IV124 Komplexní sítě

Jan Fousek, Eva Hladká

Fakulta informatiky, Masarykova univerzita

29. března 2018

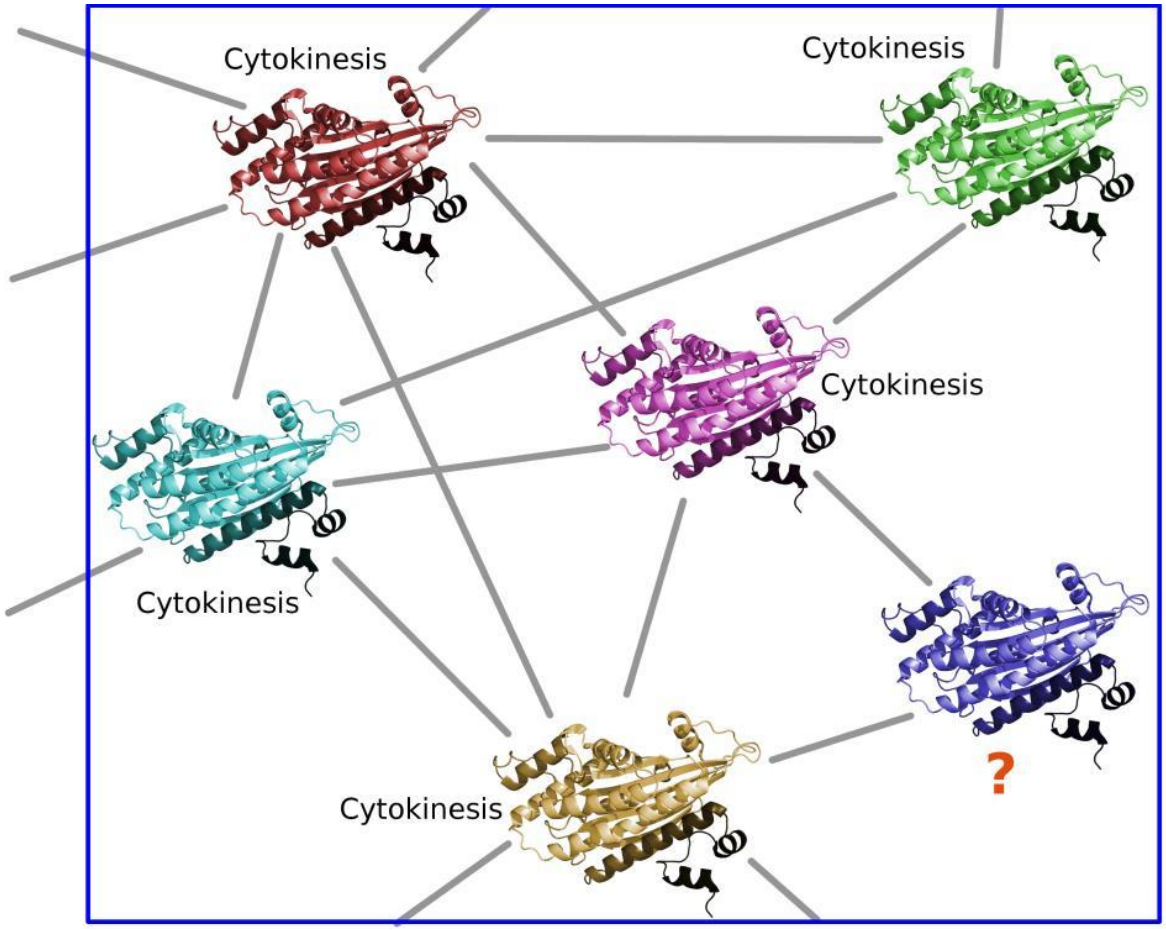
Komunitní struktura

Motivace

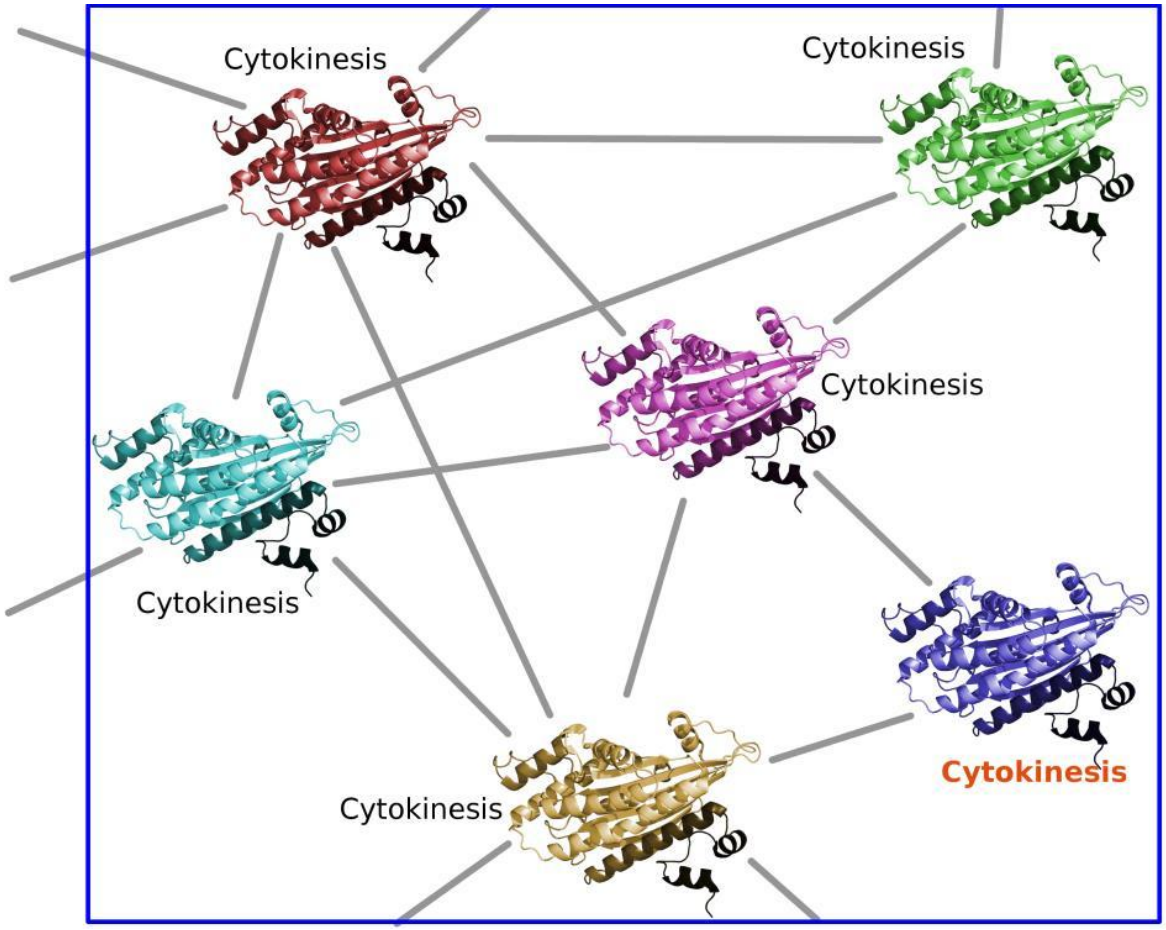
- v reálných sítích často pozorujeme formování klastrů
- těsně propojené *klastry* často odpovídají *komunitám*, které sdílí nějakou vlastnost

Přesná definice komunity/klastru závisí na povaze zkoumaného systému.

