Dobývání znalostí

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

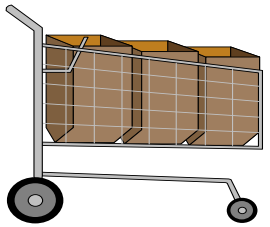
Dobývání znalostí

– Asociační pravidla –

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Analýza nákupního košíku
(MBA: Market Basket Analysis)



- **Analýza prodeje:**

  Které položky jsou v "košíku" pohromadě?

- **Výsledky:**
  - vyjádřené formou pravidel
  - lze bezprostředně použít

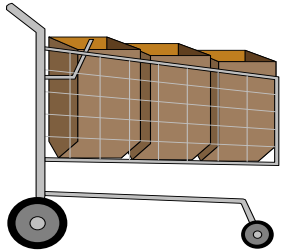
- **Použití:**
  - plánování a rozvržení obchodu
  - nabídka kupónů, omezení slev
  - "balení" produktů

Asociační pravidla

Jak spolu jednotlivé produkty navzájem souvisí?

- **Asociační pravidla by měla být:**
  - ***snadno pochopitelná:*** jakmile je nějaký vztah nalezen, lze ho snadno ověřit
  - ***použitelná:*** obsahují užitečné informace, které mohou vést k dalším intervencím
- **Asociační pravidla by neměla být:**
  - ***triviální:*** výsledky už stejně každý zná
  - ***nevysvětlitelná:*** neexistuje k nim žádné vysvětlení a nevedou k žádné akci

MBA pro porovnání obchodů

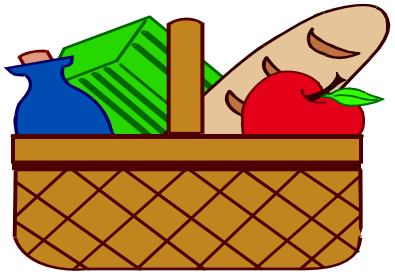


- **Virtuální položky:**
  - indikují typ obchodu, v němž proběhla daná transakce
  - neodpovídají žádnému výrobku ani službě
- **Porovnání nových a stávajících obchodů:**
  1. Pořídit data za určité období z nového obchodu
  2. Pořídit zhruba stejné množství dat ze stávajících obchodů
  3. Použít MBA k nalezení asociačních pravidel v obou sadách
  4. Uvažovat především asociační pravidla s virtuálními položkami

MBA - jak se to dělá?

- ***Položka*** - produkt nebo nabídka služeb
- ***Transakce*** obsahuje jednu nebo více ***položek***
- ***Tabulka četností***
  - udává počet výskytů libovolných dvou ***položek*** v některé z provedených ***transakcí*** (t.j. kolikrát byly tyto dva produkty zakoupeny najednou)
  - hodnoty na diagonále odpovídají ***počtu transakcí*** obsahujících příslušnou položku

MBA - příklad



- Transakce v potravinách:

Zákazník	Položky
1	chléb, máslo
2	mléko, chléb, máslo
3	chléb, káva
4	chléb, máslo, káva
5	káva, máslo

- Četnost produktů:

	chléb	máslo	mléko	káva
chléb	4	3	1	2
máslo	3	4	1	2
mléko	1	1	1	0
káva	2	2	0	3

Typ prodeje patrný z tabulky četností:

Chléb a máslo se nejspíš nakupují najednou.
Mléko se nikdy nekupuje společně s kávou.

MBA - asociační pravidla

- Pravidlo:

IF *Podmínka* THEN *Výsledek.*

(*Pravidlo_r* : IF *Položka_i* THEN *Položka_j* .)

- **Otázky:**
  - Jak dobrá jsou nalezená asociační pravidla?
    - podpora
    - spolehlivost
    - zlepšení
  - Jak hledat asociační pravidla automaticky?

