Úvod do analýzy sociálních sítí

Kde vzít sítě pro analýzu

2024-25

Zdroje

- Sítě si můžeme připravit sami (dotazníky, okolí v sociální síti apod.).

- Můžeme použít známé databáze (sklady, repozitáře), v nichž autoři sbírají sítě, které připravili výzkumníci v oblasti analýzy sítí.

- Na vytvoření sítí s dopředu známými vlastnostmi můžeme použít dostupné generátory nebo si generátory můžeme napsat sami.

Některé databáze poskytující sítě k analýze

- Sítě sesbírané Markem Newmanem
  - http://www-personal.umich.edu/~mejn/netdata/
- SNAP (Stanford Network Analysis Project) – projekt na univerzitě ve Stanfordu udržovaný Jure Leskovcem (včetně kódů v C++ a Pythonu na analýzu sítí)
  - http://snap.stanford.edu/
- Network Repository (stovky datasetů různých typů)
  - http://networkrepository.com/index.php

Co je to komunita?

- Hypotéza o propojenosti a hustotě: *Komunita je lokálně hustě propojený podgraf v síti.*

- Maximální klika: *Komunita je skupina jednotlivců, jejíž členové se navzájem znají.* Z hlediska teorie grafů to znamená, že komunita je úplný podgraf neboli klika.

  - Zatímco trojúhelníky jsou v sítích časté, větší kliky jsou vzácné.

  - Požadavek, aby komunita byla úplným podgrafem, může být příliš silný, protože vynechá mnoho dalších legitimních komunit.

Silné a slabé komunity

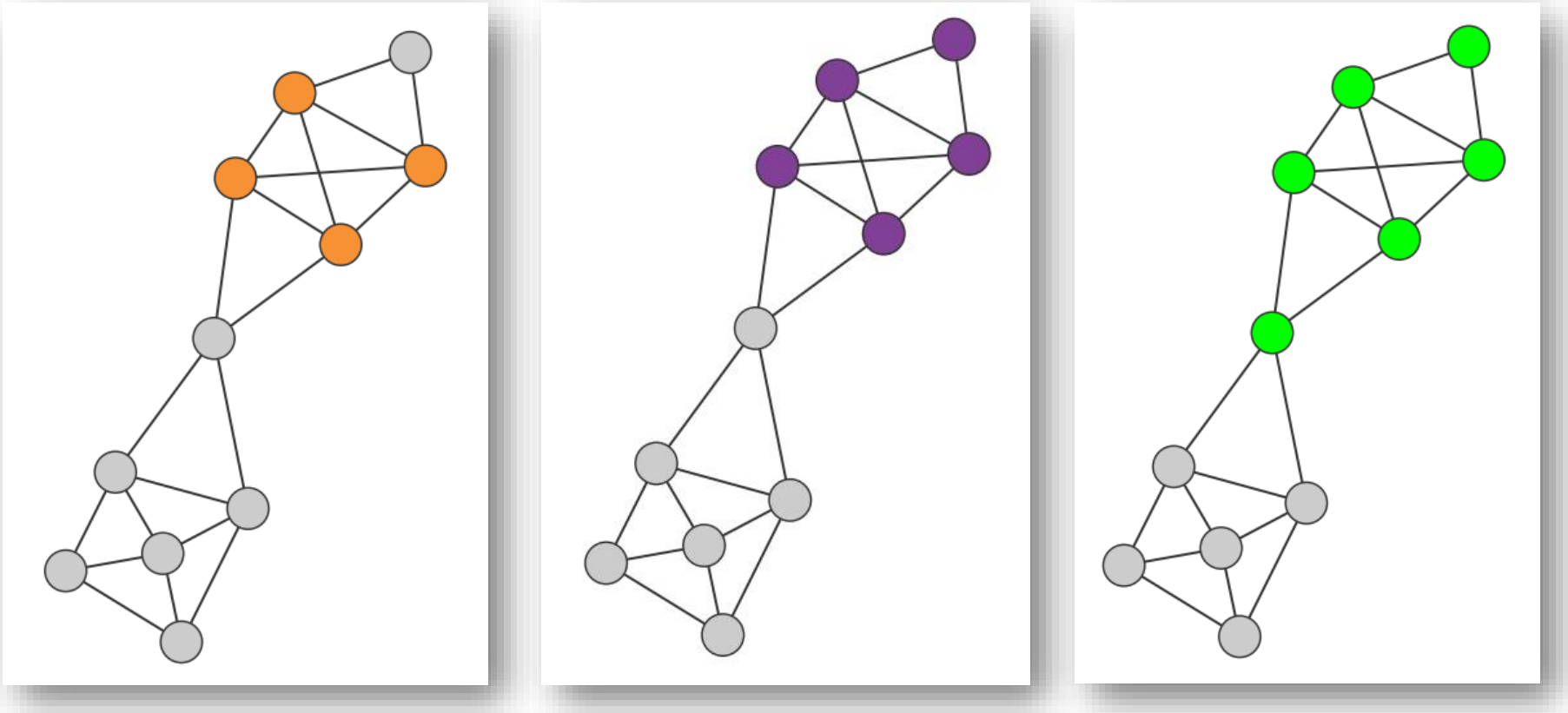
- *C* je *silná komunita,* pokud má každý vrchol v rámci *C* více vazeb uvnitř komunity než se zbytkem grafu. Konkrétně podgraf *C* tvoří silnou komunitu, jestliže pro každý uzel $i \in C$ platí:

