Katedra Informatiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

Numerická simulácia rastúcej siete

(Bakalárska práca)

Matej Juračka

Vedúci: doc. RNDr. Mária Markošová, PhD.

Bratislava, 2008

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím citovaných zdrojov.

.....

Poďakovanie

Moje úprimné poďakovanie patrí všetkým, ktorí mi pri tvorbe tejto práce pomáhali, predovšetkým mojej školiteľke doc. Markošovej za jej záujem, trpezlivosť a čas, ktorý mi pri riešení problémov venovala.

Ďakujem Peťovi Nátherovi za pomoc pri riešení problémov so simuláciami.

Ďakujem mojim rodičom za ich zhovievavosť a podporu.

Abstrakt

Témou tejto práce bude skúmanie existujúcich sietí, vlastností, ktorými sa popisujú a zachytenie najdôležitejších čŕt, ktoré majú reálne siete spoločné. Opísanie niektorých modelov a sledovanie ich vlastností bude univerzálne, konkrétnym sieťam a ilustračnej aplikácii všeobecných poznatkov budem venovať pozornosť až k úplnému záveru práce, kde popíšem vznik jazykovej siete podľa jedného z teoretických modelov.

Niektoré vlastnosti sietí sa analytickým prístupom nepodarilo odvodiť a na ich zistenie je potrebné vytvorenie siete simulovať. Preto som vytvoril aplikáciu, ktorej účelom je podľa niektorého z uvedených modelov sieť vytvoriť a zmerať jej parametre, prípadne ich priemerné hodnoty.

Modely, ktorým sa táto práca venuje sú v poradí: náhodné grafy podľa Erdősa a Rényiho, rastúce siete s náhodným zapájaním vrcholov, rastúce siete s preferenčným zapájaním vrcholov - Barabási-Albert model, zrýchlené rastúce siete - Dorogovtsev-Mendes model a nakoniec rozšírenie DM-modelu o prepájanie hrán.

Kľúčové slová

Rastúce siete, Jazyková sieť, Numerické simulácie

Obsah

Kapitola 1

Úvod

1.1 Reálne siete

Príkladov reálnych sietí je okolo nás veľké množstvo, preto ich skúmanie môže zasiahnuť do širšieho počtu vedeckých oblastí. Je až zarážajúce, že veľké množstvo diametrálne odlišných sietí, či už je to internet (na doménovej, ako aj WWW stránkovej úrovni), sieť citácií vedeckých článkov, sieť hercov hrajúcich v rovnakých filmoch, sociologické siete ľudských vzťahov, jazykové siete tvorené rôznymi prístupmi, biologické siete proteínov a ich reakcií, ako aj množstvo ďalších, má tak podobné vlastnosti a charakter. Preto skúmanie sietí na najvšeobecnejšej úrovni môže byť prínosom pre ľubovoľnú oblasť, v ktorej sa štruktúra podobnej siete objavuje. Je iba samozrejmé, že všetky reálne siete majú svoje vlastné špecifiká a model, popisujúci ich vznik, je potrebné vhodne upraviť, avšak všeobecné princípy, podľa ktorých sa tieto siete vytvárajú (a následne ich vlastnosti) budú veľmi podobné.

Spoločným znakom všetkých uvažovaných sietí bude ich veľkosť, rádovo prevažne okolo 10^4 - 10^6, prípadne vyššie, ktorá spôsobuje, že na sledovanie ich vlastností je potrebné používať štatistické metódy. Preto budeme hľadať také vlastnosti sietí, ktoré sú približne rovnaké pri rovnakom spôsobe vytváranie siete, ale ktoré spoľahlivo rozlíšia siete vznikajúce podľa rôznych pravidiel. Podľa týchto vlastností následne budeme môcť uhádnuť, akým spô-

sobom už existujúca sieť mohla vznikať.

1.2 Matematické modely

Matematickou abstrakciou, ktorú budeme pri skúmaní reálnych sietí a ich vlastností využívať, je graf. Zrejme to, aký tento graf bude, vyplýva od konkrétnej reálnej siete, ktorú má tento graf reprezentovať.

Napríklad pre spomenutú sieť sociálnych vzťahov by išlo o neorientovaný, neohodnotený graf bez násobných hrán a slučiek, čo znamená, že medzi dvomi ľuďmi buď vzťah je, alebo nie, pričom nikto nemôže byť vo vzťahu so sebou samým.

V prípade jazykového grafu, reprezentujúceho vzájomnú pozíciu slov v textoch, pôjde o neorientovaný neohodnotený graf s povolenými násobnými hranami, ktoré sa budú vyskytovať vtedy, keď sa niektoré slová objavia v spoločnom kontexte viackrát.

V nasledujúcom texte budem hovoriť vždy o neohodnotených a neorientovaných grafoch, prítomnosť násobných hrán bude závisieť od konkrétneho prípadu. Pojmy *sieť* a *graf*, ktorý ju reprezentuje, nebudem rozlišovať.

U čitateľa budem predpokladat základné znalosti teórie grafov a elementárne pojmy ako graf, vrchol, hrana, stupeň vrcholu a pod. nebudem definovať. Pekný úvod do teórie grafov sa dá nájsť v knihe [13].

1.3 Vlastnosti reálnych sietí

Na porovnanie reálnych sietí s ich matematickými modelmi budem potrebovať niekoľko vlastností, ktoré tieto siete charakterizujú. Základnou takouto vlastnosťou, ktorá popisuje štruktúru siete je distribúcia stupňov vrcholov, označovaná $P(k)$, definovaná ako relatívna početnosť vrcholov, ktoré majú stupeň k. Pri väčšine reálnych sietí sa stretávame s takzvanou *power-law* závislosťou [11], $P(k) \propto k^{-\gamma}$, kde zvyčajne $\gamma \sim 2 - 3$. Ako neskôr uvidíme, táto závislosť nie je vždy úplne presná, ale môže sa líšiť pre veľké a malé

hodnoty k [2, 8], potom hodnoty γ pre nízke k budem označovať γ_{in}, naopak pre veľké γ_{out}. Graf $P(k)$ budem vykreslovať v logaritmickej škále na oboch osiach, pretože v takomto prípade je grafom priamka so smernicou $-\gamma$, prípadne funkcia dobre aproximovaná dvomi priamkami so smernicami $-\gamma_{\text{in}}$ a $-\gamma_{\text{out}}$.

Power-law distribúciu stupňov vrcholov budem považovať za najzákladnejšiu vlastnosť, ktorú musí model, popisujúci reálnu sieť mať. Ďalšie vlastnosti, ktoré budem pre jednotlivé reálne siete a ich modely počítať a merať popíšem postupne ďalej v texte.

1.4 Simulácia vytvárania siete

Na simuláciu tvorby sietí som vytvoril program, ktorý je schopný požadovanú sieť vytvoriť, namerať jej vlastnosti a tie uložiť do súboru. Takýchto nezávislých simulácií je schopný spraviť viac a následne namerané hodnoty spriemerovať. Aby sa namerané štatistické hodnoty blížili k teoretickým, je nutné vytvoriť buď graf s dostatočne veľkým počtom vrcholov, alebo spraviť veľké množstvo meraní na malých grafoch a vyhodnotiť ich priemer, čo sa časovo ukazuje ako výhodnejšie.

Pri niektorých modeloch a ich vlastnostiach sa teoretické hodnoty hľadajú jednoducho, pri niektorých zložitejšie a niektoré sa ukazujú ako príliš zložité na analytické vyjadrenie. V takomto prípade je užitočné na ich zistenie použiť práve simulácie a merania, čím sme schopní jednoducho porovnať, či daná reálna sieť má s daným modelom kvalitatívne podobné vlastnosti, a teda či je možné, aby podľa tohoto modelu sieť vznikala.

Program, simulujúci každý z modelov, ktoré budem v mojej práci popisovať, som napísal v jazyku C++, ktorého výber zapríčinila požiadavka na čo najrýchlejšiu prácu s pamäťou a procesorom, keďže siete budú mať pomerne veľké rozmery a pamäťové zaťaženie počítača bude vysoké. Podrobnejší popis programu je uvedený v kapitole 6.

Program aj so zdrojovými súbormi je súčasťou tejto práce a je priložený

na CD. Výstupné súbory sú formátované s ohľadom na ďalšie spracovanie a vyhodnotenie v niektorom z matematických programov, napríklad GnuPlot, Matlab, Mathematica a pod. Na údaje získané zo simulácií je zvyčajne užitočné aplikovať logaritmický binning, ktorý potom na grafe s logaritmickou škálou spôsobí, že body meraní budú od seba vzdialené o rovnakú dĺžku.

Kapitola 2

Náhodný vznik grafov

2.1 Erdős-Rényi model

2.1.1 Vznik ER grafu

Jedným z najstarších modelov, ktoré popisujú veľké, náhodne vznikajúce siete, je model Erdősa a Rényiho, navrhnutý a nimi skúmaný počas rokov 1959-1961, neskôr doplnený hlavne prácami Bollobása z roku 1981, ktorý odvodil konečný vzťah pre distribúciu stupňov vrcholov.

