

Analýze sítě *Spoluaautorství* v Obecné relativitě a Kvantové kosmologii

Moravec Vojtěch

Letní semestr 2019

1 Popis datasetu

Analyzovaný dataset byl získán ze stránky *snap.stanford.edu* [1]. Dataset obsahuje seznam hran neorientovaného grafu, který prezentuje spolupráci vědců, kteří se zabývají obecnou relativitou a kvantovou kosmologií. Hrana (u, v) značí, že vědec u se podílel na článku s vědcem v . Tento dataset obsahuje některé hrany vícekrát, ale pro naši analýzu nebereme v potaz vážené hrany, kdežto smyčky, které se zde nacházení ponecháváme v síti. Celkem je v této citační síti 12 smyček.

Data pro vytvoření datasetu byla nasbírána v průběhu let od Ledna roku 1993 do Dubna 2014, jedná se tedy o celkem 124 měsíců, hrany avšak neobsahují časová razítka jejich vzniku.

2 Analýza sítě

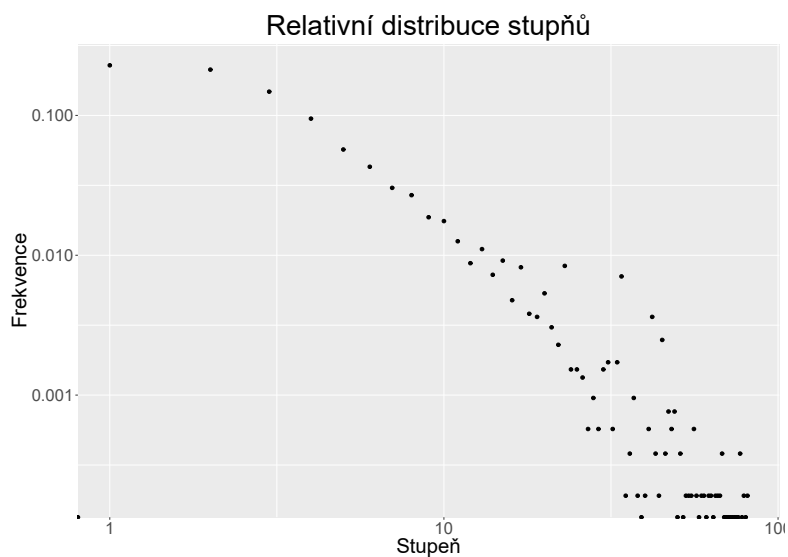
2.1 Analýza distribuce stupňů vrcholů

Citační síť se skládá z 5242 vrcholů a 14496 hran, tedy 5242 autorů a 14496 zaznamenaných spoluautorství. V tabulce 1 vidíme shrnutí číselných charakteristik stupně vrcholu. Maximální stupeň je roven 81, ten říká, že nejaktivnější autor je spoluautorem 81 článků, kdežto 1196 autorů se podílelo pouze na jednom článku, v rámci obecné relativity a kvantové kosmologie. V průměru se autoři podílejí na 5-ti až 6-ti člancích. Pokud by jsme ze sítě vyloučili smyčky, minimální stupeň by se změnil na 0.

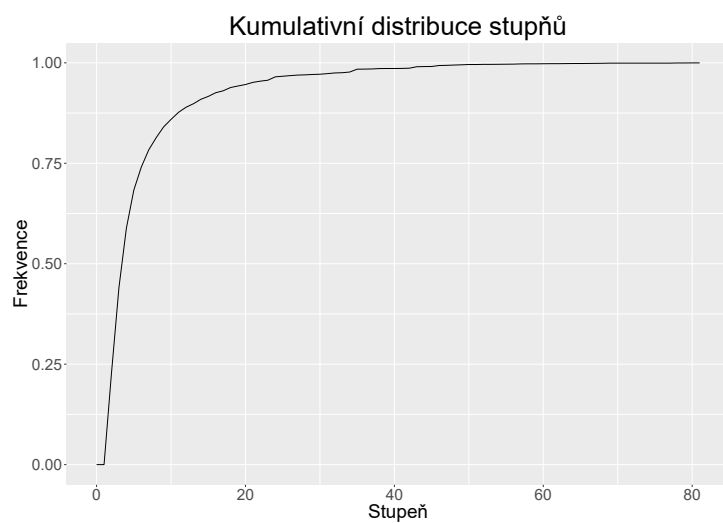
Minimum	Četnost	Maximum	Četnost	Průměr
1	1196×	81	1×	5,530

Tabulka 1: Souhrn stupňů vrcholů

V grafu na Obrázku 1 vidíme relativní distribuci stupňů a na Obrázku 2 kumulativní distribuci stupňů. Dle relativní distribuce potvrzujeme, že distribuce odpovídá mocninnému rozdělení.