$$k_i^{\text{int}}(C) > k_i^{\text{ext}}(C).$$

- *C je slabá komunita,* pokud celkový vnitřní stupeň podgrafu převyšuje jeho celkový vnější stupeň. Konkrétně podgraf *C* tvoří slabé společenství, jestliže:

$$\sum_{i \in C} k_i^{\text{int}}(C) > \sum_{i \in C} k_i^{\text{ext}}(C).$$

Příklady

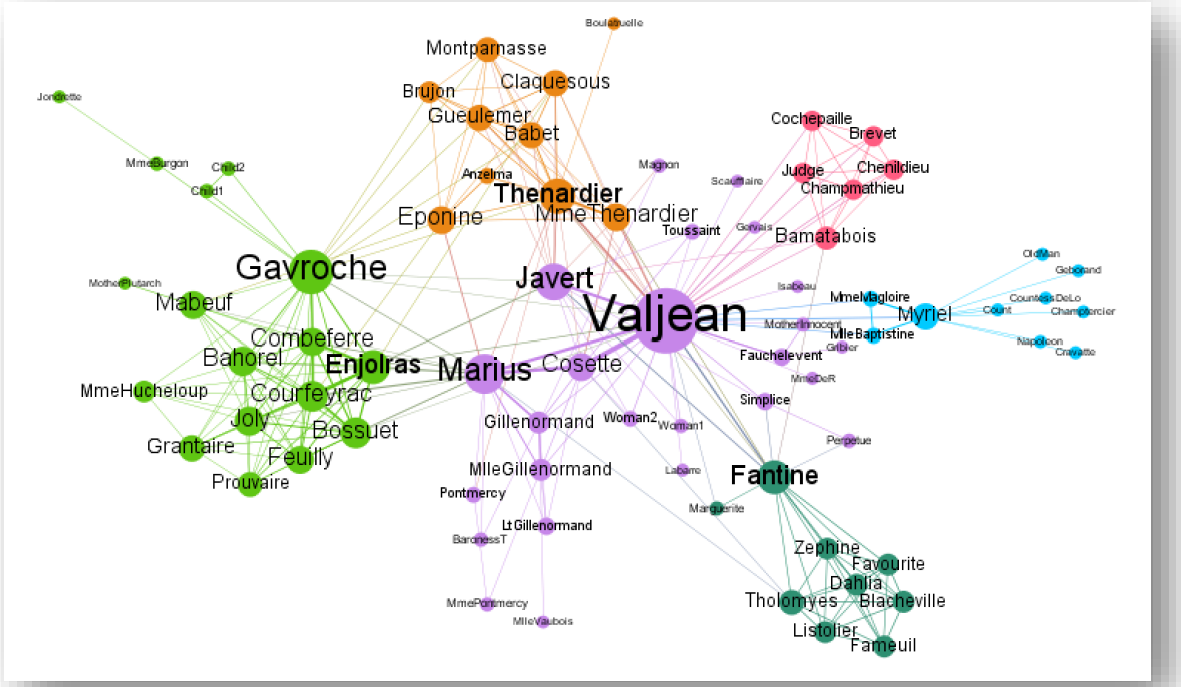


Mark Newman

- *Les Miserables*
  - neorientovaná, vážená, 77 vrcholů, 254 hran
- *Dolphin social network*
  - neorientovaná, nevážená, 62 vrcholů, 159 hran
- *Books about US politics*