Podpora a spolehlivost

Podpora: ***Jak často lze pravidlo použít?***

$$Podpora(Pravidlo\_r) = \frac{Nr\_transakcí\_obsahujících\_i\_a\_j}{Nr\_všech\_transakcí} \cdot 100\ \%$$

Spolehlivost: ***Jak moc se můžeme na výsledky pravidla spolehnout?***

$$Spolehlivost(Pravidlo\_r) = \frac{Nr\_transakcí\_obsahujících\_i\_a\_j}{Nr\_transakci\_obsahujících\_i} \cdot 100\ \%$$

Podpora a spolehlivost - příklad

Pravidlo 1: ***If zákazník kupuje chléb then zákazník kupuje také máslo.***

Pravidlo 2: ***If zákazník kupuje kávu then zákazník kupuje také máslo.***

Podpora (Pravidlo_1) = 3 / 5 • 100 % = 60 %

Podpora (Pravidlo_2) = 2 / 5 • 100 % = 40 %

Spolehlivost (Pravidlo_1) = 3 / 4 • 100 % = 75 %

Spolehlivost (Pravidlo_2) = 2 / 3 • 100 % = 66 %

Zlepšení pravidla

Zlepšení: ***Oč lepší je pravidlo při predikci použít než jeho výsledek prostě předpokládat?***

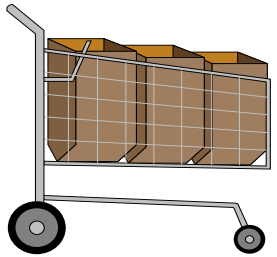
$$Zlepšení(Pravidlo\_r) = \frac{p(i\_a\_j)}{p(i) \bullet p(j)}$$

Pokud je Zlepšení < 1:

- pravidlo je při predikci horší než náhodná volba
- **NEGACE** výsledku může vést k lepšímu pravidlu

IF *Podmínka* THEN NOT *Výsledek.*

Zlepšení pravidla - příklad



Pravidlo: ***If zákazník kupuje mléko then zákazník kupuje také máslo.***

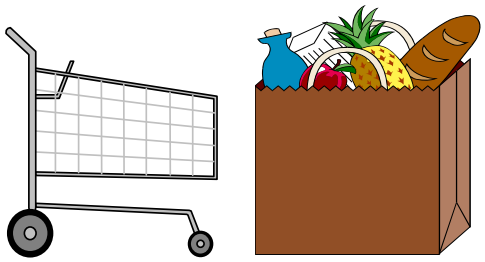
Podpora (Pravidlo_1) = 1 / 5 • 100 % = 20 %

Spolehlivost (Pravidlo_1) = 1 / 1 • 100 % = 100 %

Zlepšení (Pravidlo_1) = (1 / 5) / ((1 / 5) • (4 / 5)) =

= 5 / 4 = 1.25

Hlavní kroky MBA



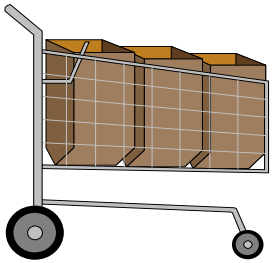
- **Zvolte** odpovídající **položky** na adekvátní úrovni
- **Vytvořte pravidla** na základě údajů z tabulky četností
  - spočítejte (podmíněné) pravděpodobnosti výskytu položek a jejich kombinací v transakcích
  - omezte prohledávání prahovou hodnotou pro podporu
- **Určete nejlepší pravidla** analýzou vypočtených pravděpodobností
  - překonat omezení daná počtem položek a jejich kombinací v "zajímavých" transakcích

MBA - volba vhodných položek

Pořízení transakčních dat:

- často horší kvalita, vyžaduje rozsáhlejší předzpracování
- relevance položek se časem může změnit
- adekvátní úroveň zpracování:
  - rostoucí počet kombinací jednotlivých položek
  - výsledky lze bezprostředně použít (specifické položky)
  - pravidla s dostatečně vysokou podporou (častý výskyt v množině dat)

Taxonomie: hierarchie kategorií



MBA - Složitost generovaných pravidel:

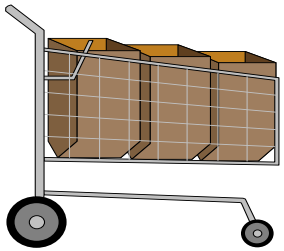
- Na začátku použít obecnější položky
- Později generovat pravidla pro specifické položky výlučně na základě transakcí, které tyto položky obsahují

MBA - Použitelné výsledky:

Položky by měly být ve zhruba stejném počtu transakcí:

- přesunout řídké položky do vyšších úrovní taxonomie (kde budou častější)
- obvyklejší položky nechat na nižších úrovních (aby pravidlům nedominovaly nejčastější položky)

Virtuální položky:
přesahují rámec tradiční taxonomie

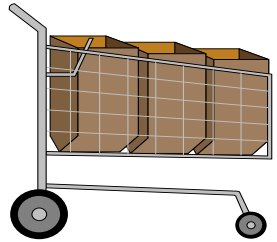


- Stírají hranice mezi jednotlivými typy výrobků u původních položek
  - např. (firemní) značky - Calvin Klein
- mohou obsahovat informatice o transakci samotné
  - ***anonymní*** (den v týdnu, čas, atd.)
  - ***signované*** (informace o zákaznících a jejich chování)
- mohou být vést k redundancím v pravidlech
  - položkám z taxonomie odpovídá jen jedna jediná virtuální položka ("*If výrobek_Škoda then Škoda.*")
  - virtuální a obecné položky se v pravidle objeví najednou ("*If výrobek_Škoda a malé auto then stan*" namísto "*If Fabia then stan*")

MBA - generování pravidel

- **Výpočet tabulky četností:**
  - udává informace o tom, které kombinace jednotlivých položek se v transakcích vyskytují nejčastěji
  - lze použít k určení základních pravděpodobností potřebných k posouzení významu generovaných pravidel
- **Získat užitečná pravidla:**
  - zlepšení by mělo být větší než 1
    - malé zlepšení lze zvětšit negací pravidel
    - negovaná pravidla mohou být hůře použitelná než ta původní
  - redukce počtu generovaných pravidel - **PROŘEZÁVÁNÍ**

Prořezávání podle minimální podpory



Eliminace méně častých položek

- podniknuté akce by se měly ***týkat dostatečného počtu transakcí***
- dvě možnosti:
  - eliminace řídkých položek (a následná eliminace příslušných asociačních pravidel)
  - použití taxonomie k vytvoření obecných položek (generalizované položky by měly vyhovovat nastaveným kritériím - prahu)
- ***variabilní minimální podpora*** - kaskádový efekt

Algoritmus APRIORI (R. Agrawal)

- Generování asociačních pravidel

  → **Hledání často se pakujících množin položek**

~ Frequent itemsets:

- Kombinace (~ konjunkce) kategorií, které dosahují předem zadané četnosti na datech **~ ´minsup´ podpora**
- Při hledání kombinací délky ***k*** využíváme znalosti kombinací délky ***k – 1*** **~ generování kombinací do šířky**
- Pro vytvoření kombinace délky ***k*** požadujeme, aby všechny její podkombinace délky ***k – 1*** splňovaly požadavek na četnost

Algoritmus APRIORI
(pokračování)

~ **Frequent itemsets (pokračování):**

- Po nalezení kombinací, které vyhovují četností, se vytvářejí asociační pravidla
  - (***četnost_nadkombinace*** $\leq$ ***četnost_kombinace***)

→ Každá kombinace ***Comb*** se rozdělí na všechny možné dvojice podkombinací ***Ant*** a ***Suc*** takové, že
Suc = Comb – Ant
($Ant \cap Suc = \{\}$ a $Ant \wedge Suc = Comb$)

→ Uvažované pravidlo ***Ant => Suc*** pak má **podporu** odpovídající četnosti kombinace ***Comb*** , jeho **spolehlivost** odpovídá podílu četností kombinací ***Comb*** a ***Ant***

Algoritmus APRIORI

Krok 1: Do L_1 přiřaď všechny kategorie, které dosahují alespoň požadované četnosti

