

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



AKO VZNIKAJÚ KONTAKTY MEDZI ĽUĎMI -
MODELOVANIE VZNIKU SOCIÁLNYCH SIETÍ

BAKALÁRSKA PRÁCA

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

**AKO VZNIKAJÚ KONTAKTY MEDZI ĽUĎMI -
MODELOVANIE VZNIKU SOCIÁLNYCH SIETÍ**

BAKALÁRSKA PRÁCA

Študijný program: Ekonomická a finančná matematika
Študijný odbor: 1114 Aplikovaná matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Vedúci práce: doc. RNDr. Beáta Stehlíková, PhD.



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Ema Löfflerová
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Ako vznikajú kontakty medzi ľuďmi - modelovanie vzniku sociálnych sietí
How are relations between people formed - modelling social networks formation

Cieľ: Sieť pozostáva z vrcholov grafu (ľudia) a hrán (dva vrcholy môžu byť spojené hranou, ktorá reprezentuje nejaký definovaný vzťah). Zaujímavou otázkou je spôsob, akým takáto sieť vzniká, teda ako postupne vznikajú hrany medzi vrcholmi. Základnými modelmi sú random network a preferential attachment. Tieto dve možnosti sa dajú kombinovať, pričom je daná váha jednotlivým spôsobom vzniku hrán. Pre takúto sieť sa dá odvodiť pravdepodobnostné rozdelenie počtu hrán, ktoré budú z vrcholov vychádzať. Na základe toho sa dajú skúmať reálne siete a môžeme sa snažiť zistiť, akým spôsobom v nich hrany vznikali. V knihe [1] je uvedený postup a ilustrovaný je na niekoľkých príkladoch.

Štruktúra práce:

- (a) Naštudovať teóriu o sieťach a ich vzniku, ilustrovať uvedené modely na vygenerovaných príkladoch, odvodiť rozdelenie počtu hrán, uviesť algoritmus odhadovania parametra pri zmiešanom modeli.
- (b) Zreprodukovať príklady použitia tohto algoritmu uvedené v [1].
- (c) Testovať algoritmus a jeho konvergenciu na simulovaných dátach.
- (d) Použiť algoritmus na niekoľkých (2-3) vlastných príkladoch pracujúcich s reálnymi a originálnymi dátami.

Literatúra: [1] M. O. Jackson, Social and Economic Networks. Princeton University Press, 2008.

Vedúci: RNDr. Beáta Stehlíková, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.
Dátum zadania: 20.10.2015

Dátum schválenia: 19.11.2015

doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Podakovanie Chcela by som sa touto cestou poďakovať svojej vedúcej bakalárskej práce doc. RNDr. Beáte Stehlíkovej, PhD, za ochotu, pomoc, odborné rady a podnetné pripomienky, ktoré mi pomohli pri písaní tejto práce. Ďakujem svojim spolužiakom, za pomoc pri zbieraní dát a taktiež ďakujem mojej rodine a priateľom za trpezlivosť a podporu.

Abstrakt

LÖFFLEROVÁ, Ema: Ako vznikajú kontakty medzi ľuďmi - modelovanie vzniku sociálnych sietí [Bakalárska práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Beáta Stehlíková, PhD, Bratislava, 2016, 50s.

Táto práca sa zaoberá modelmi opisujúcimi vznik sietí: rastúce náhodné siete, preferential attachment a hybridný model. Kombináciou prvých dvoch modelov dostaneme hybridný model, v ktorom vystupuje parameter určujúci aká časť hrán je tvorená prvým spôsobom a aká časť druhým. Uvedieme algoritmus na odhadovanie tohto parametra a spustíme ho na dvoch prebratých a dvoch vlastných príkladoch. Na základe tejto odhadnutej hodnoty zisťujeme ako dané siete vznikli. Nakoniec sa zameriame na testovanie algoritmu na simulovaných dátach, v akých prípadoch skonverguje a či sa tento odhad parametra zhoduje s hodnotou, pomocou ktorej boli tieto siete vygenerované.

Kľúčové slová: graf, rastúce náhodné siete, preferential attachment, hybridný model, odhadovanie parametra

Abstract

LÖFFLEROVÁ, Ema: How are relations between people formed - modelling social networks formation [Bachelor Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Beáta Stehlíková, PhD, Bratislava, 2016, 50p.

This thesis deals with models that describe network formation: growing random network, preferential attachment and a hybrid model. Combination of the first two models create hybrid model in which parameter determine fraction of edges formed by the first way. We introduce an algorithm for estimating this parameter and we apply it on two taken over examples and our own examples. According to this estimated value we study how the network was created. Finally, we focus on testing the algorithm on simulated data, in which cases it converges and whether this estimate matches the parameter value by which the network was generated.

Keywords: graph, growing random network, preferential attachment, hybrid model, parameter estimation

Obsah

Úvod	8
1 Základné pojmy	10
1.1 Erdős-Rényi náhodný graf	13
1.2 Vznik sietí	17
1.2.1 Rastúce náhodné siete	17
1.2.2 Preferential attachment	20
1.2.3 Hybridné modely	23
2 Odhadovanie parametra	25
2.1 Príklady použitia algoritmu	27
2.1.1 Spoluautorstvo	27
2.1.2 Citácie	30
3 Aplikácia algoritmu - vlastné príklady	33
3.1 EFM 2013-2016	33
3.2 Emailová komunikácia	35
4 Testovanie algoritmu	38
Záver	42
Literatúra	44
Príloha A	46

Úvod

Siete vyjadrujú štruktúru vzťahov medzi jednotlivými skúmanými objektmi. Tento definovaný vzťah môže predstavovať medziludské či pracovné vzťahy, ale taktiež môže ísť o určitú ekonomickú interpretáciu (preprava tovaru medzi mestami, štátne dlhy). Dôvody na skúmanie sietí sú rôzne. Vieme vďaka nim sledovať a lepšie chápať šírenie informácií medzi ľuďmi či fungovanie jednotlivých trhov. Môžeme sa zamýšľať nad tým, či štruktúra siete ovplyvňuje správanie sa jednotlivých objektov a naopak, či ich správanie ovplyvňuje jej štruktúru.

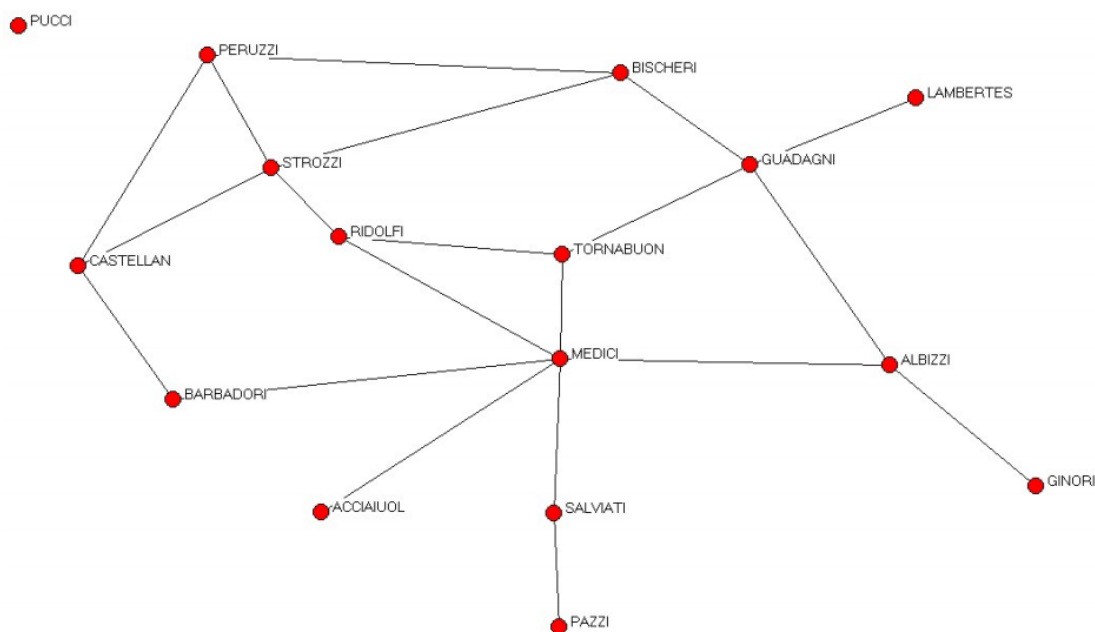
Jednou z oblastí, ktorou sa možno zaoberať je vznik týchto sietí. Existujú rôzne modely, ktoré popisujú vznik hrán medzi jednotlivými vrcholmi, ako napríklad náhodný rastúci graf, *preferential attachment* či zmiešaný model. Pre sieť potom dokážeme odvodiť pravdepodobnostné rozdelenie počtu hrán, ktoré z vrcholov vychádzajú. Pomocou toho môžeme skúmať spôsob akým vznikali reálne siete.

Cieľom tejto bakalárskej práce je naštudovať teóriu o sieťach a ich vzniku. Predstavíme čitateľovi jednotlivé modely a podrobnejšie vysvetlíme algoritmus odhadovania parametra pri hybridnom modeli. Tento algoritmus budeme postupne testovať na príkladoch uvedených v [5], príkladoch s vlastnými dátami a pozrieme sa aj na jeho konvergenciu. Táto práca je rozdelená na 4 kapitoly. V prvej kapitole si predstavíme základné pojmy týkajúce sa sietí, ktoré budeme v práci používať. Obsah tejto kapitoly sleduje výklad videoprednášok, ktoré boli súčasťou online kurzu zameraného na sociálne a ekonomické siete [7], ktoré priamo vychádzajú z knihy [8]. Ku koncu kapitoly prejdeme k trom modelom opisujúcim vznik sietí a pri každom si uvedieme tvar distribučnej funkcie rozdelenia stupňa vrcholov. V nasledujúcej kapitole sa zameriame na hybridný model, konkrétne na odhadovanie parametra, ktorý v ňom vystupuje. Algoritmus zostrojíme podľa [5] a spustíme ho na dvoch príkladoch z tohto článku. Tretia kapitola sa sústre-

duje na aplikáciu algoritmu na dvoch vlastných príkladoch. Prvá sieť bude reprezentovať študentov EFM a druhá emailovú komunikáciu zamestnancov jednej inštitúcie. V poslednej kapitole sa zameriame na konvergenciu tohto algoritmu. Spustíme ho na simulovaných dátach a budeme zisťovať za akých podmienok dokonverguje k správnej hodnote.

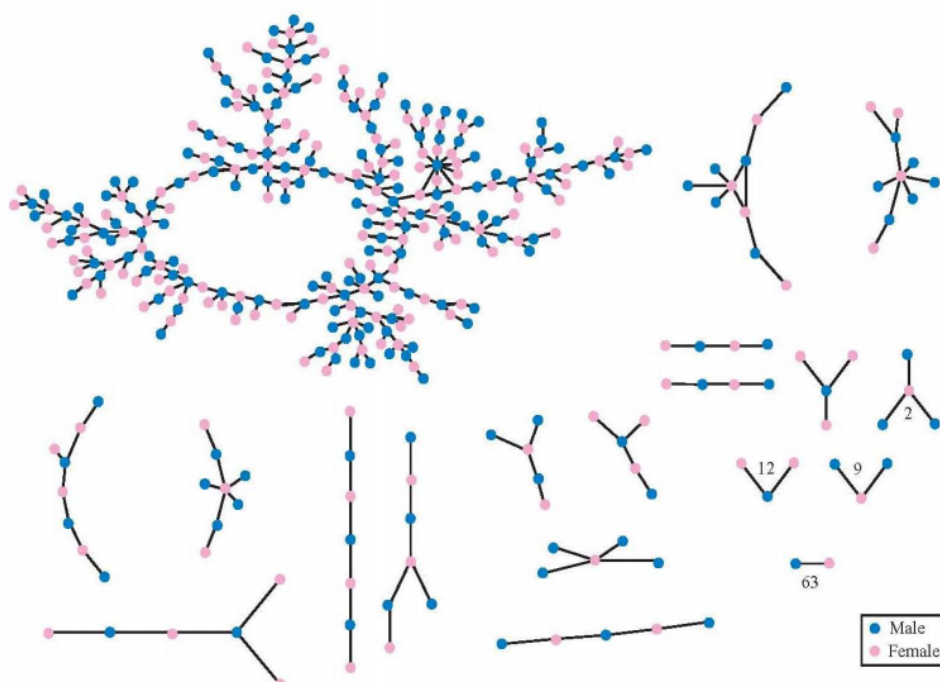
Základné pojmy

Ako prvý príklad si uvedieme manželstvá medzi florentskými rodinami začiatkom 15. storočia, ktoré môžeme vidieť na Obrázku 12. V článku [12] sa Padgett a Ansell zaoberali tým, prečo práve rodina Mediciovcov zaujala prvé miesto medzi rodinami vo Florencii aj napriek tomu, že ostatné rodiny boli bohatšie a mali väčší politický vplyv.