  - neorientovaná, nevážená, 105 vrcholů, 441 hran
- *American College football*
  - neorientovaná, nevážená, 115 vrcholů, 613 hran

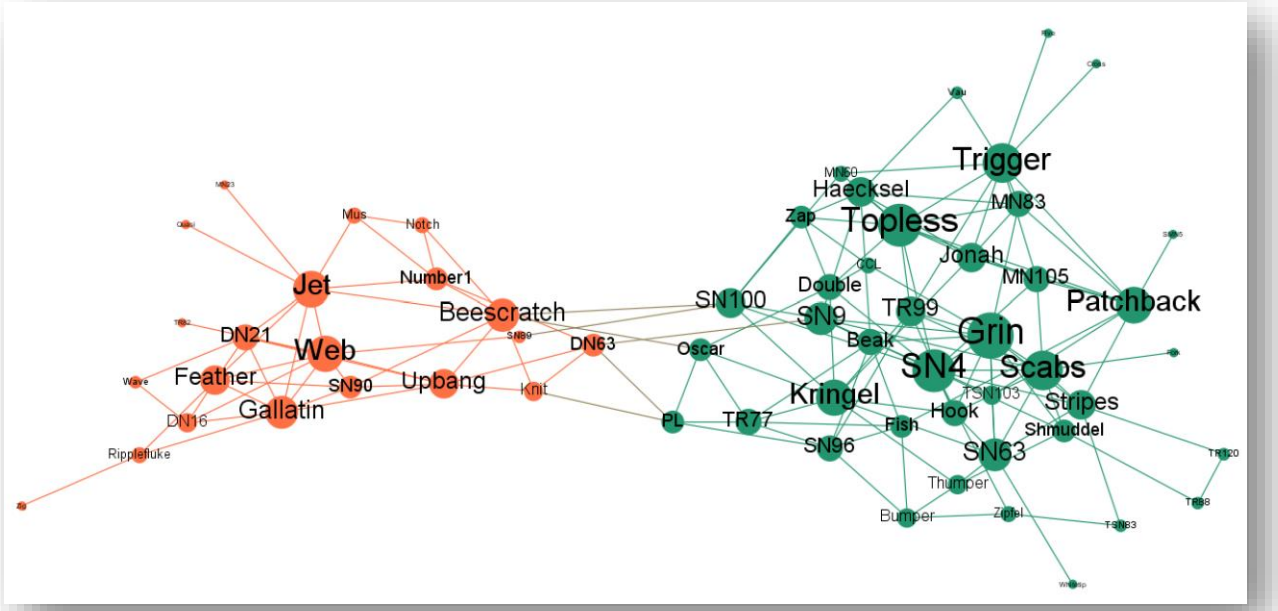
Les Miserables

- Síť vytvořená z postav románu Victora Huga Bídníci.
- Hrany jsou vážené a existují mezi postavami, které se spolu v románu potkávají. Váha odpovídá četnosti interakcí.
- Vrcholy jsou obarveny podle příslušnosti k automaticky detekovaným komunitám.
- Velikost vrcholu odpovídá jeho stupni (počtu postav, se kterými se vrchol v románu potkává).