Krok 2: Polož $k = 2$

Krok 3: Dokud $L_{k-1} \neq \{\}$

Krok 3.1: Pomocí funkce ***APRIORI-GEN*** vygeneruj na základě L_{k-1} množinu kandidátů C_k

Krok 3.2: Do L_k zařaď ty kombinace z C_k, které dosáhly alespoň požadovanou četnost

Krok 3.3: Zvětši počítadlo k

Funkce APRIORI-GEN (L_{k-1})

Krok 1: Pro všechny dvojice kombinací $Comb_p$, $Comb_q$ z L_{k-1}

Krok 1.1: Pokud se $Comb_p$ a $Comb_q$ shodují v $k - 2$ kategoriích, přidej $Comb_p \wedge Comb_q$ do C_k

Krok 2: Pro každou kombinaci $Comb$ z C_k

Krok 2.1: Pokud některá z jejich podkombinací délky $k - 1$ není obsažena v L_{k-1} , odstraň $Comb$ z C_k

MBA - Disociační pravidla

- **Pravidlo:** **IF *A* AND NOT *B* THEN *C*.**
  - Zavést nové položky inverzní k původním položkám
  - V případě, že transakce neobsahuje původní položku, bude obsahovat položku znegovanou
- **Nevýhody:**
  - dvojnásobný počet položek
  - narůstá velikost transakcí
  - negované položky se vyskytují častěji než ty původní (pravidla se všemi položkami negovanými lze hůř využít: "IF NOT *A* AND NOT *B* THEN NOT *C*.")

Analýza časových řad pomocí MBA

- **Analýza příčin a následků:**
  - informace o čase, resp. sledu událostí k určení toho, kdy transakce nastaly (jedna vzhledem ke druhé)
  - obvykle vyžaduje nějakou identifikaci zákazníka
- **Převedení problému na MBA:**
  - zavedení nových položek do transakcí před sledovanou událostí (pro *příčiny*) a po sledované události (pro *následky*); následně odstranění duplicitních položek z transakcí
  - ***okénko***: "snímek" veškerých údajů, které se vyskytly během určitého období (např. všechny transakce za uplynulý měsíc)
    - trendy pro řídké položky

Výhody MBA

- Dává jasné a srozumitelné výsledky
  - ***IF - THEN - pravidla s bezprostředním použitím***
- ***Dobývání znalostí*** (bez požadovaných výstupů)
  - důležité při zpracovávání velkého množství dat bez dalších apriorních znalostí
- Umožňuje zpracovávat ***data s variabilní délkou***
- ***Snadné a srozumitelné*** výpočty
  - Výpočetní **nároky rostou exponenciálně** s počtem položek!

Nevýhody MBA

- Exponenciální nárůst výpočetních nároků
  - potřeba taxonomií a virtuálních položek
- Omezená podpora některých položek
  - prořezávání méně použitelných obecných položek
- Obtížné určení adekvátního počtu položek
  - položky by měly mít zhruba stejnou četnost
- Znevýhodňuje řídké položky
  - variabilní hodnoty prahu při prořezávání podle minimální podpory
  - vyšší úrovně položek v taxonomii

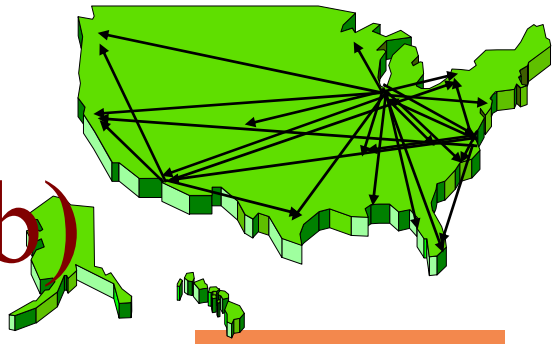
Dobývání znalostí

– Analýza linků –

Doc. RNDr. Iveta Mrázová, CSc.