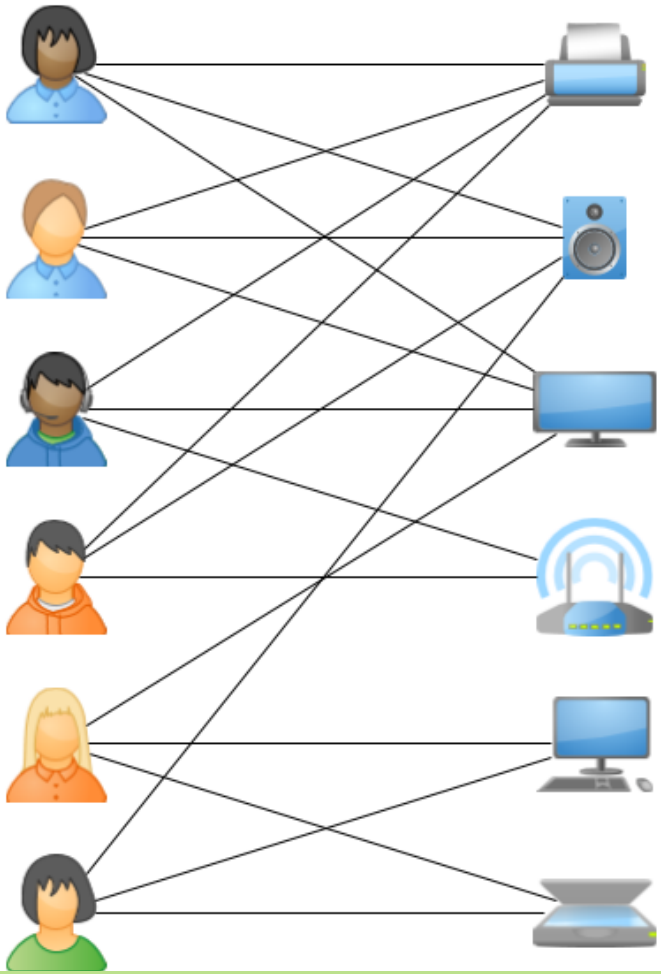
Motivační příklad: funkce proteinů



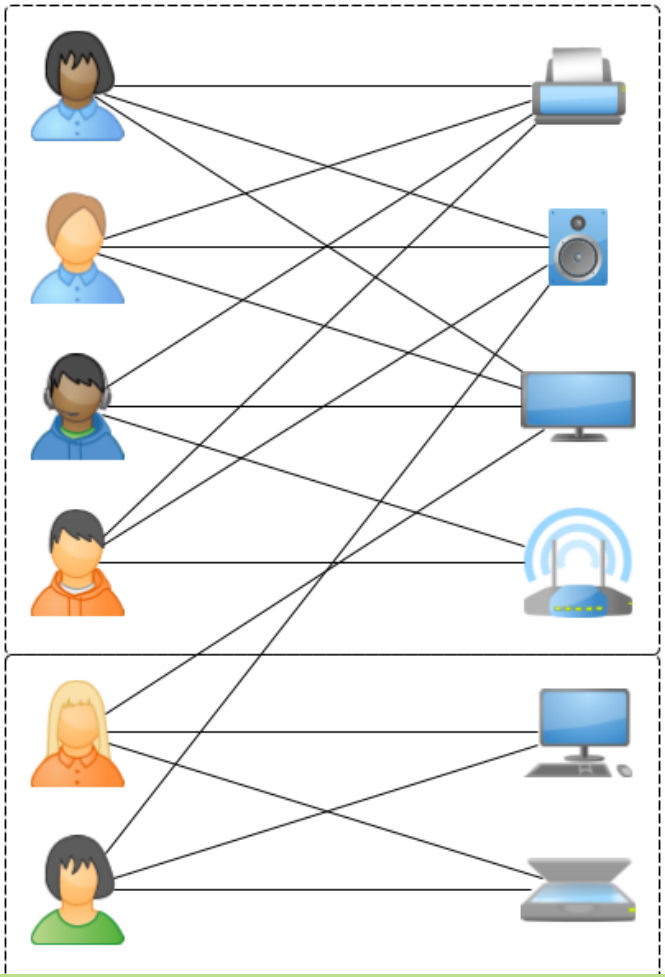
Motivační příklad: funkce proteinů



Motivační příklad: systémy doporučení



Motivační příklad: systémy doporučení



Detekce komunitní struktury

1. máme síť s konkrétní sémantikou (sociální, dopravní, biologická, ...)
2. identifikujeme klastry
3. klastry interpretujeme jako funkční celky, nebo reálné komunity