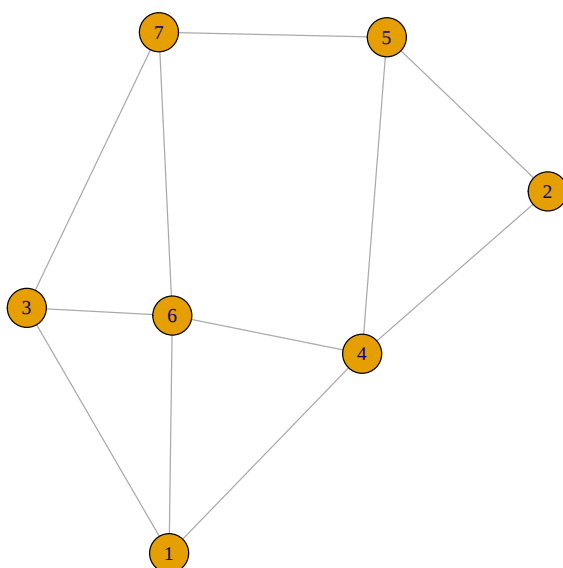
Obr. 1.1: Manželstvá vo Florencii začiatkom 15. storočia 12

V druhom príklade ide o romantické vzťahy študentov na škole Jefferson High School [2]. Graf na Obrázku 1.2 bol zostrojený na základe prieskumu National Institute of Child Health and Human Development, v ktorom požadovali od študentov, aby uviedli všetky svoje romantické známosti počas posledných šiestich mesiacov. Na základe tohto prieskumu sa snažili analyzovať prenos chorôb medzi študentmi tejto strednej školy.



Obr. 1.2: Partnerské vzťahy medzi študentmi Jefferson High School 2

V tejto kapitole sa budeme zaoberať základnými charakteristikami sietí, poskytujúcimi informácie o sieťach, ako napríklad stupne vrcholov, dĺžky ciest..., ktoré nám pomôžu pri ich porovnávaní. Pomocou týchto údajov dokážeme sledovať ako sa šíria informácie medzi ľuďmi, ako sieť ovplyvňuje správanie jednotlivých subjektov, či je normálne, že vznikajú určité zoskupenia v rámci sietí a budeme sa snažiť zistiť akým spôsobom siete vznikali. Zavedieme základné pojmy, pri ktorých sme ako zdroj použili knihu [8]. Pre lepšie porozumenie ich budeme ilustrovať na grafe na obr. 12 a obr. 1.3. Slovenskú terminológiu sme prebrali z [15].



Obr. 1.3: Graf

Vrcholom grafu nazveme naše objekty, ktoré tvoria sieť a medzi ktorými sa vzťahy vytvárajú. Na Obrázku 1.3 to sú body 1-7 a v druhom grafe sú to talianske rodiny. Množinu týchto vrcholov budeme označovať $N = \{1, \dots, n\}$. Pod pojmom *hrana* budeme rozumieť vzťah medzi dvoma vrcholmi, napr. vrcholy 1 a 4 sú spojené hranou. V jednotlivých sieťach rozlišujeme hrany s váhami, čo môže vyjadrovať intenzitu daného vzťahu, a hrany bez váh, v tomto prípade získavame len informáciu či daná hrana existuje (0 alebo 1). Ďalej hrany rozdeľujeme na orientované a neorientované, čo znamená, že môžeme mať nesymetrické vzťahy.

Informáciu o počte vrcholov a hrán vieme zaznamenať do takzvanej *matice susednosti*, ktorú budeme značiť g . Ak je zložka $g_{ij} = 1$, znamená to, že z vrcholu i ide hrana do vrcholu j . V prípade, že hrany v sieti sú neorientované, matica g bude symetrická. Pre náš graf má matica g tvar 1.1.

$$g = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix} \quad (1.1)$$

Sledom z i_1 do i_k nazveme postupnosť vrcholov $(i_1, \dots, i_k)$ a postupnosť hrán $(i_1i_2, \dots, i_{k-1}i_k)$ takých, že $i_{j-1}i_j$ patrí do g pre všetky j . Ak sa vrcholy v danom slede neopakujú budeme hovoriť o *cestě*. Postupnosť vrcholov $(1,4,6,7,3,6)$ je sled, ale nie cesta, pretože vrchol 6 sa opakuje.

Graf nazývame *súvislým*, ak medzi každými dvomi vrcholmi existuje cesta. Z Obrázku 1.3 vidíme, že tento graf je súvislý, ale graf na Obrázku 12 súvislý nie je, pretože rodina Pucciovcov nemá žiadnu hranu, tým pádom neexistuje cesta medzi nimi a ostatnými rodinami. Jednotlivé grafy môžu byť rozdelené na viacero častí - podgrafov, ktoré nie sú prepojené žiadnou hranou. Maximálny súvislý podgraf nazveme *komponent*.

Teraz uvedieme niektoré základné charakteristiky grafu. *Priemer grafu* bude predstavovať maximálnu dĺžku najkratšej cesty nachádzajúcej sa v grafe. Podobný pojem je *priemerná dĺžka cesty*, ktorý označuje priemernú najkratšiu cestu. Priemer nášho grafu je 3 a priemerná dĺžka cesty je 1,524.

Množinu indexov vrcholov, ktoré majú spoločnú hranu s vrcholom i nazveme *susedstvo* a budeme ju značiť $N_i(g)$. Potom stupeň vrcholu i je rovný mohutnosti množiny $N_i(g)$, označenie bude d_i . V našom grafe máme najvyšší stupeň $d_4 = d_6 = 4$ a najnižší stupeň $d_2 = 2$.

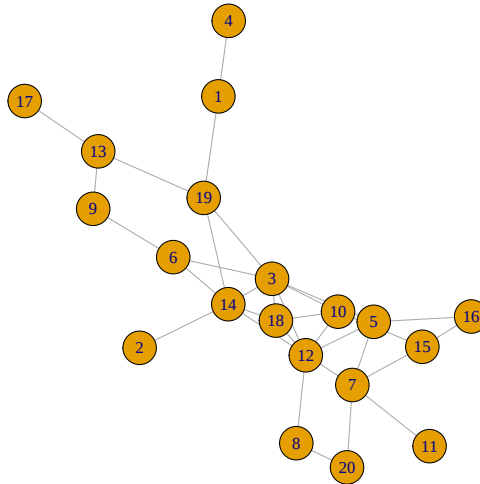
1.1 Erdős-Rényi náhodný graf

Dôležitý pojem, ktorý budeme používať je náhodný graf. V Erdős-Rényi náhodnom grafe začíname s n vrcholmi, s tým, že každá hrana vzniká nezávisle s pravdepodob-

nosťou p . Takýto graf budeme označovať $G(n, p)$. Z definície grafu vidíme, že stupne vrcholov majú binomické rozdelenie, t.j. pravdepodobnosť, že vrchol má d hrán je:

$$P[X = d] = \frac{(n-1)!}{d!(n-d-1)!} p^d (1-p)^{n-d-1}$$

V programe R sme pomocou funkcie `erdos.renyi.game(n,p)` vygenerovali graf, kde počet vrcholov bol $n = 20$ a pravdepodobnosť $p = 0,2$, ktorý môžeme vidieť na Obrázku 1.4 a v príslušnom kóde 1. Následne funkciou `degree.distribution()` vieme zistiť rozdelenie stupňov vrcholov nášho vygenerovaného grafu. Na Obrázku 1.5 je čiernou farbou znázornené rozdelenie, ktoré nám vyšlo a modrou farbou teoretické binomické rozdelenie.

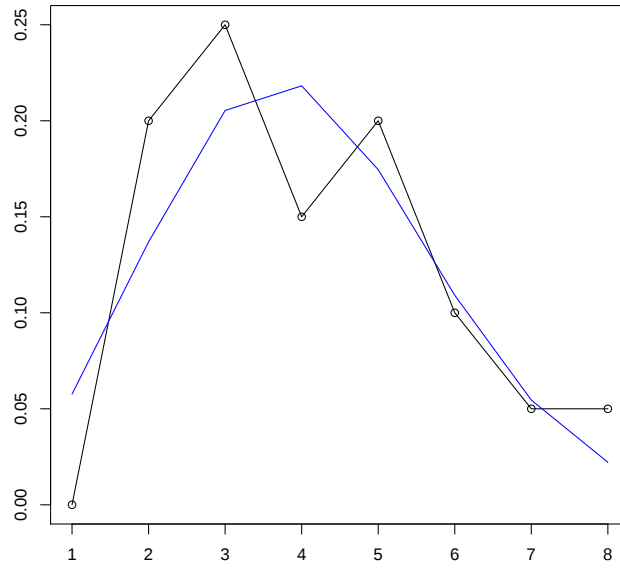


Obr. 1.4: Graf vygenerovaný `erdos.renyi.game(20, 0,2)`

Pre $n * p = \lambda$ konštantné sa dá rozdelenie počtu hrán pre $n \rightarrow \infty$ aproximovať Poissonovým rozdelením. Ide o známe odvodenie z oblasti pravdepodobnosti [9] a preto si ho tu uvedieme. Pravdepodobnosť, že vrchol má d hrán je:

$$P[X = d] = \frac{(n-1)!}{d!(n-d-1)!} p^d (1-p)^{n-d-1}. \quad (1.2)$$

Výraz $\frac{(n-1)!}{(n-1-d)!}$ môžeme prepísať na $(n-1)(n-2)\dots(n-d) = n^d + O(n^{d-1})$, kde $O(n^{d-1})$ sú mocniny $(d-1)$ -teho a nižšieho rádu.



Obr. 1.5: Porovnanie rozdelenia vrcholov Erdős-Rényiho grafu s teoretickým binomickým rozdelením

$$\lim_{n \rightarrow \infty} \frac{(n-1)(n-2) \dots (n-d)}{(n-1)^d} = \lim_{n \rightarrow \infty} \frac{n^d + O(n^{d-1})}{(n-1)^d} = 1$$

Preto môžeme limitu pre $n \rightarrow \infty$ rovnice (1.2) zapísať ako:

$$\lim_{n \rightarrow \infty} P[X = d] = \lim_{n \rightarrow \infty} \frac{((n-1)p)^d}{d!} (1-p)^{(n-1-d)} = \frac{\lambda^d}{d!} \lim_{n \rightarrow \infty} (1-p)^{(n-1-d)}.$$

p si zapíšeme ako $\frac{\lambda}{(n-1)}$ a mocninu rozdelíme na 2 časti

$$\lim_{n \rightarrow \infty} P[X = d] = \frac{\lambda^d}{d!} \lim_{n \rightarrow \infty} \left(1 + \frac{-\lambda}{(n-1)}\right)^{(n-1)} \left(1 + \frac{-\lambda}{(n-1)}\right)^{-d} \quad (1.3)$$

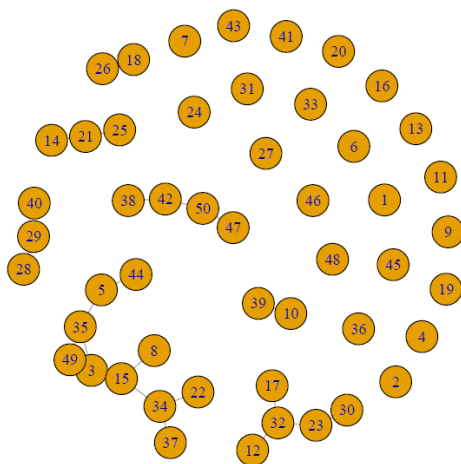
$$\lim_{n \rightarrow \infty} \left(1 + \frac{-\lambda}{(n-1)}\right)^{-(n-1)} = e^{-\lambda}$$

$$\lim_{n \rightarrow \infty} \left(1 + \frac{-\lambda}{(n-1)}\right)^{-d} = 1$$

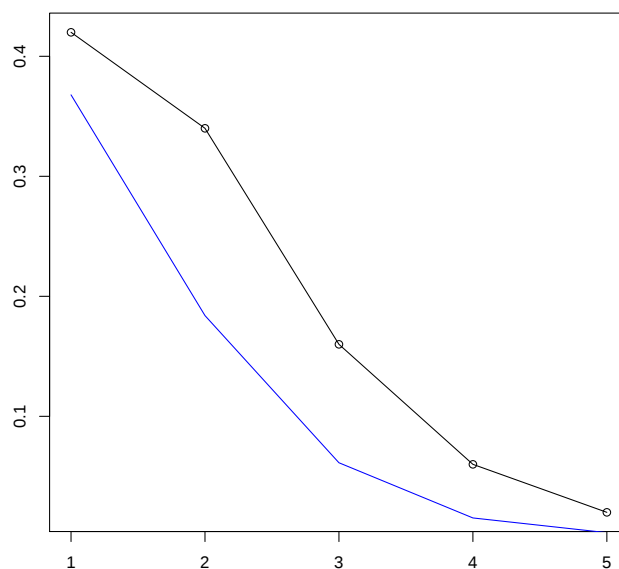
Po dosadení týchto dvoch limit do (1.3) dostávame Poissonovo rozdelenie s parametrom λ :

$$\lim_{n \rightarrow \infty} P[X = d] = \frac{\lambda^d}{d!} e^{-\lambda} \quad (1.4)$$

Na Obrázku 1.6 sa nachádza vygenerovaný graf, pre $n = 50$ a $p = 0,02$ (tj. $\lambda = 1$) a na Obrázku 1.7 porovnanie rozdelenia počtu vrcholov nami vygenerovaného grafu s teoretickým Poissonovým rozdelením.