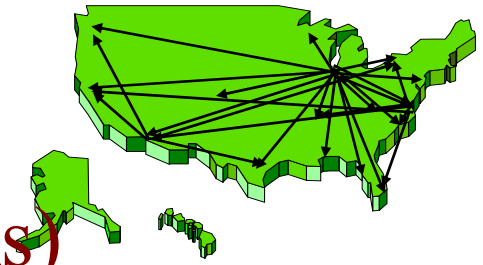
Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Analýza linků (vazeb)



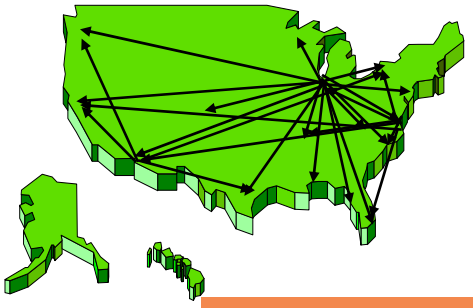
- **Cíle:**
  - ***nalezení vztahů mezi údaji***
  - ***vizualizace linků a vztahů***
- **Aplikace:**
  - telekomunikace
  - kriminalistika a právo - skupiny zločinců jsou navzájem propojené, analýza těchto vztahů je může pomoci rozkrýt
  - marketing - vztahy mezi zákazníky

SF-sítě (Scale-Free Networks)

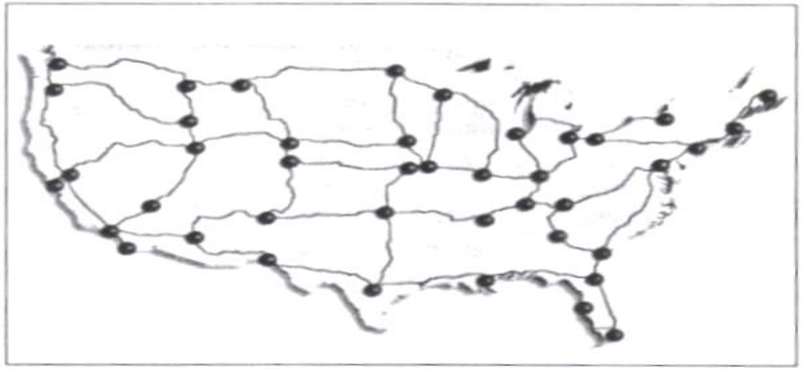


- Některé uzly mají extrémně velký počet vazeb (hran) na další uzly - **hub**
- Většina uzlů má jen několik málo vazeb na další uzly
- Odolné proti náhodným poruchám
- Zranitelné při koordinovaném útoku
- **Nové oblasti použití:**
  - ochrana před (počítačovými) viry šířenými po Internetu
  - medicína (očkování)
  - byznys (marketing)

