UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ANALÝZA REÁLNYCH SIETÍ

DIPLOMOVÁ PRÁCA

2018 Jakub RAUČINA

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ANALÝZA REÁLNYCH SIETÍ

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Doc. RNDr. Ján Boďa, CSc.

Bratislava 2018 **Jakub RAUČINA**

62544020

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Jakub Raučina
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Analýza reálnych sietí.
Network analysis.

Cieľ: Cieľom práce je získať dáta o reálnych sieťach, porovnať ich vlastnosti s náhodne generovanými bezškálovými sieťami a zároveň skúmať, ako sú reálne siete ovplyvňované okolitým svetom.

Vedúci: doc. RNDr. Ján Boďa, CSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 26.01.2017

Dátum schválenia: 27.01.2017 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Chcel by som sa touto formou úprimne poďakovať vedúcemu mojej diplomovej práce doc. RNDr. Jánovi Boďovi, CSc. za ochotu, pomoc, vecné pripomienky a rady, ktoré mi v nemalej miere pomohli pri písaní tejto práce. Tiež ďakujem svojej rodine za podporu a motiváciu.

Abstrakt v štátnom jazyku

Raučina, Jakub: Analýza reálnych sietí [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Ján Boďa, CSc., Bratislava, 2018, 43 s.

Táto práca sa zaoberá skúmaním sietí z rôznych oblastí. Porovnáva vlastnosti grafov zo skutočného života ako napr. sociálne siete, alebo sieť zamestnancov vo firme s náhodnými sieťami. Analyzuje modely použité pre simuláciu náhodných sietí a rozdiely s reálnymi sieťami.

Prvá časť je stručným úvodom do sveta terminológie sietí, rovnako ako motiváciou k ich štúdiu. Taktiež sa priblížia niektoré štatistické metódy, neskôr implementované do programu R. V druhej časti sa vyberie niekoľko príkladov skutočných sietí, rovnako ako simulácia vlastnej siete pomocou Barabási-Albertovho, Erdős–Rényiho a ďalších modelov. V závere je hodnotená úspešnosť analýzy jednotlivých sietí.

Kľúčové slová: Bezškálové siete, Barabási-Albertov model, Preferential attachment, Náhodný graf

Abstract

Raučina, Jakub: Network analysis [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Ján Boďa, CSc., Bratislava, 2018, 43p.

This thesis focuses on examining networks from various areas. Networks, that formed organically, such as social networks, or a network of actor collaborations are compared with randomly generated ones. Different models used for generation random networks are tested.

The first chapter of this thesis is focused on a brief introduction to the world of networks, as well as explaining the motivation to pursue this field. In the following chapters, several important statistical approaches are presented and later implemented in R. Subsequently, several examples of real networks are examined, as well as a random graph generated using the Barabási-Albert model, Erdős–Rényi model and a few others. Eventually, the analyses success is evaluated.

Keywords: Scale-free networks, Barabási-Albert model, Preferential attachment, Random graph

Obsah

Úvod

Siete sú fenoménom, na ktorý sa dokonale hodí prívlastok „staronový“. Základy ich skúmania boli položené už v prvej polovici 18. storočia Leonhardom Eulerom pri riešení tzv. problému siedmich mostov [13]. Práca o tomto známom príklade je považovaná za prvú v histórii zaoberajúca sa teóriou grafov, z ktorej sa samozrejme teória sietí odvíja.

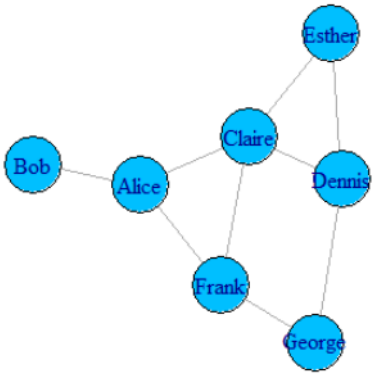
Záujem o túto oblasť však výraznejšie vzrástol až v 21. storočí, s príchodom máp sietí ako World wide web, sociálnych sietí a takisto aj databáz a výpočtovej techniky, ktorá pomáha údaje o týchto sieťach zbierať a analyzovať. Vplyv tzv. „sieťového myslenia“ na rôznorodé vedné disciplíny bol zjavný od sociálnych vied po biológiu už počas minulého storočia. Kvôli komplexnosti väčšiny skutočných sietí však k prístup k ich analýze nebol jednoznačný.

V tejto práci sa pozrieme na reálne siete bližšie. Po uvedení do ich problematiky budeme v druhej kapitole skúmať aké vlasnosti môže sieť mať a otestujeme rôzne štatistické techniky a pravdepodobnostné testy na priblíženie empirických dát tým modelovým. V tejto práci budeme vo veľkej miere využívať programovacie prostredie R, konkrétne knižnicu *igraph*, pomocou ktorej sú aj všetky vizualizácie v práci robené.

V poslednej kapitole sa pozrieme hlbšie na jednu konkrétnu sieť zamestnancov a pôjde nám o dôkladnú analýzu a spájanie dedukcie s matematickými postupmi. Navrhneme zjednodušený model šírenia informácie a budeme sledovať rozdiely medzi reálnou sieťou a generovanými náhodnými grafmi.

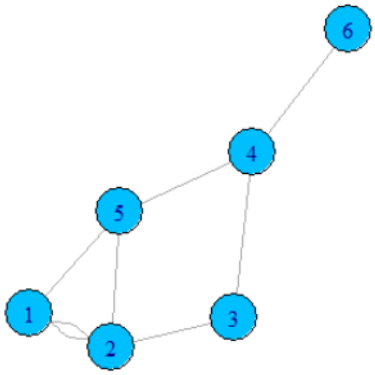
1 Úvod do grafov

Graf je definovaný ako usporiadaná dvojica $G = (V, E)$, kde V je množina, ktorej prvky sa nazývajú *vrcholy* (z angl. vertices), resp. *uzly* grafu. Označuje sa pritom $V(G)$, alebo V. E je množina *hrán* (z angl. edges) a vyjadruje prepojenia, resp. vzťahy medzi vrcholmi. Ak je graf *neorientovaný*, jednotlivé hrany sú neusporiadanými dvojicami $\{u, v\}$, kde u a v sú vrcholmi patriacimi V. Ak je graf naopak *orientovaný*, tak je dvojica vrcholov $\{u, v\}$ usporiadaná. V terminológii tejto práce budeme spomínať aj siete. Pre účely tejto práce sú siete podmnožinou grafov, pričom ich vrcholy a hrany môžu mať dodatočné vlastnosti.



Obr. 1: Známosti v divadelnom krúžku

Na obr. 1 môžme vidieť hercov v školskom predstavení, pričom vzťahy medzi hercami predstavujú, či sa poznali aj pred krúžkom alebo nie. Napríklad vrcholy Claire a Frank spája hrana, pretože sa poznajú.



Obr. 2: Nákres grafu

V 2. obrázku vidíme nákres jednoduchého, neorientovaného grafu. Má 6 vrcholov,

ktoré sú prepojené 8 hranami. Tieto údaje vieme prehľadne zaznačiť v tzv. *matici susednosti* A. V tejto štvorcovej matici, $A_{i,j} = 1$ ak medzi vrcholmi i a j existuje hrana. V opačnom prípade platí $A_{i,j} = 0$. Všimneme si tiež, že medzi vrcholmi 1 a 2 existujú 2 hrany. Preto $A_{1,2} = 2$. Graf, ktorý obsahuje aj viacero hrán medzi jednou dvojicou vrcholov sa nazýva *multigraf*. Na diagonále bude mať naša matica 0, keďže vo väčšine prípadov vrchol nemôže byť spojený sám so sebou. Keďže sa jedná o neorientovaný graf, tak matica susednosti[1] bude symetrická, a to konkrétne:

$$A = \begin{bmatrix} 0 & 2 & 0 & 0 & 1 & 0 \\ 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Každý z vrcholov v sieti má stupeň k_i, ktorý hovorí o počte hrán, ktoré z neho vychádzajú. Napríklad $k_6 = 1$, keďže jediným susedom vrcholu 6 je 4. Pomocou stupňov vrcholov vieme jednoducho vypočítať počet hrán grafu, v neorientovanom grafe to bude

$$L = \frac{1}{2} \sum k_i.$$

Priemerným stupňom tohto grafu bude potom

$$< k >= \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N},$$

čo sa v našom prípade rovná $\frac{14}{6} \approx 2.33$.

Ďalším pojmom, ktorý sa nám zíde je *cesta* medzi dvoma vrcholmi a ich *vzdialenosť* d (z angl. distance). Príkladom cesty medzi vrcholmi 1 a 6 sú napríklad $(1, 2, 3, 4, 6)$, alebo $1, 5, 4, 6$. Pritom platí, že za vzdialenosť dvoch vrcholov považujeme najkratšiu možnú cestu. Teda v našom prípade je druhá cesta tou najkratšou, a vzdialenosť medzi vrcholmi $1, 6$ sa rovná 3, resp. $d_{1,6} = 3$. *Priemerom* siete nazývame maximálnu vzdialenosť v nej, v našom prípade je to 3.

[1] Pri reálnych sieťach býva obvykle matica susednosti veľmi riedka - počet 0 výrazne prevyšuje počet 1

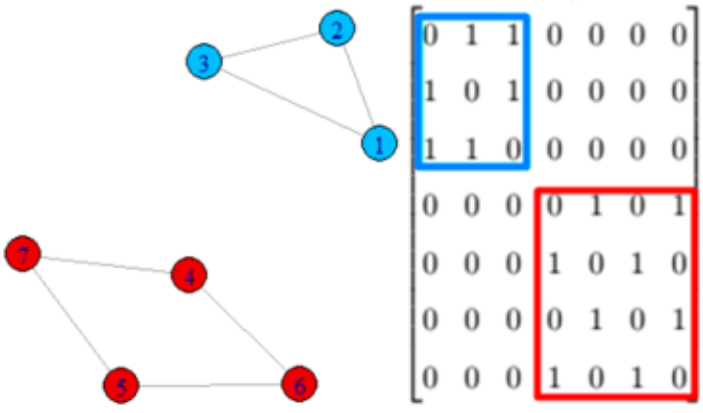
Priemerná vzdialenosť je parametrom, ktorý nám dokáže o sieťach vypovedať mnoho. Veľmi známe tvrdenie o "šiestich stupňoch separácie", ktoré hovorí, že cez 6 známostí vieme spojiť akúkoľvek dvojicu ľudí na svete, súvisí práve s týmto. Tento parameter sa počíta ako priemer vzdialeností medzi všetkými možnými dvojicami vrcholov, teda v neorientovanom grafe:

$$< d >= \frac{2}{N * (N-1)} \sum_{i \neq j} d_{i,j},$$

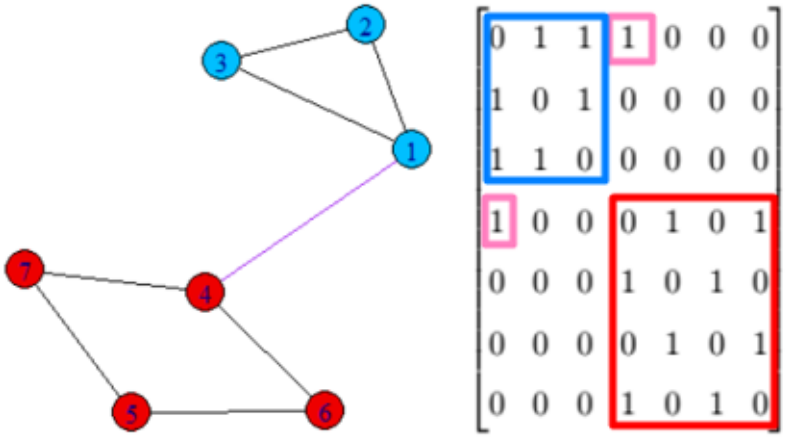
kde N je počet vrcholov v grafe. V našom grafe je priemerná vzdialenosť $< d >= \frac{5}{3} \approx 1.67$, čiže väčšina ciest v tomto grafe sa dá prejsť do 2 krokov.

Pri komplexnejších grafoch sa na zisťovanie vzdialeností používa tzv. BFS (Breadth-first search) algoritmus [8] , ktorý postupne prechádza "úrovňami"podľa vzdialenosti od počiatočného bodu a prideľuje im hodnoty. Neskôr na základe týchto pridelených hodnôt úrovní zistí vzdialenosti rýchlejšie, ako keby sme porovnávali všetky dvojice bodov postupne.

Graf sa nazýva *súvislým* v prípade, ak vieme ľubovolnú dvojicu vrcholov spojiť jednou cestou. V opačnom prípade sa jedná o *nesúvislý* graf, rozdelený na viac *komponentov*, ktoré túto vlastnosť spĺňajú. Rozdiel medzi týmito typmi môžme vidieť v obr. 3 a 4. Túto vlastnosť pozorujeme aj v matici susednosti, kde v hornom, nesúvislom prípade vidíme neprepojenosť dvoch komponentov.