Obr. 1.6: Graf vygenerovaný `erdos.renyi.game(50, 0,02)`



Obr. 1.7: Porovnanie rozdelenia vrcholov Erdős-Rényiho grafu s teoretickým Poissonovým rozdelením

1.2 Vznik sietí

V tejto časti predstavíme tri základné modely popisujúce vznik sietí - rastúce náhodné siete, *preferential attachment* a hybridné modely.

1.2.1 Rastúce náhodné siete

Už podľa názvu môžeme očakávať, že pri rastúcich náhodných sieťach sa budú hrany tvoriť s určitou pravdepodobnosťou. Na rozdiel od Erdős-Rényiho náhodného grafu sa v tomto modeli každú periódu pridá nový vrchol, ktorý si vytvorí hrany s ostatnými vrcholmi.

Budeme začínať s m kompletne prepojenými vrcholmi, t.j. medzi každými dvoma existuje hrana. V každom kole vznikne jeden nový vrchol, ktorý si vytvorí m väzieb s už existujúcimi bodmi s tým, že vrcholy si vyberá s rovnakou pravdepodobnosťou. To znamená, že bod má pravdepodobnosť $\frac{m}{t}$, že mu v danom kole pribudne nová hrana, kde t označuje počet existujúcich bodov a teda zároveň aj periódu.

Očakávaný stupeň $\bar{d}_i(t)$ vrchola i takého, že $m < i < t$ je

$$\bar{d}_i(t) = m + \frac{m}{i+1} + \frac{m}{i+2} + \dots + \frac{m}{t}$$

Pri svojom vzniku má každý vrchol stupeň m , t.j. $d_i(i) = m$. Vieme, že následne za každú jednotku času získa vrchol $\frac{m}{t}$ nových väzieb. Spravíme spojitú aproximáciu stupňa vrcholu a zapíšeme to do tvaru diferenciálnej rovnice (1.5) s počiatočnou podmienkou $d_i(i) = m$. Jej vyriešením získame aproximáciu očakávaného stupňa vrcholu.

$$\frac{dd_i(t)}{dt} = \frac{m}{t}, \quad (1.5)$$

$$dd_i(t) = \frac{m}{t} dt. \quad (1.6)$$

Integrovaním oboch strán dostaneme:

$$\int dd_i(t) = \int \frac{m}{t} dt.$$

Ďalej vypočítame jednotlivé neurčité integrály, pričom konštantu si zapíšeme v tvare aký nám vyšiel pri riešení druhého neurčitého integrálu:

$$\begin{aligned}\int dd_i(t) &= d_i(t) + \ln(c), \\ m \int \frac{1}{t} dt &= m \ln(t).\end{aligned}$$

Po dosadení do rovnosti máme:

$$\begin{aligned}d_i(t) + \ln(c) &= m \ln(t), \\ d_i(t) &= m \ln(t) - \ln(c).\end{aligned}\tag{1.7}$$

Použitím počiatočnej podmienky si môžeme vyjadriť $\ln(c)$:

$$\begin{aligned}m = d_i(i) &= m \ln(i) - \ln(c), \\ \ln(c) &= m \ln(i) - m.\end{aligned}$$

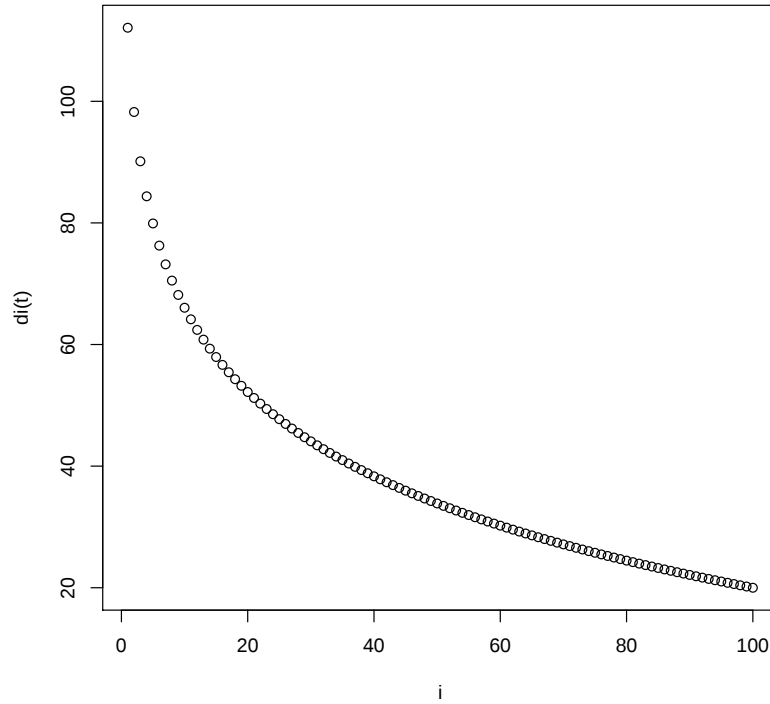
Teraz môžeme tento tvar $\ln(c)$ dosadiť do (1.7):

$$\begin{aligned}d_i(t) &= m \ln(t) - m \ln(i) + m, \\ \bar{d}_i(t) &= m \left(1 + \ln \left(\frac{t}{i} \right) \right).\end{aligned}\tag{1.8}$$

Na obrázku 1.8 môžeme vidieť závislosť očakávaného stupňa vrcholu $\bar{d}_i(t)$ od i (perióda, v ktorej vznikol). Tvar krivky si môžeme vysvetliť tak, že vrchol i , pre i malé, si na začiatku ľahko vytvorí hrany pretože je ešte málo už existujúcich vrcholov a zároveň tento vrchol žije veľa períód. Na druhej strane vrchol j , pre j veľké, si vytvorí m hrán pri svojom vzniku, ale zažije len málo períód počas ktorých vzniknú nové vrcholy a zároveň existuje už veľa vrcholov a preto stupeň tohto vrcholu bude blízky m .

Distribučnú funkciu $F_t(d)$ dostaneme pomocou nerovnice vyjadrujúcej, že očakávaný stupeň vrcholu $\bar{d}_i(t)$ je menší ako hodnota d :

$$m \left(1 + \ln \left(\frac{t}{i} \right) \right) < d.$$



Obr. 1.8: Závislosť očakávaného stupňa vrcholu $\bar{d}_i(t)$ od i , $t = 100, m = 20$

Následne si z tejto nerovnice vyjadríme i .

$$t \exp\left(-\frac{(d-m)}{m}\right) < i$$

Teraz vieme, že $t - t \exp\left(-\frac{(d-m)}{m}\right)$ nám označuje počet vrcholov, ktoré majú očakávaný stupeň menší ako d . Po vydelení t dostávame:

$$P[\bar{d}_i(t) \leq d] = F_t(d) = \frac{\left(t - t \exp\left(-\frac{(d-m)}{m}\right)\right)}{t} = 1 - \exp\left(-\frac{(d-m)}{m}\right) \quad (1.9)$$

Z distribučnej funkcie (1.9) vidíme, že očakávané stupne vrcholov majú exponenciálne rozdelenie.

1.2.2 Preferential attachment

Preferential attachment model sa snaží vylepšiť formovanie hrán z 1.2.1. Ak sa v reálnom svete objaví v komunite ľudí nový človek, nie všetci jej členovia majú rovnakú pravdepodobnosť, že sa s ním spriatelí. Osoba s veľa kamarátmi je asi komunikatívnejšia a ľahšie vychádza s ľuďmi. Preto má väčšiu pravdepodobnosť, že sa spriatelí aj s týmto novým členom.

V roku 1976 napísal Price článok [13], v ktorom skúmal citácie článkov autormi. Aj v tomto prípade platí, že článok, ktorý bol veľa krát citovaný má väčšiu šancu byť citovaný znova. Neskôr sa Barabási a Albert v článku [1] podrobnejšie zaoberali preferential attachment metódou a preto sa tento model niekedy nazýva aj Barabási-Albert (BA) model.

Preferencie vrcholov sa v tomto modeli odzrkadlia tak, že už existujúci vrchol nebude mať pravdepodobnosť, že si v danom kole vytvorí hranu rovnú $\frac{m}{t}$, ale bude to závisieť od jeho stupňa d_i .

Znova budeme začínať s m kompletne prepojenými vrcholmi. Každú periódu vznikne nový vrchol a vytvorí si m hrán s už existujúcimi bodmi, kde pravdepodobnosť vytvorenia hrany je úmerná stupňu vrcholu. Inak povedané pravdepodobnosť, že vrchol i , kde $i < t$, si v t -tom kole vytvorí hranu je $\frac{d_i(t)}{2tm}$, $d_i(t)$ je stupeň bodu i v t -tej perióde.

Každý vrchol pri svojom vzniku je stupňa m , to znamená, že $d_i(i) = m$ a neskôr každé jedno kolo získa $m \frac{d_i(t)}{2tm}$ väzieb. Zapišeme to ako diferenciálnu rovnicu (1.10) s počiatočnou podmienkou $d_i(i) = m$.

$$\frac{dd_i(t)}{dt} = m \frac{d_i(t)}{2tm} \quad (1.10)$$

$$\frac{1}{d_i(t)} dd_i(t) = \frac{1}{2t} dt \quad (1.11)$$

Obe strany integrujeme a vypočítame si oba neurčité integrály:

$$\begin{aligned} \int \frac{1}{d_i(t)} dd_i(t) &= \ln(d_i(t)) + \ln(c), \\ \int \frac{1}{2t} dt &= \ln(t^{\frac{1}{2}}). \end{aligned}$$

Následne dosadíme tieto výsledky do rovnice (1.11) a vyjadríme si $d_i(t)$:

$$\begin{aligned}\ln(d_i(t)) + \ln(c) &= \ln(t^{\frac{1}{2}}), \\ \ln(d_i(t)) &= \ln\left(\frac{t^{\frac{1}{2}}}{c}\right). \\ d_i(t) &= \frac{t^{\frac{1}{2}}}{c}\end{aligned}\tag{1.12}$$

Pomocou počiatkovej podmienky zistíme tvar konštanty c :

$$\begin{aligned}m = d_i(i) &= \frac{i^{\frac{1}{2}}}{c}, \\ c &= \frac{i^{\frac{1}{2}}}{m}.\end{aligned}$$

Po dosadení do (1.12) získavame aproximáciu očakávaného stupňa vrcholu:

$$d_i(t) = m \left(\frac{t}{i}\right)^{\frac{1}{2}}.\tag{1.13}$$

Rovnako ako v predchádzajúcej podkapitole si aj tu na obrázku 1.9 vykreslíme graf závislosti očakávaného stupňa vrcholu $\bar{d}_i(t)$ od i , pre $m = 20$ a $t = 100$. Vidíme, že v tomto prípade sú rozdiely medzi malými a veľkými i ešte výraznejšie. Vrcholy, ktoré vznikli na začiatku ľahko získali hrany v prvých periódach a pretože mali väčšie $\bar{d}_i(i)$ ako nové vrcholy, tak tie rozdiely sa zväčšovali. Inak povedané, pri tomto modeli sa bohatý z hľadiska počtu hrán stáva ešte bohatšími a chudobný chudobnejšími.

Odvodíme distribučnú funkciu rozdelenia počtu vrcholov pre preferential attachment. Zaujímá nás koľko vrcholov má očakávaný stupeň menší ako d .

$$m \left(\frac{t}{i}\right)^{\frac{1}{2}} < d$$

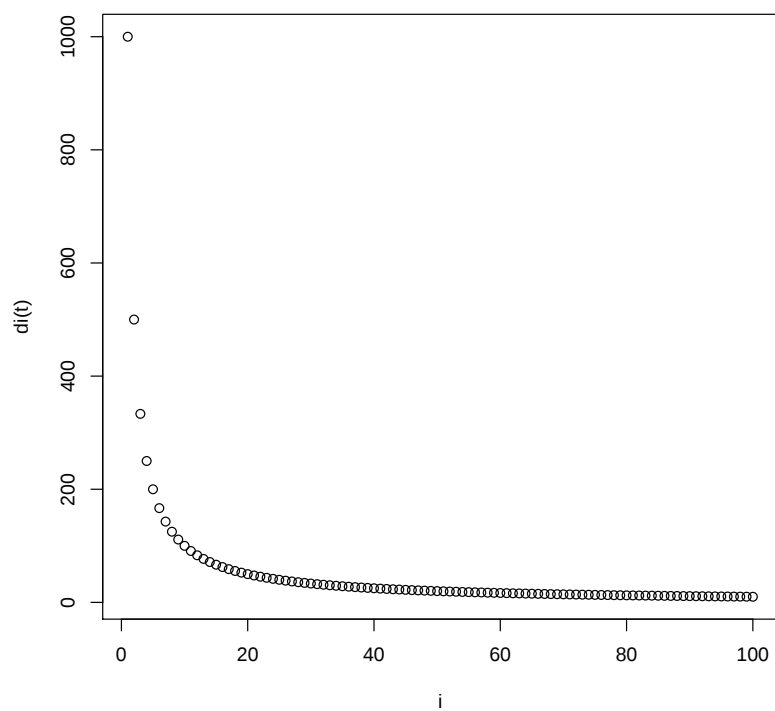
Vyjadríme si i :

$$t \left(\frac{m}{d}\right)^2 < i.$$

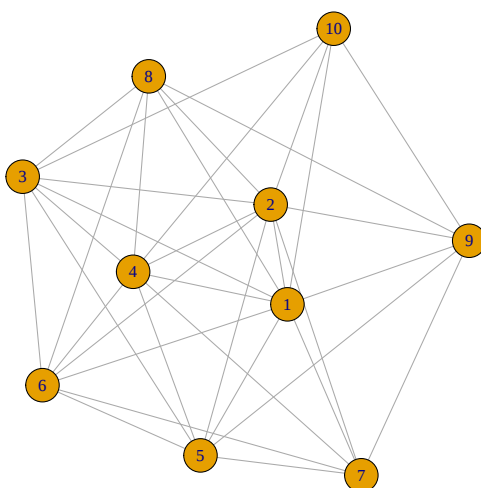
Počet vrcholov s očakávaným stupňom menším ako d je $t - t \left(\frac{m}{d}\right)^2$. Následne predeľením t dostaneme tvar distribučnej funkcie:

$$P[\bar{d}_i(t) \leq d] = F_t(d) = \frac{t - t \left(\frac{m}{d}\right)^2}{t} = 1 - \left(\frac{m}{d}\right)^2.$$

Na obrázku 1.10 máme graf vytvorený pomocou preferential attachment. Použili sme na to funkciu `barabasi.game()` v programe R, v prílohe kód 1. Graf začínal s piatimi kompletne prepojenými vrcholmi, $m = 5$ a konečný počet vrcholov je $t = 10$.