V čem je problém

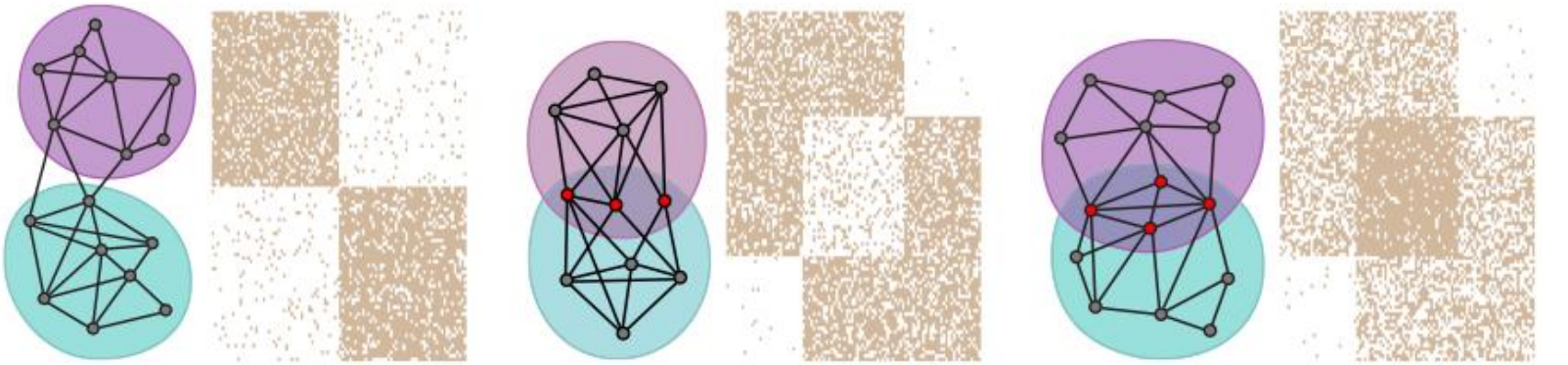
Nejasně zadaný problém

- kvalita rozdělení na klastry není jednoznačná
- interpretace nemusí být přímočará
- u většiny sítí nemáme proti čemu porovnávat výsledek

Komplikující vlastnosti sítě

- orientované hrany
- ohodnocené hrany
- hierarchická struktura
- překrývající se komunity