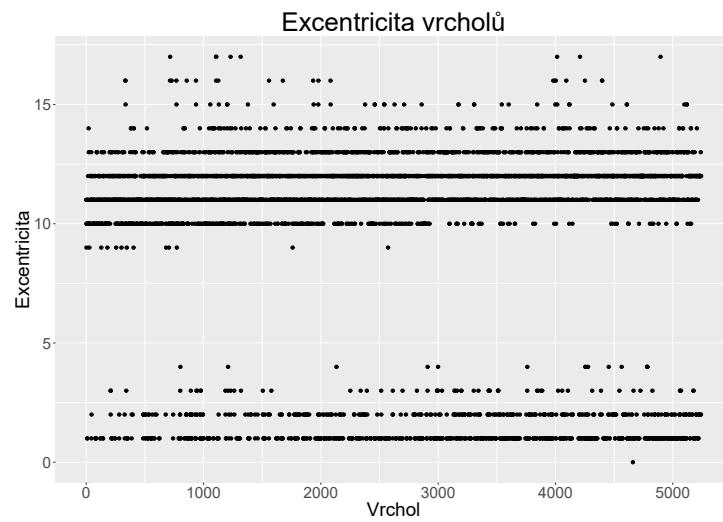
Obrázek 1: Relativní distribuce stupňů



Obrázek 2: Kumulativní distribuce stupňů

2.2 Analýza cest v síti

Průměr sítě neboli, nejdelší z nejkratších cest mezi vrcholy, je v naší síti roven 17. Průměrná délka nejkratší cesty je rovna 6,049. Autoři článku jsou tedy průměrně propojeni přes cca 6 další autorů. V grafu na Obrázku 3 můžeme vidět excentricitu jednotlivých vrcholů. Excentricita je vzdálenost nejkratší cesty k nejvzdálenějšímu vrcholu.