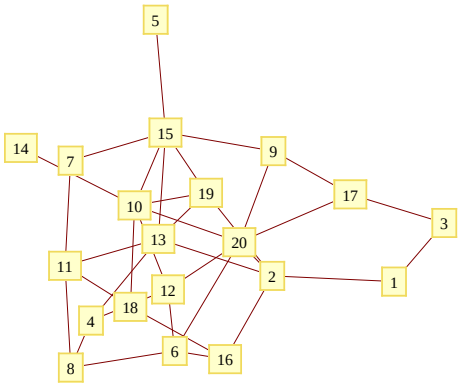
Existuje niekoľko ekvivalentných možností, ako tento model zadefinovať, jednou z nich je nasledujúca [7].

ER model náhodného grafu: Nech je daných n vrcholov grafu. Medzi každými dvoma vrcholmi tohto grafu sa hrana vyskytuje s pravdepodobnosťou p.

Z uvedenej definície zrejme vidieť, že v tomto prípade sa o existencii násobných hrán neuvažuje. Priemerný počet hrán takéhoto grafu potom bude $\frac{1}{2}n(n-1)p$, t.j. počet všetkých možných hrán medzi vrcholmi krát p. To vedie k trochu inej ekvivalentnej definícii - namiesto pravdepodobnosti výskytu hrany p sa bude uvádzať počet vrcholov n a počet hrán m, ktoré sú medzi týmito vrcholmi umiestnené náhodne. Z takto formulovanej definície vyplýva, že celkovo pre dané n a m existuje $\binom{\frac{n(n-1)}{2}}{m}$ rôznych grafov, z ktorých je vznik

každého rovnako pravdepodobný.

Obrázok 2.1 ukazuje príklad náhodného grafu pre $n = 20$ a $p = 0.2$.



Obr. 2.1: Náhodný graf, n=20, p=0.2.

2.1.2 Vlastnosti ER grafu

Distribúcia stupňov vrcholov Základnou vlastnosťou, ktorú budeme sledovať pri každom náhodne vznikajúcom grafe, je distribúcia stupňov jeho vrcholov, označovaná ako $P(k)$. Pre konkrétny graf to znamená relatívnu početnosť vrcholov v grafe, ktorých stupeň je práve k. Z definície vyplýva, že $P(k)$ sa tiež dá interpretovať ako pravdepodobnosť, že náhodne vybraný vrchol z grafu má stupeň k.

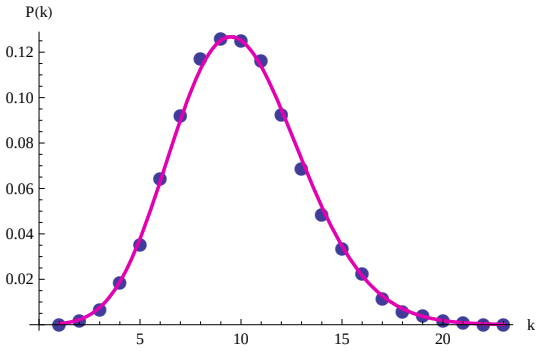
Ak vychádzame z definície, pri ktorej je daný počet vrcholov grafu n a pravdepodobnosť prepojenia dvoch vrcholov p, potom pre náhodné grafy je dobrou aproximáciou $P(k)$ vzťah [7, 11]

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{2.1}$$

ktorý pre dostatočne veľké n môže byť nahradený poissonovým rozdelením

$$P(k) \simeq e^{-pn} \frac{(pn)^k}{k!} = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{2.2}$$

kde $\langle k \rangle$ znamená priemerný stupeň vrcholu grafu, vypočítaný ako $\langle k \rangle = pn$. Na obrázku 2.2 je zobrazené porovnanie hodnôt $P(k)$ získaných zo simulácie a teoretická krivka podľa funkcie 2.2.



Obr. 2.2: Poissonova distribúcia stupňov vrcholov v náhodnom grafe, $n = 10000$, $p = 0.001$. Modré body - údaje zo simulácie, ružová krivka - teoretické hodnoty

Už podľa tvaru tejto krivky na prvý pohľad vidieť, že sa nezhoduje s požadovanou *power-law* distribúciou stupňa vrcholov, t.j. $P(k) \propto k^{-\gamma}$, ktorá je monotónna a klesajúca. V grafe sa podľa funkcie 2.2 a obrázka 2.2 nachádza preferovaný stupeň $\langle k \rangle$, ktorý (alebo približne ktorý) má väčšina vrcholov. Takýto jav sa pri reálnych sieťach neobjavuje a preto sa dá očakávať, že mechanizmus ich vzniku bude kvalitatívne iný.

Klasterizačný koeficient Ďalšou vlastnosťou, ktorá vystihuje štruktúru grafu bude jeho klasterizácia. Pod týmto pojmom chápeme niekoľko rôznych veličín odvodených od klasterizačných koeficientov jednotlivých vrcholov.

Klasterizačný koeficient $C(i)$ vrcholu i sa definuje ako pomer počtu existujúcich hrán medzi susedmi vrcholu i a počtu všetkých možných hrán medzi nimi. Neformálne ho môžeme chápať ako mieru prepojenia susedov vrcholu - ak je nízky, tak susedné vrcholy sú prepojené slabo, ak vysoký, tak naopak medzi väčšinou susedov hrana existuje. Počet všetkých možných hrán medzi k vrcholmi v neorientovanom grafe bez násobných hrán vypočítame ako

$\frac{1}{2}(k)(k-1)$. [3, 7]

Podobne môžeme zadefinovať klasterizačný koeficient C pre celý graf, ktorý bude priemerom klasterizačných koeficientov jeho vrcholov. Takto pre každý graf dostávame jedno konkrétne číslo C, vyjadrujúce jeho klasterizáciu.

Z definície $C(i)$ vyplýva, že je rovná pravdepodobnosti existencie hrany medzi dvoma náhodne vybratými susedmi vrcholu i. Keďže tá je pre náhodné grafy rovnaká pre všetky náhodne vybrané vrcholy, v náhodných grafoch platí $C(i) = p$ pre všetky vrcholy grafu. Preto celkový klasterizačný koeficient náhodného grafu bude tiež $C = p$.

Ak teda máme reálnu sieť s n vrcholmi, ktorej sme schopní spočítať priemerný stupeň vrcholov $\langle k \rangle$ a predpokladáme, že graf vznikal rovnakým spôsobom ako náhodné grafy, môžeme očakávať, že jeho klasterizačný koeficient bude rovný $C = \frac{\langle k \rangle}{n} = p$. Avšak pre reálne siete a náhodné grafy s rovnakým počtom vrcholov a $\langle k \rangle$ je klasterizačný koeficient C_{rand} rádovo menší ako C_{real}. Porovnanie pre niektoré reálne siete je uvedené v tabuľke 2.1 [1].

Tabuľka 2.1: Porovnanie klasterizačných koeficientov pre reálne siete a pre náhodné grafy s adekvátnymi parametrami

Sieť	n	$\langle k \rangle$	C_{real}	C_{rand}	referencia
Neurónová sieť	282	14.0	0.28	0.049	Watts, Strogatz (1998)
WWW	153127	3.8	0.24	0.00060	Adamic(1999)
Spolupráca biológov	1520251	15.5	0.081	0.000010	Newman(2001)
Spolupráca matematikov	253339	3.9	0.15	0.000015	Newman(2001)
Spolupráca filmových hercov	449913	113.4	0.20	0.00025	Newman *et al.* (2001)
Jazyková sieť	460902	70.1	0.44	0.00015	Cancho, Solé (2001)

Klasterizačný koeficient je teda jednou z vlastností, ktorými sa väčšina reálnych sietí líši od náhodných grafov, ktoré zaviedli Erdős a Rényi.

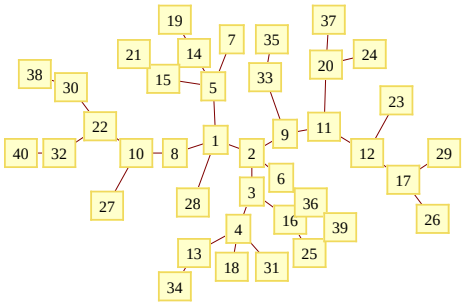
2.2 Rast siete, princíp náhodného vzniku hrán

2.2.1 Rastúce siete

Náhodný graf vznikal spôsobom, pri ktorom sa na začiatku pevne určil počet vrcholov, ktorý výsledný graf mal a potom sa medzi týmito vrcholmi vytvárali hrany. V reálnych sieťach sa však častejšie stretávame s takzvaným rastom siete, do siete vrcholy pribúdajú a zapájajú sa novými hranami k už existujúcim vrcholom. Analógiou takéhoto rastu v reálnych sieťach bude vznik nových slov v jazykových sieťach, zapojenie nového servera do siete, vytvorenie novej stránky na internete, nadviazanie kontaktu medzi ľuďmi v sociálnych sieťach a pod.

Nech na začiatku, v čase $t = 0$, tvorí sieť jeden nezapojený vrchol. Za každú jednotku času pribudne do siete nový vrchol a zapojí sa k už existujúcim vrcholom m novými hranami, pričom výber vrcholu, s ktorým sa spojenie vytvorí nech je *náhodné*. Takto vzniká sieť, ktorá v čase t má $t + 1$ vrcholov a každému vrcholu sa dá jednoznačne priradiť číslo s, vyjadrujúce čas, v ktorom vrchol do siete pribudol.

Obrázok 2.3 ukazuje príklad takto vytvorenej siete pre $t = 40$, $m = 1$.