Překrývající se komunity

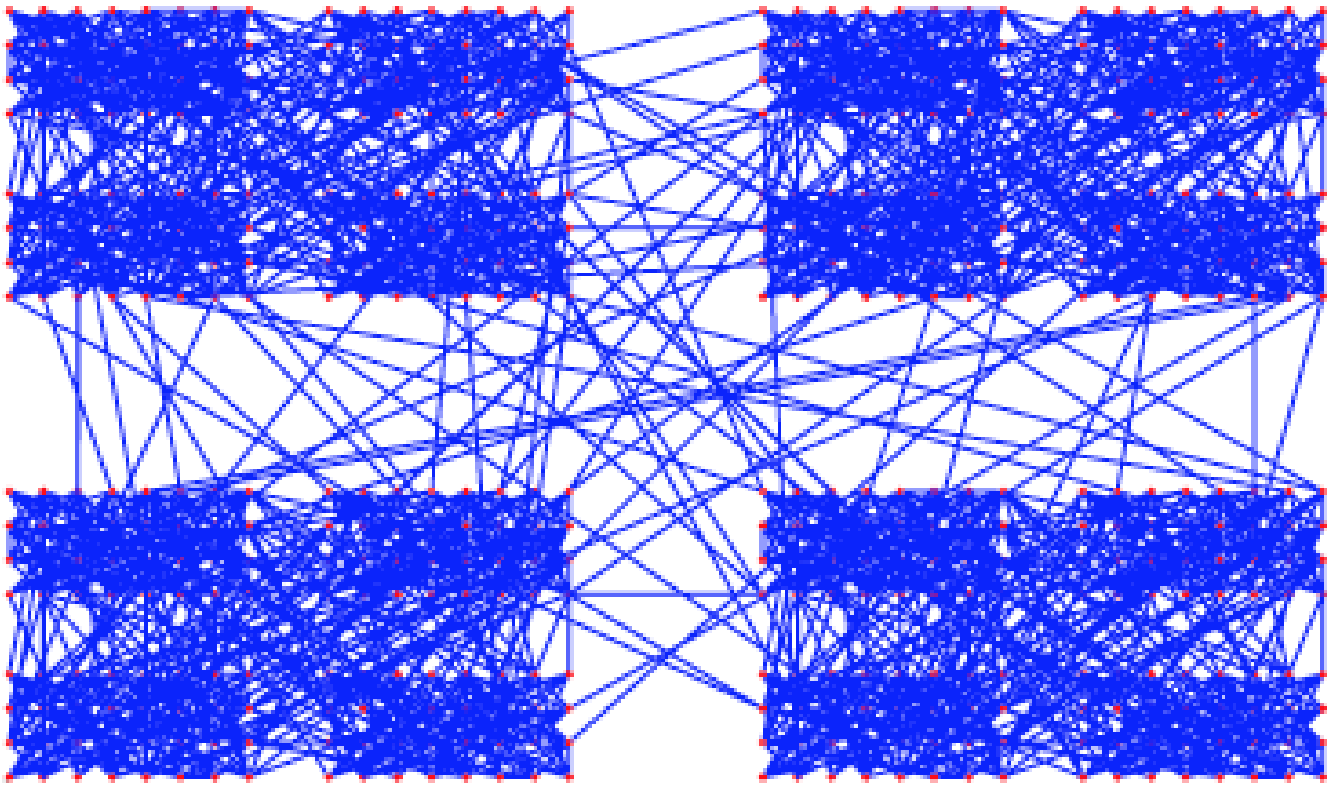


(a) No overlaps (b) Sparse overlaps (c) Dense overlaps [1]

Husté překryvy působí většině algoritmů problém.

[1] Yang &Leskovec (2014)

Hierarchická struktura



Hierarchická shluková analýza

Obecná metoda pro třídění prvků do skupin

- hierarchický systém podmnožin
- podobnostní funkce (vzdálenost)
- prvky uvnitř každé množiny jsou si podobnější mezi sebou, než s prvky vně