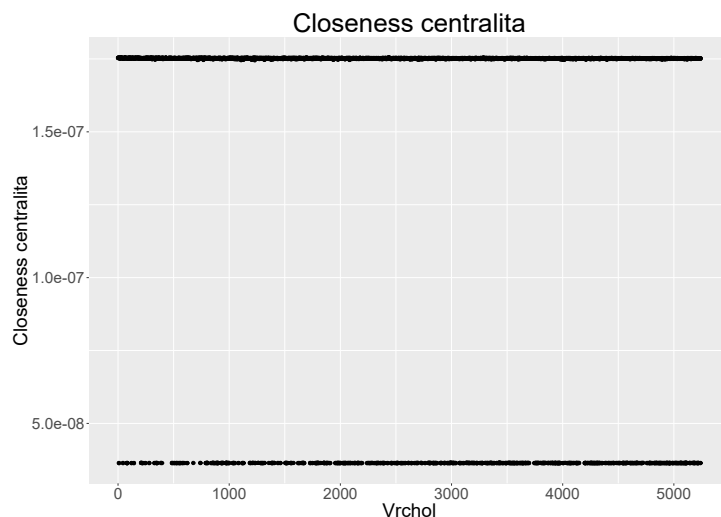


Obrázek 3: Excentricita vrcholů

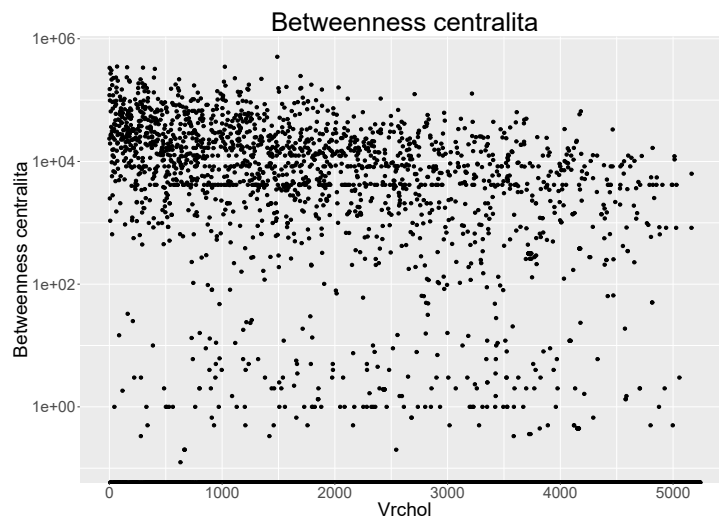
Cesta v naší síti značí posloupnost autorů, kteří spolupracovali na určitém článku. Nulovou excentricitu má pouze jeden vrchol, tento vrchol musí mít pouze smyčku sám na sebe, neboť minimální stupeň je 1 viz Tabulka 1, nikdy nespolečně pracoval s žádným dalším autorem.

2.3 Centrality v síti

Na Obrázcích 4, resp. 5 můžeme vidět closeness, resp. betweenness centralitu jednotlivých vrcholů. Closeness centralita měří, jak je vrchol centrální v dané síti, bere v potaz nejkratší cesty ke všem ostatním vrcholům. Dle closeness centrality jsou vrcholy rozděleny do dvou skupin, ačkoli ani jedna nedosahuje vysoké hodnoty této centrality. Tyto velmi nízké hodnoty jsou způsobeny tím, že daný graf není souvislý a obsahuje 355 komponent, kde největší obsahuje 4158 vrcholů, nejmenší 1 vrchol a průměrné je v jedné komponentě 14,767 vrcholů. Betweenness centralita měří poměr nejkratších cest, všech dvojic vrcholů, které procházejí přes zkoumaný vrchol. Vrcholem s vysokou betweenness centralitou prochází mnoho nejkratších cest a pokud jej odstraníme je největší šance, že dojde k rozpadu komponenty.