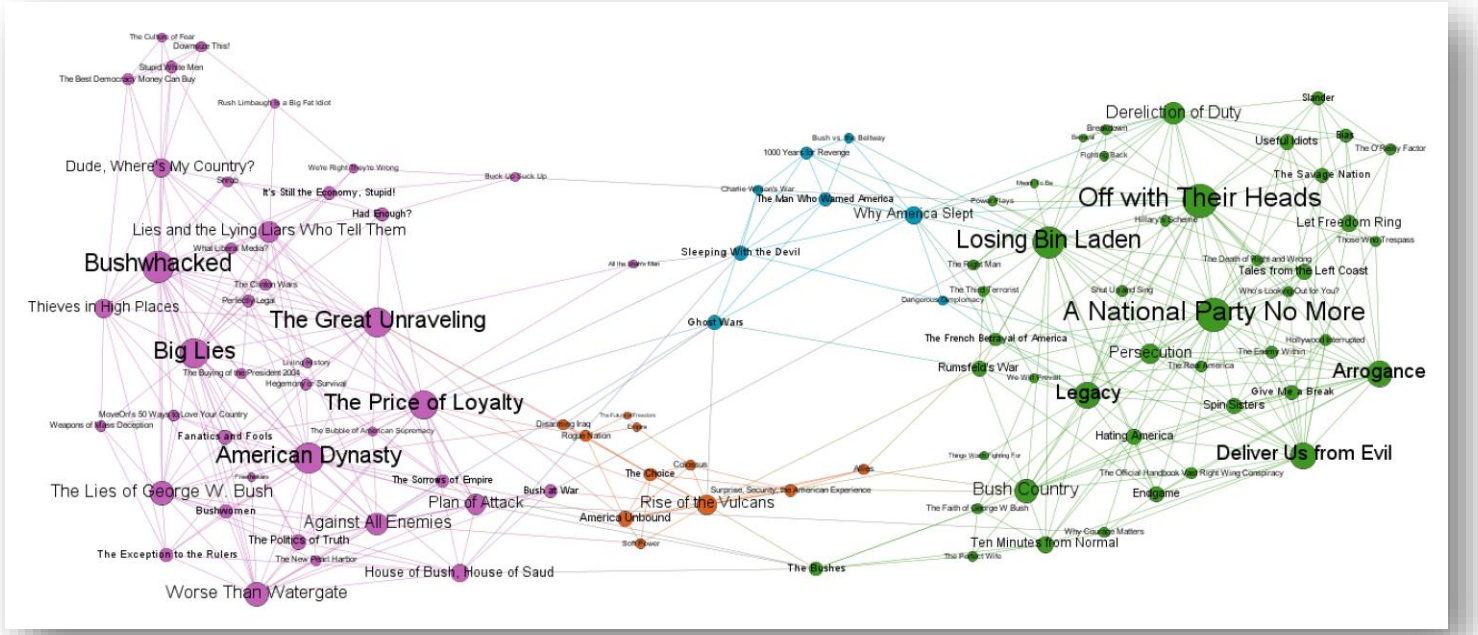


Dolphin social network

- Sociální síť mezi delfíny žijícími společně ve skupině a dlouhodobě pozorovanými v Doubtful Sound na Novém Zélandu.
- Hrana mezi dvojicemi vrcholů (delfínů) existuje v případě, že byli často pozorování společně.
- Velikost vrcholu odpovídá počtu vztahů s jinými delfíny.
- Rozložení vrcholů bylo zvoleno tak, aby byly zřetelně vidět delfíni mimo „centrum dění“ (na periferii sítě).
- Barva vrcholů odpovídá dvěma automaticky detekovaným velkým komunitám.