- reprezentujeme dendrogramem

Varianty

- aglomerativní: sjednocováním od jednotlivých prvků
- divizní: rozdělováním k jednotlivým prvkům

Hierarchická shluková analýza

V kontextu sítí je třeba definovat podobnost W_{ij}

Časté volby:

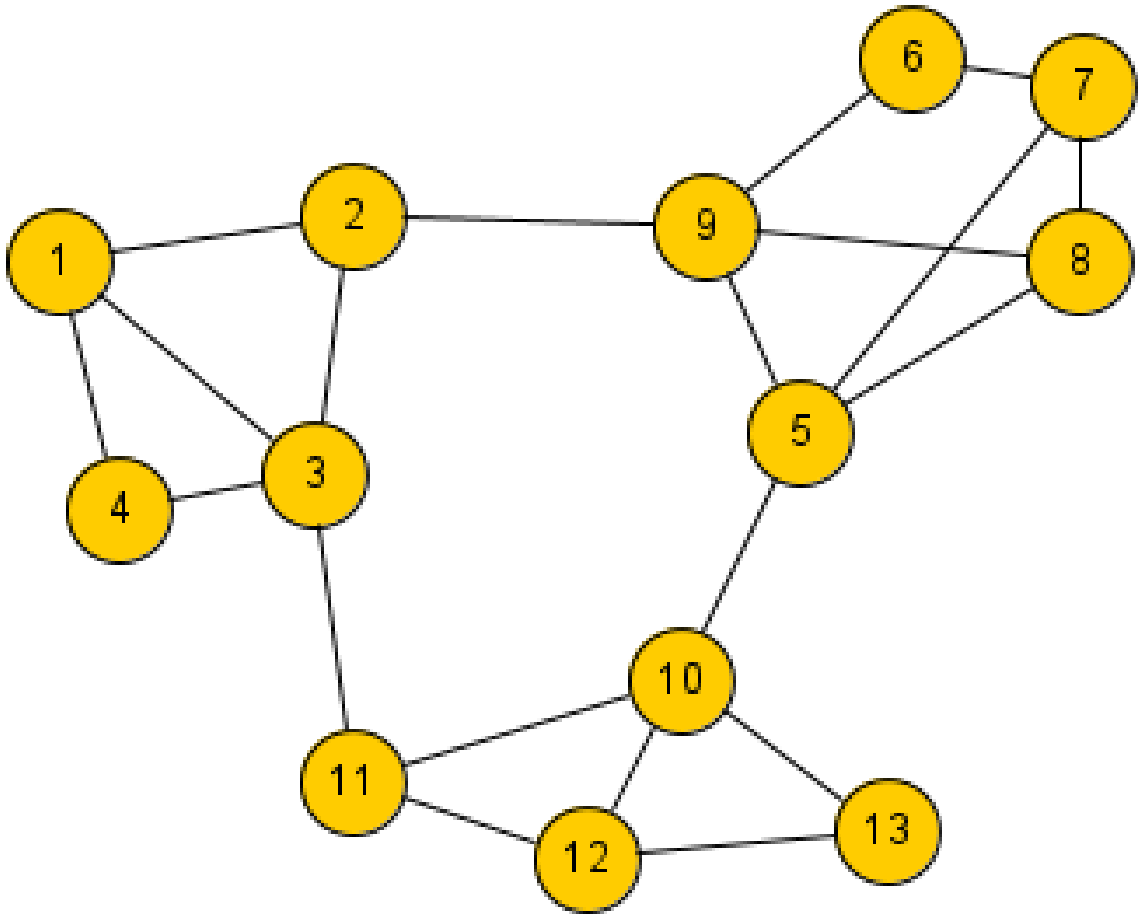
- počet po vrcholech nezávislých cest mezi i a j
  - nesmí sdílet jiné než koncové uzly
- počet po hranách nezávislých cest
  - každá hrana se může vyskytovat v nejvýše jedné cestě