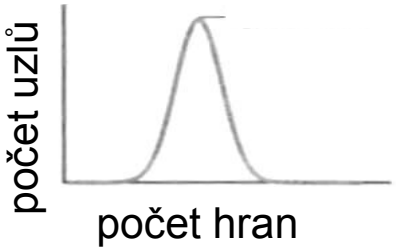
SF-sítě



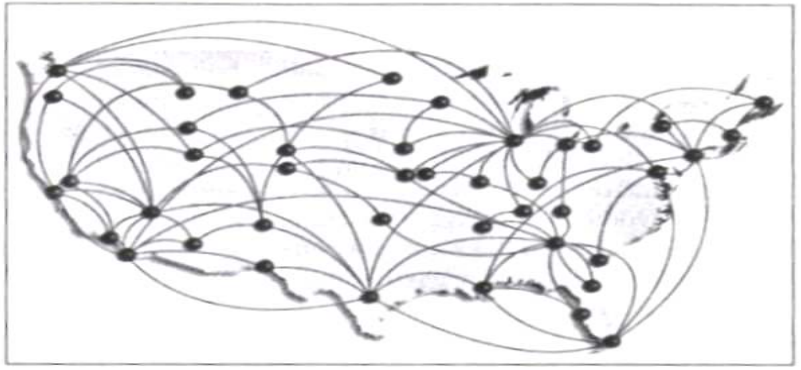
Náhodný graf



rozložení hran



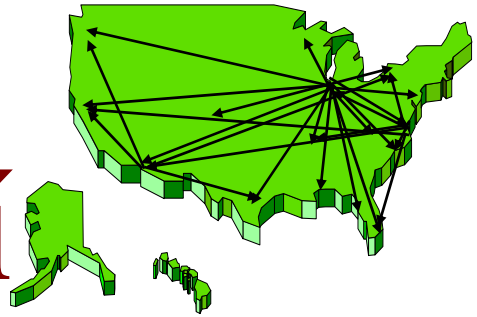
SF-síť



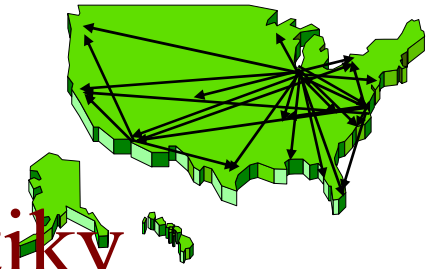
rozložení hran

Převzato z "A. L. Barabasi and E. Bonabeau: *Scale-Free Networks*, Scientific American, May 2003"

Příklady SF-sítí



- **Sociální sítě**
  - vědecká spolupráce (vědci, spoluautorstcí článků)
  - Hollywood (herci, natáčení ve stejném filmu)
- **Biologické sítě**
  - buněčný metabolismus (molekuly zůčastněné při produkci energie, účast v téže biologické reakci)
  - proteinové regulační sítě (proteiny řídící aktivitu buněk, interakce mezi proteiny)
- **Socio-technické sítě**
  - Internet (routery, optická a další spojení)
  - World Wide Web (webové stránky a URL)



SF-sítě: základní charakteristiky

Dva základní mechanizmy:

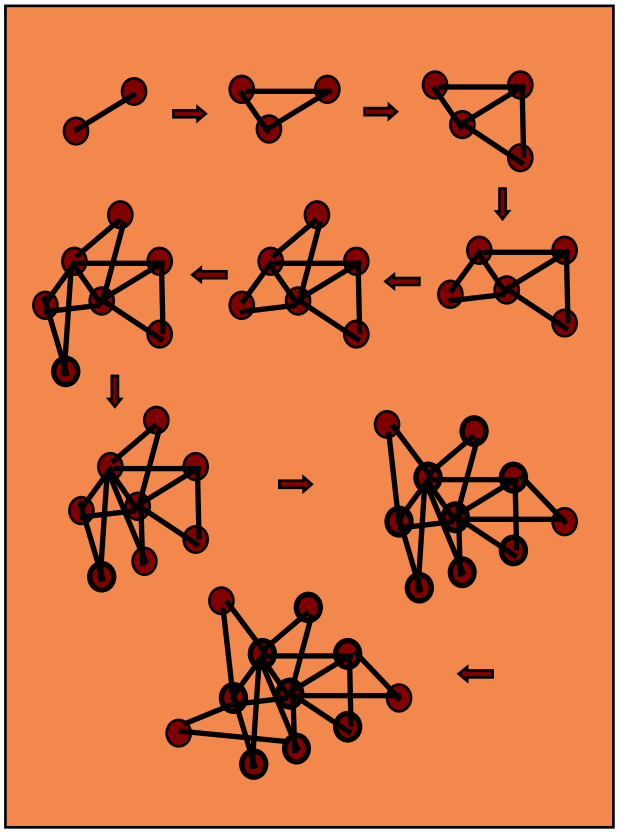
- **růst**
- **preferenční napojení**

"Bohatí bohatnou" (hubs):

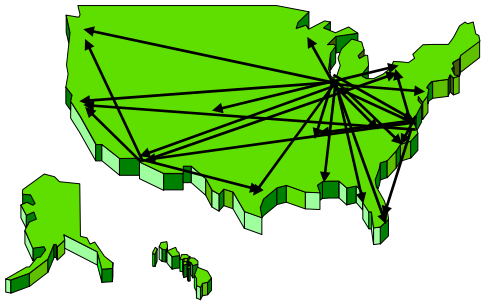
- nové uzly mají tendenci připojovat se k uzlům s větším počtem vazeb
- "populární lokality" časem získají více vazeb než jejich sousedi s menším počtem vazeb