Obr. 2.3: Príklad rastúcej siete s náhodným pripájaním vrcholov, $t = 40$, $m = 1$.

Vlastnosti siete Niektoré vlastnosti takto rastúcej siete sa vypočítajú triviálne. Počet uzlov siete v čase t je $t + 1$, počet hrán mt. Priemerný stupeň

vrcholu s v čase t budeme označovať ako $\bar{k}(s,t)$, distribúciu stupňov vrcholov rovnako, ako pri náhodných grafoch, $P(k)$. Priemerná hodnota týchto vlastností sa dá analyticky odhadnúť.

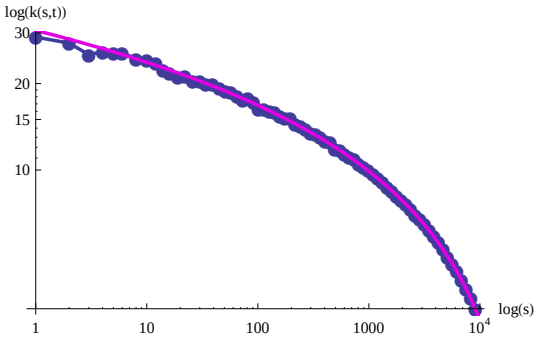
Distribúcia stupňov vrcholov Priemerný stupeň vrcholu s v čase t, t.j. $\bar{k}(s,t)$ sa dá vypočítať [4] ako

$$\bar{k}(s,t) = m(1 - ln(\frac{s}{t})) \tag{2.3}$$

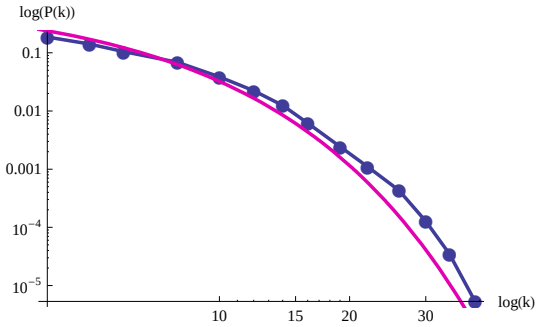
Z neho sa dá spôsobom uvedeným v nasledujúcej kapitole priamo odvodiť distribúcia stupňov vrcholov

$$P(k) = \frac{1}{m} e^{1-\frac{k}{m}} \tag{2.4}$$

Ak porovnáme takúto sieť s náhodnou sieťou, vznikajúcou spôsobom, ktorý navrhli Erdős a Rényi, zistíme, že v sieti sa už nevyskytuje žiadny preferovaný stupeň vrcholov, ale že distribúcia stupňov vrcholov exponenciálne klesá.



Obr. 2.4: Priemerný stupeň vrcholov v rastúcom grafe s náhodným zapájaním nových uzlov, simulácia a teoretická krivka, $t = 10000$, $m = 3$, logaritmická škála.



Obr. 2.5: Distribúcia stupňov vrcholov v rastúcom grafe s náhodným zapájaním nových uzlov, simulácia a teoretická krivka, $t = 10000$, $m = 3$, logaritmická škála.

Táto závislosť je však stále kvalitatívne odlišná od reálnych sietí, v ktorých platí $P(k) \propto k^{-\gamma}$. Z toho je vidieť, že takto rastúca sieť nepopisuje správanie reálnych sietí a je potrebné zmeniť spôsob jej rastu, t.j. princíp, podľa ktorého sa nové vrcholy do siete pripájajú.

Kapitola 3

Barabási-Albert model

3.1 Preferenčné pripájanie hrán

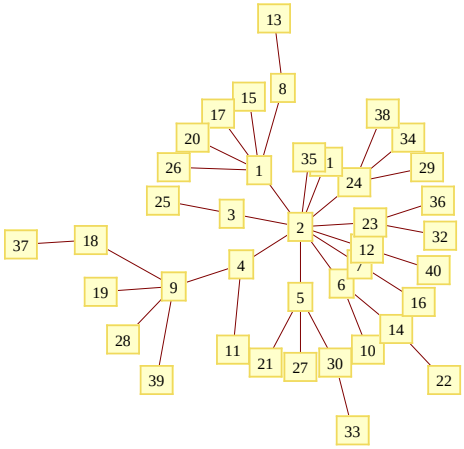
Model rastu siete, pri ktorom sa nové vrcholy pripájajú náhodne sa ukázal ako nedostačujúci na popísanie reálnych sietí. Problémom, ktorý spôsobuje výrazné rozdiely oproti reálnym sieťam, je práve náhodný výber vrcholov, ku ktorým sa nový vrchol pripája. Analógia s niektorou reálnou sieťou, (napríklad u osoby, ktorá má už veľa známych je pravdepodobnejšie, že sa zoznámi s niekým ďalším) , poskytuje intuitívne riešenie - dobre zapojené uzly siete budú nové spojenia vytvárať častejšie, ako tie, ktoré sú zapojené slabo, t.j. pravdepodobnosť, s akou sa prichádzajúci vrchol zapojí na už existujúci, by mala byť úmerná jeho stupňu. Formálne sa teda pravdepodobnosť pripojenia nového uzla t na už existujúci uzol s bude rovnať pomeru stupňa vrcholu s ku súčtu stupňov všetkých vrcholov, ktorý sa v každom grafe rovná dvojnásobku počtu hrán. Ak sa za každú časovú jednotku vytvorí m nových hrán, potom je počet hrán v grafe v čase t rovný mt. Preto pravdepodobnosť, že sa do siete prichádzajúci uzol pripojí na uzol s bude

$$p(s) \quad = \quad \frac{k(s,t)}{\sum_{i=0}^{t} k(i,t)} = \frac{k(s,t)}{2mt} \tag{3.1}$$

Tento proces vzniku siete, založený na preferenčnom pripájaní nových vrcholov k vrcholom už existujúcim, navrhli A-L. Barabási a R. Arnold [11] s

cieľom vytvoriť sieť, ktorej distribúcia stupňov vrcholov $P(k)$ bude podobná reálnym sieťam, t.j. $P(k) \propto k^{-\gamma}$. Niektoré vlastnosti takejto siete sa dajú vypočítať, podobne ako v predchádzajúcom modeli, analyticky, iné sa dajú odmerať na simulácii. V ďalšom texte budem spôsob vytvárania takéhoto grafu označovať ako *BA-model*.

Obrázok 3.1 ukazuje príklad takto vznikajúceho grafu pre $m = 1$ v čase $t = 40$. Číslo vrcholu označuje čas, v ktorom tento vrchol do siete pribudol.



Obr. 3.1: Graf vytvorený podľa BA-modelu, $m = 1$, $t = 40$

3.2 Vlastnosti BA-modelu

Niektoré vlastnosti tohoto modelu môžeme odvodiť teoretickým prístupom, ak vlastnosti, ktoré sú pre reálne grafy diskrétne, budeme pokladať za spojité a budeme pre ne vedieť zostaviť a vyriešiť diferenciálne rovnice, ktoré ich popisujú.

3.2.1 Distribúcia stupňov vrcholov

Pre zjednodušenie predpokladajme, že nový vrchol sa do existujúcej siete zapojí iba jednou novou hranou, t.j $m = 1$. Pravdepodobnosť, že vrchol s má v čase t stupeň k, označíme $p(k, s, t)$. Potom v čase $t + 1$ sa dá táto pravdepodobnosť vyjadriť ako súčet pravdepodobnosti, že vrchol mal v čase t ten istý stupeň a nový vrchol sa naňho nenapojil a pravdepodobnosti, že mal stupeň o jeden nižší a nový vrchol sa naňho napojil [9, 2]

$$p(k, s, t+1) = (1 - \frac{k}{2t})p(k, s, t) + \frac{k-1}{2t}p(k-1, s, t) \tag{3.2}$$

Táto rovnica môže byť prepísaná v tvare

$$2t[p(k, s, t+1) - p(k, s, t)] = (k-1)p(k-1, s, t) - kp(k, s, t) \tag{3.3}$$

Prechodom k spojitým premenným t a k dostávame

$$2t\frac{\partial p(k, s, t)}{\partial t} + \frac{\partial kp(k, s, t)}{\partial k} = 0 \tag{3.4}$$

Už v tomto jednoduchom prípade by však priame riešenie tejto rovnice a následné počítanie $P(k)$ pomocou pravdepodobnosti $p(k, s, t)$ pre každý uzol viedlo k závislostiam vyjadreným δ-funkciou, čo počítanie značne komplikuje. Aj pre tento, a predovšetkým pre zložitejšie modely bude praktickejšie používať prístup využívajúci priemerný stupeň uzla s v čase t, označovaný ako $\bar{k}(s, t)$.

