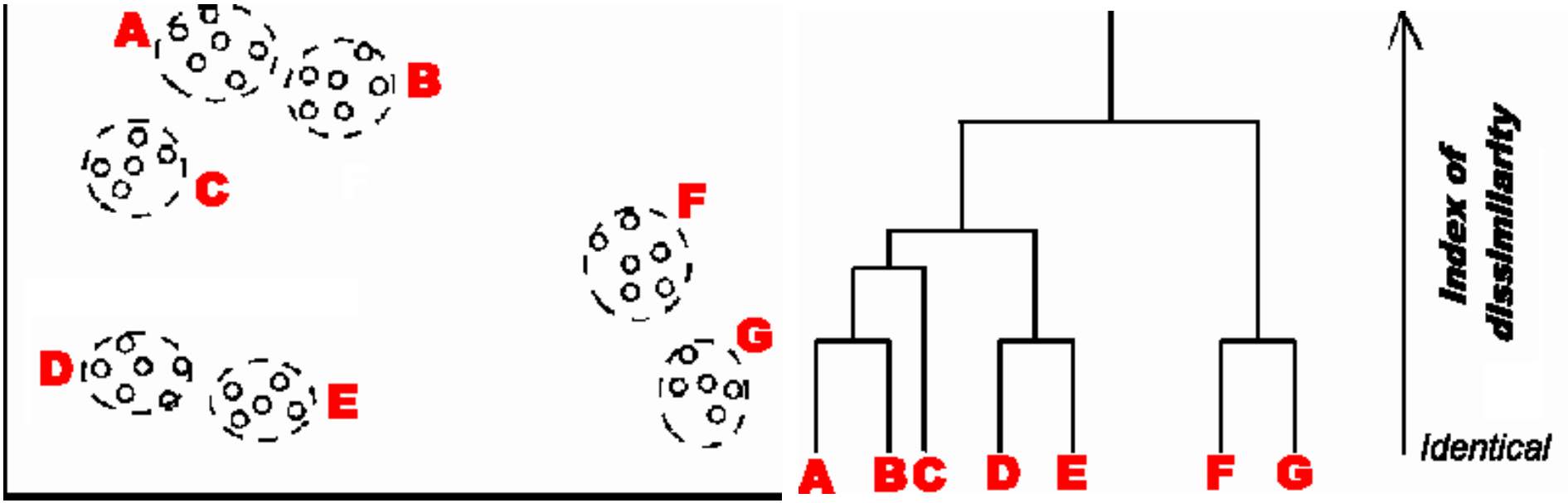
Úvod

- Shluková analýza zahrnuje několik různých algoritmů a metod pro shlukování podobných objektů do kategorií.
- Shlukování různých objektů tak, že podobnost mezi objekty je maximální, pokud patří do stejné skupiny a minimální, pokud nikoliv.
- Shlukovou analýzu lze využít pro odhalení „struktur“ v datech bez další interpretace.

Shluková analýza – metody

- Spojování (Stromové shlukování)
  - Spojování objektů do postupně větších shluků za použití nějaké metody pro určování podobnosti nebo vzdálenosti objektů
  - Typickým výsledkem tohoto přístupu je hierarchický strom – tzv. dendrogram
- Dvoucestné shlukování (Blokové shlukování)
  - V prvním kroku jsou vytvořeny shluky, které jsou v druhém kroku brány jako jednotlivé objekty, na které je aplikováno standardní hierarchické shlukování
- Shlukování metodou k průměrů
  - Vytvoření přesně k shluků, které jsou maximálně odlišné

Stromové shlukování



- Osa X reprezentuje vzdálenost spojení
- Spojování do větších shluků (algoritmus končí propojením všeho)