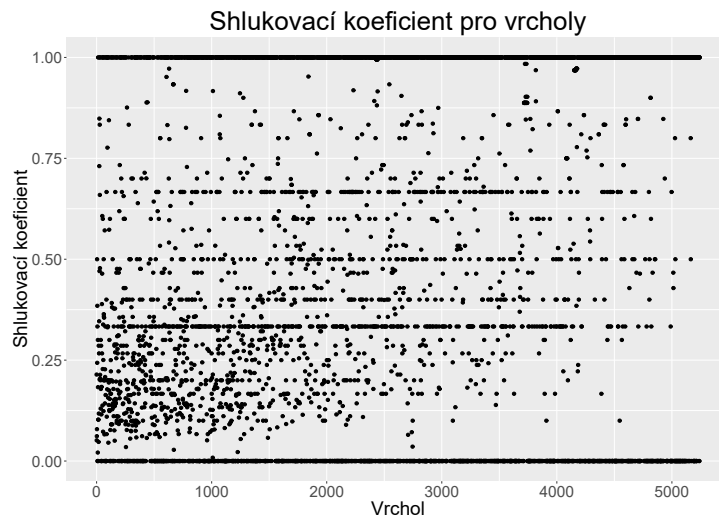
Obrázek 4: Closeness centralita vrcholů



Obrázek 5: Betweenness centralita vrcholů

2.4 Shlukování

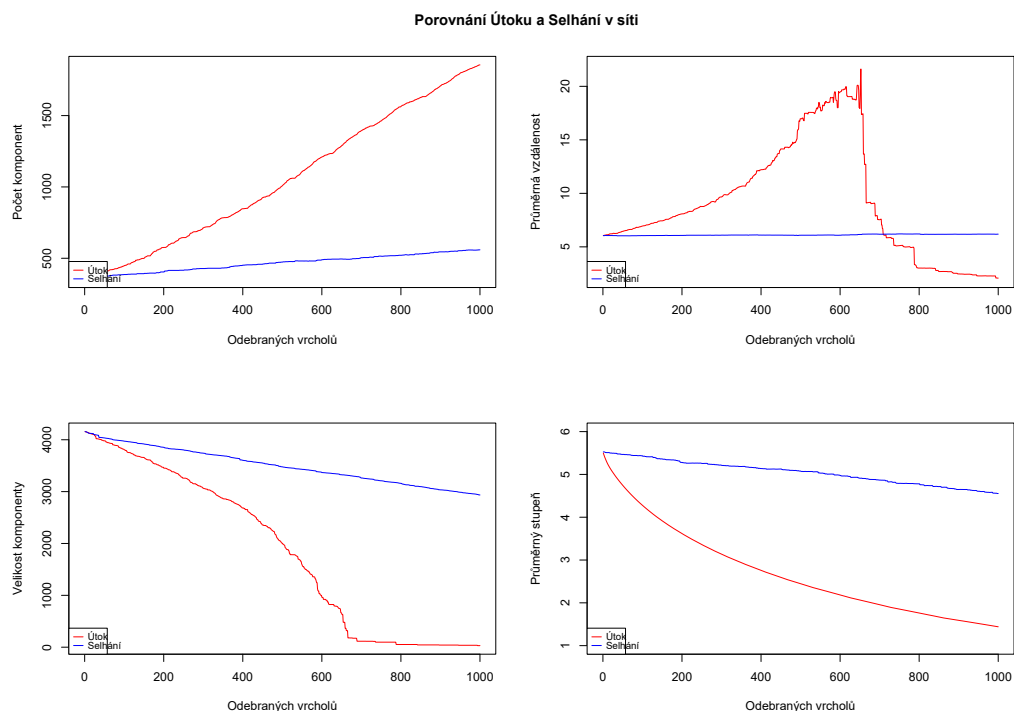
V naší síti je globální shlukovací roven 0,630. Shlukovací koeficient jednotlivých vrcholů je už velmi různorodý jak můžeme vidět v grafu na Obrázku 6. Vysoká tranzitivita vrcholů znamená, že jeho sousedé jsou s vysokou pravděpodobností taky spojeni hranou.



Obrázek 6: Shlukovací koeficient vrcholů

2.5 Odolnost sítě

V grafu na Obrázku 7 můžeme vidět jak se měnili různé charakteristiky sítě při jejím selhání resp. útoku. Při selhání je náhodně odebrán vrchol, kdežto u útoku je cíleně odebrán nejdůležitější vrchol, tedy ten s nejvyšším stupněm. Grafy popisují změnu charakteristik pro 1000 iterací, bylo tedy odebráno celkem 1000 vrcholů.



Obrázek 7: Odolnost sítě vůči selhání a útoku

Všimneme si, že při útoku rychle roste počet komponent, to způsobuje vznik menších komponent, což se odráží na průměrné vzdálenosti, která začne kolem 700-té iterace klesat. Selhání většinou kopíruje vlastnosti útoku, ale mnohem pomaleji. Průměrná vzdálenost při selhání je ale výjimkou, neboť ta se pořád drží velmi blízko originální hodnoty.

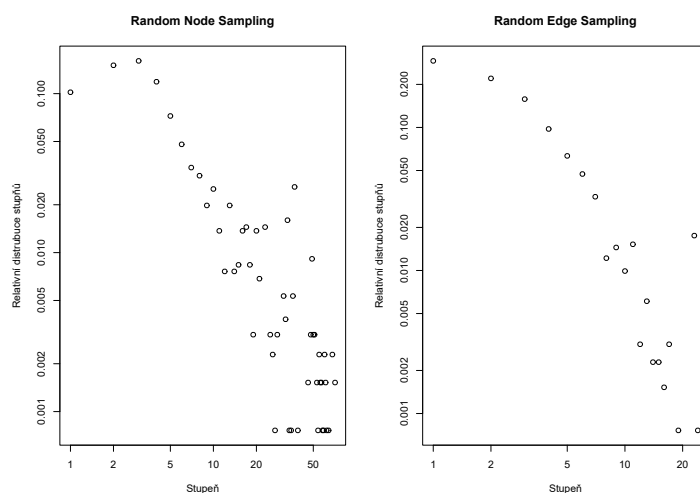
2.6 Vzorkování

Jak jsme zmínili naši síť tvoří 5242 vrcholů, rozhodli jsme se vytvořit dva vzorky, které budou 25% originální velikosti a zjistit zda jsou tyto vzorky reprezentativní, tedy jejich charakteristiky jsou velmi podobné celé síti. První vzorek byl vytvořen metodou Random Node Sampling (RNS) a druhý Random Edge Sampling (RES).