Spolehlivost

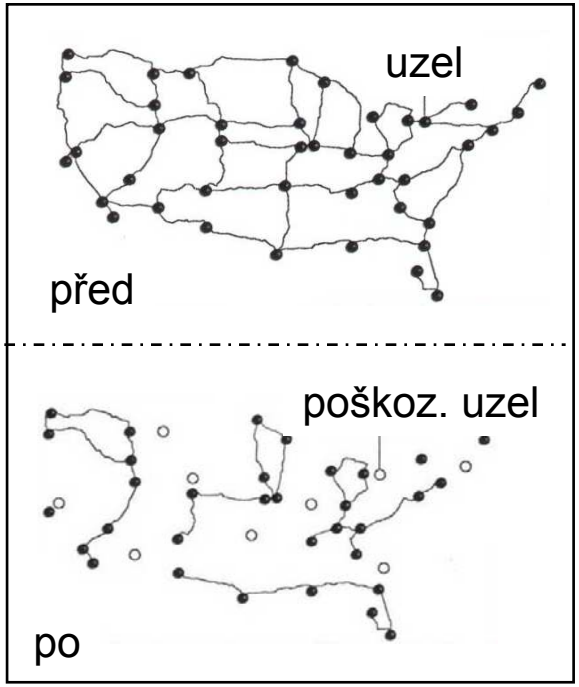
- ***náhodná selhání*** (80% *náhodně zvolených* uzlů může selhat aniž by to vedlo k fragmentaci klastru)
- ***koordinované útoky*** (eliminace 5-15% *všech hubů* může vést k selhání celého systému)