Určování vzdálenosti mezi objekty/shluky

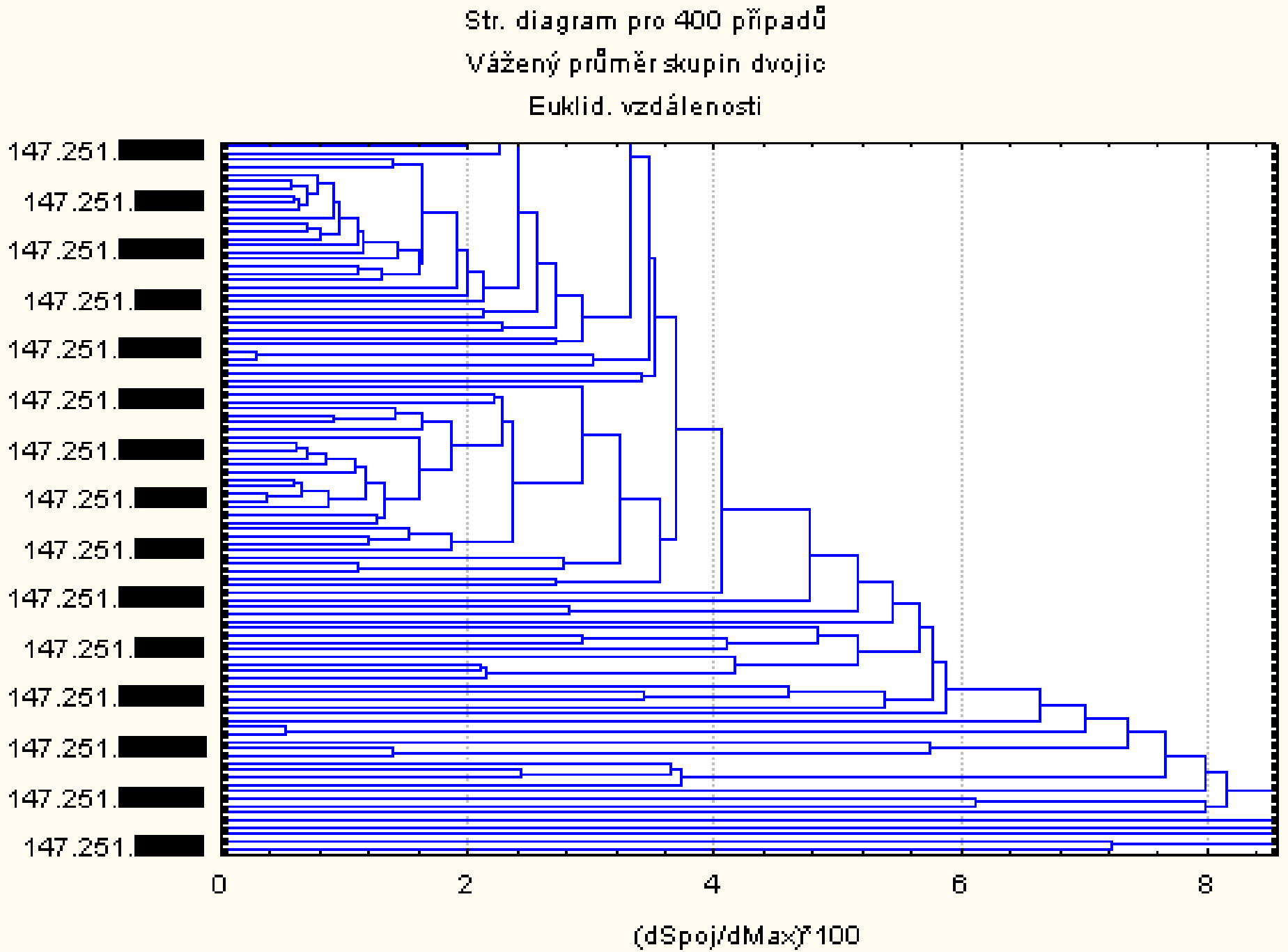
- Tvorba shluků
- Kritéria pro shlukování a odlišení objektů
- Existují metody pro jedno- i více-dimenzionální objekty
- Nejpoužívanější metoda pro více-dimenzionální objekty je *Euklidovská vzdálenost*