	Počet vrcholů	Počet hran	Prům. stupeň	Prům. vzdálenost	Shlukovací koef.	Prům. excentricita
Originál	5242	14496	5,530	6,049	0,630	9,557
RNS	1313	6197	9,440	5,608	0,813	8,750
RES	1312	2371	3,614	6,646	0,679	7,813

Tabulka 2: Charakteristiky vzorků

Souhrn charakteristik originální sítě i dvou vzorků vidíme v Tabulce 2, relativní distribuce vzorků pak na Obrázku 8. Obě metody vytvořili vzorky téměř stejné velikosti, ale metodou RES jsme přišli o cca $3\times$ více hran, což se projevuje na průměrném stupni vrcholu. Průměrné vzdálenosti obou vzorku jsou velmi podobné originálu a taktéž shlukovací koeficient a méně pak i průměrná excentricita. Bohužel podle relativní distribuce stupňů nemůžeme jistě říci, že by vzorky měli distribuci podle mocninného rozdělení, ikdyž náznak tam vidíme.



Obrázek 8: Relativní distribuce stupňů ve vzorcích

2.7 Hledání komunit

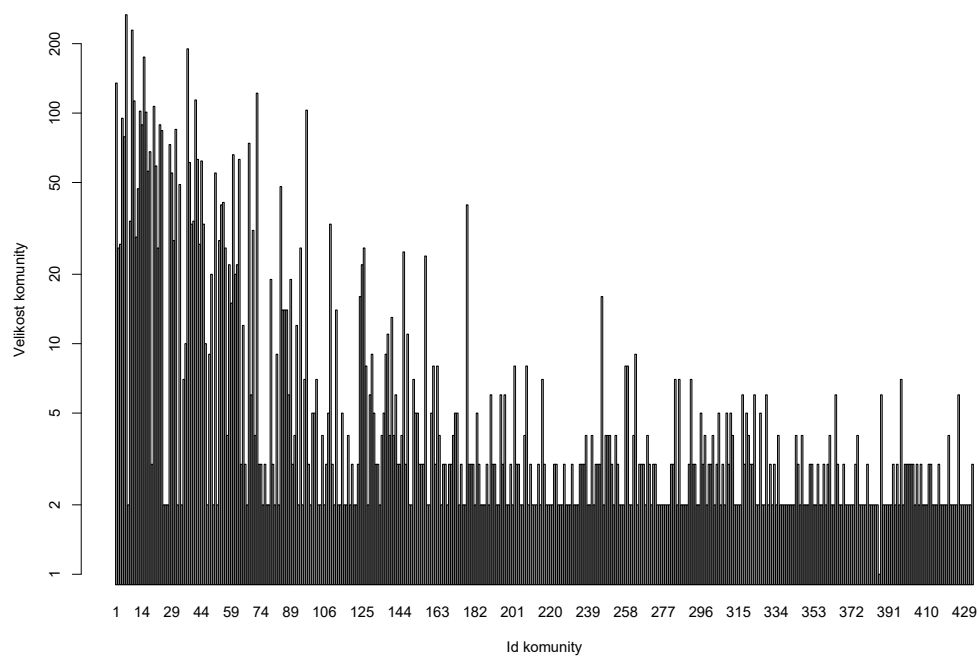
Jak jsme již zmínili naše síť není souvislá a skládá se z 355 komponent, je tedy nemožné aby jsme našli méně než 355 komunit. Pro hledání komunit jsme využili 4 různých metod, které jsou dostupné v R balíčku *igraph*. Kvalita nalezených komunit je hodnocena podle jejich modularity.