Obr. 1.9: Závislosť očakávaného stupňa vrcholu $\bar{d}_i(t)$ od i , $t = 100$, $m = 20$



Obr. 1.10: Vygenerovaný graf, `barabasi.game(n=10,power=1,m=5, directed=FALSE)`

1.2.3 Hybridné modely

Pri hybridných (zmiešaných) modeloch ide o kombináciu rastúcich náhodných sietí a preferential attachment. V [5] Jackson a Rogers predstavujú jednoduchý hybridný model. Nový vrchol si časť hrán vytvorí tak, že všetky už existujúce vrcholy majú rovnakú pravdepodobnosť vytvorenia hrany, a zvyšok hrán si bude tvoriť len s vrcholmi patriacimi do susedstiev vrcholov, s ktorými má spoločnú hranu (bude spoznávať priateľov jeho priateľov). Spôsob vytvárania hrán typu 2 je totožný s *preferential attachment*, t.j. pravdepodobnosť, že vrchol i získa hranu v t -tom kole, kde $i < t$, je $\frac{d_i(t)}{2tm}$.

Stupeň každého vrcholu je na začiatku m , preto $d_i(i) = m$. Ďalšie kolo vrchol získa $\frac{am}{t} + \frac{(1-a)d_i(t)}{2t}$, kde prvý zlomok predstavuje hrany tvorené náhodne s rovnakou pravdepodobnosťou a druhý zlomok sú hrany vytvorené preferential attachment. Môžeme si to zapísať ako diferenciálnu rovnicu (1.14) s počiatočnou podmienkou $d_i(i) = m$, kde a vyjadruje aká časť hrán je typu 1:

$$\frac{dd_i(t)}{dt} = \frac{am}{t} + \frac{(1-a)d_i(t)}{2t}, \quad (1.14)$$

$$\frac{1}{2am + (1-a)d_i(t)} dd_i(t) = \frac{1}{2t} dt. \quad (1.15)$$

Obe strany rovnice zintegrujeme:

$$\begin{aligned} \int \frac{1}{2am + (1-a)d_i(t)} dd_i(t) &= \frac{1}{(1-a)} \ln(2am + (1-a)d_i(t)) + \ln(c), \\ \int \frac{1}{2t} dt &= \frac{1}{2} \ln(t) = \ln(t)^{\frac{1}{2}} dt. \end{aligned}$$

Dosadíme to do (1.15) a vyjadríme si z toho $d_i(t)$:

$$\frac{1}{(1-a)} \ln(2am + (1-a)d_i(t)) + \ln(c) = \ln(t)^{\frac{1}{2}}, \quad (1.16)$$

$$d_i(t) = \frac{1}{(1-a)} \left((ct^{\frac{1}{2}})^{(1-a)} - 2am \right). \quad (1.17)$$

$$(1.18)$$

Teraz použijeme začiatočnú podmienku pre vyjadrenie konštanty c :

$$\begin{aligned} m = d_i(i) &= \frac{1}{(1-a)} \left((ci^{\frac{1}{2}})^{(1-a)} - 2am \right), \\ c &= \frac{1}{i^{\frac{1}{2}}} (2am + m(1-a))^{\frac{1}{(1-a)}}. \end{aligned}$$

Tento tvar našej konštanty dosadíme do (1.17):

$$\bar{d}_i(t) = \frac{1}{(1-a)} \left[\left(\frac{t}{i} \right)^{\frac{(1-a)}{2}} [2am + m(1-a)] - 2am \right] \quad (1.19)$$

$$= \left(m + \frac{2am}{1-a} \right) \left(\frac{t}{i} \right)^{\frac{1-a}{2}} - \frac{2am}{1-a}. \quad (1.20)$$

Teda aproximácia očakávaného stupňa vrcholu má tvar (1.20).

Distribučnú funkciu získame rovnakým spôsobom ako pri predchádzajúcich modeloch. Vyjadríme si vrcholy, ktoré majú očakávaný stupeň menší ako d s tým, že $x = \frac{2}{1-a}$:

$$(m + xam) \left(\frac{t}{i} \right)^{\frac{1}{x}} - xam < d.$$

Následne si z toho vyjadríme i :

$$\left(\frac{m + xam}{d + xam} \right)^x < \frac{i}{t}.$$

Z tejto nerovnice vieme, že počet vrcholov, ktoré majú očakávaný stupeň menší ako d je $t - \left(\frac{m+xam}{d+xam} \right)^x$. Z toho dostávame:

$$P[\bar{d}_i(t) \leq d] = F_t(d) = \frac{(t-i)}{t} = 1 - \left(\frac{m + xam}{d + xam} \right)^x \quad (1.21)$$

Kapitola 2

Odhadovanie parametra

Tvar distribučnej funkcie hybridného modelu sme si odvodili ako rovnicu (1.21) v predchádzajúcej kapitole z čoho po zlogaritmovaní oboch strán dostávame rovnicu (2.1).

$$\ln(1 - F(d)) = x \ln(m + xam) - x \ln(d + xam) \quad (2.1)$$

$$\ln(1 - F(d)) = c - x \ln\left(d + \frac{2am}{1-a}\right) \quad (2.2)$$

Výraz $x \ln(m + xam)$ nezávisí od stupňa vrcholu, preto si to označíme ako konštantu c a dostávame rovnicu (2.2). Hodnoty d a $F(d)$ poznáme z našich dát. Počet hrán m , ktoré si vrchol pri narodení vytvorí vypočítame z celkového počtu hrán nachádzajúcich sa v grafe, čo je $t * m$, kde t označuje celkový počet vrcholov. V prípade neorientovaného grafu je potom súčet všetkých stupňov vrcholov $2 * t * m$. Teraz vidíme, že po vydelení $2 * t * m$ počtom vrcholov t získame našu hodnotu m . Pre orientované grafy je celkový počet hrán $t * m$ zároveň aj súčet všetkých stupňov vrcholov. Počet pridávajúcich sa hrán m preto vypočítame ako celkový počet hrán vydelený počtom vrcholov.

Jediná neznáma, ktorá nám ostáva je parameter a . Vyskytuje sa v logaritme a taktiež je zahrnutý v $x = \frac{2}{1-a}$. Keďže odhadujeme parameter a , tak naším cieľom je minimalizovať rozdiel medzi pravou a ľavou stranou rovnice (2.2). Jeden zo spôsobov ako odhadnúť tento parameter je pomocou iteračného procesu. Na začiatku si zvolíme a_0 , ktoré bude vystupovať v logaritme na pravej strane:

$$\ln(1 - F(d)) = c - x \ln\left(d + \frac{2a_0m}{1-a_0}\right). \quad (2.3)$$

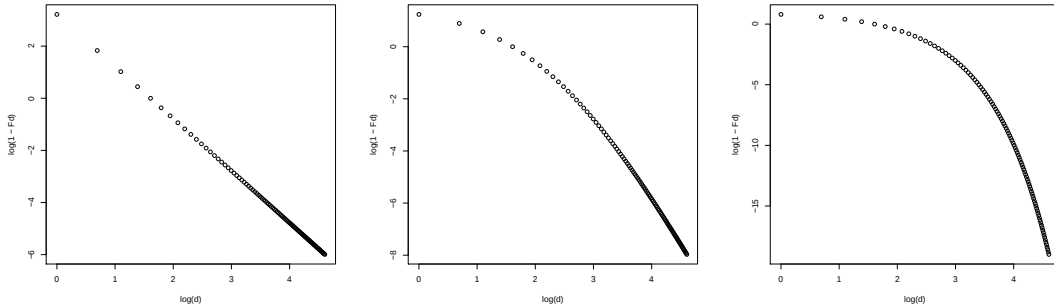
Keď spravíme lineárnu regresiu (2.5) s dvomi parametrami β_0 a β_1 , tak nové a_1 dostaneme zo vzťahu $\beta_1 = -x = -\frac{2}{1-a}$ a preto $a_1 = 1 + \frac{2}{\beta_1}$.

$$y_d = \beta_0 - \beta_1 x_d + \epsilon_d \quad d=1, \dots, \text{maximálny stupeň vrcholu} \quad (2.4)$$

$$\ln(1 - F(d)) = \beta_0 - \beta_1 \ln\left(d + \frac{2a_0 m}{1 - a_0}\right) + \epsilon_d \quad (2.5)$$

Všetky naše získané dáta, na ktorých sme testovali algoritmus boli zapísané ako hodnoty pravdepodobnostného rozdelenia stupňa vrcholov. Preto prvý krok, hneď po načítaní dát, bolo získanie hodnôt distribučnej funkcie. Po vyradení posledných dát s $F(d) = 1$, ktoré by nám spôsobovali v regresii problémy kvôli $\ln(1 - F(d))$ môžeme pristúpiť k samotnému odhadovaniu parametra a . Naš algoritmus sme testovali pre rôzne štartovacie a_0 , konkrétne pre 0, 0,5, 0,15, ..., 0,95. Vnútri tohto cyklu sa nachádzal jeden while cyklus, v ktorom sme vždy s aktuálnym a spravili regresiu podľa rovnice (2.5). Následne sme si pomocou druhého parametra regresie vypočítali nové a . While cyklus sa zastaví, keď rozdiel starého a nového parametra a je menej ako 0,001 alebo ak sa parameter dostane mimo intervalu (0,1). Kód tohto algoritmu sa nachádza v príloha ako kód 2.

Na obrázku 2.1 vidíme 3 grafy závislosti $\ln(1 - F(d))$ od $\ln(d)$ zľava pre $a = 0$, $a = 0,5$ a pre $a = 1$. Tieto hodnoty sme vypočítali pomocou rovnice (1.21) s tým, že m sme zvolili 5 a počet vrcholov 100. Vďaka týmto grafom budeme mať pri uvedení príkladov sietí predstavu o tom, aký by nám mal vychádzať parameter a .



Obr. 2.1: Porovnanie závislosti $\ln(1-F(d))$ od $\ln(d)$ pre $a=0$, $a=0.5$, $a=1$

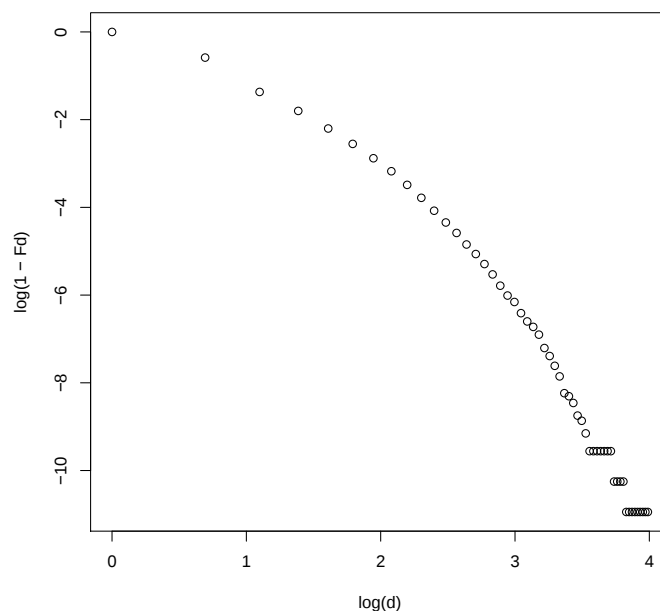
2.1 Príklady použitia algoritmu

V článku [5] počítali Jackson a Rogers parameter a pre 6 rôznych sietí. V našej práci tento algoritmus otestujeme na dvoch príkladoch, na súbore dát spoluautorstvá a citácie, ktoré sme získali z [6].

