Uzorkovanje na kompleksnim mrežama metodom snježne grude

Batur, Josip

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET

FIZIČKI ODSJEK

Josip Batur

UZORKOVANJE NA KOMPLEKSNIM MREŽAMA METODOM SNJEŽNE GRUDE

Diplomski rad

Zagreb, 2019.

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET

FIZIČKI ODSJEK

INTEGRIRANI PREDDIPLOMSKI I DIPLOMSKI SVEUČILIŠNI STUDIJ
FIZIKA; SMJER ISTRAŽIVAČKI

Josip Batur

Diplomski rad

UZORKOVANJE METODOM SNJEŽNE GRUDE NA KOMPLEKSNIM MREŽAMA

Voditelj diplomskog rada: prof. dr. sc. Hrvoje Štefančić

Ocjena diplomskog rada: _____

Povjerenstvo: 1. _____

2. _____

3. _____

Datum polaganja: _____

Zagreb, 2019.

Zahvaljujem se mentoru Hrvoju Štefančiću na strpljivom vođenju, prijedlozima i savjetima koji su doprinijeli nastanku ovog rada. Hvala komentoru Davoru Horvatiću na prepravkama i savjetima.

Hvala mojoj divnoj obitelji, roditeljima Zlatku i Dragici te sestri Anđeli na bezrezervnoj podršci i ljubavi kroz sve ove godine. Mojim babama, Luci i Iki, kojim će biti ovo biti posebno drago.

Dragom prijatelju i kolegi Antoniu Supini, na pomoći i beskrajnim kavama na kojima smo najčešće razglabali ni o čemu.

Luci, Martini, Luji, Josi i Eni, jer su najbolji prijatelji koje čovjek može imati i jer sam od njih naučio više nego što su i sami svjesni.

Svim ostalim prijateljima i rodbini, koji su bili, ili još uvijek jesu, dio mog života.

Sažetak

Veliki broj vrlo raznorodnih kompleksnih sustava koje proučavanju fizika, kemija, biologija, sociologija, ekonomija, računalne i druge znanosti mogu se prikazati kao kompleksne mreže. Istraživanje strukture i dinamike kompleksnih mreža, kao i dinamike procesa koji se odvijaju na kompleksnim mrežama predstavlja jedno od najpropulzivnijih interdisciplinarnih područja moderne znanosti. U mnogim situacijama je od važnosti uzorkovanjem istraživati populaciju za koju su odnosi među jedinkama predstavljeni kompleksnom mrežom. Cilj teme diplomskog rada je istražiti svojstva postupka uzorkovanja metodom snježne grude na kompleksnim mrežama te ispitati da li se i pod kojim uvjetima metoda snježne grude može približiti po svojim svojstvima slučajnom uzorkovanju na kompleksnim mrežama. Naime, metoda snježne grude se standardno klasificira kao neprobabilistička metoda uzorkovanja kod koje iz uzorka nije moguće pouzdano zaključivanje o populaciji. S druge strane, uzorkovanje metodom snježne grude je često jedini realistični pristup uzorkovanju ukoliko je ciljanoj populaciji onemogućen ili otežan pristup, pa tako i probabilističke metode uzorkovanja poput slučajnog uzorkovanja. Specifični cilj diplomskog rada je kvantifikacija razlike slučajnog uzorkovanja i uzorkovanja metodom snježne grude na kompleksnim mrežama. Istraživanje će se provoditi simulacijama uzorkovanja na računalno generiranim kompleksnim mrežama.

Ključne riječi: kompleksne mreže, uzorkovanje, raspodjela, stupanj čvora

Snowball sampling method on complex networks

Abstract

A large number of various complex systems in physics, chemistry, biology, sociology, economics, computer and other sciences can be represented as complex networks. Studies of structure and dynamics of complex networks, as well as dynamics of processes that take place on complex networks represent one of the most propulsive interdisciplinary areas of modern science. In complex systems, where relations and interactions between its constituent units are represented by complex network, there are many situations when sampling of population is of great importance. The goal of this theses is to examine properties of snowball sampling on complex networks and to find out under which conditions (if any) can snowball sampling assessments get close to random sampling on complex networks. Particularly, snowball sampling method is classified as non-probabilistic, meaning it is not possible to reliably conclude about some property distribution in population through snowball sample. On the other hand, snowball sampling is often the only realistic sampling method, particularly when there is limited or no approach to the population of interest. Specific task of this thesis is quantification of difference between random sampling and snowball sampling on complex networks. Research will be conducted by simulations of sampling on computer generated complex networks.

Keywords: complex networks, sampling, distribution, node degree

Sadržaj

1 Uvod

U modernoj znanosti od izuzetne je važnosti analiza kompleksnih sustava. Sam termin obuhvaća sustave čije je ponašanje teško modelski opisati zbog ovisnosti, povezanosti, sukoba i drugih interakcija između njihovih sastavnih komponenti. Kompleksni sustavi imaju jedinstvena svojstva koja proizlaze iz interakcija među komponentama, kao što su spontano uređenje, nelinearnost i sl. Isti mogu biti predmetom proučavanja raznorodnih znanstvenih djelatnosti, od prirodnih (fizika, kemija, biologija...) do društvenih (geografija, sociologija, politika).