- Girvan-Newmanova metoda - divizivní metoda, kde při každé iteraci odebíráme hranu s největší betweenness centralitou, je velká šance, že tato hrana spojuje silně propojené komponenty, potencionální komunity.
- Fast-Greedy metoda - optimalizační metoda, maximalizující modularitu výběrem hustých podgrafů.
- Louvainova metoda - optimalizační metoda, hledání nejlepších komunit pomocí přesouvání vrcholů z jedné komunity do druhé. Vrchol bude vložen do komunity tak ať je maximalizována modularita.
- Walktrap metoda - hledá komunity pomocí náhodných procházek, v našem případě délky 5. Myšlenkou je, že krátké procházky zůstanou v komunitě daného vrcholu.

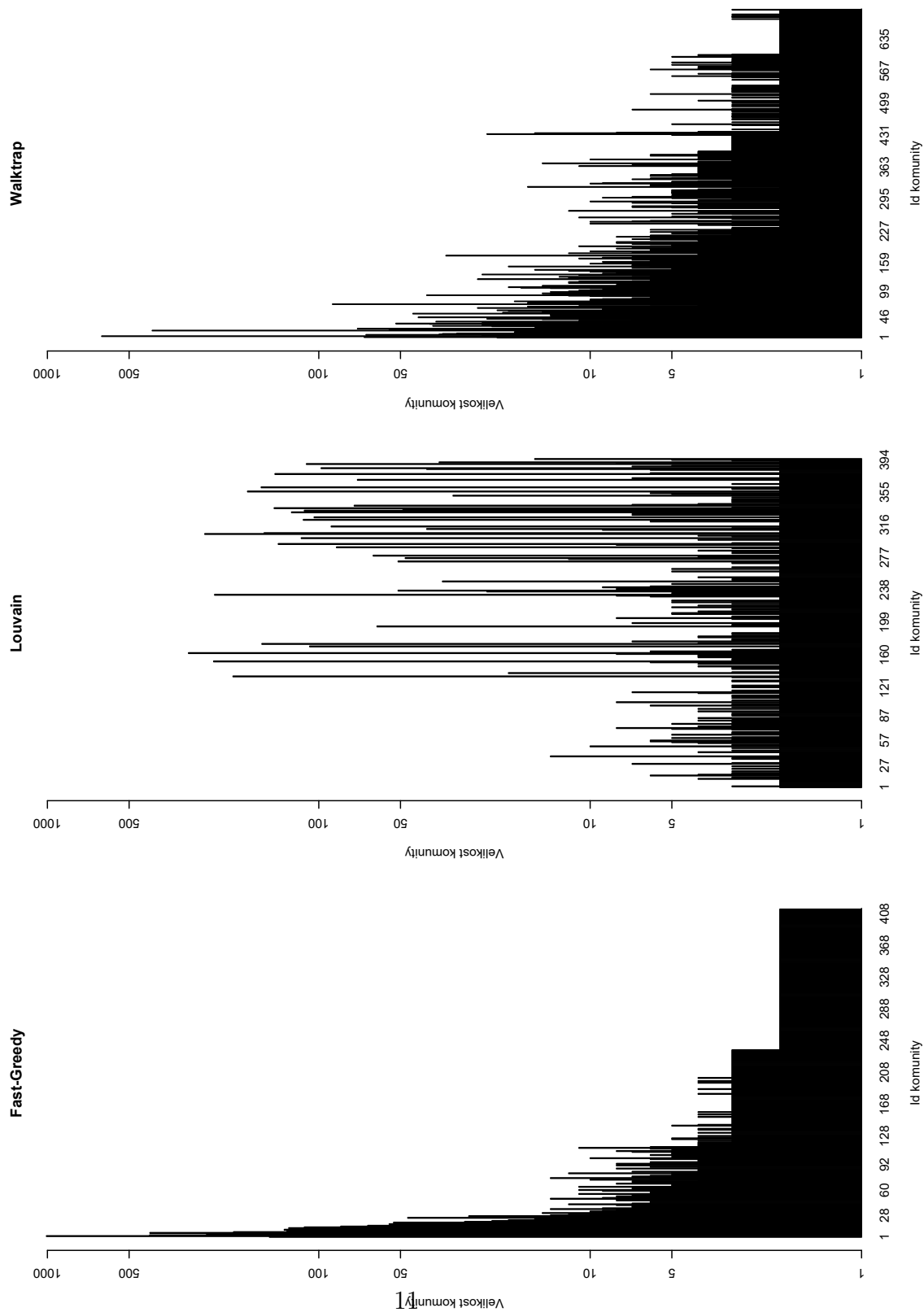
Algoritmus	Počet komponent	Modularita
Girvan-Newman	433	0,849
Fast-Greedy	414	0,820
Louvain	395	0,861
Walktrap	698	0,790