2.1.1 Spoluautorstvo

Dáta o spoluautorstvách pochádzajú z článku [4], v ktorom autori Goyal, van der Leij a Moraga-González analyzujú vývoj izolovanosti, alebo naopak prepojenosti ekonómov. Táto sieť mapuje interakcie medzi ekonómami v rokoch 1970-2000. Vrcholy siete sú reprezentované jednotlivými publikujúcimi autormi a hrany predstavujú ich spoluprácu, teda či v priebehu sledovaných tridsiatich rokov publikovali spolu článok.

Sieť obsahujúca tieto dáta je tvorená 56639 vrcholmi s tým, že $m = 1,198$ je polovica priemerného stupňa vrcholov, keďže je to neorientovaný graf. Na obrázku 2.2 môžeme vidieť graf závislosti $\ln(1 - F(d))$ od $\ln(d)$.



Obr. 2.2: Závislosť $\ln(1 - F(d))$ od $\ln(d)$ pre sieť spoluautorstiev

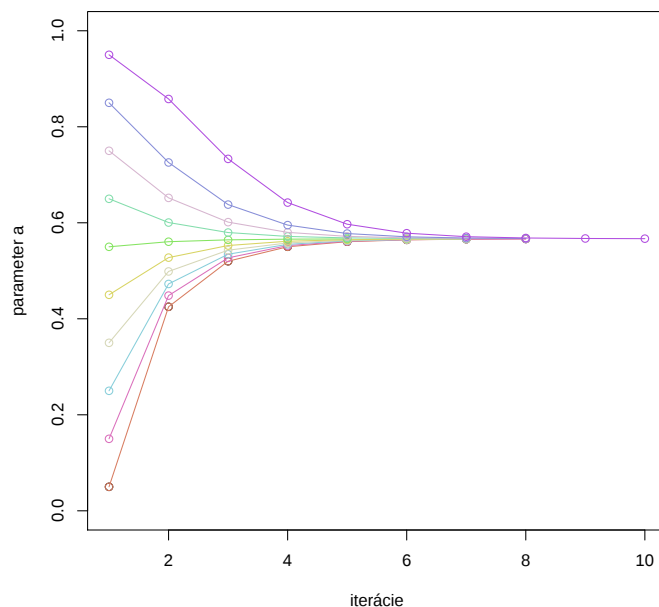
V tabuľke 2.1 sa v stĺpci a_0 nachádzajú hodnoty štartovacieho bodu, v ktorom sme začínali náš algoritmus. Vo vedľajšom stĺpci sú hodnoty parametra a , ku ktorým dokonvergoval cyklus a v poslednom stĺpci je zaznamenaný počet iterácií, ktorý bol na to potrebný. Odhad parametra a nám vyšiel 0,566, čo sa zhoduje s hodnotou uvádzanou v [7]. Algoritmus konverguje aj pre štartovacie hodnoty, ktoré boli vzdialenejšie od 0,566 a vo všetkých prípadoch to trvalo menej ako 10 iterácií.

	a_0	a	iterácie
1	0,05	0,566	7
2	0,15	0,566	7
3	0,25	0,566	6
4	0,35	0,566	6
5	0,45	0,566	6
6	0,55	0,566	4
7	0,65	0,567	6
8	0,75	0,567	7
9	0,85	0,567	7
10	0,95	0,567	9

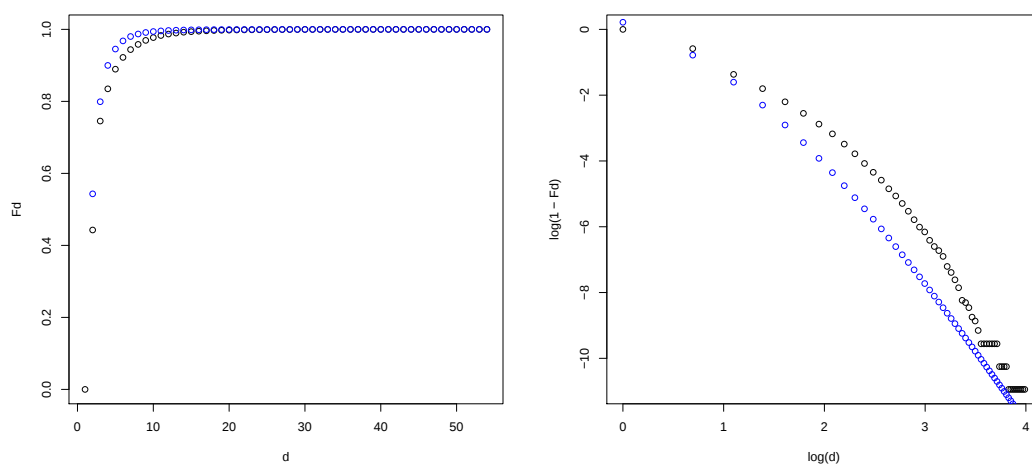
Tabuľka 2.1: Tabuľka štartovacích a_0 , odhadnutých hodnôt parametra a a počtu iterácií

Na obrázku 2.3 môžeme vidieť graficky znázornené hodnoty parametra a počas iterácii pre rôzne štartovacie a_0 . Rovnako ako v tabuľke 2.1 aj tu vidíme, že najviac iterácií bolo potrebných pri štarte z $a_0 = 0,95$ a najmenej keď sme začínali z $a_0 = 0,55$. Na prvom obrázku 2.4 vidíme porovnanie distribučných funkcií dvoch sietí, prvá je zo získaných dát a druhá je jej naše najlepšie odhadnutie. Môžeme si všimnúť, že už od hodnoty 20 sú prakticky identické. Na druhom obrázku sa nachádza porovnanie závislosti $\ln(1 - F(d))$ od $\ln(d)$ pre našu sieť s reálnymi dátami a pre simulované dáta.

Pre odhadnutý parameter má platíť, že je to pevný bod funkcie 2.2, t.j. po dosadení tejto hodnoty a do regresie máme po upravení koeficientu opäť dostať našu odhadnutú hodnotu. Na obrázku 2.5 sú čiernou farbou zobrazené hodnoty parametra a , ktoré vyšli



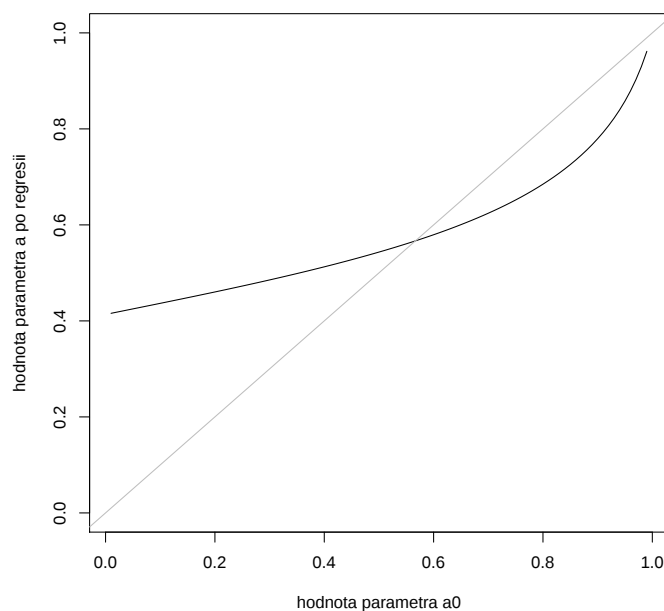
Obr. 2.3: Hodnoty parametra a počas iterácii pro $a_0 = 0.05, 0.15, \dots, 0.95$



Obr. 2.4: (i) Porovnanie distribučnej funkcie spoluautorstiev s distribučnou funkciou (1.21) pre $a = 0.566, m = 1, 198$. (ii) Porovnanie závislosti $\ln(1 - F(d))$ od $\ln(d)$ pre naše dáta a pre vypočítanú distribučnú funkciu, kde $a = 0.566, m = 1, 198$. Čierna farba - reálne dáta, modrá farba - teoretické hodnoty.

po dosadení do regresie a šedou farbou je zvýraznená priamka $y = x$. Priesečník týchto dvoch kriviek označuje náš pevný bod, ktorý nám vyšiel $a = 0,567$. Vidíme, že táto

hodnota sa líši od nášho odhadu parametra len o 0,001 a tak vieme, že náš algoritmus dokonvergoval k správnej hodnote.



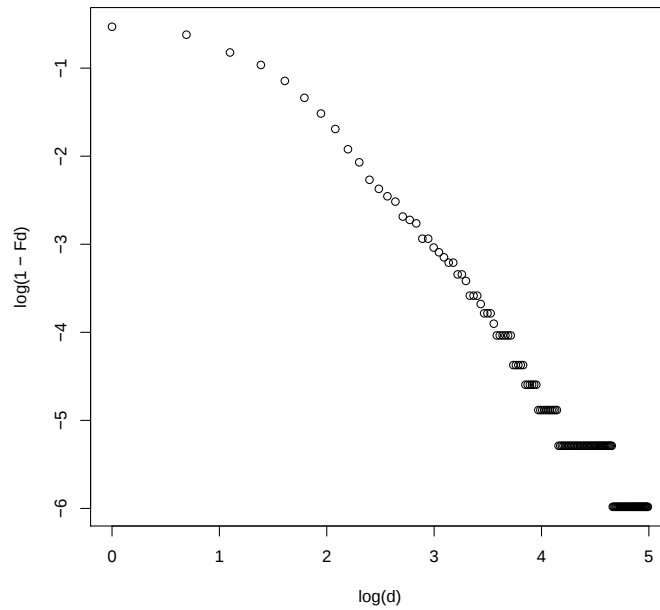
Obr. 2.5: Hodnoty parametra a , ktoré vyšli z regresie v závislosti od a , ktoré vstupovalo do regresie

2.1.2 Citácie

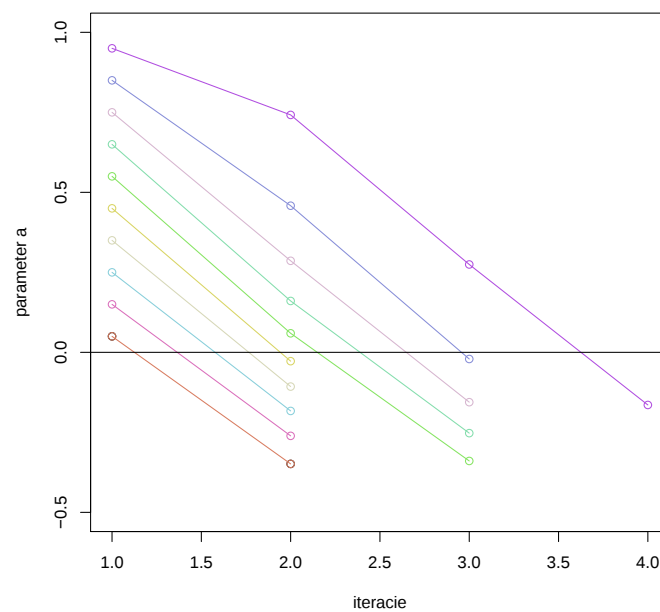
Druhá sada dát popisuje graf citácií. Medzi vrcholy boli zaradené všetky články, ktoré mali v použitej literatúre zaradený článok [11] alebo sa v ich názve vyskytoval výraz „smallworld“. Stupeň vrcholu potom predstavoval počet článkov, v ktorých bol daný vrchol citovaný. Dáta pochádzajú zo systému HistCite, ktorý uľahčuje získavanie dát pre bibliometrické analýzy. Tento systém vytvoril Eugene Garfield, ktorý je považovaný za jedného zo zakladateľov bibliometrie [3].

Sieť bola tvorená 396 vrcholmi a keďže tento graf je orientovaný, tak priemerný stupeň vrcholu bol $m = 5,02$. Na obrázku 2.6 sa nachádza graf závislosti $\ln(1 - F_{(d)})$ od $\ln(d)$.

Obrázok 2.7 zobrazuje hodnoty parametra počas iterácií pri rôznych štartovacích a_0 . Môžeme si všimnúť, že bez ohľadu na to na akej hodnote sme štartovali, parameter a sa dostal do záporných hodnôt a teda mimo intervalu $(0, 1)$.