Obr. 3: Nesúvislý graf a jeho matica susednosti



Obr. 4: Súvislý graf a jeho matica susednosti

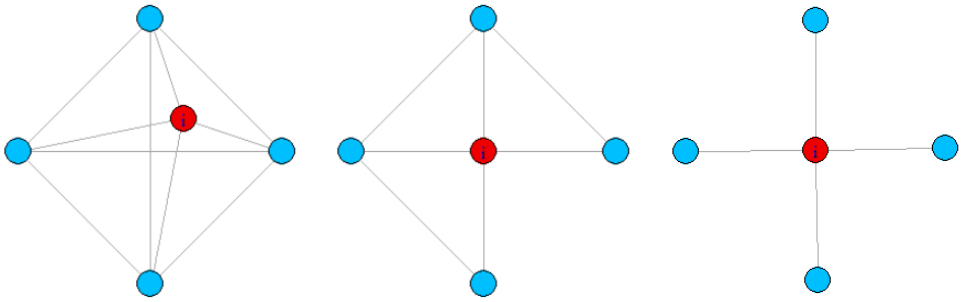
1.1 Zhluky

Ďalším kľúčovým ukazovateľom je miera zhlukovania, ktorá opisuje, ako veľmi sú vrcholy siete zhromaždené. Tento ukazovateľ sa dá počítať dvoma rôznymi spôsobmi a to lokálne, pre každý z vrholov, alebo globálne pre celú sieť.

Ako prvý sa pozrime na lokálny *zhlukový koeficient* (z angl. cluster), ktorý je pre vrchol i definovaný ako

$$C_i = \frac{2L_i}{k_i(k_i - 1)},$$

pričom L_i je počet hrán medzi susedmi vrcholu i, ktorých je k_i (stupeň). V tejto definícii teda C_i vyjadruje pravdepodobnosť, že náhodná dvojica susedov vrcholu i bude prepojená. Názorne to vidno na obr. 5. V strednom prípade má vrchol i 4 susedov, ktorí majú medzi sebou dokopy 3 prepojenia. Teda $C_i = \frac{2\times3}{4\times3} = \frac{1}{2}$.



$C_i = 1$ $C_i = 1/2$ $C_i = 0$

$C = 1$ $C = 9/14$ $C = 0$

Obr. 5: Okolie vrcholu i a hodnoty zhlukových koeficientov

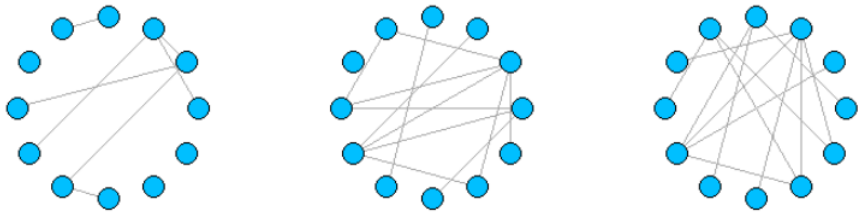
Druhým typom je globálny zhlukový koeficient (tiež nazývaný tranzitivita), ktorý sa nepozerá na jednotlivé vrcholy, ale na sieť ako celok. Konkrétne skúma pomer trojuholníkov v sieti a počet spojených trojíc, ktoré ale nemusia nutne byť uzavreté. Teda napríklad v strednom príklade z obrázku 5 je AiC trojuholníkom, v ktorom sa ukrývajú 3 "netrojuholníkové"trojice, ale AiB, alebo AiD sú iba trojicami. Konkrétny výpočet je teda

$$C = \frac{3 \times \text{počet trojuholníkov}}{\text{počet prepojených trojíc}}.$$

Poznamenávame pritom, že toto číslo sa môže líšiť od priemeru všetkých lokálnych zhlukových koeficientov jednotlivých vrcholov. Ten by sa napríklad v strednom prípade obrázku 5 rovnal $< C >= \frac{1}{5}\sum_{j=1}^{N} C_j = \frac{1}{5}(\frac{1}{2} + \frac{2}{3} + \frac{2}{3} + 1 + 1) = \frac{23}{30}$.

1.2 Náhodné grafy

Aby sme mohli vytvoriť náhodný graf, musíme použiť model, na základe ktorého budeme postupovať. Máme niekoľko možností, najprv si ukážeme Gilbertov modelo náhodného grafu $G(N, p)$. Do tohto modelu zadáme počet vrcholov N a pravdepodobnosť p, že sa medzi dvoma vrcholmi vytvorí hrana. Hrany pritom vznikajú nezávislo.



Obr. 6: Tri realizácie náhodnej siete pre $N = 12$ a $p = 1/6$

V obrázku 6 môžme vidieť, že náhodné siete vytvorené pomocou Gilbertovho modelu môžu mať rôzne vlastnosti ako počet hrán (7, 12, 13) a stupne jednotlivých vrcholov aj ak sú vstupné parametre (N, p) totožné. Z podstaty tohto modelu môžme vidieť, že jednotlivé vrcholy budú mať práve k hrán s pravdepodobnosťou

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

$N - 1$ je počet možných hrán tohto vrcholu, p^k pravdepodobnosť vytvorenia k hrán a $(1-p)^{N-1-k}$ je pravdepodobnosť, že so zvyšnými vrcholmi sa hrana nevytvorí.

Podobnú úvahu môžme použiť aj z hľadiska celkového počtu hrán v sieti. Pravdepodobnosť, že zo všetkých $\binom{N}{2}$ možných dvojíc utvorí hranu práve L je potom

$$p_L = \binom{\binom{N}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}.$$

Vidíme, že obe tieto premenné majú binomické rozdelenie. Ich stredné (očakávané) hodnoty ľahko vypočítame po dosadení do všeobecného vzorca pre výpočet strednej hodnoty $E(x) = np$. Keď teda do tohto vzorca dosadíme premenné L, k dostaneme

$$E(L) = p\frac{N(N-1)}{2},$$

$$E(k) = p(N-1).$$

Po dosadení hodnôt z druhej realizácie Erdős-Rényiho modelu v obr. 6, teda $N = 12, p = 1/6$ si tieto hodnoty vypočítame: $E(L) = 11$ a $E(k) = 11/6 \approx 1.83$. Pre ilustráciu sa môžme pozrieť na niekoľko pravdepodobnosti pre rôzne stupne vrcholov, resp. počty hrán pre túto sieť.

k	p_k
0	0.1122
1	0.2692
2	0.2961
3	0.1974
4	0.0888
5	0.0284

L	p_L
7	0.0592
8	0.0874
9	0.1126
10	0.1284
11	0.1307
12	0.1198

1.3 Aproximácia binomického rozdelenia

Binomické rozdelenie závisí na niekoľkých parametroch, čo byť miestami nepraktické. Pokúsime sa ukázať, že vo väčšine skutočných sietí sa dá analyticky zjednodušiť a aproximovať iným rozdelením.

Začneme binomickým rozdelením charakterizujúcim náhodnú sieť, teda

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

Prvý člen tohto súčinu rozpíšeme ako

$$\binom{N-1}{k} = \frac{(N-1)(N-1-1)(N-1-2)...(N-1-k+1)(N-1-k)!}{k!(N-1-k)!} = \frac{(N-1)^k}{k!},$$

a posledný člen upravíme:

$$ln[(1-p)^{(N-1)-k}] = (N-1-k)ln(1-p) = (N-1-k)ln(1-\frac{\langle k\rangle}{N-1}).$$

Zároveň na tento člen využijeme rozvoj Taylorovho radu

$$ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - ..., \forall x : |x| \leq 1,$$

čím dostaneme

$$ln[(1-p)^{N-1-k}] \cong (N-1-k)\frac{-\langle k\rangle}{N-1} = -\langle k\rangle(1-\frac{k}{N-1}) \cong -\langle k\rangle,$$

čo platí pre zlomok $\frac{k}{N-1}$ blížiaci sa k 0, podmienku splnenú pre $N >> k$, teda aj pre väčšinu reálnych sietí. Ak túto aproximáciu doplníme do predošlých vzťahov dostávame

$$(1-p)^{(N-1)-k} = e^{-\langle k\rangle}$$

Vlastnosť	Binomické Rozdelenie	Poissonovo Rozdelenie
Hustota	$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$	$p_k = e^{\langle k \rangle} \frac{\langle k \rangle}{k!}$
Stredná hodnota	$\mu = p(N-1)$	$\mu = \langle k \rangle$
Variancia	$\sigma^2 = p(1-p)(N-1)$	$\sigma^2 = \langle k \rangle$

Tabuľka 1: Porovnanie vlastností rozdelení

a následne

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} = \frac{(N-1)^k}{k!} p^k e^{-\langle k \rangle} = \frac{(N-1)^k}{k!} (\frac{\langle k \rangle}{N-1})^k e^{-\langle k \rangle} = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!},$$

čo je definícia Poissonovho rozdelenia. Môžme si všimnúť, že toto rozdelenie nezávisí na počte vrcholov N, teda siete s rovnakým priemerným stupňom $\langle k \rangle$ budú mať rovnaké p_k, aj ak budú rôznych veľkostí. Napriek tomu, že Poissonovo rozdelenie je iba aproximáciou, ho budeme používať vzhľadom na jeho jednoduchosť pri analýze a výpočtoch. Zároveň je toto rozdelenie diskrétne, čo treba brať do úvahy pri jeho využití.

1.4 Vznik a rast sietí

Aby sme mohli siete analyzovať čo najpresnejšie, je potrebné pochopiť, ako prebieha proces ich vzniku, a čo ho ovplyvňuje. Ukážeme si Barabási-Albertov model, ktorý využíva dva hlavné princípy: Rast siete a preferential attachment.

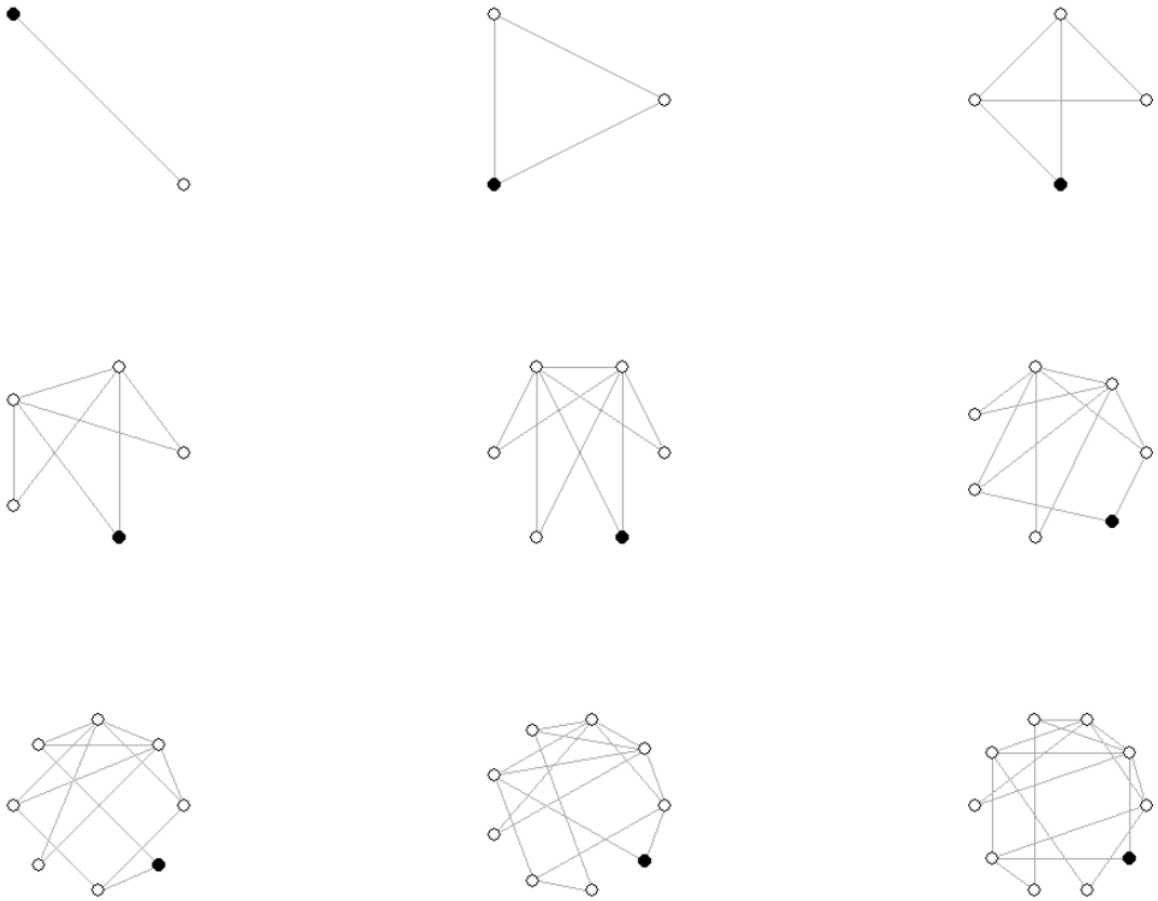
Sieť začne s m_0 vrcholmi, ktoré sú prepojené náhodne, avšak každý má aspoň jednu hranu. V tejto sieti vznikne každú periódu nový vrchol vytvárajúci m väzieb s už existujúcimi bodmi, pričom $m \leq m_0$. Toto sa nazýva rast siete.

Druhým konceptom je preferential attachment, teda princíp vytvárania týchto hrán. Majme dvoch ľudí na večierku. Peter, ktorý je extrovert a pozná tu takmer každého a introvert Milan, ktorý tu pozná iba svoju priateľku. Pri príchode nového človeka na večierok je väčšia šanca, že sa spozná s Petrom, keďže ten sa pohybuje vo viacerých skupinách a rád spoznáva nových ľudí.

Tento princíp teda znamená, že pravdepodobnosť P, že nový vrchol vytvorí väzbu s existujúcim vrcholom i závisí na jeho stupni k_i a ráta sa ako

$$P(k_i) = \frac{k_i}{\sum_j k_j}.$$

V sieti, ktorá vznikla na základe tohto modelu, je teda po t periódach $N = t + m_0$ vrcholov a $L = m_0 + mt$ hrán.