Pomocou $p(k, s, t)$ ho môžeme vyjadriť ako

$$\bar{k}(s, t) = \sum_{k=1}^{\infty} kp(k, s, t) = \int_0^{\infty} kp(k, s, t)\mathrm{d}k \tag{3.5}$$

Zmenu $\bar{k}(s, t)$ v čase t vypočítame ako

$$\frac{\partial \bar{k}(s, t)}{\partial t} = \int_0^{\infty} k\frac{\partial p(k, s, t)}{\partial t}\,\mathrm{d}k \tag{3.6}$$

Keď teraz obe strany rovnice 3.4 prenásobíme k a zintegrujeme s $\int_0^{\infty} \mathrm{d}k$ po úprave *per partes* dostávame vzťah

$$\frac{\partial \bar{k}(s, t)}{\partial t} = \frac{1}{2t}\bar{k}(s, t) \tag{3.7}$$

Takýmto spôsobom sa dostávame k diferenciálnej rovnici, ktorej význam je intuitívne zrejmý. Zmena stupňa vrcholu s v čase t je rovná pravdepodobnosti, že sa naňho pripojil nový, do siete prichádzajúci vrchol. Tento prístup sa ukazuje ako oveľa praktickejší pre odvodenie vlastností rastúcich sietí. Pre väčšinu nasledujúcich modelov budeme schopní napísať a vyriešiť podobnú diferenciálnu rovnicu oveľa jednoduchšie, ako riešiť diferenčné rovnice popisujúce $p(k, s, t)$. Tento prístup však so sebou prináša aj riziká, predpokladá totiž spojitosť v premenných t a k. Spojitosť v t sa zabezpečí predpokladom, že siete a teda aj konečný čas budú dostatočne veľké. So spojitosťou v k môžu nastať problémy, ktorým je pri zložitejších modeloch potrebné venovať pozornosť.

Riešením rovnice 3.7 so začiatočnou podmienkou $\bar{k}(s, s) = 1$ je funkcia

$$\bar{k}(s, t) = \left(\frac{s}{t}\right)^{-\frac{1}{2}} \tag{3.8}$$

Pomocou vypočítaného $\bar{k}(s, t)$ teraz vieme vyjadriť distribúciu stupňov vrcholov v čase t ako

$$P(k, t) = \frac{1}{t}\int_0^t \delta(k - \bar{k}(s, t))\,\mathrm{d}k = -\left[t\frac{\partial \bar{k}(s, t)}{\partial s}\right]^{-1} \tag{3.9}$$

kde s môžeme vyjadriť z rovnice $k = \bar{k}(s, t)$. Dosadením 3.8 do 3.9 dostávame

$$P(k, t) = 2k^{-3} \tag{3.10}$$

A keďže vidíme, že $P(k, t)$ je nezávislé na čase, ďalej platí

$$P(k) = 2k^{-3} \tag{3.11}$$

Z tohoto postupu vidieť, že ak predpokladáme, že $\bar{k}(s, t) \propto s^{-\beta}$, potom dosadením a vyrátaním $P(k)$ dostávame vzťah $P(k) \propto k^{-\gamma}$, pričom platí

$$\beta(\gamma - 1) = 1 \tag{3.12}$$

Ak sa vrátime k trochu všeobecnejšiemu prípadu, kedy sa nový vrchol zapojí m novými hranami, diferenciálna rovnica popisujúca funkciu $\bar{k}(s, t)$

bude mať tvar

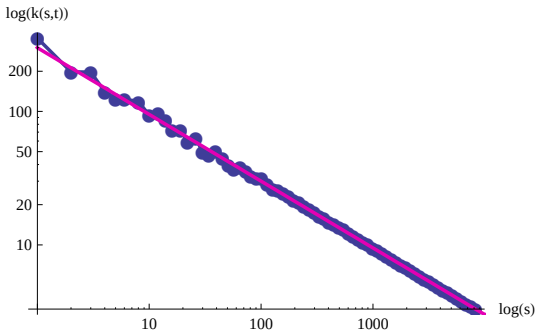
$$\frac{\partial \bar{k}(s,t)}{\partial t} = m\frac{\bar{k}(s,t)}{2mt} = \frac{1}{2t}\bar{k}(s,t) \tag{3.13}$$

$$\bar{k}(s,s) = m \tag{3.14}$$

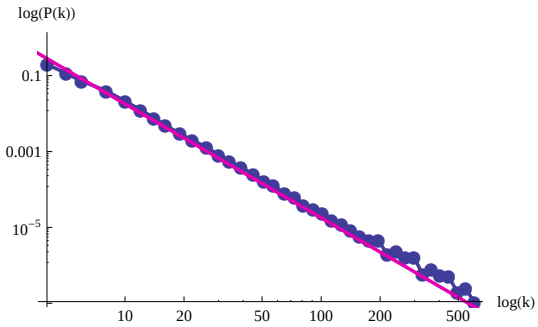
Riešením tejto rovnice je funkcia $\bar{k}(s,t) = m\sqrt{\frac{t}{s}}$. Ak sa na funkciu $\bar{k}(s,t)$ pozeráme ako na čiastočnú funkciu premennej s, t.j. t pokladáme za konštantné, dostávame závislosť $\bar{k}(s,t) \propto s^{-\frac{1}{2}}$ a následne podľa 3.12 $P(k) \propto k^{-3}$. Tu vidieť, že počet hrán, ktorými sa nový vrchol do siete pripája nezmení exponenty β a γ, iba koeficienty ktorými sú $\bar{k}(s,t)$ a $P(k)$ násobené. Presné vyjadrenie $P(k)$ v tomto prípade bude

$$P(k) = 2m^2k^{-3} \tag{3.15}$$

Presnosť tohoto odvodenia je uspokojujúca. Pre dostatočne veľké t, alebo pre dostatočne veľký počet priemerovaných simulácií je zhoda s nameranými hodnotami $\bar{k}(s,t)$, ako aj $P(k)$ relatívne vysoká.



Obr. 3.2: BA-model, závislosť $\bar{k}(s,t)$ od s, $t = 10000$, $m = 3$, počet priemerovaných meraní = 30, teoretická krivka a simulácia



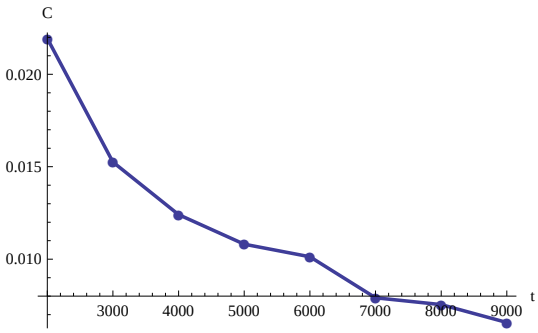
Obr. 3.3: BA-model, závislosť $P(k)$ od k, $t = 10000$, $m = 3$, počet priemerovaných meraní = 30, teoretická krivka a simulácia

3.2.2 Klasterizačné koeficienty

Klasterizačné koeficienty, ktoré boli zavedené v kapitole 2.1.2 sa dajú podobným spôsobom skúmať aj v tomto prípade. Jediným problémom je existencia násobných hrán. Ak by sme tieto povolili, tak možných existujúcich hrán medzi všetkými susedmi ľubovoľného vrcholu, ak ich má tento aspoň dvoch, je nekonečne veľa. To vyriešime tým, že násobné hrany vo vzniknutom grafe budeme ignorovať, t.j. vytvoríme nový graf, ktorý bude mať rovnaké vrcholy ako pôvodný a medzi dvomi vrcholmi bude hrana práve vtedy, ak bola v pôvodnom grafe medzi týmito vrcholmi aspoň jedna hrana a klasterizačné koeficienty budeme počítať pre takýto graf.

Klasterizačný koeficient je jedna z vlastností, ktoré sa analyticky počítajú ťažko. Preto na jeho vypočítanie použijem sériu simulácií a jeho hodnoty určím štatisticky. Z nameraných hodnôt je vidieť, že klasterizačné koeficienty su veľmi nízke a ani rádovo nedosahujú priemerné hodnoty C reálnych sietí. S postupným rastom siete klasterizačný koeficient dokonca postupne klesá (obr. 3.4), čo sa s vysokou klasterizáciou reálnych sietí, ktoré sú zvyčajne veľkých rozmerov, nezhoduje.

Ak poznáme klasterizačný koeficient každého z vrcholov siete, môžeme skúmať jeho ditribúciu vzhľadom na stupeň jednotlivých vrcholov, t.j. priemerný klasterizačný koeficient vrcholov so stupňom k .



Obr. 3.4: Závislosť klasterizačného koeficientu C od času, rast siete spôsobuje jeho klesanie.

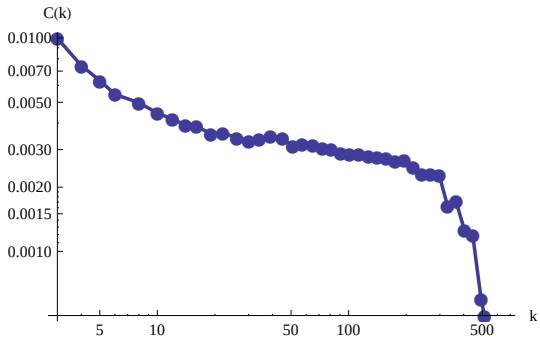
Vo väčšine rastúcich sietí platí, že vrcholy s podobným stupňom majú podobné aj klasterizačné koeficienty. Distribúcia klasterizačných koeficientov pre reálne siete môže byť rôzna, pre väčšinu z nich nebola meraná a zverejnená. Pre jazykovú sieť, uvedenú v kapitole 5, boli namerané klasterizačné koeficienty pre nízke k vysoké a konštantné, pričom za určitou hranicou k začali klesať k nule.

Keďže celkový klasterizačný koeficient je v BA-modeli siete nízky, dá sa očakávať, že aj jeho distribúcia v závislosti od stupňa vrcholov sa bude pohybovať v nízkych hodnotách a podľa obr. 3.5 je pre nízke stupne trochu vyššia, na prevažnej časti siete približne konštantná a pre uzly s vysokými stupňami klesá.

