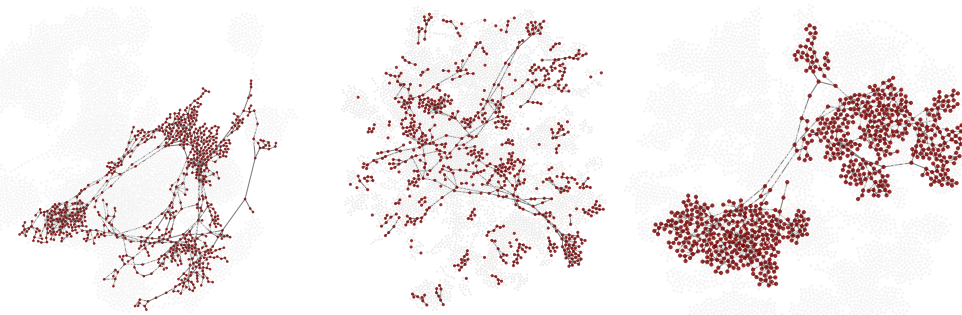


Vizualizace vzorků vytvořených
metodou Random Walk, TDS a
Forest Fire (zleva)



Motivace

Analýza tzv. velkých dat je současným fenoménem, kterým se zabývají nejenom vědci, ale také firmy, které mohou díky analýze dat reprezentující chování zákazníků plánovat své budoucí strategie. Pokud je dat velké množství, můžeme místo celé datové kolekce pracovat jen s jejím reprezentativním vzorkem.

Jedním z typů dat jsou i **síťová data**. Výpočty nebo simulace nad sítí, která má velký počet vrcholů, jsou výpočetně i časově velmi náročné. **Vzorkování sítí** poskytuje jednoduché, ale účinné řešení, při kterém je ze sítě vybrán reprezentativní podgraf, který zachovává vlastnosti původní sítě. Při analýze se následně pracuje se **vzorkem mnohem menší velikosti**, který zastupuje původní síť.

Cíle práce

- **popsat** vzorkovací metody
- vytvořit **aplikaci implementující popsání vzorkovací metody**, která umožňuje vytvářet vzorky a porovnávat úspěšnost jednotlivých metod
- vyhodnotit experimenty nad vybranými datovými sadami a **určit nejlepší metody** pro vzorkování

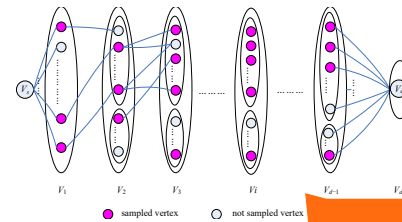
Návrh a implementace

Jádrem aplikace je knihovna zajišťující práci se sítěmi, jejich vzorkování a výpočet vlastností. Nad tím je postaveno GUI umožňující:

- načtení grafu ze souboru nebo vygenerování sítě podle několika dostupných modelů
- **vygenerování vzorků** podle implementovaných metod
- **evaluace** vzorkovacích metod

Vzorkovací metody

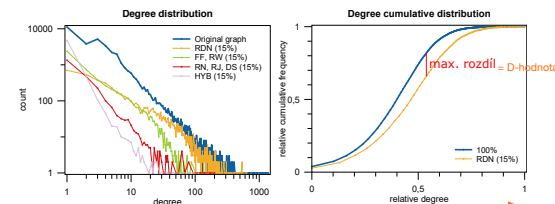
- založené na výběru vrcholů
- založené na výběru hran
- založené na procházení grafu (např. **Forest Fire**, **multi-dimenzionální Random Walk**, nebo **Topologically Divided Stratus (TDS)**)



Princip metody TDS

Evaluační technika

Jedna z otázek při vzorkování sítí se zabývá tím, jak **kvalitní** je vzorek a jak lze **změřit** kvalitu vzorku. Cílem evaluace je určit nejlepší vzorkovací metodu pro daný druh sítě, přičemž by vzorek měl zachovávat co nejvíce vlastností původní sítě.



Každá distribuce (např. distribuce **stupňů**, **hop-plot**, **singulární hodnoty** matice sousednosti nebo **shlukovací koeficient**) vlastností vzorku je porovnávána s distribucí vlastností původního grafu pomocí dvou výběrového **Kolmogorova-Smirnovova testu**, jehož **D-hodnota** je použita pro

určení podobnosti dvou distribucí. Podobnost distribucí je měřítkem úspěšnosti metody.

Výsledky a závěr

- Byly generovány vzorky o velikosti 15% původní velikosti sítě. Podle D-hodnot si nejlépe vedly metody založené na procházení grafu - **Divided Stratus (TDS)**, **Forest Fire** a **Random Walk**.
- Optimální velikost vzorku byla **20% až 25%**. Při rozsáhlejších sítích se optimální velikost snižuje.
- Topologickou strukturu nejlépe zachovává metoda **Divided Stratus (TDS)** (viz vizualizace vzorků, kde je vidět, že vzorek obsahuje vrcholy ze všech částí původní sítě).

Podařilo se implementovat nástroj umožňující vytváření vzorků a zjišťování globálních vlastností a distribuce vlastností. Distribuce lze přímo v aplikaci porovnávat jak statisticky, tak i vizuálně. Výsledky experimentů lze jednoduše uložit v podobě grafů a tabulek pro další použití. Svým rozsahem je aplikace vhodná pro další použití všude tam, kde je potřeba zkoumat vlastnosti sítí a jejich vzorků.

Rozsáhlá experimentální část zkoumala celkem 6 reálných sítí a 10 vzorkovacích metod. Experimenty byly zjištěny metody, které nejlépe vytvářejí vzorky s ohledem na zachování co nejvíce vlastností původní sítě. Zjistily se např. optimální parametry některých metod nebo se prozkoumaly vzorky z pohledu zachování topologie sítě.

Distribuce stupňů různých vzorků a měření podobnosti sítí pomocí KS testu

Ukázka D-hodnot (pro jednu testovací síť)

	deg	sec	clust	hops	hops on k	avg-sec	avg-val	average D
REN	0.401	1.000	0.183	0.042	0.042	0.175	0.062	0.281
RE	0.797	1.000	0.180	0.231	0.065	0.744	0.210	0.802
RDN	0.092	1.000	0.071	0.068	0.068	0.014	0.048	0.194
HYB	0.853	1.000	1.000	0.372	0.192	0.807	0.208	0.654
HYB	0.209	0.000	0.178	0.034	0.034	0.062	0.092	0.114
RL	0.103	1.000	0.132	0.128	0.129	0.127	0.260	0.211
MHWR	0.118	0.000	0.275	0.213	0.213	0.756	0.226	0.257
FF	0.117	0.000	0.091	0.112	0.112	0.219	0.302	0.183
FS	0.100	1.000	0.085	0.017	0.017	0.038	0.164	0.203
FS	0.291	1.000	0.120	0.070	0.070	0.119	0.250	0.274

Ukázka prostředí