Obr. 7: Vývoj siete v Barabási-Albertovom modely, čierna výplň označuje nový vrchol

Aby sme lepšie porozumeli vývoju siete v tomto modely, pozrime sa ako sa mení stupeň k_i vrcholu i v závislosti od času. S každým novým vrcholom s m hranami má vrchol i m šancí na zvýšenie počtu hrán. Jeho vývoj v čase je teda daný vzorcom

$$\frac{dk_i}{dt} = mP(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j},$$

pričom suma v menovateli sčítava stupne všetkých vrcholov, okrem práve vzniknutého, teda

$$\sum_{j=1}^{N-1} k_j = 2mt - m.$$

Spojením týchto vzťahov dostávame

$$\frac{dk_i}{dt} = \frac{k_i}{2t-1}.$$

Pre dostatočne vysokú periódu nebude hrať (-1) v menovateli významnú rolu, môžme teda aproximovať na

$$\frac{dk_i}{k_i} = \frac{1}{2}\frac{dt}{t}.$$

Riešenie tejto rovnice po integrácii bude

$$k_i(t) = c\sqrt{t}.$$

Využitím faktu, že $k_i(t_i) = m$, teda vrchol vzniknutý v perióde t má m hrán (vlastnosť rastu) dostávame

$$k_i(t_i) = m = c\sqrt{t_i} \Rightarrow c = \frac{m}{\sqrt{t_i}}.$$

Teda výsledný vzorec bude mať tvar

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{1/2}.$$

Z tohto vzťahu vidíme niekoľko dôsledkov pre tento model:

- Exponent je konštantný, teda stupeň každého vrcholu rastie rovnakou mierou.
- Nárast stupňov je sublineárny, keďže exponent $\frac{1}{2} < 1$), teda narozdiel od Erdos-Rényiho modelu sa priemerný stupeň $\langle k \rangle$ nezvyšuje s pridanými vrcholmi.
- Čím skôr bol vrchol i pridaný, tým väčší je jeho stupeň $k_i(t)$. Centrá siete (vrcholy s vysokým stupňom) teda sú centrami spravidla preto, že boi vytvorené skôr ako ostatné vrcholy. Tento fenomén je v biznise a marketingu známy ako *first mover advantage*.

2 Analýza empirických dát sietí

Problém s dátami, ktoré nepochádzajú z teoretických rozdelení je, že môže byť neúmerne komplikované pozorovať ich vlastnosti. Naopak, ak sa dá ukázať, že skúmaný dataset sa správa (aspoň do istej miery) akoby boli jeho prvky z teoretického rozdelenia, môže to značne urýchliť proces analýzy. Taktiež týmto spôsobom vieme nájsť a poukázať na všeobecné vlastnosti podobných datasetov.

Preto sa v tejto kapitole bližšie pozrieme na empirické dáta sietí z reálneho sveta a ich charakteristiky. Zároveň budeme pozorovať, či sa dajú tieto dáta aspoň približne modelovať existujúcimi distribučnými modelmi generovania náhodných grafov.

2.1 Fitovanie modelov

Často je možné sa pri fitovaní štatistických modelov na dáta stretnúť s tzv. grafickou metódou. Aj keď počiatočný odhad dát môže slúžiť na zvolenie vhodného kandidáta na modelovanie, stále bude potrebné overiť, či dané hodnoty testovanému rozdeleniu môžu patriť. Keďže budeme v tejto práci narábať s viacerými štatistickými rozdeleniami, ako hlavnú pomôcku pri overovaní fitovania modelov si zvolíme Kolmogorov-Smirnov test, ktorý je v tomto ohľade všestranný. Jeho testovacia štatistika má hodnotu

$$D_n = \sup_x |F_n(x) - F(x)|,$$

kde $\sup_x$ je suprémom párov vzdialeností medzi empirickou distribučnou funkciou $F_n(x)$ a kumulatívnou distribučnou funkciou $F(x)$ teoretického rozdelenia, na ktorú empirické dáta fitujeme.

Priebeh procesu fitovania vzorky dát na distribučný model sa dá zhrnúť do 3 krokov:

1. Zvoliť vhodný teoretický model na základe vlastnosti dát.
2. Odhadnúť parametre modelu pre vzorku našich dát.
3. Určiť hladinu významnosti - zistiť ako dôveryhodne model s odhadnutými parametrami vystihuje pôvodné dáta.

Na odhad parametrov jednotlivých modelov budeme využívať metódu maximálnej vierohodnosti (MMV) [7], pri ktorej je pravdepodobnosť (likelihood) toho, že náhodný

výber dát $X_1, X_2, .., X_n$ (pričom X_i je vzorka z funkcie hustoty $f(X_i|\theta)$) sa správajú podľa parametra θ:

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta).$$

Túto funkciu maximalizujeme vzhľadom na parameter θ, teda $\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta)$. Vo finálnom tvare budeme hľadať riešenie rovnice

$$log(L(\theta)) = \sum_{i=1}^{n} log f(X_i|\theta).$$

Riešenia tejto rovnice a ich odvodenia pre niektoré distribúcie sa dajú nájsť v [7].

Najväčšia výzva nastane v treťom bode, a síce pri určovaní hladiny významnosti. Jedna z limitácii Kolmogorovho-Smirnovho testu je, že nemožno veriť hladine významnosti, keď sú pri testovaní použité parametre odhadnuté z pôvodnej dátovej vzorky. V tomto prípade môže byť totiž p-hodnota testu výrazne nadhodnotená. Tento problém sa dá vyriešiť pomocou generovania nových, syntetických datasetov zo skúmaného teoretického rozdelenia avšak s použitím parametrov odhadnutých z dátovej vzorky. Pri každom z týchto syntetických datasetov určíme novú testovaciu štatistiku Ds, pričom p-hodnotu určíme ako percento syntetických testovacích štatistík Ds väčších ako testovacia štatistika pôvodných dát D, alebo tiež

$$p = \frac{1}{k} \sum_{i=1}^{k} [Ds_i \geq D],$$

pričom k je počet synteticky vygenerovaných datasetov. Takto odhadnutá p-hodnota je už spoľahlivejšia.

Kód testovania tejto metódy v R pomocou balíku MASS pre lognormálne rozdelenie môžme vidieť nižšie, pričom sa dá jednoducho upraviť pre množstvo iných rozdelení. Myšlienku tejto metódy čerpáme z [3], kde je odvodené, že pre nanajvýš 2500 syntetických datasetov by mala byť výsledná hodnota p-hodnoty presná na dve desatinné miesta.

```
1 \library(MASS)
2 lognormal = function(d, limit=2500) {
3   #odhadnutie parametrov metodou MMV
4   meanlog = fitdist(d, "lnorm","mle")$estimate[1];
5   sdlog = fitdist(d, "lnorm","mle")$estimate[2];
6   #testovacia statistika pre povodne data
7   D = ks.test(d, "plnorm", meanlog = meanlog, sdlog = sdlog)
```

```
  #vypocitanie p-hodnoty
  count = 0;
  for (i in 1:limit) {
    #vytvorenie syntetickho datasetu
    syn = rlnorm(length(d), meanlog = meanlog, sdlog = sdlog);
    meanlog2 = fitdist(syn, "lnorm","mle")$estimate[1];
    sdlog2 = fitdist(syn, "lnorm","mle")$estimate[2];
    #testovacia statistika pre syntetick data
    Ds = ks.test(syn, "plnorm", meanlog = meanlog2, sdlog = sdlog2);
    #porovnanie testovacej statistiky syntetickeho datasetu s povodnym
    if(Ds$stat >= D$stat) {count = count + 1};
  }
  return(list(meanlog=meanlog, sdlog=sdlog, stat = t$stat, p = count/limit, KSp = D$p)
      );
}
```

2.2 Power-law distribúcia

Jedným zo zaujímavejších typov sietí sú bezškálové siete, ktorých stupne rozdelení k sa riadia tzv. power-law distribúciou. Matematicky povedané, vzorka dát x sa riadi zákonom power-law, ak pre jej pravdepodobnostnú hustotu platí

$$p(x) \propto x^{-\alpha},$$

pričom (kladná) α je tzv. *škálovací* parameter (resp. exponent) a obvykle sa pohybuje v rozmedzí $2 < \alpha < 3$. V praxi je veľmi obmedzený počet empirických datasetov [3], ktoré by sa riadili power-law rozdelením ako celok. Vo väčšine prípadov dáta vyhovujú tomuto rozdeleniu až od istého bodu x_{min}. V tomto prípade hovoríme, že power-law rozdelením sa riadi *chvost* datasetu x.

Spojité power-law rozdelenie (ktorým sa budeme zaoberať) pre skúmanú veličínu x je opísané hustotou pravdepodobnosti $p(x)$ takou, že

$$p(x)dx = Pr(x \leq X < x + dx) = Cx^{-\alpha}dx,$$

kde X je pozorovaná hodnota a C normalizačná konštanta. Z tejto definície je zjavné, že toto rozdelenie pre $x \to 0$ diverguje, teda táto rovnica nebude platiť pre $x \geq 0$. Bude preto existovať dolná hranica x_{min} ohraničujúca, kde sa "správanie"podľa power-law

začína. Definičná rovnica sa po integrovaní za predpokladu, že $\alpha > 1$ rovná

$$p(x) = C\frac{x^{1-\alpha}}{1-\alpha}.$$

Keďže v bode $x = x_{min}$ bude hodnota určíte patriť rozdeleniu, a teda dostávame podmienku

$$p(x_{min}) = 1 = \frac{C}{1-\alpha}x_{min}^{1-\alpha},$$

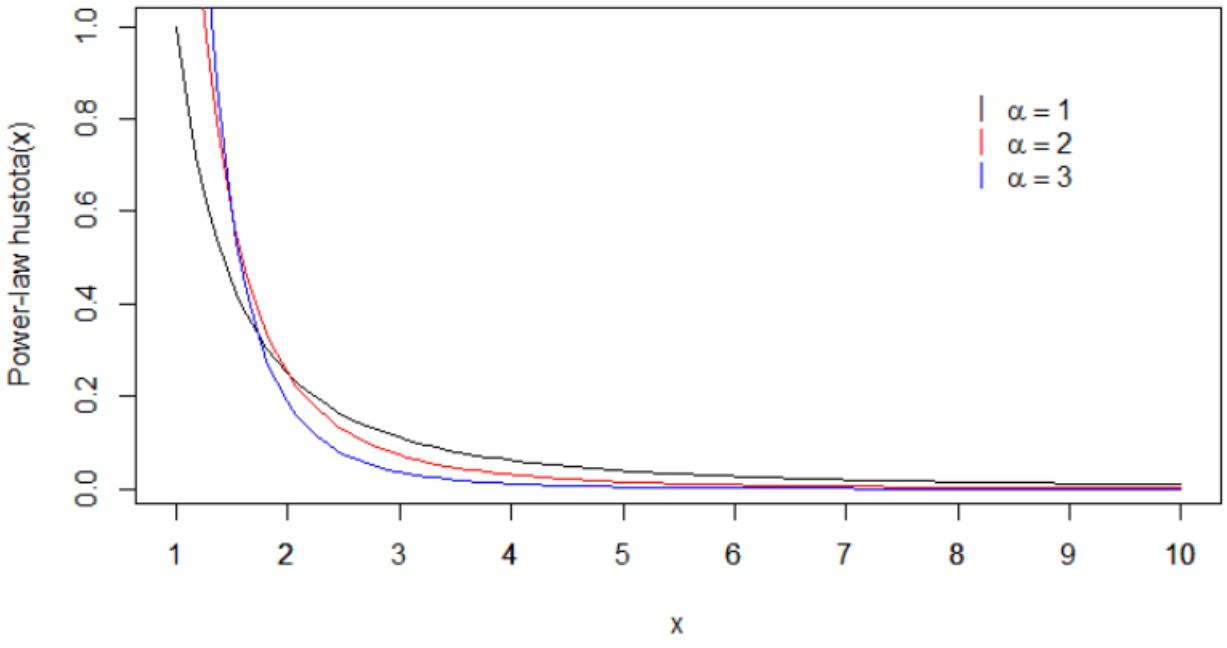
z ktorej vyjadríme normalizačnú konštantu jednoducho ako

$$C = (1-\alpha)(x_{min})^{\alpha-1}.$$

Skombinovaním týchto dvoch faktov a miernou úpravou na kompaktnejší tvar dostávame finálny stav rovnice hustoty pravdepodobnosti

$$p(x) = \frac{\alpha-1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha}$$

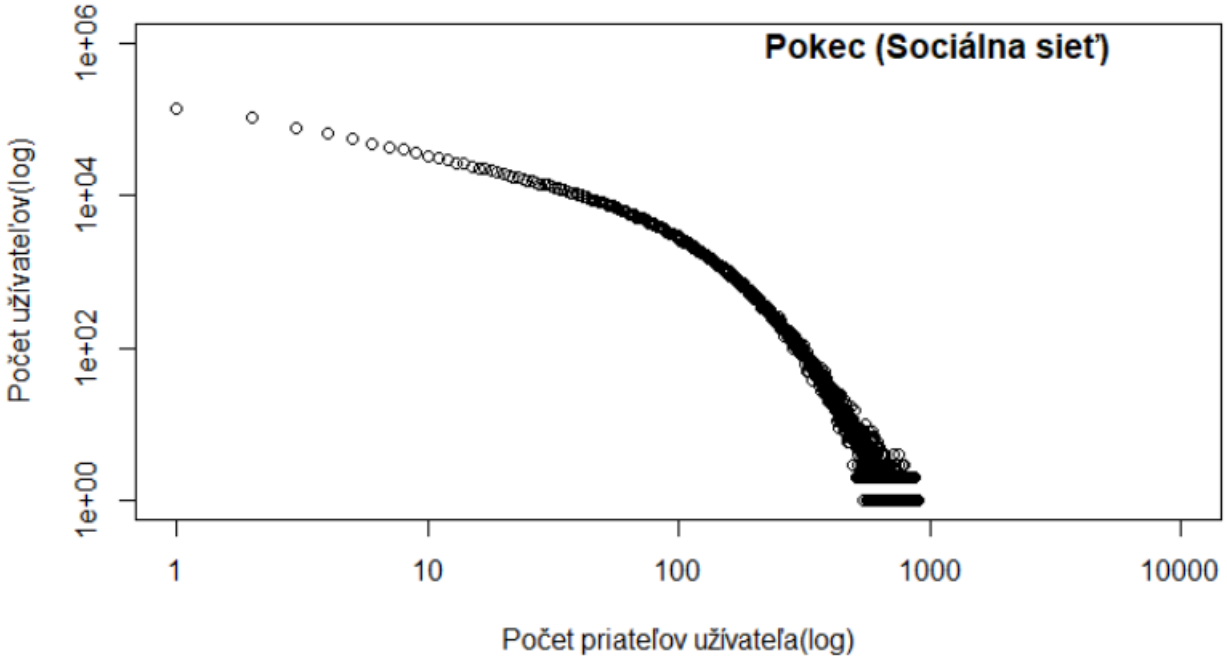
pre $x \geq x_{min}$. Teoretické hustoty power-law rozdelenia pre rôzne hodnoty α môžme vidieť v obr. 8.