Kako bi definicija bila jasnija, oprimjerimo je sa nekoliko sustava. Kao prvi primjer navedimo World Wide Web, kojeg, pojednostavljeno, možemo promatrati kao kolekciju web stranica, a njihove interakcije kroz povezanost neke stranice sa drugima u kolekciji. Nadalje, spomenimo ljudski živčani sustav, sastavljen od neurona povezanih sinaptičkim vezama. Sva naša promišljanja zahtjevaju koherentnu aktivnost i povezanost milijardi neurona. Naša biološka postojanost, pak, ovisi o neprimjetnom mehanizmu uputa izmjenjenih između tisuća gena i metabolita u našem tijelu. Spomenimo za kraj električnu mrežu generatora koja dostavlja energiju gotovo svoj modernoj tehnologiji.

Svi ovi sustavi imaju nešto zajedničko: njihovo kolektivno ponašanje jako je teško opisivo ako poznajemo samo njihove sastavne komponente. Uzevši u obzir ulogu koju u našim svakodnevnim životima, znanosti i ekonomiji igraju kompleksni sustavi, poznavanje njihovog matematičkog opisa, strukture, povezanosti i interakcija u istima predstavlja jedan od najvećih izazova znanosti 21.st.

Kompleksne mreže predstavljaju jednu od najjednostavnijih reprezentacija kompleksnih sustava. Pogodne su jer, u principu, pojednostavljuju izuzetno kompliciran sustav na samo njegove diskretne djelove (čvorove) i poveznice između tih dijelova. Od svojih matematičkih počela u teoriji nasumičnih grafova [1] [2], preko prvih primjena modela i fenomena *malog svijeta* (engl. *small world phenomenon*) [3] [4], kompleksne mreže svoj zvjezdani status dobivaju modelima *mreža bez skale* mreža [5] [6] i njihovima primjenama [7] [8].

Jednu od novih i uzbudljivih podgrana istraživanja kompleksnih mreža predstavlja uzorkovanje na kompleksnim mrežama. Želimo li doznati karakteristike i svojstva pojedinih sustava kada nam je kompletan sustav nedostupan ili je njegova analiza ne-

isplativa, to možemo napraviti uzimanjem uzoraka, tj. izdvajanjem određenog broja komponenti kojima izvrjednjujemo karakteristike od interesa za čitav sustav. Npr, želimo li saznati političke preference i šanse pojedinih političkih stranaka na izborima, to možemo učiniti tako što ćemo ispitati preference M broja ljudi u populaciji države veličine N.

U prvom dijelu ovog rada predstavit ćemo detaljnije kompleksne mreže i neka njihova najbitnija svojstva, kao što su stupanj vezanja čvora i matrica susjednosti. Predstavit ćemo i *Dijkstra* algoritam traženja najkraćeg puta između dvaju čvorova u mreži.

Nadalje, izložit ćemo temeljne karakteristike Erdős - Rényi i Barabási - Albert kompleksnih mreža, na kojima smo računalnim putem provodili uzorkovanje. Potom ćemo ukratko opisati metodologiju dva tipa uzorkovanja koja ćemo provoditi na gore navedenim kompleksnim mrežama: nasumično uzorkovanje (engl. *random sampling*) i uzorkovanje metodom snježne grude (*snowball sampling*).

U zadnjem dijelu rada predstavit ćemo rezultate i usporedbu dva tipa uzorkovanja na računalno generiranim Barabási - Albert i Erdős - Rényi mrežama, za dva tipa distribucije svojstava. Za kraj, predstavit ćemo "*random-snowball* sukob", svojevrsnu "utrku" između dva tipa uzorkovanja gdje je "pobjednik" ono uzorkovanje koje za danu kvadratnu kompleksnu mrežu i "pravila igre" prvo postigne *perkolaciju* na mreži, što može biti od velikog interesa u raznim područjima fizike.

2 Kompleksne mreže

Kompleksnu mrežu kao reprezentaciju kompleksnih sustava predočavamo čvorovima (engl. *nodes*) koji predstavljaju diskretne komponente takvih sustava i poveznicama (engl. *links*) koji predstavljaju interakcije među diskretnim komponentama.

Budući da mogu predstavljati raznorodne kompleksne sustave, kompleksne mreže su po svojoj tvorbi raznovrsne, tako da čvorovi i poveznice u različitim reprezentacijama predstavljaju različite komponente i interakcije. Tako, npr., čvorovi metaboličke mreže čovjeka predstavljaju sitne molekule, dok poveznice predstavljaju interakcije među istima vođene zakonima kemije i kvantne mehanike. One su nastale kao rezultat milijuna godina evolucije. Čvorovi World Wide Weba su web dokumenti čije su poveznice specifični lokatori (URL-ovi). Ovaj sustav nastao je suradnjom i interakcijom milijuna korisnika, kako individualnih tako i institucionalnih.