Betweenness clustering

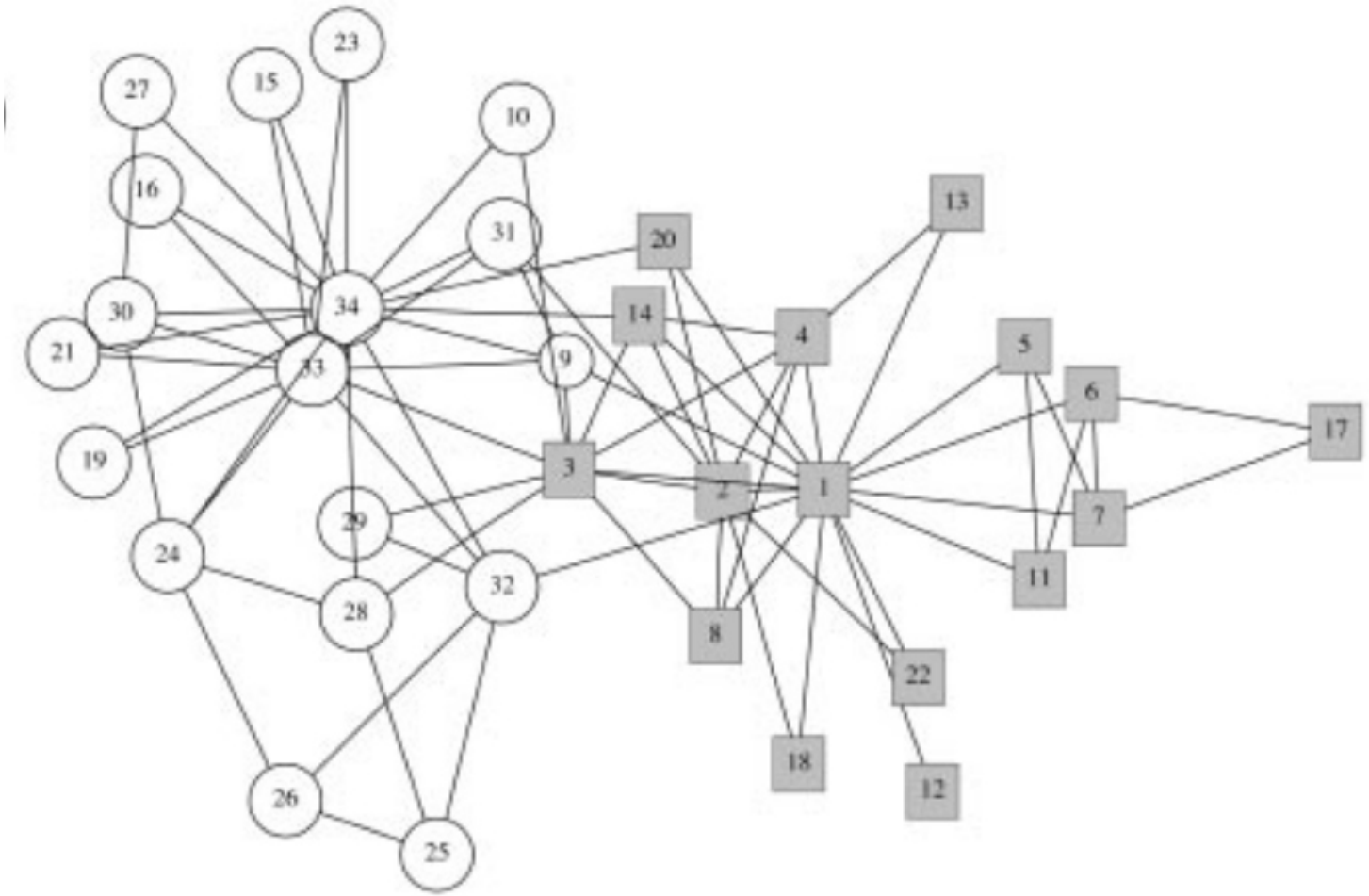
Hlavní myšlenka

- hrany s vysokou betweenness považujeme za mosty mezi komunitami
- postupně odstraňujeme hrany od nejcentrálnějších
- vznikající komponenty považujeme za komunity