Obr. 8: Teoretické hustoty rozdelenia pre rôzne hodnoty α

Avšak pri skutočných dátach sa power-law rozdelenia obvykle vizualizujú ako distribúcia frekvencií pre jednotlivé hodnoty dát. V obr. 9 sú to počty priateľstiev používateľov na Pokeci v log-log grafe. Vidno, že používateľ s najväčším počtom prepojení

ich má cez 20 000, naopak ľudí s iba jedným priateľom je cez 130 000. Stupne k vrcholov tejto siete zjavne nie sú všetky z rozdelenia power-law - to by tento graf musel byť priamkou. Avšak pre $x_{max} \approx 100$ sa zdá, že by mohli hodnoty zo začiatku tohto datasetu patriť power-law rozdeleniu.



Obr. 9: Dáta o vzťahoch medzi užívateľmi slovenskej sociálnej siete Pokec, zdroj dát: [12]

2.2.1 Odhadovanie spodnej hranice

Power-law sa líši od ostatných rozdelení v tejto práci práve existenciou spodnej hranice x_{min}, ktorej odhadovanie hrá dôležitú rolu pri fitovaní tohto modelu na empirické dáta. Preto sa bližšie pozrieme na metódu jej odhadu predstavenú *Clausetom* v [4].

Myšlienka tejto metódy spočíva vo zvolení hodnoty odhadnutého $\hat{x}_{min}$ tak, aby bolo pravdepodobnostné rozdelenie dátovej vzorky a fit power-law modelu čo najpodobnejšie pre hodnoty presahujúce $\hat{x}_{min}$. Teda ak zvolíme $\hat{x}_{min}$ vyššie ako skutočné x_{min}, tak si zmenšujeme dátovú vzorku, čím sa medzera medzi rozdeleniami zväčší kvôli štatistickým fluktuáciám. A naopak, ak zvolíme $\hat{x}_{min}$ nižšie ako skutočné x_{min}, rozdelenia budú odlišnejšie kvôli využití dát zásadne odlišnými oproti modelu, ktorým ich opisujeme. Z týchto dôvodov bude náš odhad ak zvolíme $\hat{x}_{min}$ vyššie ako skutočné $\hat{x}_{min}$, ležať medzi týmito dvoma prípadmi.

Na testovanie rozdielu medzi spomínanými rozdeleniami použijeme opäť Kolmogorov-

Smirnovu testovaciu štatistiku majúcu tvar

$$D = \max_{x \geq x_{min}} |S(x) - P(x)|,$$

kde $S(x)$ a $P(x)$ sú kumulatívne distribučné funkcie. $S(x)$ reprezentuje hodnoty aspoň x_{min} a $P(x)$ sa týka power-law modelu najlepšie fitujúcemu dáta spĺňajúce $x > x_{min}$. Našim odhadom $\hat{x}_{min}$ bude hodnota x_{min} minimalizujúca testovaciu štatistiku D.

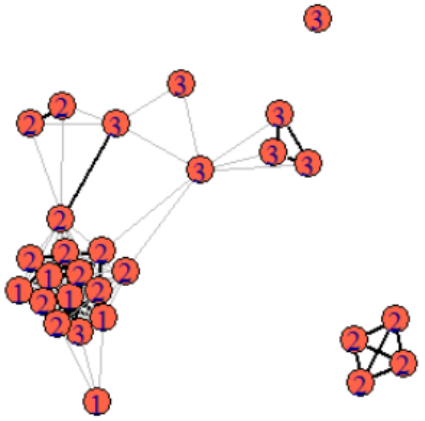
Teoretické odvodenie odhadu hodnoty α je dostupné napríklad v [3], zatiaľ čo implementáciu odhadov oboch týchto parametrov v Rku si predstavíme neskôr.

2.3 Zebry a ich interakcie

V prvej dátovej vzorke sa budeme zaoberať relatívne skromnou a jednoduchou sieťou, ktorá sa zaoberá 28 zebrami z africkej Keni (dáta z 11). Vrcholy V sú jednotlivé zebry, zatiaľ čo hrany E medzi nimi vyjadrujú interakciu. Ak sa dve zebry z_i a z_j počas štúdie stretli, tak platí $\{z_i, z_j\} = 1$. Ak sa spolu stretávali signifikantne často, tak hodnota ich hrany je $\{z_i, z_j\} = 2$. V prípade, že k stretnutiu nedošlo má hrana hodnotu 0. Ďalší údaj, ktorým disponujeme, je reprodukčné štádium v ktorom sa zebra nachádza. Tie sú rozdelené do troch typov:

1. Dojčiaca samica,
2. Nedojčiaca samica,
3. Samec.

Tieto typy sú znázornené spolu s prehľadom o interakciách v obr. 10.



Obr. 10: Interakcie medzi zebrami v rôznych štádiách reprodukcie, hrubé čiary vyjadrujú častejšie socializovanie.

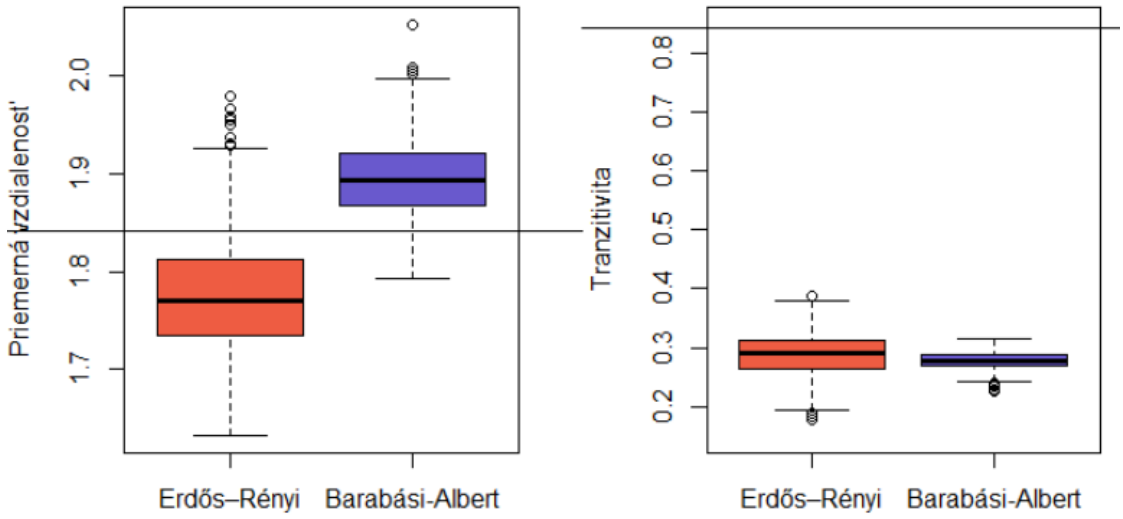
Ako prvé vyskúšame, ako úspešné budú pri odhade základných charakteristík dva základné modely predstavené v prvej kapitole: Erdős–Rényiho $G(n, p)$ a Barabási-Albertov model. Funkcie v R s vhodnými parametrami budú mať tvar:

```
\library(igraph)
gl <- vector('list', 1000)
for(i in 1:1000){
  gl[[i]] <- erdos.renyi.game(n = gorder(zebry), p.or.m = edge_density(zebry), type =
      "gnp")
}
md_erdos <- lapply(gl, mean_distance, directed = FALSE)
trans_erdos <- lapply(gl, transitivity)

for(i in 1:1000){
  gl[[i]] <- barabasi.game(n = gorder(zebry), m = gsize(zebry)/gorder(zebry))
}
md_barabasi <- lapply(gl, mean_distance, directed = FALSE)
trans_barabasi <- lapply(gl, mean_distance)
```

V obr. 11 môžme vidieť boxploty vyjadrujúce priemernú vzdialenosť a globálny zhlukový koeficient (tranzitivitu) pre oba tieto modely, ako aj pre pôvodné dáta (vodorovná čiara). Aj keď priemernú vzdialenosť $\langle d \rangle$ jeden z modelov podhodnotil, a druhý nadhodnotil, tranzitivitu oba modely mnohonásobne podcenili so skutočnou hodnotou na úrovni 85%. Toto pripisujeme faktu, že v skutočných dátach sú až 3 komponenty, ktoré navyše obsahujú až 10-členné "cliques", čo sú susedstvá vrcholov, v ktorých je každá dvojica prepojená. Tieto dva ukazatatele sú veľmi nezvyčajné pre tak malú sieť, čomu

aj pripisujeme nepresný odhad modelmi náhodných grafov.



Obr. 11: Porovnanie simulácii náhodných grafov so skutočnými hodnotami parametrov.

Z nákresu siete sa tiež zdá, že zebry sa skôr socializujú s rovnakými typmi ako sú ony samé. Túto teóriu otestujeme porovnaním socializovania so svojím typom oproti všeobecnosti.

Typ zebry	Dojčiaca	Nedojčiaca	Samec
Dojčiaca	7	35	5
Nedojčiaca	35	42	13
Samec	5	13	9

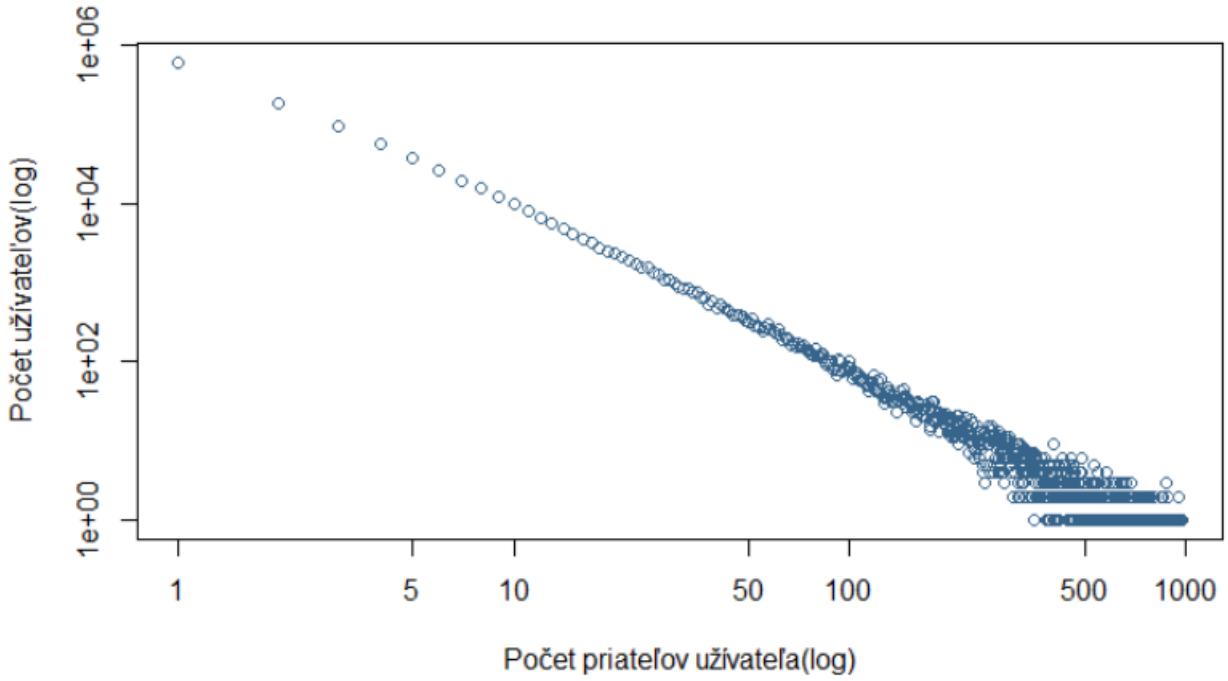
Tabuľka 2: Počet pút medzi jednotlivými typmi zebry

Dvojice typov zebry z tabuľky podrobíme Fisherovmu exaktnému testu. Jedinú dvojicu typov, pri ktorej jeho p-hodnota zamieta H_0, sú dojčiace a nedojčiace samice, čo znamená, že tieto dva typy sa radšej socializujú so samicami v rovnakom štádiu reprodukcie. Tento výsledok majú prevažne na svedomí nedojčiace samice, ktorých je v datasete podstatne viac (15 oproti 5).

Nakoniec sa pozrieme na to, či hrá rozdiel medzi typmi pri vytváraní pevnejších pút: Fisherovým exaktným testom tentokrát porovnávame kategorické premenné, pričom p-hodnota vychádza na úrovni 8%, čo je tesne nad hranicou, teda nezamietame hypotézu, že k signifikantnému počtu interakcií môže prísť medzi zebrami rôznych typov.