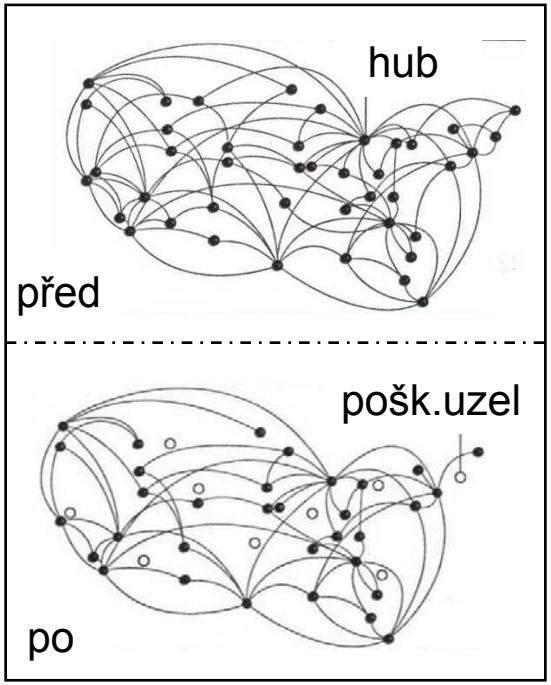
SF-sítě



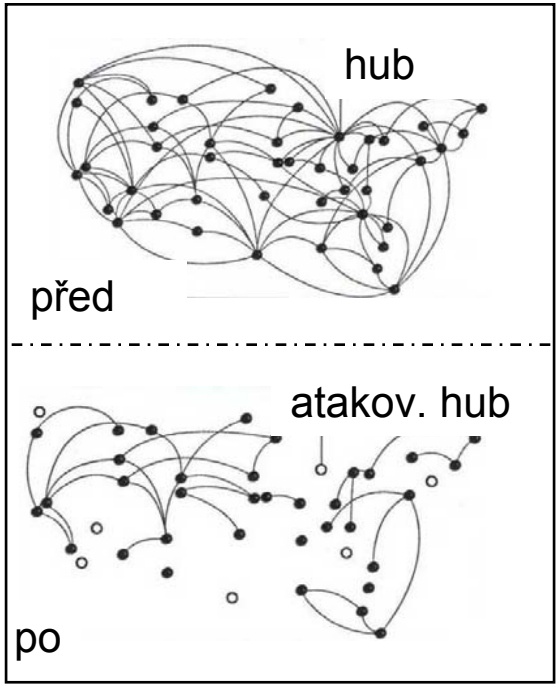
Náhodná síť: selhání náhodného uzlu



SF-síť: selhání náhodného uzlu

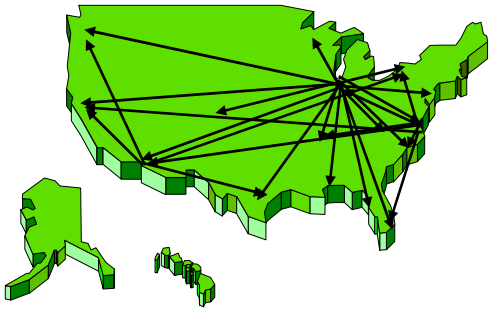


SF-síť: koordinovaný útok na huby



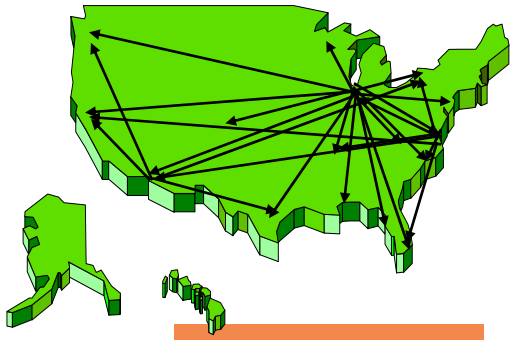
převzato z "A. L. Barabasi and E. Bonabeau: *Scale-Free Networks*, Scientific American, May 2003"

Využití SF-sítí



- **Computing**
  - sítě se SF-architekturou
- **Medicína**
  - očkovací kampaně a nové léky
- **Byznys**
  - kaskádové finanční krachy
  - marketing

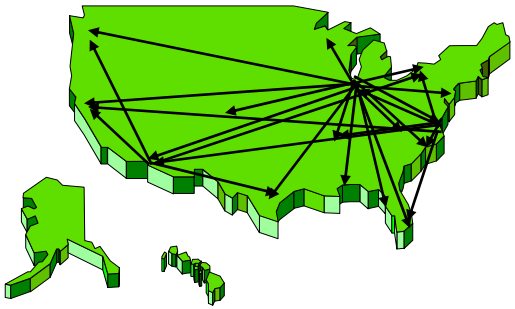
Využití SF-sítí



Computing

- počítačové sítě se SF-architekturou (např. WWW)
  - extrémně odolné vůči náhodným selháním
  - velmi zranitelné při koordinovaném útoku nebo sabotáži
- vymýcení internetových virů je v podstatě nemožné

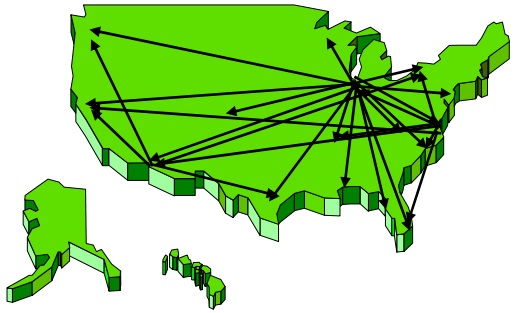
Využití SF-sítí



Medicína

- očkovací kampaně zacílené na huby
  - lidé s mnoha kontakty a styky
  - obtížná identifikace těchto lidí
- nové léky zacílené na klíčové molekuly (huby) zúčastněné v příslušných chorobách
- ovlivnění vedlejších účinků léků prostřednitvím zmapovaných sítí uvnitř buněk

Využití SF-sítí



Byznys

- **finanční krachy**
  - pochopení vzájemných vazeb mezi společnostmi, průmyslem a ekonomikou
  - monitorování a eliminace kaskádových finančních krachů
- **marketing**
  - studium šíření nákazy v SF-sítích
  - efektivnější reklama pro nové výrobky