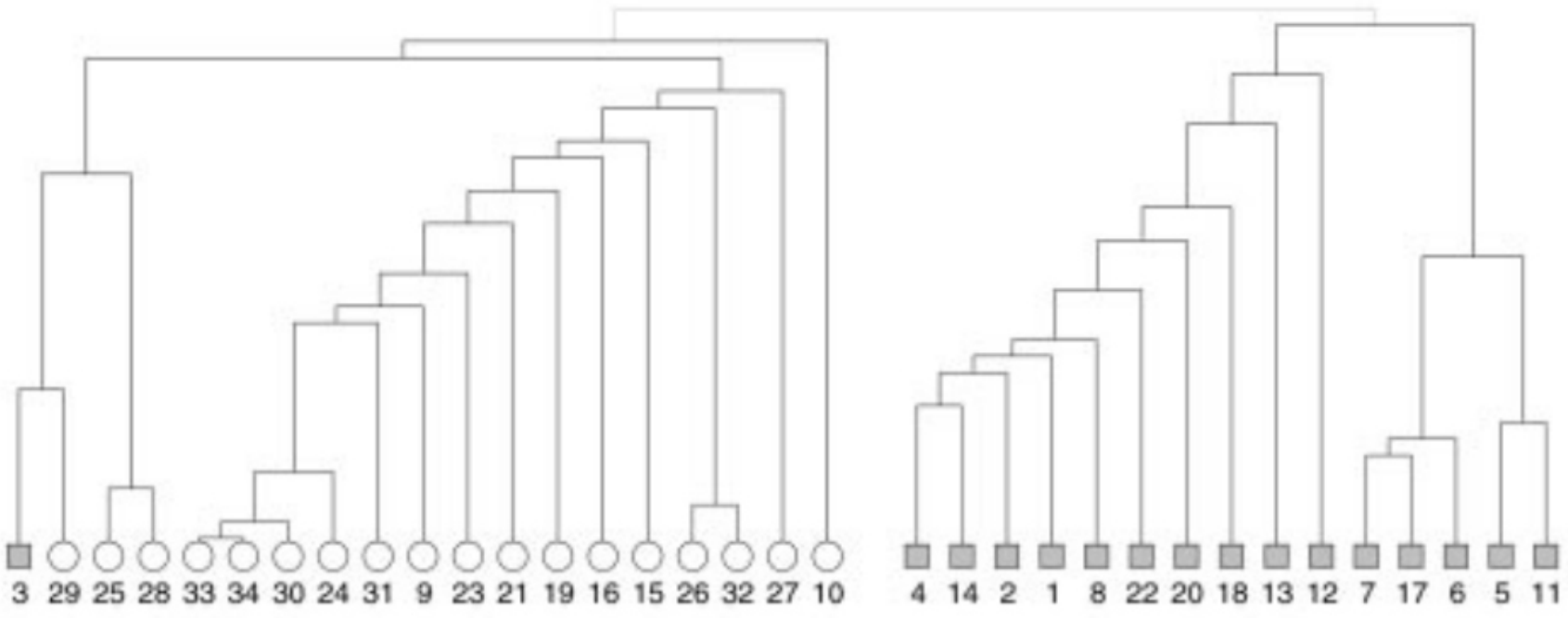
Betweenness clustering



Příklad: Zacharyho karate klub [2]



Příklad: Zacharyho karate klub [3]



[3] Girvan, M., & Newman, M. E. (2002)

Modularita

Hlavní myšlenka

- vytvoříme rozdělení uzlů do skupin C
- rozdělení ohodnotíme funkcí $Q(C)$
- hledáme maximum pro Q

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

- kde $P_{ij} = \frac{k_i k_j}{2m}$ je pravděpodobnost hrany mezi i a j
- $\delta(a, b) = 1 \iff a = b$

Modularita: vlastnosti

- Q udává míru separace komunit
- pro náhodnou síť $Q = 0$
- výpočetně náročné, NP úplný problém
- optimalizační heuristiky (např. simulované žíhání)

Modularita: efektivní algoritmus[4]

Hladový přístup:

- začneme s izolovanými uzly
- postupně spojujeme dvojice klastrů tak, že ΔQ je maximální
- konec, pokud nelze spojením dvou klastrů Q zlepšit

Úspěšně nasazeno na sítích s $|V| > 400k$ (např. související položky na Amazonu).

[4] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6), 066111.

Modularita: rozlišení

Hlavní problém

- nulový model je *globální*: $\frac{k_i k_j}{2m}$
- ve velké síti mají komunity spíše lokální charakter
- problémy s komunitami řádově různých velikostí

Řešení: limit rozlišení

- $Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \gamma P_{ij}) \delta(C_i, C_j)$
- malé γ upřednostňuje více malých komunit
- velké γ upřednostňuje méně velkých komunit

Lokální optimalizace[5]

Hodnotící funkce klastru

- $f(C) = \frac{k_{int}}{(k_{ext}+k_{int})^{\alpha}}$
- k_{int} suma vnitřních stupňů klastru
- k_{ext} suma vnějších stupňů klastru
- α je parametr rozlišení

[5] Lancichinettiet al., Detecting the overlapping and hierarchical community structure in complex networks, New Journal of Physics, 2009

Lokální optimalizace[6]

Postup detekce

- začneme s jedním uzlem
- připojujeme sousedy tak, že Δf je maximální
- v každém kroku testujeme, zda odstraněním uzlu nemůžeme zvýšit f
- klastr uzavřen, pokud nemůžeme přidáním sousedícího uzlu zvýšit f
- začíname odznovu s nezařazeným uzlem

[6] Lancichinettiet al., Detecting the overlapping and hierarchical community structure in complex networks, New Journal of Physics, 2009

Testování klastrovacích algoritmů

Posouzení kvality konkrétního algoritmu je obtížné

- zobecnitelnost vs. přesnost v konkrétním případě
- obtížně se získávají trénovací data se známou komunitní strukturou
- Yang, Jaewon, and Jure Leskovec. *Defining and evaluating network communities based on ground-truth.* Knowledge and Information Systems 42.1 (2015): 181-213.

Testování klastrovacích algoritmů

LFR Benchmark

- sada syntetických sítí s komunitní strukturou
- různé distribuce velikosti klastrů, stupně a dalších vlastností sítě
- umožňuje srovnávat jednotlivé algoritmy na obecných sítích
- Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. Physical review E, 78(4), 046110.

Ukázka: formování názoru

...