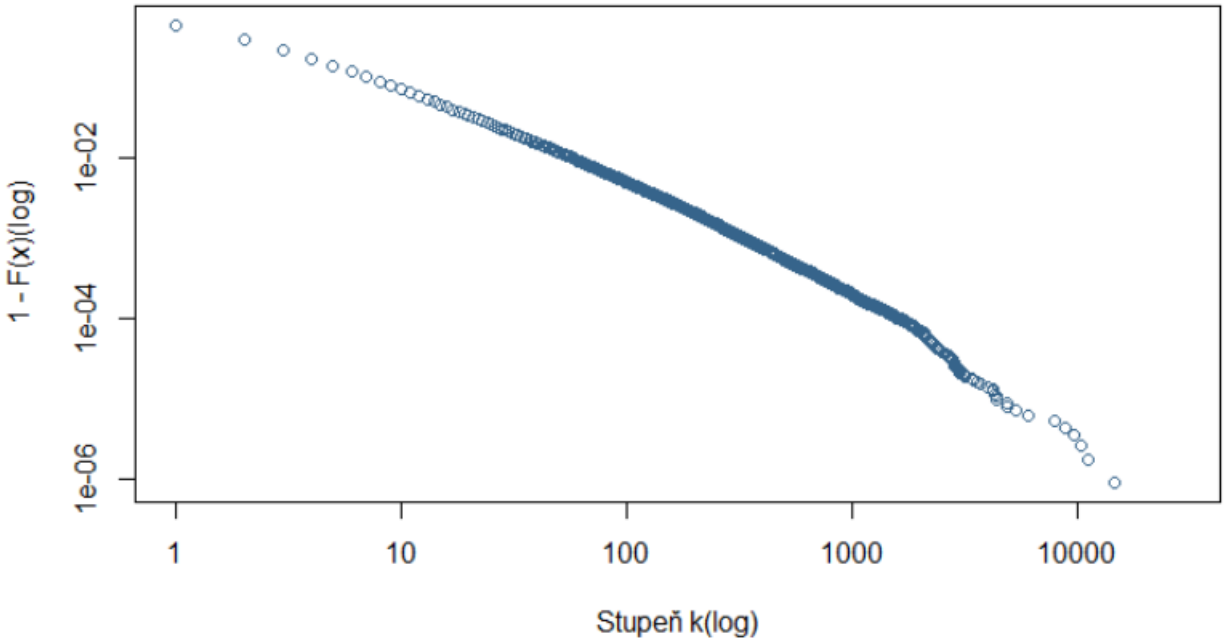
2.4 Sociálna sieť

Po malej sieti v predošlej časti sa teraz zameriame na pravý opak. Pozrieme sa na sieť priateľstiev na Youtube (dáta z [14]), ktorá má 1134890 používateľov, resp. vrcholov V a 2987624 priateľstiev, teda hrán E. Pri sieti takejto veľkosti je takmer nemožné ju zrozumiteľne načrtnúť, preto sa pozrieme na jej rozdelenie.



Obr. 12: Hustota rozdelenia priateľstiev medzi používateľov

Z obr. 12 sa ukazuje, že táto sieť je veľmi dobrým kandidátom na power-law rozdelenie. Až 602 000 používateľov (viac ako polovica) má iba jedného priateľa, čo opäť potvrdzuje nerovnováhu tohto rozdelenia.



Obr. 13: Distribučná funkcia rozdelenia priateľstiev medzi používateľov

Pomocou implementácie metódy z [3], ktorú sme si predstavili v tejto kapitole, do programu R odhadneme MLE parametre pre fitovanie týchto dát na power-law model.

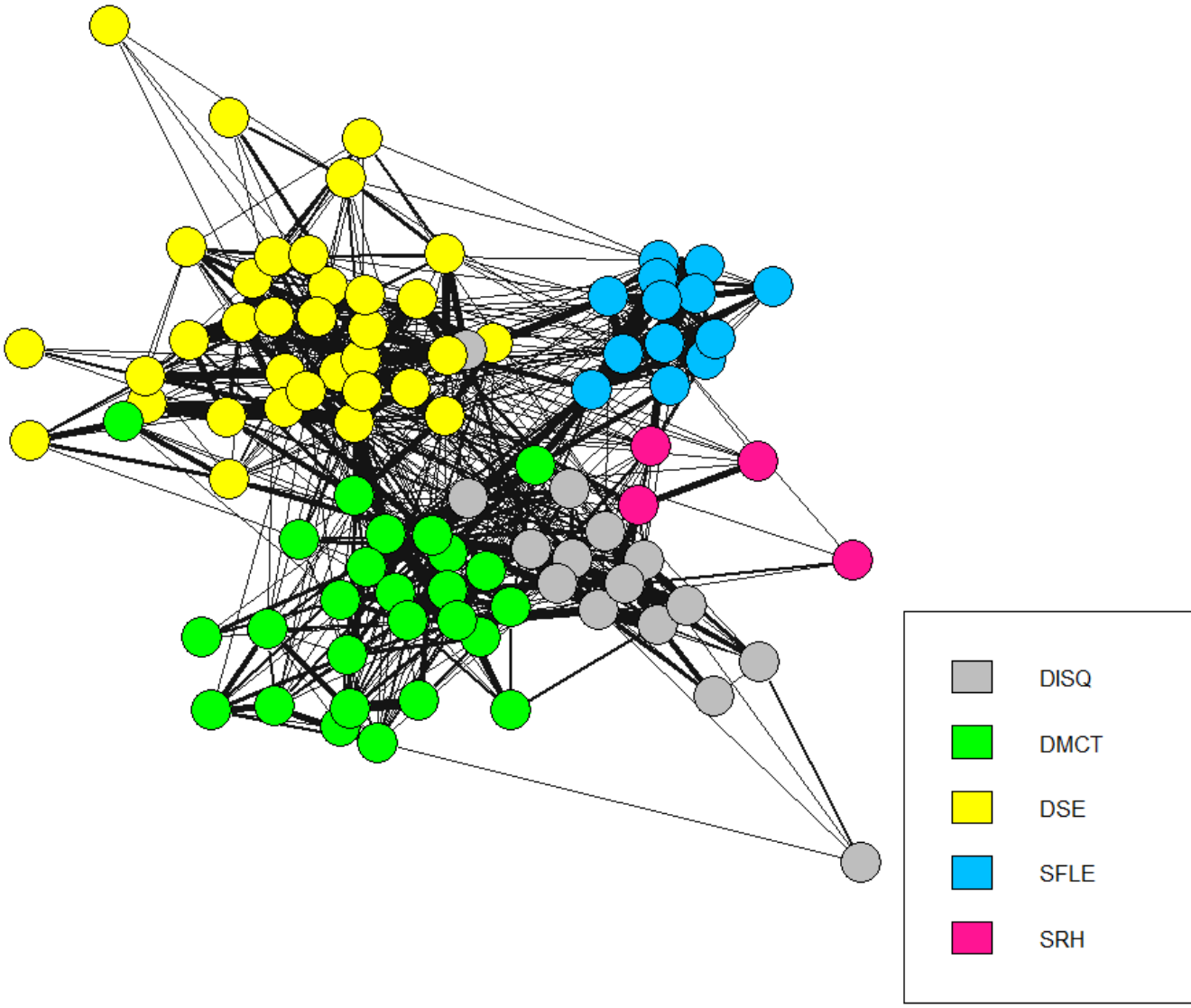
```
1
2  \library(igraph)
3  \library(MASS)
4  #definovanie funkcii
5  dpareto=function(x, shape=1, location=1) shape * location^shape / x^(shape + 1);
6  ppareto=function(x, shape=1, location=1) (x >= location) * (1 - (location / x)^shape);
7  qpareto=function(u, shape=1, location=1) location/(1 - u)^(1 / shape);
8  rpareto=function(n, shape=1, location=1) qpareto(runif(n), shape, location);
9
10 location = function(d, max=100)  {
11   stat = NULL;
12   p = NULL;
13   mstat = 1;
14   mp = NULL;
15   mlocation = NULL;
16   mshape = NULL;
17   for (location in 1:max) {
18     d = d[d >= location];
19     # MLE for shape parameter (discrete case)
20     shape = 1 / mean(log(d / (location - 0.5)));
21     t = ks.test(d, "ppareto", shape = shape, location = location);
22     stat[location] = t$stat;
23     p[location] = t$p;
24     if (stat[location] < mstat) {
```

```
25        mstat = stat[location];
26        mp = p[location];
27        mlocation = location;
28        mshape = shape;
29      }
30    }
31    return(list(stat=mstat, p=mp, shape=mshape, location=mlocation, statvec=stat, pvec=p
          ));
32  }
33
34  pareto = function(d, max=100, limit=2500) {
35
36    par = location(d, max);
37    shape = par$shape;
38    location = par$location;
39    t = ks.test(d[d>=location], "ppareto", shape = shape, location = location);
40
41    ntail = length(d[d>=location]);
42    head = d[d < location];
43    count = 0;
44    for (i in 1:limit) {
45      syn = rpareto(ntail, shape = shape, location = location);
46      syn = c(head, syn);
47      par2 = location(syn, max);
48      shape2 = par2$shape;
49      location2 = par2$location;
50      t2 = ks.test(syn[syn >= location2], "ppareto", shape = shape2, location =
            location2);
51      if(t2$stat >= t$stat) {count = count + 1};
52    }
53
54    return(list(shape=shape, location=location, stat = t$stat, p = count/limit, KSp = t$
          p));
55  }
```

Pomocou tohto kódu dostávame odhad power-law exponentu $\alpha = 2.141 s x_{min} = 8$, so simulovanou p-hodnotou= 0.84 môžme usúdiť, že rozdelenie stupňa k z pôvodných dát skutočne zodpovedá power-law distribúcii.

3 Šírenie informácií v sieti

V tejto kapitole sa budeme zaoberať sieťou stretnutí spolupracovníkov v pobočke Francúzskeho inštitútu pre verejné zdravie (InVS) . Jedná sa o graf s 92 vrcholmi, ktoré značia jednotlivých zamestnancov spoločnosti označenými ID kódmi a spolu až 9827 hranami, pričom každá hrana označuje stretnutie dvoch kolegov na chodbe z obdobia 2 pracovných týždňov, teda 10 dní. Tieto dáta pochádzajú z [5] a použijeme ich na analýzu chovania týchto zamestnancov a jeho dopad na šírenie nepravdivých informácii. Budeme pritom pracovať s programom R, konkrétne knižnicou igraph.



Obr. 14: Vizualizácia siete zamestnancov, Fruchterman-Reingoldov model usporiadania vrcholov

Označenie	Počet vrcholov	Hustota hrán	Tranzitivita
DISQ	15	11.371	0.684
DMCT	26	4.415	0.597
DSE	34	5.027	0.518
SFLE	4	6	0
SRH	13	33.025	0.822
Celá sieť	92	2.348	0.368

Tabuľka 3: Vlastnosti podgrafov oddelení

Na obr. 14 môžme vidieť rozdelenie pracovníkov do oddelení, pričom hrúbka hrany medzi dvoma vrcholmi reprezentuje počet[2] (resp. celkovú dĺžku) stretnutí počas sledovaného obdobia. Pre ďalší postup je dôležité ukázať správanie zamestnancov vrámci týchto oddelení.

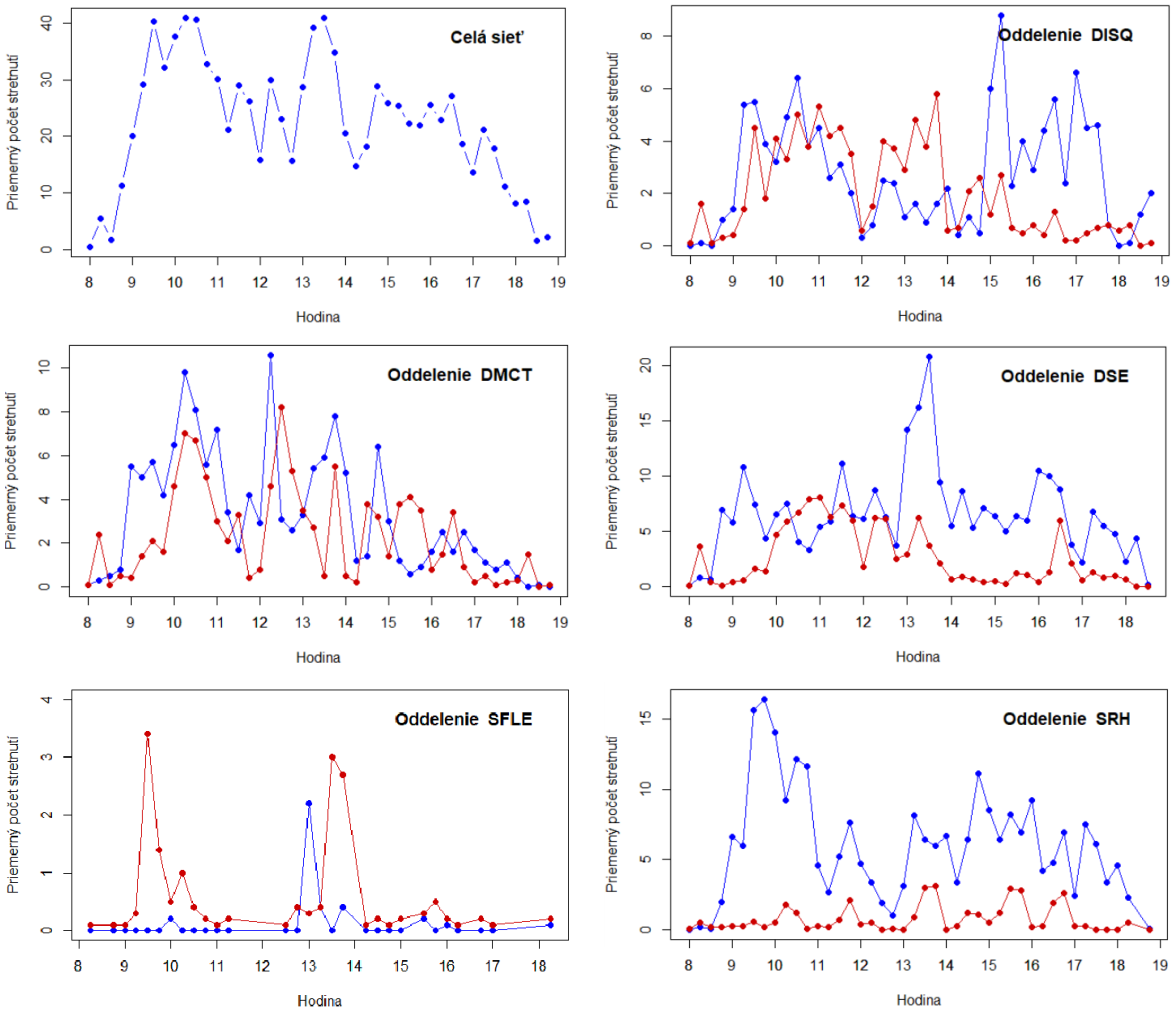
Z tabuľky 1 vidno, že podgrafy jednotlivých oddelení majú značne vyššie hodnoty hustoty a tranzitivity, z čoho vyplýva, že vrámci oddelení sú vrcholy prepojenejšie ako v kontexte celej siete. Toto môžme pozorovať aj v obr. 14 a dáva to zmysel aj v kontexte častejšej spolupráce kolegov z rovnakého oddelenia.