Pre niektoré reálne siete je distribúcia klasterizačných koeficientov dôležitým parametrom vystihujúcim prítomnosť hierarchie medzi jednotlivými uzlami tejto siete, ktorá je určená závislosťou $C(k) \propto k^{-\alpha}$. Podrobnejší popis tejto vlastnosti je nad rámec tejto práce a dá sa s ním zoznámiť v [12, 3, 5].

3.3 Zhrnutie

Model, ktorý navrhli Barabási a Albert je najjednoduchším známym modelom rastúcej siete, ktorého distribúcia stupňov vrcholov má power-law charakter, konkrétne $P(k) \propto k^{-3}$. Preto jeho preferenčný princíp rastu sa pova-



Obr. 3.5: Distribúcia klasterizačných koeficientov v závislosti od stupňa vrcholov, $t = 8000$, $m = 3$, počet meraní $= 200$

žuje za základný prvok, ktorý by mal model popisujúci reálnu sieť mať. Jeho jednoduchosťou je však spôsobené, že niektoré ďalšie jeho vlastnosti, ako napríklad klasterizácia grafu, sú pre popísanie väčšiny reálnych sietí nedostačujúce, a preto je namieste rozmýšľať o rôznych rozšíreniach a korekciách tohoto modelu.

Kapitola 4

Zrýchlenie rastu siete

4.1 Model Dorogovtsev-Mendes

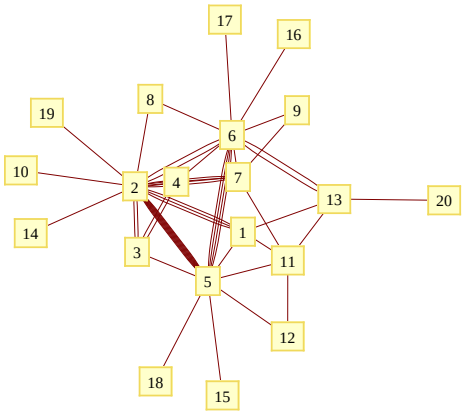
Ako som spomenul v závere predchádzajúcej časti, niektoré vlastnosti BA modelu neboli dostačujúce na popis reálnych sietí. Predovšetkým graf distribúcie stupňov vrcholov v niektorých reálnych sieťach nie je v logaritmickej škále lineárny, ale mierne zakrivený, čo sa dá vyjadriť ako smernica dotyčnice v tejto logaritmickej škále na začiatku a na konci grafu, t.j. γ_{in} a γ_{out}. V tabuľke sú porovnané niektoré namerané hodnoty $\gamma_{\text{in,out}}$ reálnych sietí [2].

Tabuľka 4.1: Hlavné vlastnosti reálnych sietí, veľkosť t, klasterizačný koeficient C, smernice γ_{in} a γ_{out}

Sieť	t	γ_{in}	γ_{out}
stránky WWW, Altavista, Oct 1999	2.711×10^8	2.1	2.7
mapa domény nd.edu	325729	2.1	2.45
Filmoví herci	212250	2.3	2.3
Slová - synonymá	22311	2.8	2.8
Spolupráca matematikov	70975	2.5	2.5

Druhým problémom, ktorý je potrebné riešiť je stále nízka hodnota klasterizačných koeficientov. Čiastočné riešenie oboch týchto problémov navrhli Dorogovtsev a Mendes rozšírením BA-modelu [2]. Sieť v tomto prípade bude vznikať rovnako ako pri BA modeli, ale v každom čase t sa v sieti preferenčne

vytvorí $c_1 t$ nových hrán medzi pôvodnými vrcholmi, kde $c_1 \ll 1$ je určená konštanta. To spôsobí, že závislosť počtu hrán od počtu vrcholov bude v grafe kvadratická, pričom nové hrany budú najčastejšie vznikať medzi dobre zapojenými vrcholmi. Tie sa tým stanú priamo spojenými a vzdialenosti medzi nimi sa tým výrazne skrátia, čo pri BA modeli nefungovalo, pretože dobre zapojené vrcholy s vysokým stupňom mohli byť spojené pomocou dlhšej cesty slabo zapojených vrcholov. Tým, že do grafu pribudnú nové hrany sa dá očakávať zvýšenie klasterizačného koeficientu pre celý graf a tým, že sa tieto hrany zapoja preferenčne sa zvýši klasterizačný koeficient najvýraznejšie pre slabo zapojené vrcholy. To je spôsobené tým, že pôvodne boli tieto vrcholy najpravdepodobnejšie zapojené na vrcholy s vysokým stupňom, ktoré však prepojené medzi sebou neboli, takže klasterizácia bola nízka. Vznikom týchto spojení medzi vrcholmi s vysokým stupňom sa klasterizačný koeficient vrcholov na ne napojených výrazne zvýši.



Obr. 4.1: Príklad siete vznikajúcej podľa DM modelu, $t = 20$, $m = 1$, $c_1 = 0.2$

4.2 Vlastnosti DM-modelu

4.2.1 Distribúcia stupňov vrcholov

Podobne, ako v prechádzajúcom prípade, najjednoduchšou metódou, akou sa dopracovať k priemernému stupňu vrcholu s v čase t, je riešiť diferenciálnu rovnicu, ktorá ho popisuje. [2] Zmena stupňa vrcholu s v čase t sa dá vyjadriť ako súčet pravdepodobnosti, že sa k nemu pripojí prichádzajúci uzol napájajúci sa svojimi m novými hranami preferenčne, a pravdepodobnosti, že sa medzi ním a iným uzlom v sieti preferenčne vytvorí jedna z $c_1 t$ nových hrán. Preto platí diferenciálna rovnica

$$\frac{\partial \bar{k}(s,t)}{\partial t} = \frac{m + 2c_1 t}{\int_0^t \bar{k}(s,t)\,\mathrm{d}s}\bar{k}(s,t) = \frac{m + 2c_1 t}{2mt + c_1 t(t-1)}\bar{k}(s,t) \tag{4.1}$$

ktorá sa dá riešiť aj presne, ale pre kratšiu formu výsledného vzťahu je výhodné rovnicu zjednodušiť na tvar

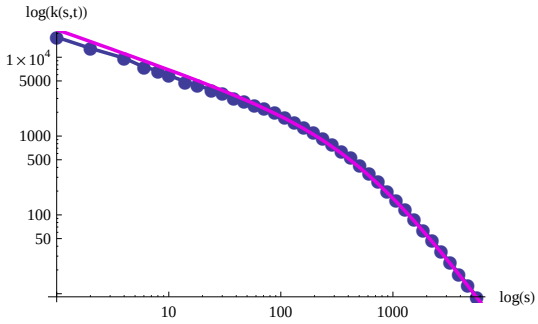
$$\frac{\partial \bar{k}(s,t)}{\partial t} = \frac{m + 2c_1 t}{2mt + c_1 t^2}\bar{k}(s,t) \tag{4.2}$$

Približný vzťah $t(t-1) \sim t^2$ je pre veľké t akceptovateľným zjednodušením a výsledné riešenie rovnice 4.2 (so začiatočnou podmienkou $\bar{k}(s,s) = m$) bude

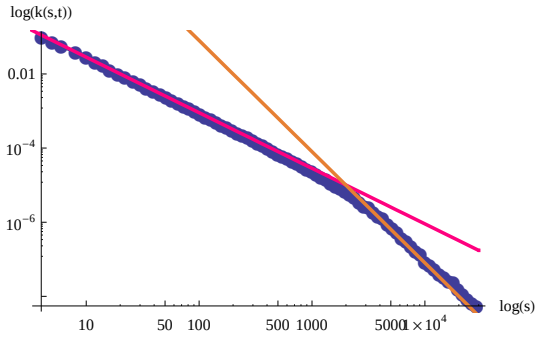
$$\bar{k}(s,t) = m\sqrt{\frac{t}{s}}\left(\frac{2m + c_1 t}{2m + c_1 s}\right)^{\frac{3}{2}} \tag{4.3}$$

a zjednodušene $\bar{k}(s,t) \propto s^{-\frac{1}{2}}$ pre malé hodnoty s a naopak, $\bar{k}(s,t) \propto s^{-\frac{1}{2}-\frac{3}{2}}$ pre $s \to t$. Z tohoto vzťahu sa nám však nepodarí jednoducho vyjadriť $s = s(k,t)$ a preto presnú závislosť pre $P(k,t)$ nepoznáme. Ak ale vieme približné koeficienty $\beta_{\text{in}} = \frac{1}{2}$ a $\beta_{\text{out}} = 2$, vieme pomocou nich vypočítať aj $\gamma_{\text{in}} = 1.5$ a $\gamma_{\text{out}} = 3$. Je dôležité si všimnúť, že uzly s vysokou hodnotou s majú nízku hodnotu stupňa k, takže β_{in} je vzťahom 3.9 späté s γ_{out} a opačne. Preto môžeme $P(k)$ aproximovať dvoma krivkami $P(k) = a_1 k^{-1.5}$ pre $k \to 0$ a $P(k) = a_2 k^{-3}$ pre veľké k (obr. 4.3). Koeficienty a_1 a a_2 však z teórie nevieme zistiť, preto museli byť odhadnuté až podľa hodnôt zo simulácie. Je však vidieť, že smernice týchto kriviek na logaritmickej škále popisujú namerané

$P(k)$ pomerne presne. Graf $\bar{k}(s,t)$ na obrázku 4.2 taktiež ukazuje zhodu teórie so simuláciami.