Obr. 2.6: Závislosť $\ln(1 - F(d))$ od $\ln(d)$ pre sieť citácií



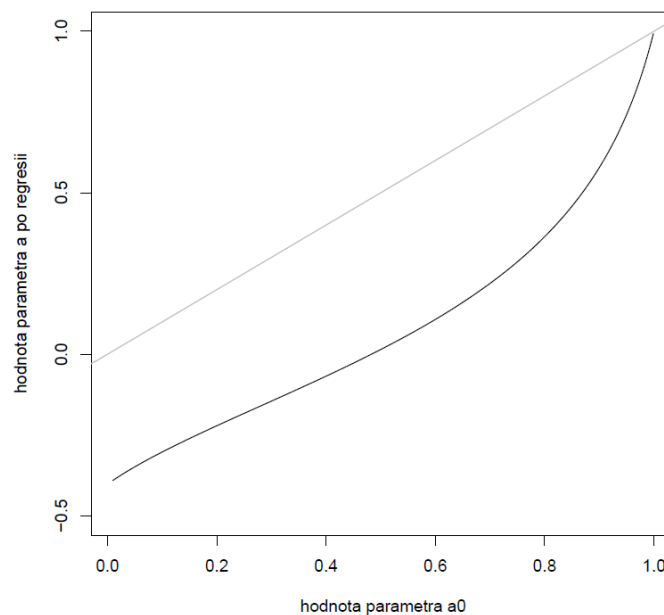
Obr. 2.7: Hodnoty parametra a počas iterácii pre $a_0 = 0,05, 0,15, \dots, 0,95$

V tabuľke 2.2 sú zapísané odhadnuté hodnoty parametra a , po ktorých sa algoritmus zastavil. Vidíme, že najviac 3 iterácie trvalo algoritmu kým sa dostal s parametrom

	a_0	a	iterácie
1	0.05	-0.349	1
2	0.15	-0.261	1
3	0.25	-0.183	1
4	0.35	-0.107	1
5	0.45	-0.028	1
6	0.55	-0.339	2
7	0.65	-0.252	2
8	0.75	-0.156	2
9	0.85	-0.021	2
10	0.95	-0.164	3

Tabuľka 2.2: Tabuľka štartovacích a_0 , odhadnutých hodnôt parametra a a počtu iterácií

mimo intervalu $(0, 1)$. Pevným bodom regresie je $a = 0,993$, čo môžeme vidieť na obrázku 2.8 ako priesečník dvoch kriviek. V tomto prípade náš algoritmus nedokázal skonvergovať ani pre štartovaciu hodnotu $a_0 = 0,95$, ktorá bola najbližšie pevnému bodu.



Obr. 2.8: Hodnoty parametra a , ktoré vyšli z regresie v závislosti od a , ktoré vstupovalo do regresie

Kapitola 3

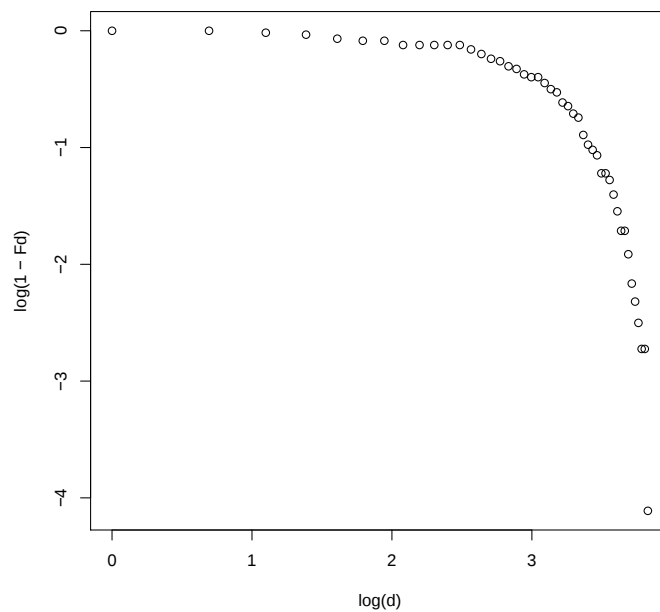
Aplikácia algoritmu - vlastné príklady

3.1 EFM 2013-2016

Ako prvý vlastný príklad, na ktorom sme sa rozhodli otestovať odhadovanie parametra a a tým pádom zistiť ako táto sieť vznikala, sme si vybrali graf opisujúci priateľstvá vo facebookovej skupine EFM 2013-2016. Ide o uzavretú skupinu ľudí, ktorí začali študovať odbor EFM v roku 2013 alebo ktorí mali s nimi nejaký spoločný predmet.

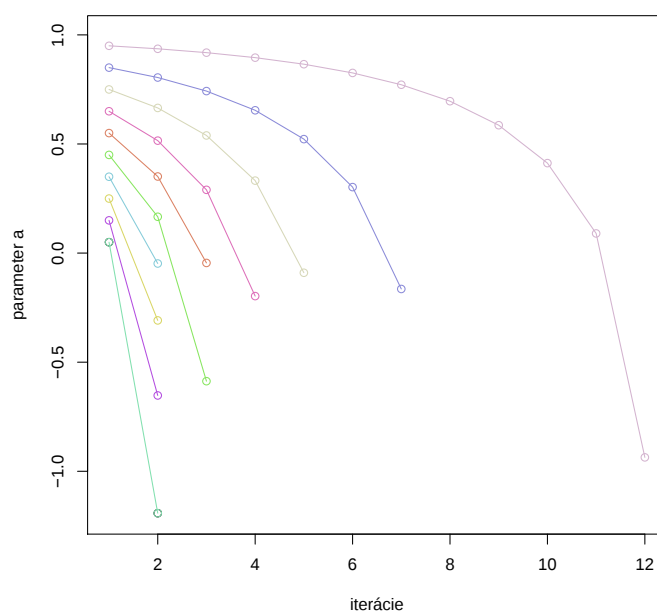
Informácie od jednotlivých študentov sme získavali formou zdieľaného dokumentu, do ktorého každý člen skupiny napísal počet svojich priateľov na Facebooku, ktorí sú tiež členmi tejto skupiny. Graf, ktorý sme získali je súvislý, neorientovaný a je tvorený 61 študentmi, minimálny počet priateľov je 2 a maximálny 46. Polovica priemerného stupňa vrcholov, ktorá pri našom odhadovaní predstavovala hodnotu m bola 12,754. Na obrázku 3.1 môžeme vidieť závislosť $\ln(1 - F(d))$ od $\ln(d)$.

Obrázok 3.2 predstavuje vývoj parametra a počas iterácii. Vidíme, že pre ani jednu štartovaciu hodnotu a_0 algoritmus neskonvergoval a hodnoty skončili mimo intervalu $(0, 1)$, konkrétne išlo o záporné čísla. Parameter a sme nakoniec dopočítali ako pevný bod regresie (2.5). Na Obrázku 3.3 je čiernou farbou znázornená krivka hodnôt a , ktoré vyšli po ich použití v regresii a priamka $a_0 = a$, ktorá je zobrazená šedou farbou. Pevný bod vyšiel $a = 0,998 \doteq 1$ čo znamená, že hrany v tomto hybridnom modeli neboli tvorené na princípe preferential attachment, ale pri vzniku nového vrcholu, mali

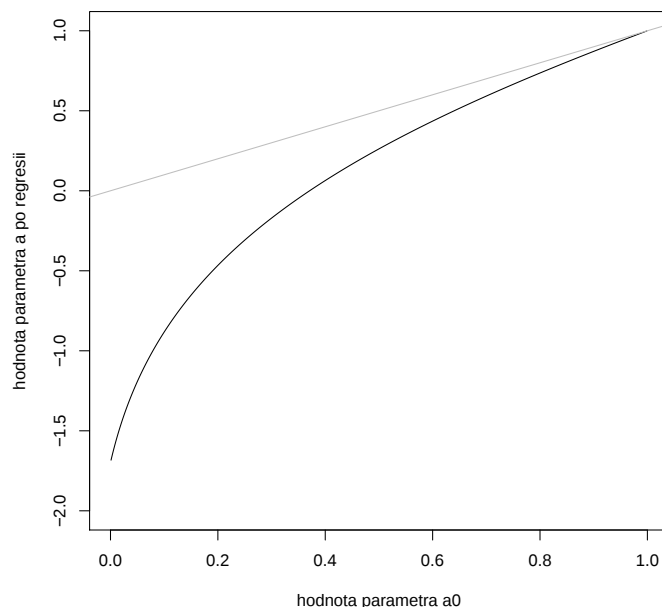


Obr. 3.1: Závislosť $\ln(1 - F(d))$ od $\ln(d)$ pre sieť EFM

všetky existujúce vrcholy rovnakú pravdepodobnosť vytvorenia hrany.



Obr. 3.2: Hodnoty parametra a počas iterácii pre $a_0 = 0.05, 0.15, \dots, 0.95$

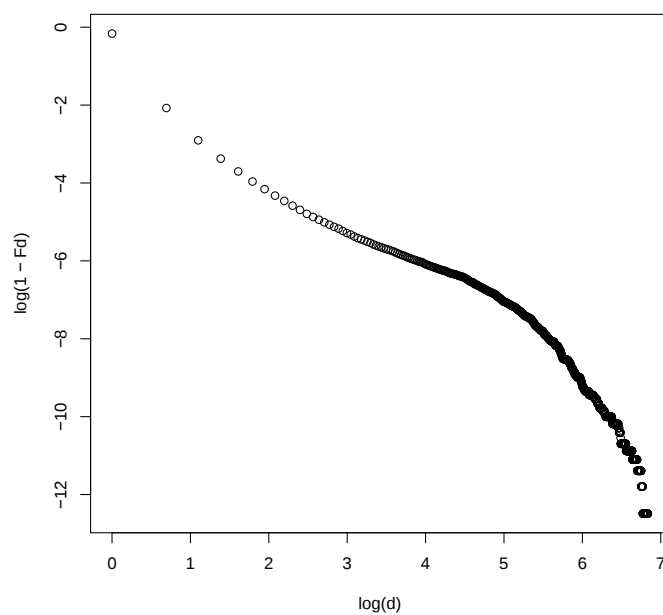


Obr. 3.3: Hodnoty parametra a , ktoré vyšli z regresie v závislosti od a , ktoré vstupovalo do regresie

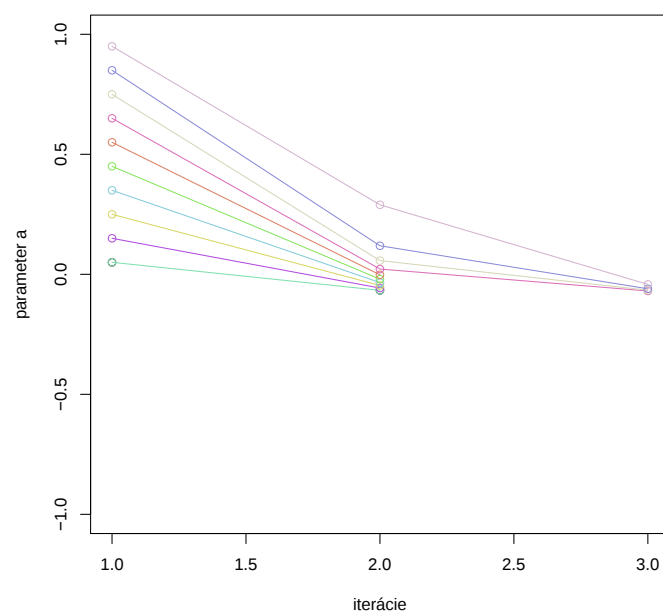
3.2 Emailová komunikácia

Druhý nami vybraný príklad je sieť predstavujúca emailovú komunikáciu jednej európskej výskumnej inštitúcie. Ide o dáta, ktoré zaznamenávajú obdobie od októbra 2003 po máj 2005, čo predstavuje 18 mesiacov. Tieto dáta sú voľne dostupné na [14] a boli použité v článku [10]. Graf je orientovaný a je tvorený 265214 vrcholmi, medzi ktorými je 420045 hrán, t.j. poslaných emailov. Polovica priemerného stupňa vrcholov je 1,584, minimálny počet emailov je 0 a maximálny 930. Závislosť $\ln(1 - F(d))$ od $\ln(d)$ je zobrazená na Obrázku 3.4.

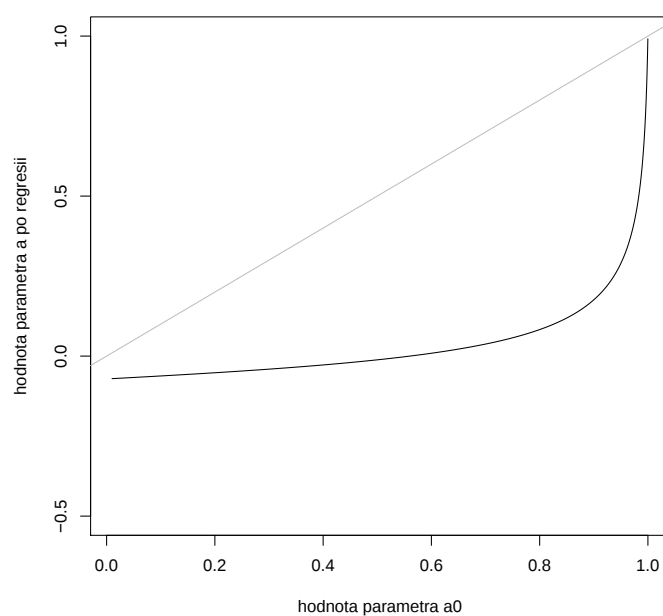
Na obrázku 3.5 je zobrazený parameter a počas iterácii. Rovnako ako v predchádzajúcom podkapitole 3.1 náš algoritmus neskonvergoval pre ani jednu štartovaciu hodnotu parametra a a vo všetkých prípadoch sme sa dostali do záporných hodnôt. Dopočítali sme hodnotu parametra a ako pevného bodu 2.2. Na Obrázku 3.6 je zobrazená závislosť parametra a , ktorý vyšiel po dosadení do regresie od hodnôt, ktoré boli dosádzané. Priesečník priamky $a_0 = a$ s krivkou je v hodnote $a = 1$.