Pozrieme sa preto bližšie na archetypy zamestnancov. Zo všetkých hrán je až 82% interných - teda v rámci jedného oddelenia. Z tohto sa dá usúdiť, že zamestnanci, ktorí prichádzajú do styku s ľuďmi z rôznych oddelení, a teda ich prepájajú, budú mať kľúčovú rolu pri šírení informácií.

Závislosť množstva aktivity od hodiny dňa sledujeme na obrázku 15, kde je tiež zjavná dominancia interných stretnutí (modrou) nad externými (červenou). Výnimkou potvrdzujúcou pravidlo je pritom miestami oddelenie SFLE (logistika), u ktorého je však časťou dôvodu fakt, že má len 4 pracovníkov. Ďalej pozorujeme hodiny, v ktorých dochádza k najviac kontaktom (okolo 10. a 13. hodiny), čo môže znamenať zvýšené riziko šírenia.

Túto hypotézu porovnáme s ďalším koeficientom, ktorým sa často dôležitosť vrcholov analyzuje. Ten sa nazýva stredová medzipoloha (z angl. Betweenness centrality,

[2]Takáto vizualizácia je výrazne prehľadnejšia ako zobrazovať každú hranu zvlášť.

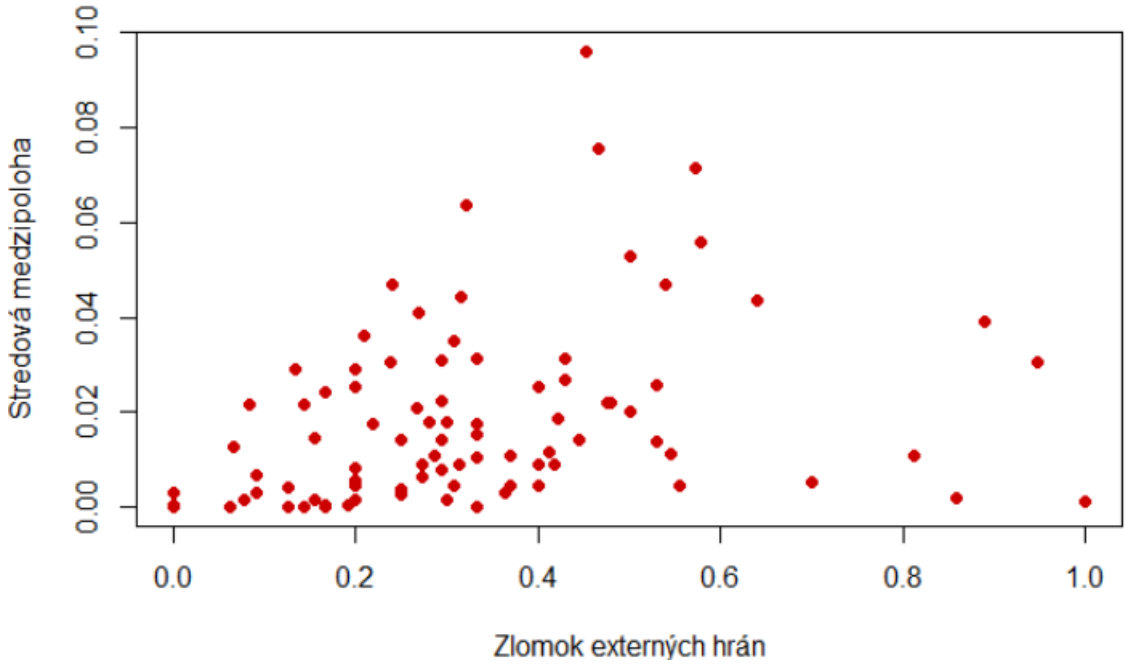


Obr. 15: Priemerná aktivita v priestoroch spoločnosti - 15 minutové intervaly

slovenský preklad z [10]) a počíta sa ako

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

pričom σ_{st} je počet najkratších ciest z vrcholu s do vrcholu t a $\sigma_{st}(v)$ je podmnožina týchto ciest prechádzajúca vrcholom v. Teda vystihuje, ako kritický je tento vrchol v v zmysle prepájania grafu.



Obr. 16: Vzťah stredovej medzipolohy vrcholu a zlomku jeho externých (prechodných) hrán

Na obr. 16 môžeme pozorovať koreláciu medzi "dôležitosťou"vrcholov a tým, koľko percent ich hrán smeruje do iného oddelenia. Zároveň si všimneme tri základné archetypy správania zametnancov. Ako sa dalo očakávať, *rezidenti* - teda vrcholy ktoré sa zdržujú prevažne vo svojom oddelení majú stredovú medzipolohu nízku. Zaujímavé však je, že ani *tuláci* - zamestnanci s vysokým percentom prechodov do iných oddelení nemajú tento koeficient najvyšší. Najdôležitejšími pri prepájaní budú tzv. *spojky*, teda vrcholy, pri ktorých sa percento externých hrán pohybuje okolo 50%. Z tohto sa dá predpokladať, že spojky budú mať kritickú úlohu pri šírení falošných správ.

3.1 Modelový problém

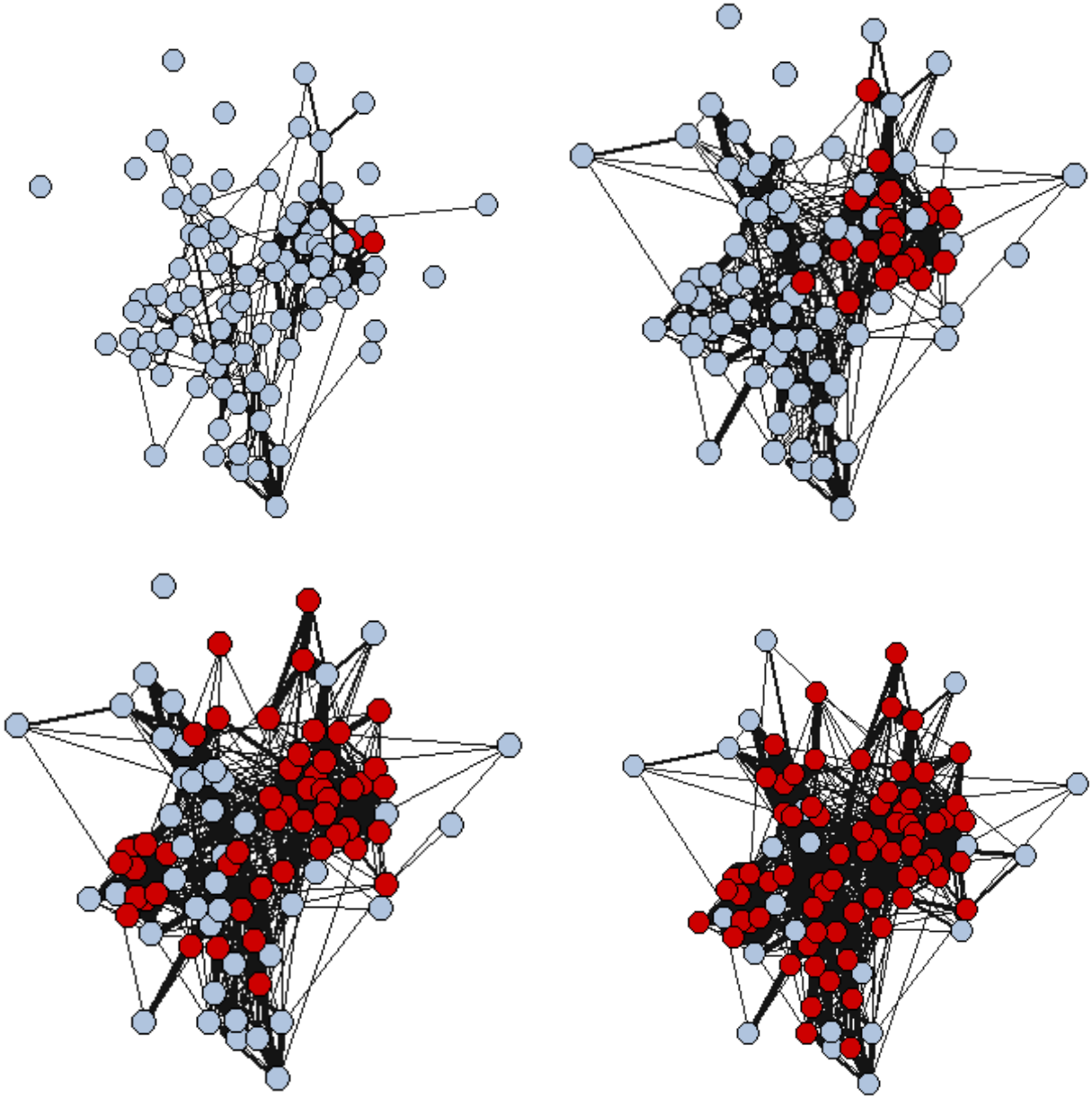
Ide o postupne rastúcu sieť a teda je možné sledovať jej vývoj v čase. Vďaka tejto vlastnosti máme možnosť pomocou programu R simulovať šírenie falošnej správy sieťou.

Budeme pracovať so zjednodušeným modelom, na ktorom bude našim cieľom ukázať dôležitosť vlastností vrcholov.

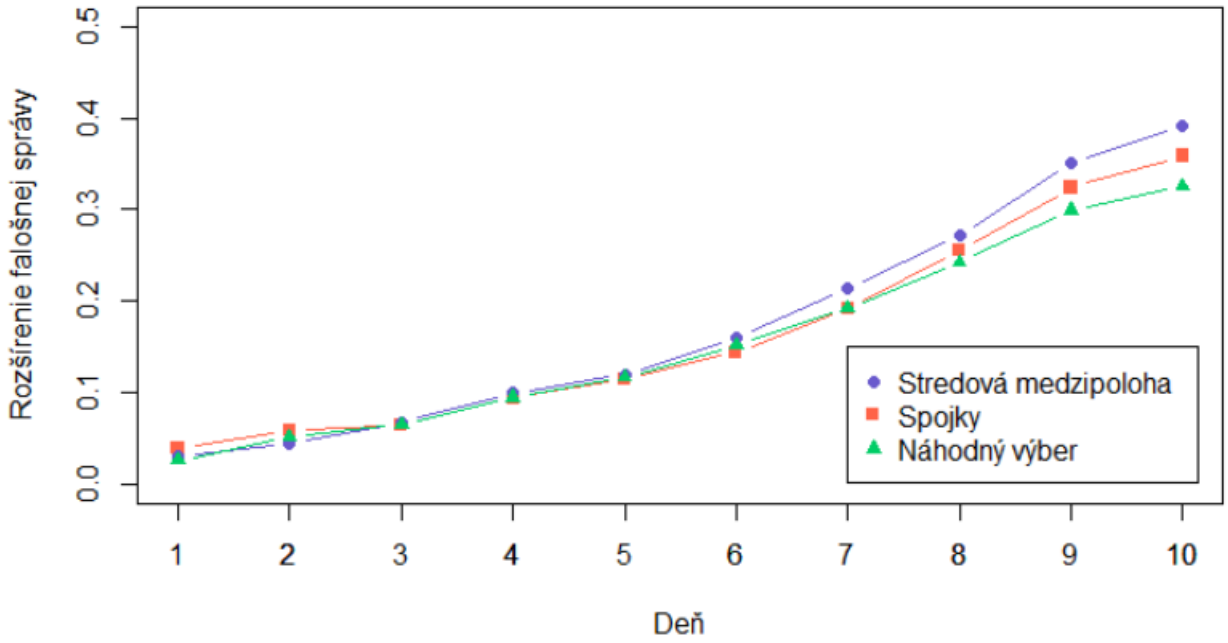
```
\library(igraph)
p<-0.03
zac<- sample(1:gorder(g),1)
g<-set_vertex_attr(gsub,"veri",value=0)
V(g)[zac]$veri <- 1
for (i in 1:gsize(g)){
    akt_edge<-ends(g,E(g))[i,]
    if (V(g)[akt_edge[1]]$veri==1 & V(g)[akt_edge[2]]$veri==0){
        V(g)[akt_edge[2]]$veri<-rbinom(1,1,p)
    }  else
    if (V(g)[akt_edge[1]]$veri==0 & V(g)[akt_edge[2]]$veri==1){
        V(g)[akt_edge[1]]$veri<-rbinom(1,1,p) } }
```

Falošná správa sa dostane prvý deň k jednému zo zamestnancov. Tento je öznačenýä bude túto novinku zdieľať s každým, s kým príde do kontaktu. Existuje pritom 3% pravdepodobnosť, že ho presvedčí a teda získa ďalšieho pomocníka pri šírení tejto správy. Na obr. 17 môžme vidieť vývoj tejto siete a postupnú ïnfekciu"nepravdivou novinou.