Obr. 4.2: Priemerný stupeň vrcholu s, $t = 10000$, $m = 3$, $c_1 = 0.01$, počet meraní = 100, porovnanie s teoretickou krivkou 4.3



Obr. 4.3: Distribúcia stupňa vrcholov $P(k)$, $t = 10000$, $m = 3$, $c_1 = 0.01$, počet meraní = 100, aproximované dvoma krivkami $a_1 k^{-1.5}$ a $a_2 k^{-3}$

Pridaním vytvárania nových spojení medzi existujúcimi hranami v grafe sme teda schopní znížiť smernicu distribúcie stupňa vrcholov pre nízke hodnoty k na hodnotu -1.5, pre vysoké hodnoty k ostáva táto smernica rovná -3.

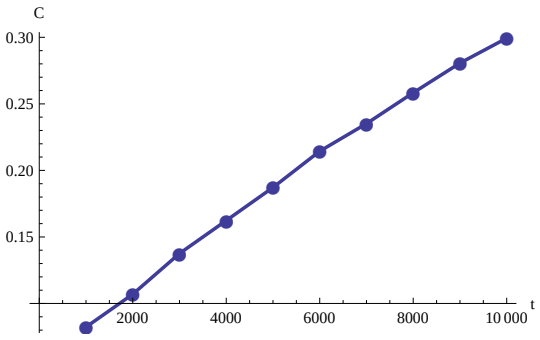
Ďalším efektom, ktorý pozorujeme, je výrazné zvýšenie počtu hrán v grafe. Namiesto lineárnej závislosti pri BA-modeli sa dostávame ku závislosti kvadratickej, ktorá je pre rastúce siete charakteristickejšia, pretože počet všetkých možných hrán v grafe s n vrcholmi je $\mathcal{O}(n^2)$. Potom pomer počtu

hrán v sieti vznikajúcej podľa DM-modelu ku všetkým možným hranám v tomto grafe bude konvergovať ku konštante c_1 pre dostatočne veľký čas rastu siete. Preto jednou z možností ako približne určiť c_1 pre sieť, pri ktorej sa predpokladá, že vznikala takýmto spôsobom, bude zistiť počet jej vrcholov t a vydeliť počet jej hrán číslom $\frac{1}{2}t(t-1)$, čo bude približná hodnota c_1. Ak by hrany do siete pribúdali rýchlejšie, ako $\mathcal{O}(n^2)$, potom by sa sieť preplnila a väčšina vrcholov by mala spojenie so všetkými ostatnými, takže rozmýšľať o ešte rýchlejšom pribúdaní hrán nemá pre väčšinu reálnych sietí praktický zmysel.

4.2.2 Klasterizačné koeficienty

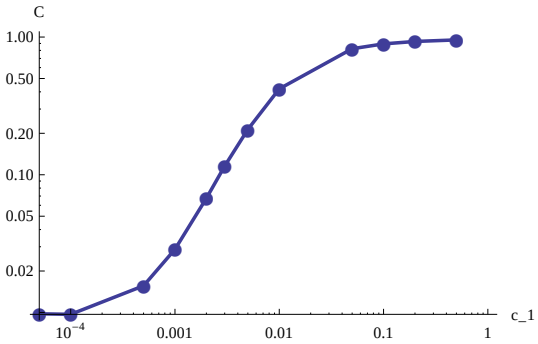
Ďalšou vlastnosťou, ktorej výraznú zmenu oproti BA-modelu očakávame, je správanie sa klasterizácie grafu. Tým, že sa graf hranami výrazne zahustí a to vhodným - preferenčným spôsobom, očakávame, že klasterizácia bude stúpať.

Podobne ako v predchádzajúcich prípadoch, ani teraz nevieme klasterizačný koeficient analyticky vyjadriť. V úvode som navrhol intuitívne vysvetlenie toho, prečo očakávame že sa jeho hodnota aj distribúcia v závislosti od stupňa vrcholov zmení. Prvou zmenou je závislosť klasterizačného koeficientu od rastu siete. Táto bola pri BA-modeli klesajúca (obr. 3.4). Obrázok 4.4 ukazuje, že v tomto prípade bude klasterizačný koeficient, naopak, rásť.



Obr. 4.4: Zmena klasterizačného koeficientu počas rastu DM-siete, $c_1 = 0.005$, $m = 3$

Ďalšou možnosťou, ako ovplyvniť klasterizačný koeficient, je meniť veľkosť konštanty c_1. Logickým očakávaním bude, že čím vyššia konštanta c_1, tým vyššie klasterizačné koeficienty bude sieť mať. To je znázornené na obrázku 4.5.

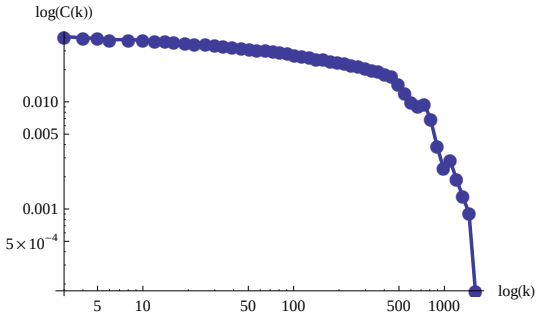


Obr. 4.5: Závislosť klasterizačného koeficientu od parametra c_1, $t = 6000$, $m = 3$, počet priemerovaných meraní na každý bod = 30, logaritmická škála

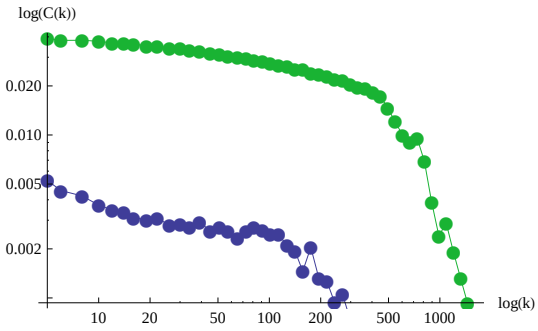
A poslednou vlastnosťou, ktorú uvediem a ktorá sa týka klasterizácie, bude opäť distribúcia klasterizačných koeficientov v závislosti od stupňa k. Tá bola pri BA-modeli nízka, približne konštantná, iba pre nízke hodnoty k sa trochu zvýšila a naopak, pre vysoké k klesala k nule. Na obrázku 4.6 vidieť, že v tomto prípade je závislosť podobná, rozdiel je predovšetkým v tom, že všetky hodnoty $C(k)$ sú položené vyššie, ale ich pomalé klesanie (opäť na logaritmickej škále, na lineárnej by toto klesanie bolo oveľa prudšie) je zachované. Porovnanie grafov $C(k)$ pre DM-model a BA-model je na obrázku 4.7

4.3 Zhrnutie

Model Dorogovtseva a Mendesa popísaný v tejto kapitole je význačný predovšetkým dvomi zmenami oproti BA-modelu, zakrivením priamky (na logaritmickej škále) distribúcie stupňa vrcholov v grafe a výrazným zvýšením klasterizačných koeficientov, ktoré sú závislé na veľkosti siete, t.j. času za ktorý táto sieť rástla, ako aj na parametri c_1. Taktiež tento model pomohol



Obr. 4.6: Distribúcia klasterizačného koeficientu podľa k, $t = 10000$, $m = 3$, $c_1 = 0.001$, počet priemerovaných meraní = 100, logaritmická škála



Obr. 4.7: Porovnanie distribúcie klasterizačných koeficientov podľa k, modrá krivka BA-model, zelená krivka DM-model, $t = 10000$, $m = 3$, v prípade DM-modelu $c_1 = 0.001$, logaritmická škála

stabilizovať hustotu hrán v grafe, pri BA-modeli pomer počtu hrán v grafe ku všetkým možným hranám konvergoval pri ľubovoľných parametroch pre $t \to \infty$ k nule. Teraz sa tento pomer stabilizuje ku konštante c_1.

Aj tento model má však mierne nedostatky, ktoré pre jednotlivé siete bude potrebné napraviť, čomu sa budem venovať v nasledujúcej kapitole, kde sa tento model pokúsim aplikovať na konkrétnu jazykovú sieť.

Kapitola 5

Jazyková sieť

Na záver mojej práce by som rád uviedol jeden príklad reálnej siete, na ktorej sa získané poznatky pokúsim aplikovať. Toutou sieťou bude jazyková sieť, popisujúca vzťahy medzi jednotlivými slovami v jazyku. Týmto jazykom bude v tomto prípade angličtina, ale jeho výber kvalitatívne vlastnosti siete neovplyvní. Sieť sa vytvárala z dostatočne dlhých, čo najvšeobecnejších textov, konkrétne z viacerých anglických prekladov biblie.