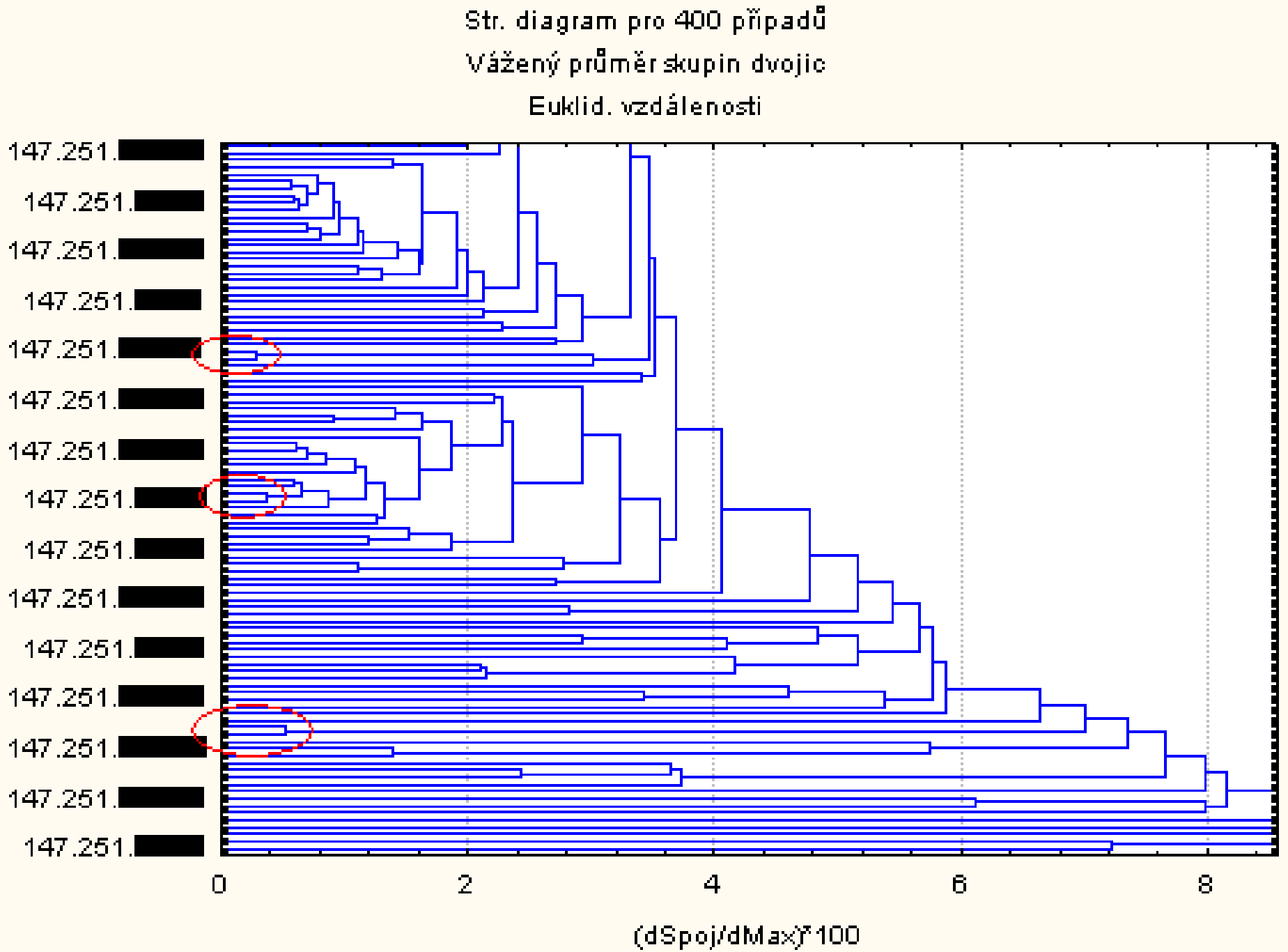
$$distance(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

- Některé další metody (např. větší důraz na izolované objekty atp.)
  - squared euclidean distance, chebychev distance, city-block (Manhattan)
- Hierarchické shlukování – určování vzdálenosti mezi shluky
  - Metoda nejbližšího souseda – minimální vzdálenost objektů různých shluků
  - Metoda nejvzdálenějšího souseda – maximální vzdálenost objektů různých shluků
  - UPMGA – vzdálenost je určena jako průměrná vzdálenost mezi všemi dvojicemi objektů v rámci dvou shluků

Použití metody shlukování – netflow MU

- Databáze logů se záznamy komunikace v síti MU
  - cca 180 miliónů záznamů za každý den
  - zdrojová/cílová IP; protokol; porty; čas; počet přenesených bytů...
- Omezující podmínky – pouze určitá část sítě; konkrétní port (Fakulta informatiky a koleje; port 80)
  - z logu vyhledáme určité množství nejnavštěvovanějších IP adres
  - pro zdrojové IP zjistíme počty „hitů“ na cílové IP
- Výsledkem je matice zdrojových vs. cílových adres a počty hitů
  - získali jsme vektory popisující „chování“ zdrojových IP adres
  - vstupní data pro shlukovou analýzu
  - pozn. použitý SW pro shlukovou analýzu – Statistica 7


Str. diagram pro 400 případů
Vážený průměr skupin dvojic
Euklid. vzdálenosti
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
0
2
4
6
8
(dSpoj/dMax)*100


Str. diagram pro 400 případů
Vážený průměr skupin dvojic
Euklid. vzdálenosti
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
147.251.
0
2
4
6
8
(dSpoj/dMax)*100

Dendrogram a tabulka popisující chování

	dst 1	dst 2	dst 3	dst 4	dst 5	dst 6	...
147.251.aaa.bbb	5	16	3	20	14	0	...
147.251.ccc.ddd	0	0	0	0	0	0	...
147.251.eee.fff	8	14	10	88	4	0	...
147.251.ggg.hhh	0	0	0	0	0	13	...
147.251.iii.jjj	120	0	0	0	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Další práce

- Vhodné omezení rozsahu vstupních dat
  - poměr počtu parametrů a počtu případů
  - entropie jednotlivých parametrů
- Použité informace z logu – usage based; frequency based; viewing-time based; visiting-order based
- Použití jiných algoritmů pro tvorbu shluků
- Další automatizace procesu shlukování
  - především automatické zpracování výsledků z programu Statistice
  - matice spojování
- Interpretace výsledků shlukové analýzy a diskuze vlivu různých kontextových informací na výsledek shlukování
- Hledání další vhodné aplikace popsaného přístupu

Otázky?

Děkuji za pozornost!