Ďalej sa pokúsime zistiť, či sú spojky pri šírení správy naozaj efektívnejšie ako ostatné vrcholy, teda pozorovaním, aké percento zamestnancov správe uverí. Toto otestujeme nasledovne: proces budeme simulovať opakovane s rôznymi začiatočnými vrcholmi. 50 krát pre vrcholy spojky, 100 krát pre náhodný vrchol a 50 krát pre vrcholy s najvyšším koeficientom stredovej medzipolohy. Výsledky týchto simulácií sú na obr. 18.



Obr. 17: Ukážka šírenia falošnej správy sieťou, dni 1,4,7 a 10



Obr. 18: Zlomok zamestnancov veriacich falošnej správe v závislosti od typu začiatočného vrcholu

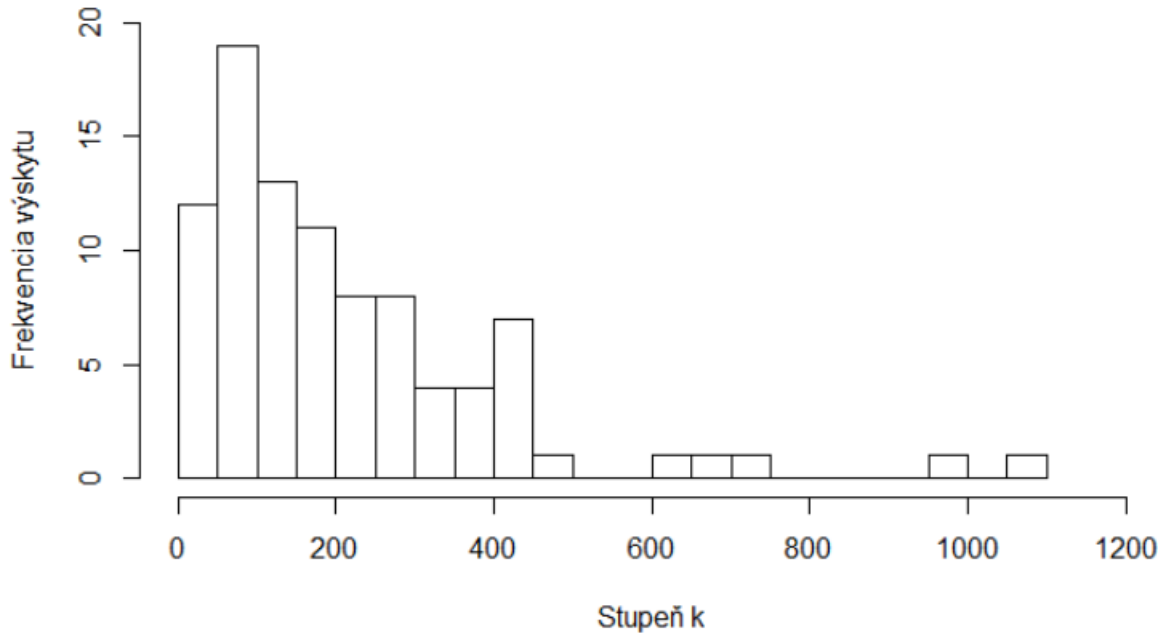
Simulácia našu hypotézu potvrdzuje - pre začiatočný vrchol z náhodného výberu sa správa šíri pomalšie ako pre výber z množiny spojkových vrcholov. Ukazuje sa, že by teda pri prípadnom boji s nepravdivými informáciami v tejto spoločnosti mohlo mať zmysel zameriavať sa na spojky, ktoré prepájajú viaceré oddelenia a majú tak veľký vplyv na ostatné vrcholy v sieti[3]. Aj keď vrcholy s vysokým koeficientom stredovej medzipolohy majú vyššiu mieru šírenia, sústredenie sa na archetypy správania vrcholov tiež prináša výsledky a umožňilo nám lepšie pochopiť dynamiku v sieti. Takisto poznamenávame, že rozdiel medzi spojkami a náhodným výberom sa prejavil hlavne v neskorších fázach, čo prisudzujeme schopnostiam spojok preniknúť do viacerých oddelení, v ktorých sa už potom správa šíri jednoducho.

3.2 Skúmanie vlastností siete

Zatiaľ sme si utvorili vcelku detailný obraz o tom, ako skúmaná sieť vyzerá, a ako sa v nej šíria informácie. V tejto časti sa pozrieme na to, ako by náš model fungoval v

[3]Tieto výsledky sú dosiahnuté za pomoci dostupných dát, teda s väčším časovým intervalom ako 10 dní by sme možno dosiahli presnejšie výsledky

iných, simulovaných sieťach, a či existuje model vytvorenia náhodného grafu, ktorý by dosahoval podobné výsledky ako sieť z reálnych dát. Za bežných podmienok by sme začali Erdős-Rényiho modelom, avšak ten je v tomto prípade nepoužiteľný, keďže sa ním nedajú modelovať grafy s násobnými hranami.



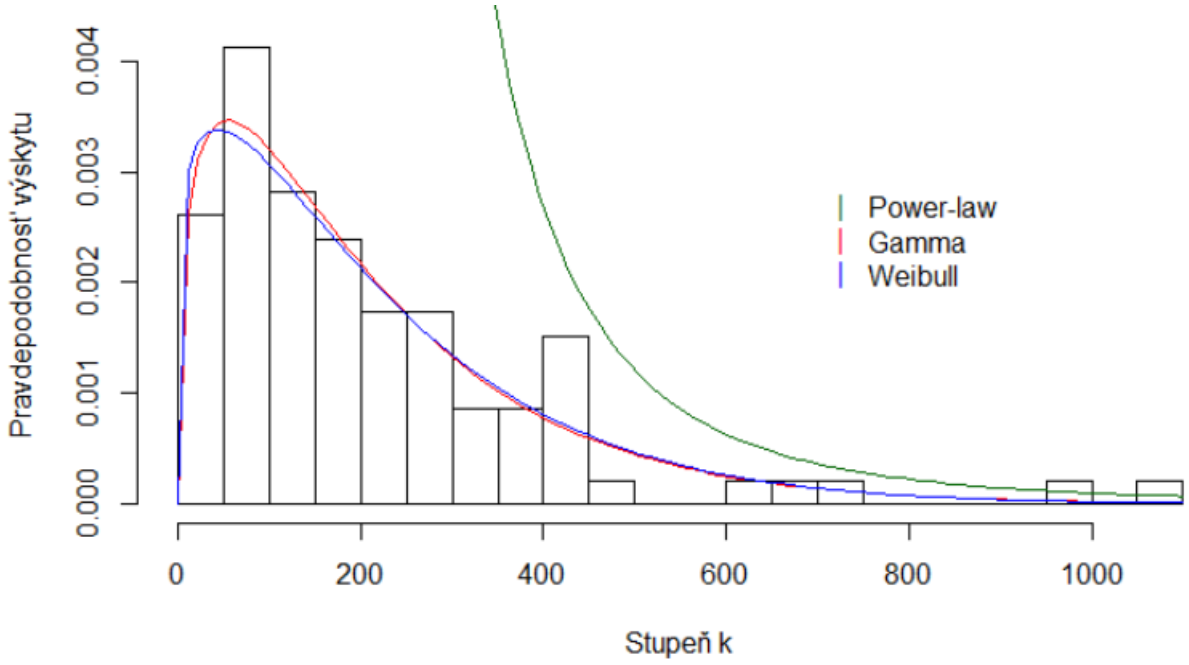
Obr. 19: Rozdelenie stupňov k vrcholov pôvodných dát

V obr. 19 vidíme distribúciu stupňov našich dát, ktorá nám poslúži na zvolenie modelov na simuláciu. Keďže naša sieť nevzniká postupným pridávaním vrcholov, ale len sa medzi už existujúcimi vrcholmi vytvárajú hrany, vylúčime aj Barabási-Albertov model. Postupom vysvetleným v predošlej kapitole otestujeme, ako dobre naše dáta fitujú niektoré známe rozdelenia.

V tabuľke 2 vidíme odhadnuté parametre a p-hodnoty vyrátané pomocou funkcií z druhej kapitoly. Táto tabuľka naše domnenie po zhliadnutí histogramu potvrdila - dáta určite fitovať na normálne rozdelenie nebude reálne. Z obr. 20 vidno, že rozdelenia Gamma aj Weibull sú na tom relatívne podobne. Naopak power-law distribúcia mierne fituje rozdelenie iba pre hodnoty $x > x_{min}$, čo z tabuľky 2 vieme, že je $x > 285$. Do tejto kategórie teda spadá 28% vrcholov, teda sa určite nedá povedať, že naša sieť je bezškálová.

Rozdelenie	Hustota	1. parameter	2. parameter	p-hodnota
Normálne	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\hat{\mu} = 213.63$	$\hat{\sigma} = 194.76$	0
Weibull	$\begin{cases} 1 - e^{-(x/\lambda)^k}, & \text{pre } x \geq 0 \\ 0, & \text{pre } x < 0 \end{cases}$	$\hat{k} = 1.17$	$\hat{\lambda} = 226.12$	0.79
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\hat{\alpha} = 1.35$	$\hat{\beta} = 0.0063$	0.88
Power law	$p(x) = \frac{\alpha-1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}$	$\hat{\alpha} = 2.57$	$\hat{x}_{min} = 285$	0.506

Tabuľka 4: Vlastnosti podgrafov oddelení



Obr. 20: Fitovanie hustôt rozdelení na histogram stupňov vrcholov k

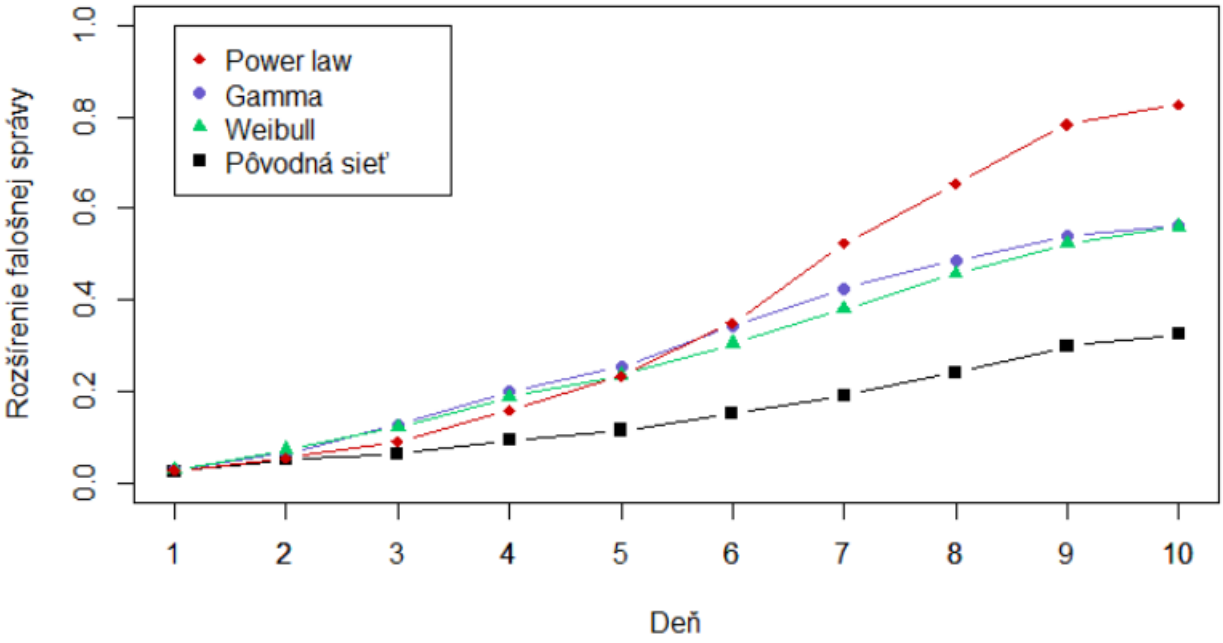
3.3 Simulácia pre náhodné grafy

V tejto časti zopakujeme model šírenia falošnej informácie medzi zamestnancami, pričom budeme sledovať ako tento proces prebieha pri náhodných grafoch. Inšpirujeme sa tabuľkou vyššie a zostavíme siete s distribúciou stupňa k z troch rôznych rozdelení:

Weibullovo, Gamma a Power-law. Aj keď tretie z nich nijak presne nesimulovalo stupeň pôvodného grafu, bude zaujímavé sledovať jeho správanie pri tejto simulácii. N