5.1 Vytvorenie jazykovej siete

Vrcholy tejto siete tvorili jednotlivé slová vyskytujúce sa v texte a hrana medzi dvomi slovami vznikla práve vtedy, keď boli tieto slová priamymi, alebo o jedno slovo vzdialenými susedmi vo vete. Ilustrované na príklade :

> My name is absurd too: Malachi Mulligan, two dactyls. [1]

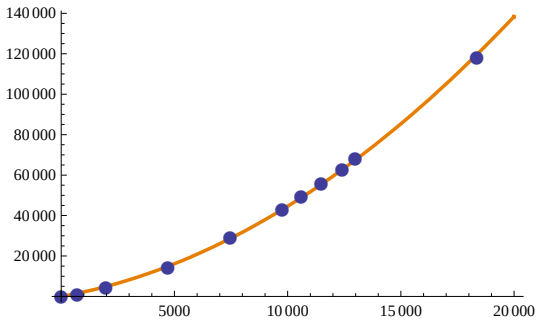
Do grafu sa v takomto prípade pridajú slová *my*, *name*, *is*, *absurd* atď.; hrany sa vytvoria medzi slovami (*my*, *name*), (*my*, *is*), (*name*, *is*) ... Takýmto spôsobom sa prejde celý text a vytvorí sa jazykový graf. Je dôležité zdôrazniť, že nie je podstatné, v akom poradí sa slová do grafu pridávajú, skutočný jazykový graf je takýmto spôsobom iba rekonštruovaný a predpokladá sa, že

[1] James Joyce : Ulysses

vznikal prirodzeným vývojom. Takto zrekonštruovaný graf teda zachytáva syntaktické vzťahy v jazyku medzi jednotlivými slovami a takisto intenzitu týchto vzťahov, čo vyjadruje násobnosť hrán.

Program, ktorý daný text spracuje a jazykovú sieť pomocou neho vytvorí a vyhodnotí vytvoril Peter Náther.

Jednou z možností, ako odhadnúť veľkosti parametrov pre vytváranie siete je odmerať závislosť počtu hrán od počtu vrcholov. Tá je zobrazená na obrázku 5.1 a ako obrázok naznačuje, táto závislosť vyzerá byť kvadratická a dá sa fitnúť parabolou. Zmeraním jej koeficientov sa tak dá odhadnúť približné celočíselné m a reálne c_1 ako $m = 2$ a $c_1 = 0.000492131526$. Tieto koeficienty nie sú presné, pretože sieť, z ktorej som merania robil, bola vytvorená z krátkeho textu - History of the United States (Charles A. Beard, Mary R. Beard) z projektu Gutenberg [15]. Tento príklad je iba ilustráciou, ktorá ukazuje, aké veľkosti parametrov môžeme pre jazykovú sieť očakávať a akým spôsobom sa ku nim môžeme dopracovať.

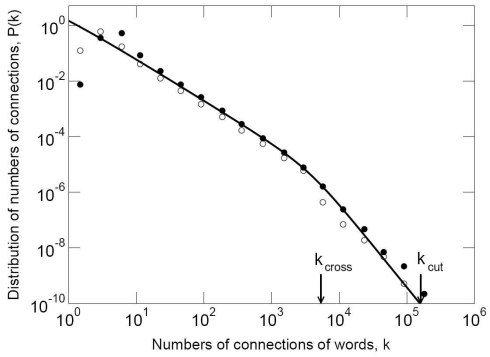


Obr. 5.1: Závislosť počtu hrán od počtu vrcholov v častiach textu. Modré body - namerané údaje, žltá krivka - fit parabolou

Vlastnosti takto vytvorenej jazykovej siete vieme namerať a porovnať s teoretickými (alebo simulovanými) vlastnosťami DM-modelu siete. Základné vlastnosti, popisujúce preferenčne rastúce grafy, sú $P(k)$ a $\bar{k}(s,t)$. Pri $\bar{k}(s,t)$ sa však vyskytuje už vyššie spomenutý problém, že sieť sa v skutočnosti nevytvára, ale rekonštruuje, čo spôsobí, že poradie vrcholov pribúdajúcich do siete, t.j. ich s nepoznáme. Mohli by sme skúsiť predpokladať monotónnosť

tejto vlastnosti, ale je otázne, ako prijateľný tento predpoklad je. Ekvivalentná vlastnosť $P(k)$ v sebe tento problém neskrýva, jej teoretické odvodenie pre DM-model však nie sme schopní presne určiť, iba ho aproximovať dvoma krivkami so smernicami γ_{in} a γ_{out}.

Preto pre jazykové siete budeme tieto dve smernice považovať za dostatočne popisujúce distribúciu stupňov vrcholov grafu. Pre veľké siete vytvorené z Biblií boli veľkosti týchto koeficientov odmerané ako $\gamma_{\text{in}} = 1.32$ a $\gamma_{\text{out}} = 2.21$ [5]. Nameranú distribúciu $P(k)$ pre jazykové grafy ukazuje obrázok 5.2.



Obr. 5.2: Distribúcia stupňov vrcholov $P(k)$ pre jazykový graf [8]

Slová v strmšej časti krivky, ktoré sú do siete zapojené veľmi silno, t.j. majú vysoký stupeň, ale je ich relatívne málo oproti celkovému počtu slov, tvoria *jadro* jazyka. V tomto jadre sa nachádzajú slová najbežnejšie používané a tvoriace základ vetných konštrukcií. Ich počet sa s rastom siete mení iba minimálne a rádovo sa pohybuje v hodnotách $\sim 10^3$, čo je spôsobené posunom bodu zlomu k_{cross}, v ktorom prechádza smernica grafu z γ_{in} do γ_{out}. [8, 2]

5.2 Rozšírenie DM-modelu o prepájanie hrán

Ako vidieť z nameraných koeficientov, ich podoba s DM-modelom je výrazná, ale nie úplná. Koeficienty v oboch častiach grafu sú vychýlené, čo dáva pries-

tor na rozširovanie modelu a snahu o ich posun.

Jedným z možným prístupov, ako upraviť DM-model pre účely popísania jazykového grafu, je zaviesť do procesu tvorby siete prepájanie hrán. Model upravíme nasledovným spôsobom: za každú jednotku času vyberieme m_r náhodných vrcholov a ku každému vybranému vrcholu jednu s ním koincidentnú hranu, ktoré z grafu odstránime a vytvoríme m_r nových hrán z vybraných vrcholov, ktoré do grafu pripojíme preferenčne [5, 6]. Tento jav v jazykovom grafe bude ekvivalentný zániku významu slova a jeho používaniu v inom kontexte, čo sa v procese vývoja jazyka nepochybne deje. Takto vzniknutý model popisuje diferenciálna rovnica

$$\frac{\partial \bar{k}(s,t)}{\partial t} = \frac{(m + 2c_1 t + m_r)}{\int_0^t \bar{k}(i,t)\,\mathrm{d}i}\bar{k}(s,t) - \frac{m_r}{t} = \frac{(m + 2c_1 t + m_r)}{2mt + c_1 t(t-1)}\bar{k}(s,t) - \frac{m_r}{t} \tag{5.1}$$

ktorej riešenie po zjednodušení (pre kratší tvar výsledného vzťahu a jeho jednoduchšiu interpretáciu) bude

$$\bar{k}(s,t) \propto \left(\frac{t}{s}\right)^{\frac{m+m_r}{2m}} \left(\frac{2m + c_1 t}{2m + c_1 s}\right)^{2-\frac{m+m_r}{2m}} \tag{5.2}$$

Z tohto riešenia sa dá odvodiť rovnakým spôsobom, ako v predchádzajúcom modeli $\beta_{\text{in}} = 2$, $\beta_{\text{out}} = 2 - \frac{m+m_r}{2m}$a $\gamma_{\text{in}} = 1.5$, $\gamma_{\text{out}} = 2 + \frac{m-m_r}{m+m_r}$. Voľbou vhodne veľkého $m_r < m$, v tomto prípade napríklad $m_r = 1$ dostávame $\gamma_{\text{out}} = 2.33$.

5.3 Zhrnutie

Reálnym procesom zmeny významu slov a jeho implementáciou sa podarilo dosiahnuť posun teoretických hodnôt distribúcie stupňa vrcholov k meraniam. Zhoda modelu s experimentálnymi dátami však stále nie je úplná a preto by sa dalo uvažovať o niektorých jeho ďalších rozšíreniach. Taktiež je potrebné venovať sa skúmaniu niektorých jeho ďalších vlastností, predovšetkým klasterizačným koeficientom, ktoré vyzerajú byť veľmi silnou vlastnosťou popisujúcou vlastnosti sietí ako takých.

Kapitola 6

Program na simuláciu vytvárania siete

V tejto kapitole podrobnejšie popíšem program, ktorý som vytvoril a používal na simulácie vytvárania všetkých spomenutých sietí.

6.1 Základné vlastnosti

Primárnym účelom tohoto programu je podľa nastavených parametrov náhodne vytvoriť požadovanú sieť, zmerať jej vlastnosti a tie uložiť do výstupných súborov. Parametre, ktorými sa sieť definuje sú nasledovné:

- int *VertexCount* = počet vrcholov v grafe, pri rastúcich sieťach sa toto číslo rovná času t, za ktorý sa sieť bude tvoriť.
- int *RepeatCount* = počet simulácií, ktoré prebehnú a z ktorých sa následne vypočítajú priemerné výstupné hodnoty.
- int *MultipleEdges* = konštanta Graf::MUL_NO, Graf::MUL_ALLOW, alebo Graf::MUL_IGNORE, definujúca vytváranie násobných hrán. Ak v simulácii násobné hrany vytvárať nechceme, zvolíme Graf::MUL_NO, čo spôsobí, že hrana sa zapája iba na voľné miesta v grafe. Graf::MUL_IGNORE

spôsobuje, že vznikajúcej hrane sa určí miesto, kam sa má zapojiť, rovnako, ako pre grafy s násobnými hranami, ale ak je toto miesto už obsadené, tak sa hrana odignoruje a graf sa vytvára ďalej.