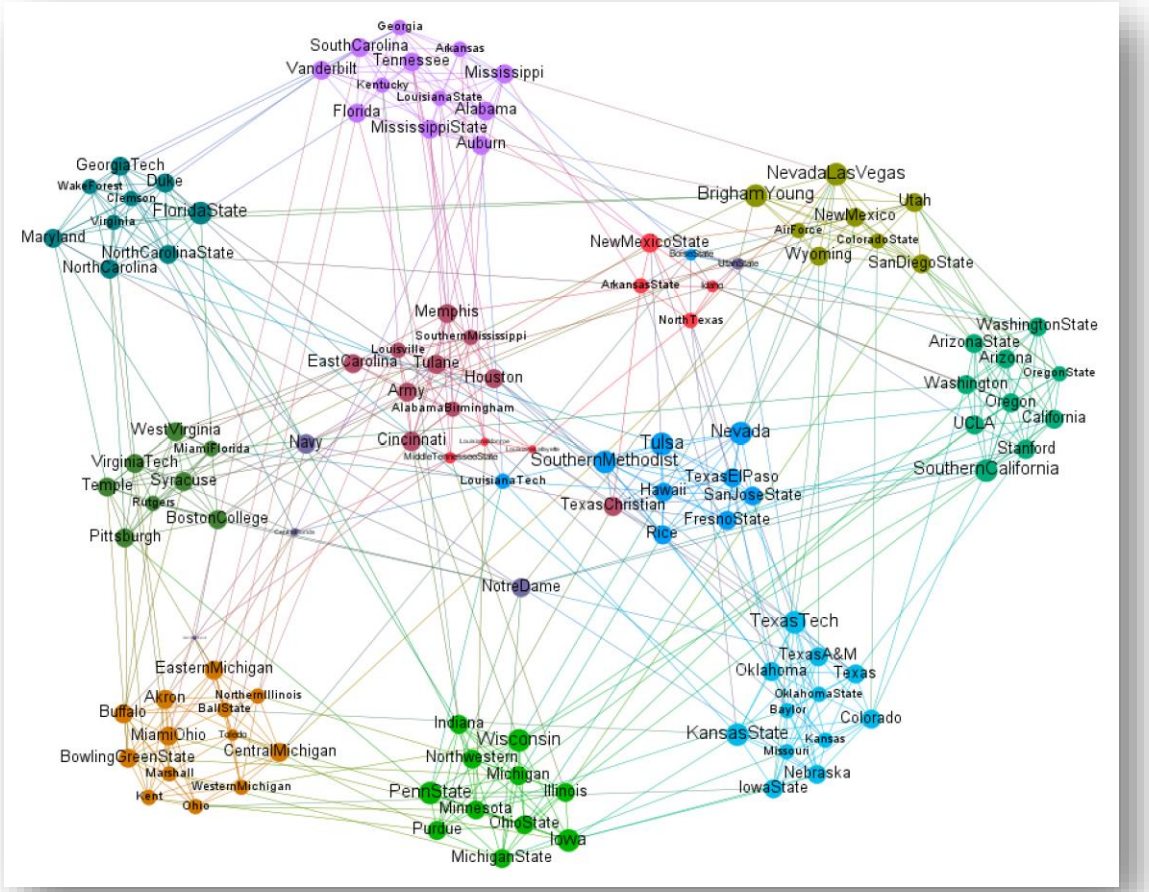
Books about US politics



- Síť knih o politice v USA publikovaná v době prezidentských voleb v roce 2004 a prodávaných online na Amazonu.
- Hrany mezi vrcholy (knihami) představují časté kupování knih společně jedním zákazníkem.
- Velikost vrcholů odpovídá knihám, které byly nejčastěji kupovány s jinou knihou.
- Barva vrcholů odpovídá čtyřem skupinám knih kupovaným nejčastěji společně.

American College Football

- Vrcholy představují týmy, mají jména jejich univerzit.
- Hrany představují hry pravidelná utkání mezi dvěma týmy v sezóně 2000.
- Síť je zajímavá tím, že obsahuje známou komunitní strukturu - týmy jsou rozděleny do konferencí, z nichž každá obsahuje přibližně 8–12 týmů.
- Utkání jsou častější mezi členy stejné konference než mezi členy různých konferencí, přičemž týmy hrály v sezóně 2000 v průměru asi sedm konferenčních utkání a čtyři mezi-konferenční utkání.
- Týmy, které jsou si geograficky blízké, ale patří k různým konferencím, hrají častěji, než týmy oddělené velkými geografickými vzdálenostmi.
- Barva vrcholu (týmu) označuje konferenci (parametr „value“), velikost vrcholu pak počet týmů, se kterými vrchol hrál.



Úkoly

- Pokuste se v Gephi docílit vizuálního rozložení čtyř sítí podobně, jak je to v této prezentaci.
- Textově (a s co největší mírou detailu) popište některé zajímavé situace, které se v sítích opakují nebo jsou naopak odlišují některou síť ostatních.
- Vyberte si alespoň jednu jinou síť z odkazů z této přednášky a analyzujte ji (včetně výstižné vizualizace).