Obr. 3.4: Závislosť $\ln(1 - F(d))$ od $\ln(d)$ pre sieť emailov



Obr. 3.5: Hodnoty parametra a počas iterácii pre $a_0 = 0.05, 0.15, \dots, 0.95$



Obr. 3.6: Hodnoty parametra a , ktoré vyšli z regresie v závislosti od a , ktoré vstupovalo do regresie

Kapitola 4

Testovanie algoritmu

V predchádzajúcej kapitole sme sa pri odhadovaní parametra a stretli s problémom, keď náš algoritmus nekonvergoval. Parameter a vybiehal z intervalu $(0, 1)$, konkrétne išiel do záporných čísel. Preto sa pozrieme bližšie na to, kedy takýto prípad nastáva.

Vygenerovali sme si umelé dáta pre rôzne hodnoty parametra a a m , na ktorých sme testovali konvergenciu algoritmu a jej rýchlosť. Presnosť riešenie sme zmenili vo while cykle na 10^{-6} . V našich testovacích súboroch dát bolo $a = 0, 1, \dots, 0, 9$ a začiatočný stupeň vrcholu $m = 10, \dots, 90$, čo predstavuje 90 sietí. Vo všetkých grafoch bolo $t = 100$. Postupne sme pre každý súbor dát odhadovali parameter a rovnakým spôsobom ako v kapitole 2, to znamená, že sme iteračný cyklus spustili vždy pre desať rôznych štartovacích hodnôt a_0 . Príslušný kód sa nachádza v prílohe ako kód 3.

Tabuľka 4.1 nám ukazuje počet štartovacích a_0 , pri ktorých iterácie skonvergovali. Hodnoty v prvom stĺpci nám zadávajú a , pri ktorom bola sieť vygenerovaná a prvý riadok určuje aké m bolo použité. Vidíme, že len v 14 prípadoch náš algoritmus skonvergoval. V tabuľke 4.2 sa nachádzajú odhadnuté hodnoty parametra a . Vo všetkých prípadoch nám vyšli hodnoty takmer rovnaké, $a \doteq 1$. S výnimkou posledného riadku, kde $a = 0, 9$ sa naše odhadnuté a nezhoduje s hodnotou parametra, pomocou ktorej boli dáta vygenerované.

Na základe týchto výsledkov to vyzerá, že ak algoritmus skonverguje, tak odhadnutá hodnota parametra vyjde $a = 1$. Lenže v kapitole 2 sme v prípade spoluautorstiev dostali hodnotu $a = 0.566$, preto vieme, že náš algoritmus je funkčný. Konvergenca

	10	20	30	40	50	60	70	80	90
a=0.1	0	0	0	0	0	0	0	0	0
a=0.2	0	2	0	0	0	0	0	0	0
a=0.3	7	2	3	0	0	0	0	0	0
a=0.4	0	1	0	0	0	0	0	0	0
a=0.5	0	0	2	0	0	0	0	0	0
a=0.6	0	2	0	0	0	0	0	0	0
a=0.7	7	6	1	0	0	0	0	0	0
a=0.8	3	1	0	0	0	0	0	0	0
a=0.9	2	1	0	0	0	0	0	0	0

Tabuľka 4.1: Tabuľka zaznamenávajúca koľko bolo konvergencií pre dáta s $a = 0, 1, \dots, 0, 9$ a $m = 10, 20, \dots, 90$

	10	20	30	40	50	60	70	80	90
0.1	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.2	NA	0,997	NA	NA	NA	NA	NA	NA	NA
0.3	0,995	0,998	0,999	NA	NA	NA	NA	NA	NA
0.4	NA	0,997	NA	NA	NA	NA	NA	NA	NA
0.5	NA	NA	0,999	NA	NA	NA	NA	NA	NA
0.6	NA	0,998	NA	NA	NA	NA	NA	NA	NA
0.7	0,999	1	0,997	NA	NA	NA	NA	NA	NA
0.8	0,997	0,996	NA	NA	NA	NA	NA	NA	NA
0.9	0,993	0,996	NA	NA	NA	NA	NA	NA	NAe

Tabuľka 4.2: Odhady paramtera a

algoritmus môže závisieť od pomeru počet dát a hodnota m , preto sme vygenerovali nové siete. Parameter a sme zachovali rovnaký ako v predchádzajúcom teste, t.j. $a = 0, 1, \dots, 0, 9$ a taktiež aj počet hrán, ktoré si vrchol pri vzniku vytvorí sme nechali $m = 10, \dots, 90$. Jediná hodnota, ktorú sme zmenili je počet vrcholov $t = 1000$.

	10	20	30	40	50	60	70	80	90
a=0.1	0	0	9	0	0	7	0	0	0
a=0.2	0	0	0	0	0	0	0	0	0
a=0.3	9	0	0	0	0	0	7	0	2
a=0.4	8	0	0	0	7	0	5	0	5
a=0.5	10	7	7	8	0	7	5	0	0
a=0.6	8	8	8	0	6	0	3	6	0
a=0.7	10	9	7	7	4	8	0	4	5
a=0.8	10	10	6	4	5	5	4	0	4
a=0.9	10	8	8	7	8	4	2	4	5

Tabuľka 4.3: Tabuľka zaznamenávajúca koľko bolo konvergencií pre dáta s $a = 0, 1, \dots, 0, 9$ a $m = 10, 20, \dots, 90$

V tabuľke 4.3 máme zapísaný počet štartovacích hodnôt a_0 , pre ktoré algoritmus skonvergoval. Oproti testovaniu pre $t = 100$, kde konvergencia bola len v 14 prípadoch, nám tento krát algoritmus skonvergoval 46 krát, čo je pre viac ako polovicu sietí. Pre menšie hodnoty m konvergoval pri väčšom počte štartovacích a_0 ako pre $m = 90$. V tabuľke 4.4 sú priemery odhadnutých hodnôt parametra a , ktoré sme dostávali pre rôzne a_0 . Všimnime si, že teraz máme aj hodnoty, ktoré sa nerovnajú 1. Pre $m = 10$ odhadnuté parametre zodpovedajú a , pre ktoré boli siete vygenerované. Keď zvyšujeme m , tak počet prípadov, ktoré dokonvergovali k hodnote $a \doteq 1$ narastá, až napokon pre $m = 70, 80, 90$ je odhadnutý parameter $a \doteq 1$ v 13 zo 14 sietí.

Teraz sa pozrieme na závislosť konvergenzie od a generujúceho sietí. Pre hodnoty $a \leq 0,4$ nám algoritmus skonvergoval deväť krát a len v jednom prípade k hodnote zodpovedajúcej pôvodnému a . Pre $0,5 \leq a \leq 0,7$ sme v siedmich prípadoch nemali konvergenciu, ale vidíme, že pri väčšine sietí sme dokonvergovali k hodnote $a \doteq 1$. Pre $a \geq 0,8$ sme mali konvergenciu v sedemnástich prípadoch. Odhad nám vychádzal $a \doteq 1$ vo všetkých sieťach okrem troch.

Na základe testovania algoritmu so 162 sieťami môžeme tvrdiť, že jeho konvergencia závisí od počtu vrcholov t , hodnoty m a skutočnej hodnoty a . Pre veľkú sieť, s malým

	10	20	30	40	50	60	70	80	90
	10	20	30	40	50	60	70	80	90
a=0.1	NA	NA	0.541	NA	NA	0.814	NA	NA	NA
a=0.2	NA	NA	NA	NA	NA	NA	NA	NA	NA
a=0.3	0.231	NA	NA	NA	NA	NA	0.998	NA	0.883
a=0.4	0.245	NA	NA	NA	0.91	NA	0.985	NA	0.999
a=0.5	0.601	0.447	0.695	0.952	NA	0.998	0.997	NA	NA
a=0.6	0.5	0.699	0.907	NA	0.953	NA	0.961	0.999	NA
a=0.7	0.831	0.874	0.892	0.996	0.927	0.999	NA	0.998	0.999
a=0.8	0.969	0.999	0.872	0.941	0.997	0.998	0.998	NA	0.999
a=0.9	0.995	0.997	0.999	0.999	0.999	0.998	0.99	0.999	0.999

Tabuľka 4.4: Odhady parametra a

počtom vznikajúcich hrán dostávame najpresnejšie odhady. Pre m , ktoré je rádovo desatina a viac oproti celkovému počtu vrcholov sa v prípade konvergenzie dostávame k hodnotám $a \doteq 1$, ktoré nezodpovedajú skutočnému a . Pre malé hodnoty parametra a nám algoritmus nekonverguje a pre veľké a dostávame hodnoty $a \doteq 1$.

Záver

Táto práca bola venovaná grafom a modelom opisujúcim ich vznik. Čitateľa sme postupne oboznámili so základnými pojmami súvisiacimi s grafmi a vysvetlili sme mu ako vznikajú hrany pri jednotlivých modeloch. V druhej kapitole sme uviedli algoritmus na odhadovanie parametra a pri hybridných modeloch, vysvetlili sme ako funguje a použili sme ho na dvoch zreprodukovaných príkladoch, v prílohe kód 2. Pre sieť spoluautorstiev nám algoritmus skonvergoval k hodnote $a = 0,566$, pre ktorú platí, že je to pevný bod regresie 2.5. Ako druhý príklad sme uviedli sieť citácií, pri ktorej nám algoritmus neskonvergoval pre žiadne a_0 a preto hodnotu parametra $a = 0,993$ sme získali ako pevný bod regresie. V príkladoch s vlastnými dátami nám opäť algoritmy neskonvergovali a hodnoty a boli 0,998 pre EFM dáta a 1 pre emailovú komunikáciu. Jedným z cieľov práce bolo testovať algoritmus na simulovaných dátach. V kapitole 4 sme zistili, že jeho konvergencia ešte nezaručila správny odhad parametra a , keďže vo všetkých prípadoch pre $t = 100$ sme dostali hodnotu $a \doteq 1$ aj keď a , pomocou ktorého boli siete vygenerované sa nerovnálo 1. Pri zmene celkového počtu vrcholov na $t = 1000$ hodnoty odhadnutých parametrov viac zodpovedali skutočnému a . Na základe výsledkov z tejto kapitoly môžeme tvrdiť, že veľkosť siete, jej priemerný stupeň vrcholu a skutočná hodnota a ovplyvňujú konvergenciu algoritmu. V troch príkladoch, ktoré sme uvádzali sa odhadnutý parameter dostal mimo intervalu $(0, 1)$ a tým pádom algoritmus neskonvergoval. Pre dáta citácií bolo $m = 5,02$ a $t = 396$. Parameter a nám ako pevný bod vyšiel rovný 0,993. Hoci algoritmus neskonvergoval, podľa kapitoly 4 nemal mať problém kvôli hodnotám a, m a t . Pri dátach EFM 2013-2016 bolo $m = 12,75$, čo sa zdá ako príliš veľké v pomere k $t = 61$. Poslednou sieťou boli emailové komunikácie, ktoré mali $m = 1,584$ a $t = 265214$. Dopočítané a sa rovnalo 1, čiže algoritmus nemal mať problém s konvergenciou.

Hlavným prínosom pre autora bolo zlepšenie programovacích zručností v programe R a práca s dátami, či už hľadanie alebo samotné ich spracovanie. Táto bakalárska práca by sa dala rozšíriť o opakované simulácie dát v kapitole 4 a následné odhadovanie parametra. Ďalej by sme sa mohli zamerať na zavedenie ďalšieho kritéria, na základe ktorého by sme odhadovali parameter a ako napríklad priemerná dĺžka cesty v grafe alebo zoskupovanie vrcholov.