- int M = počet hrán, ktorými sa nový vrchol pripája k sieti.
- double *C1* = koeficient c_1, za každú časovú jednotku preferenčne vznikne medzi starými vrcholmi v sieti $c_1 t$ nových hrán.
- int *Mr* = počet hrán M_r, ktoré sa za jednotku času náhodne odpoja a pripoja preferenčne (kap. 5)
- double *C2* = koeficient c_2, za každú časovú jednotku sa $c_2 t$ hrán náhodne odpojí a pripojí preferenčne
- double *Pr* = pravdepodobnosť p, ktorú má definovať zmysel iba pri ER-modeli vytvárania siete.
- bool *PrefR* = určuje, či sa budú pri rastúcom grafe hrany vyberať preferenčne (*true*), alebo náhodne (*false*).

Tieto parametre sú jednoducho nastaviteľné v súbore *Masina.cpp*. Po kompilácii a následnom spustení programu sa sieť vytvorí (určený počet krát) a namerané hodnoty sa poukladajú do jednotlivých textových súborov s príponou *.dat*. Vzhľadom na čisto vedecký účel tohoto programu som používateľské prostredie alebo definovanie parametrov za behu programu nepovažoval za dôležité. Pre používateľa sa teda predpokladá aspoň základná znalosť OOP programovania a syntaxe jazyka C++.

Následne sa vytvorí buď rastúca sieť volaním metódy *Graf::NapracVrcholmi()*, alebo náhodná sieť podľa Erdősa a Rényiho volaním *Graf::NapracVrcholmiRand(double Pr)*

To, čo všetko program namieria a exportuje sa deklaruje hneď za určenými parametrami.

- g→ExportToFilePk(int i); - Vyráta pre graf g distribúciu $P(k)$ a uloží ju do dočasného súboru *Temp_pk_*i.*dat*

- g→ExportToFileKst(int i); - Vyráta pre graf g stupne jeho vrcholov $k(s,t)$ a uloží ich do dočasného súboru *Temp_kst_*i.*dat*

- g→ExportToFileCkCs(int i); - Vyráta pre graf g klasterizačné koeficienty jednotlivých vrcholov $C(s)$ a ich distribúciu $C(k)$ a uloží ich do dočasných súborov *Temp_Cst_*i.*dat* a *Temp_ck_*i.*dat*

- g→ExportDataBin(string s); - Uloží maticu susednosti vrcholov grafu (t.j. celý graf) do binárneho súboru *s*.

- g→ExportData(string s); - Uloží maticu susednosti vrcholov grafu do textového súboru *s*.

Nakoniec sa zavolajú funkcie *ReaderKst()*, *ReaderPk()*, *ReaderCk()* a *ReaderC()*, ktoré z dočasných súborov pre jednotlivé grafy vytvoria priemerné hodnoty a tie uložia do výsledných súborov.

6.2 Implementácia

Program tvoria dve základné triedy - *Masina* a *Graf*. Trieda *Masina* spolu s jej metódami bola popísaná v predchádzajúcom odstavci. V hlavnej funkcii *Main* sa vytvorí jej inštancia a spustí sa jej hlavná metóda *Masina::Count()*.

Trieda *Graf* je zodpovedná za vytvorenie grafu. V konštruktore sa vytvoria iba defaultné hodnoty jednotlivých parametrov, ktoré sú verejné a teda môžu byť prepísané mimo tejto triedy. Zavolaním metódy *Graf::NapracVrcholmi()* sa sieť postupne podľa nastavených parametrov vytvorí a uloží do matice susedností, ktorou je reprezentovaná. Následne môžu byť volané už spomínané metódy *Graf::Export**, ktoré vypočítajú a exportujú požadované veličiny siete.

Ako náhodný generátor čísiel som použil voľne dostupnú triedu implementujúcu algoritmus *MersenneTwister*, ktorá je šíriteľná pod licenciou BSD a stiahnuteľná na stránke autora [14].

6.3 Úpravy programu

Pri písaní programu som sa snažil, aby jednotlivé jeho časti boli od seba nezávislé a teda aby jeho rozšírenie o niektoré nové pravidlá, merania vlastností, a pod. boli čo najjednoduchšie. V budúcnosti je teda možné program rozšíriť, implementovať nové modely a sledovať iné ich vlastnosti podobným spôsobom.

Kapitola 7

Záver

V tejto práci som zhrnul a popísal najzákladnejšie princípy, podľa ktorých sa môžu siete vytvárať a následne rásť. Tieto princípy mali väčšinou motiváciu prebranú z reálnych sietí, ktorých spôsob rastu sa snažili napodobniť. Charakteristické vlastnosti týchto modelov, ktoré som v jednotlivých kapitolách popísal, nám umožňujú tieto modely rozlišovať a následne podľa nameraných vlastností konkrétnej siete predpokladať, podľa ktorého z týchto modelov mohla sieť vznikať. Ako ilustračný príklad som použil sieť vznikajúcu zo slov anglického jazyka, zachytávajúcu vzťahy medzi jednotlivými slovami. Aj na tomto príklade bolo vidieť, že niektoré vlastnosti reálnych sietí sa dajú odvodiť iba na základe jednoduchých pravidiel, podľa ktorých sa tieto siete vytvárajú, aj keď samotné tieto siete môžu byť veľmi zložité, ako to pri uvedenom jazykovom grafe určite je.

Niektoré vlastnosti sietí sa teoreticky odvodiť nepodarilo, preto je užitočné používať na ich určenie numerické simulácie. Na ich vytvorenie som vytvoril program, ktorý tieto vlastnosti na simuláciách odmieria, prípadne takýchto meraní spraví väčší počet a vypočíta ich priemerné hodnoty. Problém pri takomto prístupe je predovšetkým v technických možnostiach, pretože vytvárané siete musia byť na dostatočné ustálenie meraných parametrov buď veľmi veľké, alebo sa ich musí vytvoriť veľké množstvo. Preto je časová aj pamäťová náročnosť tohoto programu pomerne vysoká a dalo by sa pracovať

na jeho optimalizácii, čo by nám umožnilo vytvárať siete väčších rozmerov.

Možné rozšírenia Modely, ktoré som v tejto práci popisoval, patrili medzi najzákladnejšie. Existuje množstvo ich rozšírení, zavedení nových princípov, úprav a zmien, ktoré budú viesť k vystihnutiu niektorých ďalších charakteristických vlastností reálnych sietí. Taktiež som sa v tejto práci venoval iba elementárnym vlastnostiam sietí, a to predovšetkým distribúcii stupňov vrcholov a niektorým parametrom týkajúcich sa klasterizácie siete. Práve vlastnosti spojené s klasterizačnými koeficientami sa v poslednom čase ukazujú ako dobre charakterizujúce štruktúru a hierarchiu siete a preto je vhodné ich ďalšie skúmanie. Tiež by bolo potrebné preskúmať niektoré ďalšie vlastnosti sietí, napriklad spektrum vlastných hodnôt matice susednosti vrcholov[7], priemernú dĺžku cesty medzi vrcholmi a ďalšie.

Literatúra

[1] Bornholdt, S., Schuster, H. G. : *Handbook of Graphs and Networks: From the Genome to the Internet* Wiley-VCH ISBN-13: 978-3527403363 (2003)

[2] Dorogovtsev, S. N., Mendes, J. F. F. : *Evolution of networks* Journal-ref: Adv. Phys. 51, 1079 (2002)

[3] Almaas, E., Barabási, A-L. : *Power laws in biological networks* Review article, to appear in "Power laws, scale-free networks and genome biology"

[4] Markošová, M. : *Language as a graph*

[5] Markošová, M. : *Modely jazyka ako dynamickej siete*

[6] Markošová, M. : *Network model of human language* Journal-ref: Physica A 387 661-666 (2008)

[7] Barabási, A-L., Albert, R. : *Statistical mechanics of complex networks* Journal-ref: Reviews of Modern Physics 74, 47 (2002)

[8] Dorogovtsev, S. N., Mendes, J. F. F. : *Language as an Evolving Word Web.* Proc. Royal Soc. London B, 268, 2603 (2001)

[9] Dorogovtsev, S. N., Mendes, J. F. F. : *Scaling properties of scale-free evolving networks: Continuous approach* Journal-ref: Phys. Rev. E 63, 056125 (2001)

[10] Watts, D. J., Strogatz, S.H. : *Nature* 393, 440 (1998)

[11] Barabási, A-L., Albert, R. : *Emergence of Scaling in Random Networks* Journal-ref: Science 286, 509 (1999)

[12] Ravasz, E., Barabasi, A-L. : *Hierarchical Organization in Complex Networks* Journal-ref: Phys. Rev. E 67, 026112 (2003)

[13] Matoušek, J., Nešetřil, J. : *Kapitoly z diskrétní matematiky* Karolinum (2007) ISBN: 978-80-246-1411-3

[14] http://www-personal.umich.edu/∼wagnerr/MersenneTwister.html

[15] http://www.gutenberg.org