Tabulka 3: Zhodnocení detekce komunit dle metody

Podle Tabulky 3 hodnotíme jako nejlepší Louvainův algoritmus, který dosáhl největší modularity s nejmenším počtem komunit. Celkově všechny algoritmy dosahují podobných modularit. Nejhuř vychází algoritmus Walktrap s velkým počtem komunit. Velikosti jednotlivých komunit dle algoritmů můžeme vidět v Obrázcích 9, 10.



Obrázek 9: Komunity podle Girvan-Newmanova algoritmu



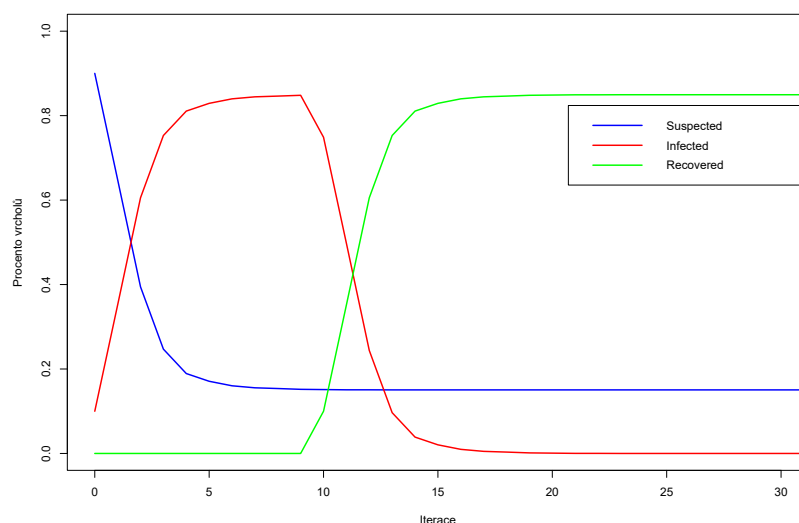
Obrázek 10: Komunity podle Fast-Greedy algoritmu

2.8 Šíření jevů

Šíření jevů, např. nákazy v síti se dá simulovat SIR modelem, kde v každém momentu je vrchol v jednom ze stavu:

- S - nenakaženy, může být nakažen
- I - nakažený, může nakazit sousedy
- R - uzdraven, už nemůže být znovu nakažen

Na začátku se vybere náhodně určitý počet nakažených vrcholů, ty s určitou pravděpodobností nakazí své sousedy. Nakažený vrchol je uzdraven po určeném počtu kroků. Uzdravený vrchol již nemůže být nakažen. Pokud bude na počátku vybrán vrchol s vysokým stupněm tak se bude epidemie šířit rychle, pokud naopak s nízkým stupněm tak pomaleji. Navíc, tím že síť je rozdělena do komponent mohou být určité komponenty zcela imunní pokud nebude nakažen žádný vrchol, který do nich patří.



Obrázek 11: Simulace epidemie dle SIR modelu

V grafu na Obrázku 11 můžeme vidět simulaci epidemie právě podle SIR modelu. Na začátku bylo nakaženo 10 % vrcholů, tedy 524, pravděpodobnost nákazy byla nastavena na 70 % a čas potřebný pro celkové uzdravení 10. V prvních iteracích dochází k rychlému šíření díky 70-ti % pravděpodobnosti nákazy. Všimneme si, že po 30 iteracích je epidemie zcela vyhlazena. Velká odolnost této sítě vůči epidemii je způsobena velkým množstvím komponent,

neboť pokud ze začátku není vybrán náhodný vrchol v komponentě tak je tato komponenta zcela imunní vůči epidemi.

Reference

- [1] J. Leskovec and A. Krevl, “SNAP Datasets: Stanford large network dataset collection.” <http://snap.stanford.edu/data>, June 2014.