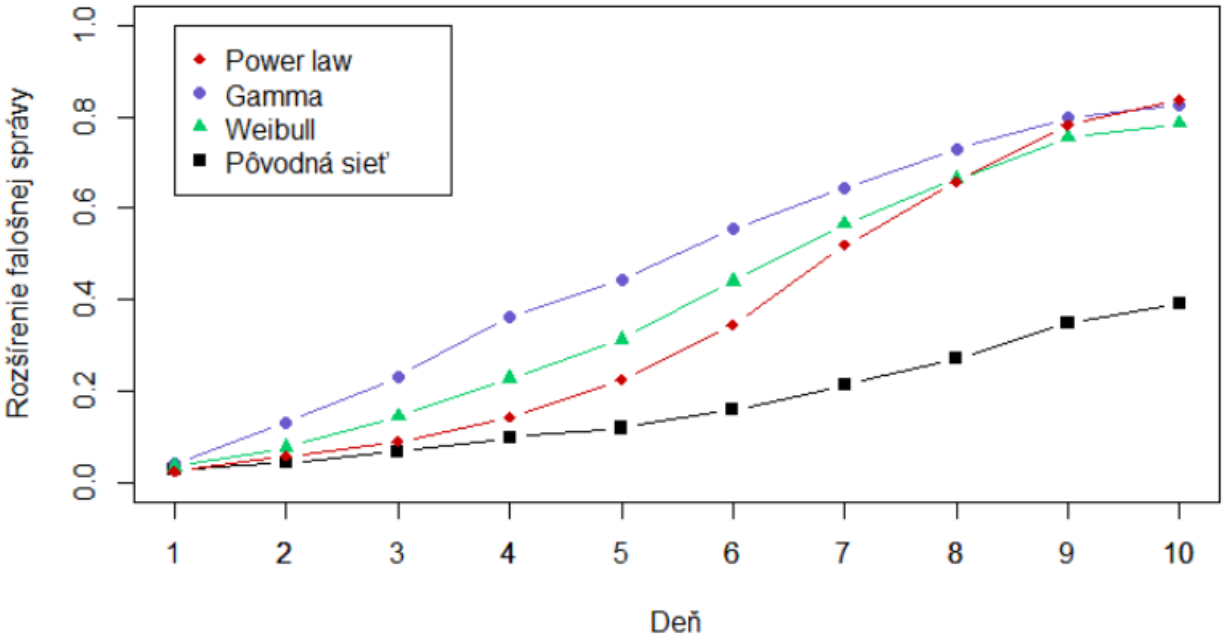
Po odhadnutí ich parametrov pre naše dáta (v tabuľke) vygenerujeme 92 hodnôt (rovnaký počet vrcholov ako v pôvodnej sieti), ktoré následne dosadíme do funkcie *sample_degseq* z knižnice *igraph* v R[4].

Dostávame náhodný graf, ktorého stupne sú zo žiadaného rozdelenia. V tomto grafe prevedieme rovnakú simuláciu ako na pôvodných dátach, pričom budeme sledovať postupný vývoj šírenia na 10 rôznych náhodných sieťach - v každej z nich 15 simulácii s náhodnými počiatočnými šíriteľmi pre každé rozdelenie. Potom rovnaký postup zopakujeme, ale miesto náhodných bodov zvolíme 15 najdôležitejších - teda vrcholov s najvyššími hodnotami stredovej medzipolohy.



Obr. 21: Podiel rozšírenosti pre skúmané rozdelenia k, - náhodný počiatočný šíriteľ

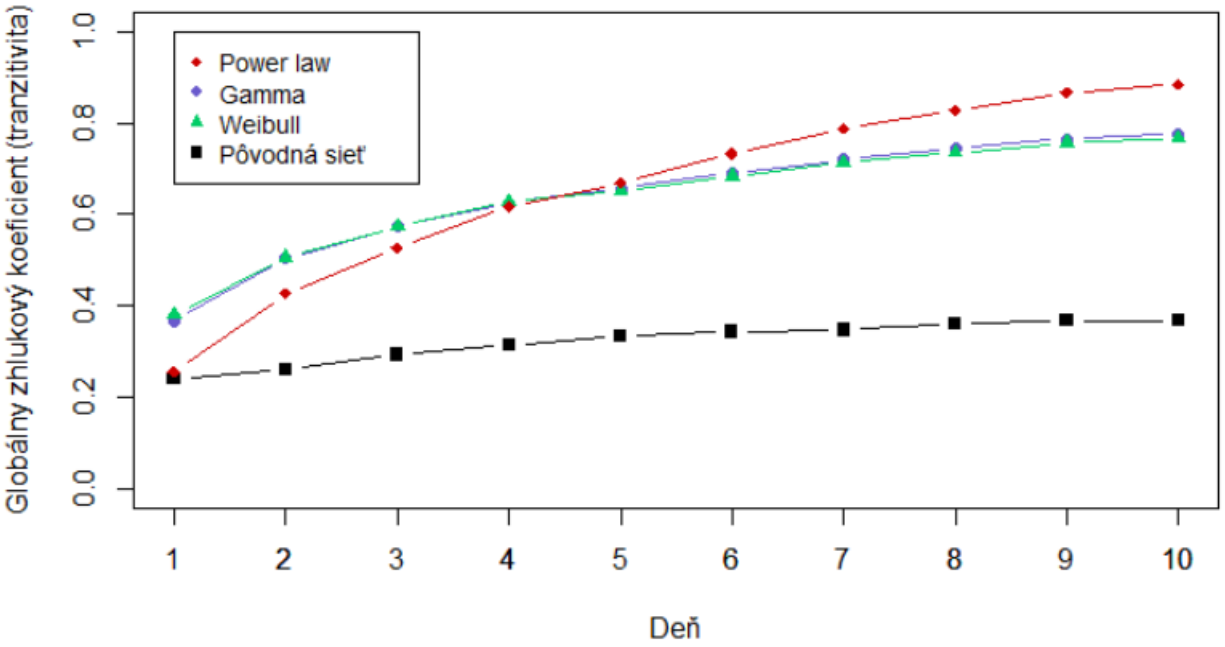
[4]Hodnoty je nutné zaokrúhliť a zabezpečiť párnosť ich súčtu - v neorientovanom grafe je súčet stupňov vždy párny. Tieto opatrenia rozdelenie významne nenarušia.



Obr. 22: Podiel rozšírenosti pre skúmané rozdelenia k, - dôležitý počiatočný šíriteľ

V obr. 21 a 22 vidno, že pre náhodný výber počiatočného šíriteľa sa ani jednému z rozdelení nedarí kopírovať model šírenia dosiahnutý v pôvodných dátach. Už vo 4. dni vidíme náskok náhodných grafov pred skutočnými dátami, ktorý sa vcelku lineárne zväčšuje až do konca skúmanej doby. Tiež sledujeme, že grafy s k z power-law distribúcie nedosiahli takmer nijaký rozdiel pri zmene spôsobu výberu počiatočného šíriteľa informácie - toto nasvedčuje, že stredová medzipoloha týchto grafov je rozdelená rovnomernejšie ako pri ostatných dvoch náhodných rozdeleniach, v ktorých mala táto zmena výrazný dopad na výsledné rozšírenie. Aj vďaka tomu sú pri 2. grafe všetky 3 simulované rozdelenia v okruhu 80%, zatiaľ čo empirické dáta sú falošnou správou infikované iba na 32%.

Táto analýza naznačuje, že grafy zo zvolených rozdelení sú prepojené užšie ako pôvodná sieť. O tomto svedčí aj tranzitivita, teda globálny zhlukový koeficient z obr. 23. Tu vidíme postupný rast tranzitivity, teda prepojenosti trojuholníkov v grafoch jednotlivých rozdelení. Opäť sledujeme výrazný rozdiel medzi empirickými dátami a náhodnými grafmi. Postupným pridávaním hrán rastie aj rozdiel v tranzitivite pomerne výrazne (narozdiel od pôvodnej siete).



Obr. 23: Podiel rozšírenosti pre skúmané rozdelenia k, - dôležitý počiatočný šíriteľ

Po porovnaní modelov šírenia informácie a tranizitivity sa zdá, že sú tieto hodnoty značne korelované. Pearsonov test ([7]) túto domnienku potvrdzuje, korelácia rozšírenosti informácie a tranzitivty v sieti je 82.5% s p-hodnotou takmer rovnou 0, vďaka čomu môžme vyvodiť, že tieto modely sú do istej miery korelované.

Usudzujeme, že skúmané náhodné grafy nedokážu spoľahlivo modelovať šírenie informácii z našich pôvodných dát. Za hlavné dôvody považujeme veľkú špecifickosť našich hrán, keďže niektoré dvojice zamestnancov mohli byť v častom kontakte, čo tieto zjednodušené modely nemohli spoľahlivo simulovať, rovnako ako faktor oddelení. Následkom bola väčšia prepojenosť vrcholov náhodných grafov a tranzitivita, ktoré mali za následok rýchlejší prenos falošnej správy ako tomu bolo v simulácii pre empirické dáta stretnutí zamestnancov. Za faktorý prispievajúce k neúspechu pri hľadaní vhodného modelu považujeme neortodoxnosť dát (multigraf s veľkým počtom hrán) a zjednodušený model šírenia. Napriek tomu veríme, že tento rozbor môže slúžiť ako zaujímavá ukážka analýzy postupne sa vyvíjajúcej siete a zmien jej vlastností v čase.

Záver

V tejto práci sme sa zamerali na problematiku skúmaní sietí a komplikácie pri ich analýze, ktoré boli donedávna považované za nepodstatné. Zameranie tejto práce bolo na postupné prenikanie hlbšie do analýz teórie sietí.

V prvej kapitole boli stručne ale predstavené grafy a siete ako koncept, terminológia s nimi spojená, a boli položené teoretické základy pre zvyšok práce. Tá sa v druhej kapitole zaoberala najprv priblížením a neskôr aj implementáciou štatistických metód. Tiež v nej boli predstavené 2 veľmi rozdielne siete, na ktorých sme tieto rôzne metódy aplikovali. Aj keď pri prvej z týchto sietí nebolo modelovanie pomocou Erdős–Rényiho a Barabási-Albertovho modelu úspešné, uvedomili sme si dôvod (veľmi špecifické pôvodné dáta) a dokázali sme v nej nájsť iné zákonitosti.

Ďalším výhodou tejto práce je ukážka všestrannosti knižnice *igraph* v R, pomocou ktorej boli takmer všetky grafy v tejto práci zobrazované. Táto knižnica veľmi uľahčila našu prácu s dátami.

V tretej kapitole sme sa zaoberali jedinou sieťou zamestnancov a ich stretávaniu sa na chodbách. Vymysleli sme vlastný (zjednodušený) model šírenia informácie a skúmali, aké majú rôzne vlastnosti pôvodnej, a neskôr aj generovaných, sietí vplyv na rýchlosť šírenia prípadnej falošnej správy. Aj keď sa nám nepodarilo úplne splniť naše pôvodné očakávanie nájsť model, ktorý by ich presne vystihoval, tak považujeme túto prácu za prínos v podobe predstavenia zaujímavých metód, ktoré sa pri analyzovaní sietí zo skutočného sveta dajú využiť.

Zoznam použitej literatúry

[1] BARABÁSI, A. L.: *Network Science*, 2016, Cambridge University Press (2016).

[2] BARABÁSI, A. L., ALBERT, R.: *Emergence of Scaling in Random Networks*, Science 286 (1999), 509-512.

[3] CLAUSET, A., SHALIZI, C. R., NEWMAN, M. E. J.: *Power-Law Distributions in Empirical Data*, SIAM Review, Volume 51, Issue 4 (2009), Dostupné na internete (7.5.2018): `https://arxiv.org/pdf/0706.1062.pdf`

[4] CLAUSET, A., YOUNG, M., GLEDITSCH, K., S.: *The Developmental Dynamics of Terrorist Organizations*, Journal of Conflict Resolution 51, 58 (2012), Dostupné na internete (8.5.2018): `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0048633`

[5] GÉNOIS, M. et al.: *Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers*, Network Science 3, 326 (2015), Dostupné na internete (9.5.2018): `https://doi.org/10.1017/nws.2015.10`

[6] HAZEWINKEL, M. (ed.):*Encyclopaedia of Mathematics*, Supplement III, Kluwer, (2002).

[7] JANKOVÁ, K., PÁZMAN, A.: *Pravdepodobnosť a štatistika*, Univerzita Komenského, Bratislava, 2011.

[8] MOORE, E.F.: *The Shortest Path Through a Maze*, Bell Telephone System 3523 (1959).

[9] NEWMAN, M. E. J.: *Power Laws, Pareto Distributions and Zipf's Law*, Arxiv (2006).

[10] ONDREJÁKOVÁ, Z.: *Centralita vrcholov v sociálnej sieti*, FMFI UK (2017), Dostupné na internete (10.5.2018): `http://www.iam.fmph.uniba.sk/institute/stehlikova/BC/2017-ondrejakova.pdf`

[11] SUNDARESAN, S. R., et al.: *Defining and Evaluating Network Communities based on Ground-truth*, ICDM (2012), Dostupné na internete (9.5.2018): `https://arxiv.org/abs/1205.6233v3`

[12] TAKAC, L., ZABOVSKY, M.: *Data Analysis in Public Social Networks*, International Scientific Conference (2012), Dostupné na internete (7.5.2018): `http://snap.stanford.edu/data/soc-pokec.pdf`

[13] TAYLOR, P.: *Whatever Happened to Those Bridges?*, Australian Mathematics Trust (2000), Dostupné na internete (10.1.2018): `https://web.archive.org/web/20120319074335/http://www.amt.canberra.edu.au/koenigs.html`

[14] YANG, J., LESKOVEC, J.: *Defining and Evaluating Network Communities based on Ground-truth*, ICDM (2012), Dostupné na internete (9.5.2018): `https://arxiv.org/abs/1205.6233v3`