Literatúra

- [1] Barabási, A., Albert, R.: *Emergence of Scaling in Random Networks*, Science 289 (1999), 509-512
- [2] Bearman, P. S., Moody, J., Stovel .: *Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks*, American journal of sociology 110 (2004), 44-91
- [3] Godin, B.: *On the Origins of Bibliometrics*, Project on the History and Sociology of STI Statistics, Working Paper no. 33 (2006), dostupné na internete (16.5.2016): http://www.csiic.ca/PDF/Godin_33.pdf
- [4] Goyal, S., van der Leij, M. J., Moraga-González, J. L.: *Economics: An Emerging Small World*, Journal of Political Economy 114(2) (2006), 403-412
- [5] Jackson, M. O., Rogers, B.W.: *Meeting Strangers and Friends of Friends: How Random are Socially Generated Networks?*, American Economic Review 97 (2007), 890-915
- [6] Jackson, M. O., osobná stránka, dostupné na internete (25.4.2016): <http://web.stanford.edu/~jacksonm/Data.html>
- [7] Jackson, M. O.: *Social and Economic Networks: Models and Analysis*, dostupné na internete (14.11.2015): <https://www.coursera.org/learn/networksonline>
- [8] Jackson, M. O.: *Social and Economic Networks*, Princeton University Press, 2008
- [9] Janková, K., Pázman, A.: *Pravdepodobnosť a štatistika*, Univerzita Komenského, Bratislava, 2013

- [10] Leskovec, J., Kleinberg, J., Faloutsos, Ch.: *Graph evolution: Densification and shrinking diameters*, ACM Transactions on Knowledge Discovery from Data (ACM TKDD) 1(1) (2007)
- [11] Milgram, S.: *The Small-World Problem*, Psychology Today 1(1) (1967), 61-67
- [12] Padgett, J. F., Ansell, Ch. K.: *Robust Action and the Rise of the Medici, 1400-1434*, American journal of sociology 98 (1993), 1259-1319
- [13] Price, D. J. de S.: *A General Theory of Bibliometric and Other Cumulative Advantage Processes*, Journal of the American Society for Information Science 27 (1976), 292-306
- [14] Stanford Large Network Dataset Collection, dostupné na internete (9.6.2016): <http://snap.stanford.edu/data/>
- [15] Znám, Š.: *Kombinatorika a teória grafov*, Univerzita Komenského, Bratislava, 1982

Príloha A

```
1 library(igraph)
2 set.seed(12345) # kvoli reprodukovateľnosti
3
4 #nazorný graf
5 ukazka <- erdos.renyi.game(n=7, p=0.5) #vygenerovanie grafu
6 pdf("ukazka.pdf")
7 average.path.length(ukazka)
8 plot(ukazka)
9 dev.off()
10
11
12 #rozdelenie stupnov vrcholov-binomicke rozdelenie
13 graf2 <- erdos.renyi.game(n=20, p=0.2) #vygenerovanie grafu
14 pdf("renyi0.pdf")
15 plot(graf2)
16 dev.off()
17 pdf("renyi1.pdf")
18 d1 <- length(degree.distribution(graf2))
19 #graf distribucnej funkcie stupnov vrcholov
20 plot((1:d1),degree.distribution(graf2), xlab="□", ylab="□")
21 lines((1:d1),degree.distribution(graf2), col="black")
22 #teoreticke hodnoty pri binomickom rozdeleni
23 lines((1:d1), dbinom((1:d1), 20, 0.2), col="blue")
24 dev.off()
25
26 #rozdelenie stupnov vrcholov-Poissonovo rozdelenie
27 graf2 <- erdos.renyi.game(n=50, p=0.02) #vygenerovanie grafu
28 pdf("renyi20.pdf")
29 plot(graf2)
30 dev.off()
31 pdf("renyi2.pdf")
32 d1 <- length(degree.distribution(graf2))
33 #graf distribucnej funkcie stupnov vrcholov
34 plot((1:d1),degree.distribution(graf2), xlab="□", ylab="□")
35 #teoreticke hodnoty pri Poissonovom rozdeleni
36 lines((1:d1),degree.distribution(graf2), col="black")
37 lines((1:d1), dpois((1:d1), 1), col="blue")
38 dev.off()
```

Listing 1: Kód 1

```

1 library(randomcoloR)
2
3 coauthor <- read.table(file="coauthor.txt", header=TRUE)
4 head(coauthor)
5 attach(coauthor)
6 x <- coauthors
7 plot(x, degree)
8 sum(x)
9 m <- sum(degree*x)/sum(x)/2 #priemerny stupen vrchola
10
11 vsetky <- NA
12 for (i in 1:length(x)) vsetky <- c(vsetky, rep(degree[i], x[i]))
13 Fd.funkcia <- ecdf(vsetky)
14 d <- seq(from=0, to=max(degree), by=1)
15 Fd <- Fd.funkcia(d)
16 # uprava - kvoli tomu ze ecdf ma  $F(x)=Pr(X\leq x)$ ,
17 #my chceme podla sladkov  $Pr(X<x)$ 
18 pocet <- length(d)
19 d <- d[-1]
20 Fd <- Fd[-pocet]
21
22 #vymazanie prvkov, kde Fd=1, pretoze by vznikol problem
23 #pri  $\log(1-Fd)$  v regresii
24 if (Fd[length(Fd)]==1) {
25   Fd <- Fd[-which(Fd==1)]
26   d <- d[1:length(Fd)]
27 }
28
29 #pdf("coauthor.pdf")
30 plot(log(d), log(1-Fd), type="p") #  $\log(1-Fd)$  v závislosti o  $\log(d)$ 
31 #dev.off()
32
33
34 #iteracny cyklus
35 #start s a0, ktore vstupi do regresie
36 #nasledny vypocet noveho a1, ktore si vyjadrieme pomocou
37 #koeficientou regresie
38 k <- NA #bude pocitat iteracie pre jednotlivu a0
39 a <- NA
40 z <- seq(from=0.05, to=0.95, by=0.1)
41 for (i in z) {
42   a0 <- 0
43   a1 <- i
44   iter <- 0
45   a.vektor <- NA
46   while (abs(a1-a0)>0.001)
47   {
48     a0 <- a1
49     if (a0<0 || a0>1) break
50     parameterMODEL <- lm(formula = log(1-Fd)~1+log(d+2*a0*m/(1-a0)))
51     x <- -unname(parameterMODEL$coeff[2])
52     a1 <- 1-(2/x)
53     iter <- iter+1
54     a.vektor[iter] <- a1
55   }
56   a.vektor <- t(a.vektor)
57   if (i==0.05) {
58     A.matica <- matrix(a.vektor, ncol=(length(a.vektor)), nrow=1)
59   } else {
60     rozmer <- abs(length(a.vektor)-ncol(A.matica))
61     if (length(a.vektor) > ncol(A.matica)) {
62       A.matica <- cbind(A.matica, matrix(rep(NA,
63         nrow(A.matica)*rozmer), ncol=(rozmer), nrow=(i-0.05)*10))
64     } else {
65       a.vektor <- cbind(a.vektor, matrix(rep(NA, rozmer),

```

```

66     ncol=rozmer, nrow=1)) }
67     A.matica <- rbind(A.matica, a.vektor)
68   }
69   a[(i+0.05)*10] <- a1
70   k[(i+0.05)*10] <- iter
71 }
72 A.matica <- cbind(z, A.matica)
73
74 cl <- distinctColorPalette(10)
75
76 cl <- rainbow(15, start=0.1, end=0.9)
77
78 pdf("iteracieCA.pdf")
79 plot(seq(1,sum(!is.na(A.matica[1,]))),
80      A.matica[1,1:sum(!is.na(A.matica[1,]))], xlab="□", ylab="□",
81      xlim=c(1,ncol(A.matica)), ylim=c(0,1))
82 for (i in 1:10) {
83   lines(seq(from=1, to=sum(!is.na(A.matica[i,])), by=1),
84         A.matica[i,1:sum(!is.na(A.matica[i,]))], col=cl[i])
85   points(seq(from=1, to=sum(!is.na(A.matica[i,])), by=1),
86          A.matica[i,1:sum(!is.na(A.matica[i,]))], col=cl[i])
87 }
88 title(xlab="iteracie", ylab="parameter□a")
89 dev.off()

```

Listing 2: Kód 2

```

1  #simulovane data, neorientovane grafy
2  library(gdata)
3
4  Afor <- seq(from=0.1, to=0.9, by=0.1)
5  Mfor <- c(10,20,30,40,50,60,70,80,90)
6  #Afor <- seq(from=0.6, to=0.9, by=0.1)
7
8  h <- 0
9  #matice pre zapisovanie konvergencii
10 maticaA0 <- matrix(rep(NA,2430), ncol=243, nrow=10)
11 maticaAD <- matrix(c(NA,Afor,rep(NA,90)), ncol=10, nrow=10,
12                    byrow=FALSE)
13 maticaAD[1,] <- c(NA, Mfor)
14 odhadA <- maticaAD
15 #t <- 1000
16 t <- 100
17
18 #cykly, v ktorych sa pre rozne a,m vygeneruju siete
19 #a nasledne odhadne parameter a
20 for (i in Afor) {
21   for (j in Mfor) {
22     h <- h+1
23
24     a <- i
25     m <- j
26     maticaS <- matrix(rep(NA,t*t), ncol=t, nrow=t)
27     priatel1a <- NA
28     pria2 <- NA
29     HMdata <- NA
30     HMroz <- NA
31
32     #generovanie siete
33     for (r in 1:m) {
34       maticaS[1:m,r] <- rep(1,m)
35       maticaS[r,r]=NA
36     }
37
38     for (r in (m+1):t) {

```

```

40  #nepreskalovane vahy
41  priatelial1 <- NA
42  for (ii in 1:(r-1)) priatelial1[ii]=sum(!is.na(maticas[ii,]))
43  priatelial2 <- 1:(r-1)
44
45  #vytvorenie nahodnych a*m hran
46  if (a!=0) {
47    hrany <- sample(1:(r-1), a*m)
48    maticas[r,hrany] <- rep(1,length(hrany))
49    maticas[hrany,r] <- rep(1,length(hrany))
50  }
51
52  if (a!=1) {
53    #hľadanie (1-a)*m hran cez preferential attachment
54    priatelial1 = priatelial1[-hrany] # nepreskalovane vahy
55    priatelial2 = priatelial2[-hrany] # vrcholy indexy
56    priatelial1 = priatelial1/sum(priatelial1) # preskalovane vahy
57
58    hrany2 <- resample(x=priatelial2, size=(1-a)*m, prob=priatelial1)
59
60    maticas[r,hrany2] <- rep(1,length(hrany2))
61    maticas[hrany2,r] <- rep(1,length(hrany2))
62  }
63 }
64
65 #HMroz je pravdepodobnostne rozdelenie stupna vrcholov
66 for (r in 1:t) {
67   HMdata[r] <- sum(!is.na(maticas[r,]))
68 }
69 HMroz <- matrix(rep(NA,2*(max(HMdata)+1)), ncol=2,
70 nrow=max(HMdata)+1)
71 for (r in 0:max(HMdata)) {
72   HMroz[r+1,] <- c(r,sum(HMdata==r))
73 }
74
75 #vytvorenie distribucnej funkcie
76 x <- HMroz[,2]
77 degree <- HMroz[,1]
78 m <- sum(degree*x)/sum(x)/2 #priemerny stupen vrchola
79
80 vsetky=HMdata
81 Fd.funkcia <- ecdf(vsetky)
82 d=degree
83 Fd <- Fd.funkcia(d)
84 # uprava - kvoli tomu ze ecdf ma F(x)=Pr(X<=x), my chceme podla
85 #sladjov Pr(X<x)
86 pocet <- length(d)
87 d <- d[-1]
88 Fd <- Fd[-pocet]
89
90 #vymazanie prvkov, kde Fd=1, pretoze by vznikol problem pri
91 #log(1-Fd) v regresii
92 if (Fd[length(Fd)]==1) {
93   Fd <- Fd[-which(Fd==1)]
94   d <- d[1:length(Fd)]
95 }
96
97 #iteracny cyklus
98 #start s a0, ktore vstupi do regresie
99 #nasledny vypocet noveho a1, ktore si vyjadrieme pomocou
100 #koeficientov regresie
101 k <- NA #bude pocitat iteracie pre jednotlivé a0
102 a <- NA
103 z <- seq(from=0.05, to=0.95, by=0.1)
104 for (r in z) {
105   a0 <- 0

```

```

106 | a1 <- r
107 | iter <- 0
108 | a.vektor <- NA
109 | #for (jj in 1:20)
110 | while (abs(a1-a0)>0.0001)
111 | {
112 |   a0 <- a1
113 |   if (a0<0 || a0>1) break
114 |   parameterMODEL <- lm(formula = log(1-Fd)~1+log(d+2*a0*m/(1-a0)))
115 |   x <- -unname(parameterMODEL$coeff[2])
116 |   a1 <- 1-(2/x)
117 |   iter <- iter+1
118 |   a.vektor[iter] <- a1
119 | }
120 | a.vektor <- t(a.vektor)
121 | if (r==0.05) { A.matica <- matrix(a.vektor, ncol=length(a.vektor),
122 |                                   nrow=1)
123 | } else {
124 |   rozmer <- abs(length(a.vektor)-ncol(A.matica))
125 |   if (length(a.vektor) > ncol(A.matica)) {
126 |     A.matica <- cbind(A.matica, matrix(rep(NA,
127 |                                             nrow(A.matica)*rozmer), ncol=rozmer, nrow=(r-0.05)*10))
128 |   } else {
129 |     a.vektor <- cbind(a.vektor, matrix(rep(NA, rozmer),
130 |                                           ncol=rozmer, nrow=1))
131 |     A.matica <- rbind(A.matica, a.vektor)
132 |   }
133 |   a[(r+0.05)*10] <- a1
134 |   k[(r+0.05)*10] <- iter
135 | }
136 | A.matica <- cbind(z, A.matica)
137 |
138 | #zapisovanie ci pre tuto sadu dat bola nejaka konvergencia
139 | if (sum(a>=0)!=0) {
140 |   maticaAD[i*10+1, which(Mfor==j)+1] <- sum(a>=0)
141 |   odhadA[i*10+1, which(Mfor==j)+1] <- round(mean(a[a>=0]),3)
142 | } else maticaAD[i*10+1, which(Mfor==j)+1] <- 0
143 | #zapisovanie odhadov a a poctu iteracii
144 | maticaA0[, (h*3)-2] <- z
145 | maticaA0[, (h*3)-1] <- a
146 | maticaA0[, (h*3)] <- k
147 |
148 | }
149 | }
150 |
151 | write.table(maticaAD, "maticaADd2.txt")
152 | write.table(odhadA, "odhadAd2.txt")
153 | write.table(round(maticaA0,3), "maticaA0d.txt")

```

Listing 3: Kód 3