Različitosti u prirodi, veličini, povijesti i nastanku pojedinih kompleksnih sustava navodi na pomisao kako su kompleksne mreže koje ih predstavljaju uvelike različite. No, elegancija i upotrebljivost koncepta kompleksnih mreža dolazi od empirijski utvrđene činjenice da je arhitektura mreža u raznorodnim područjima znanosti ista ili vrlo slična. Posljedično možemo koristiti zajednički skup matematičkih svojstava i njime opisati raznorodne kompleksne sustave.

U ovom poglavlju predstavit ćemo ključne koncepte koji definiraju kompleksnu mrežu i neke korisne alate koje smo koristili u istraživanju.

2.1 Mrežna svojstva

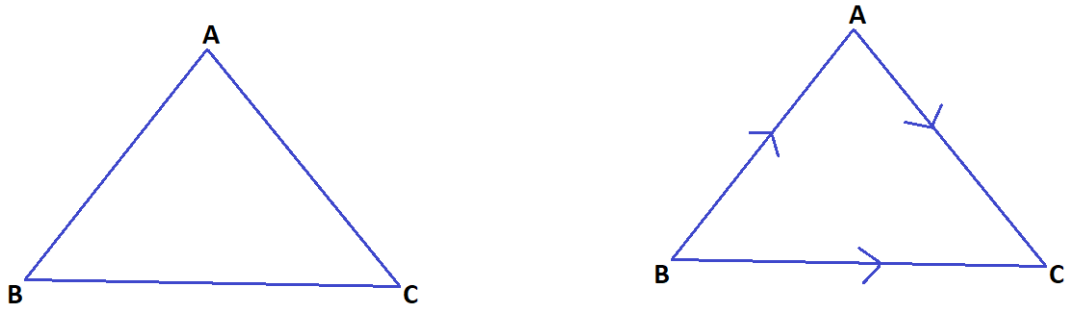
Kako bismo imali dobre temelje za analizu kompleksnih mreža, potrebno je predstaviti neke osnovne parametre kojima definiramo kompleksne mreže. Kvantificirajmo za početak gradivne jedinice kompleksne mreže :

- Broj čvorova u mreži označavamo sa *N*. Ovaj parametar opisuje veličinu mreže, tj. broj diskretnih elemenata koje imamo na raspolaganju. Također, služi nam kako bismo numerirali i samim time razlikovali čvorove ($i = 1, 2, 3...N$).

- Broj poveznica u kompleksnoj mreži *L* govori nam o ukupnom broju interakcija u kompleksnom sustavu. Njih obično ne označavamo posebno, budući da je pojedinu poveznicu dovoljno definirati početnim i krajnjim čvorom koje povezuje.

Tako, npr. poveznica (*l*,*k*) je ona koja spaja *l*-ti s *k*-tim čvorom.

Poveznice u kompleksnoj mreži, zavisno od interakcija koje predstavljaju, mogu biti *usmjerene* i *neusmjerene* (Slika 2.1.). Neusmjerene mreže podrazumijevaju uzajamnu interakciju dvaju čvorova, dakle poveznice nemaju preferirani *smjer*. Poveznica između dva čvora u ovakvoj mreži ima svoj ”izvor” i ”ponor” u oba čvora koja povezuje. Promatramo li, npr. društvenu mrežu Facebook, korisnici predstavljaju čvorove dok su poveznice između njih *prijateljstva* (engl. *friendship*). *Prijateljstvo* na ovoj društvenoj mreži je uzajamna interakcija, budući da ne možete biti nekome *prijatelj* tko nije *prijatelj* vama.



Slika 2.1: Grafički prikaz neusmjerene (lijevo) i usmjerene (desno) kompleksne mreže s $N=3$ čvora.

Usmjerene mreže, pak, imaju jednosmjerne interakcije i samim time jednosmjerne poveznice. Možemo reći kako poveznica ”izvire” u jednom čvoru i ”ponire” u drugom. Uzmimo opet Facebook kao primjer gdje jednu od interakcija opisuje *praćenje* (engl. *following*). *Praćenje* je često jednosmjerna interakcija, budući da osoba koju vi *pratite*, ne mora *pratiti* vas.

Za neusmjerenu mrežu, ukupan broj poveznica može biti izražen kao suma poveznica po pojedinom čvoru :

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i \,, \tag{2.1}$$

gdje k_i predstavlja *stupanj vezanja čvora* (dalje u tekstu kao ”*stupanj čvora*”). Faktor $\frac{1}{2}$ dolazi zbog činjenice da smo u sumi poveznice prebrojali dvaput, što je direktna

posljedica neusmjerenosti grafa. Od interesa je i definirati srednji stupanj čvora $\overline{k}$:

$$\overline{k} = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}\,. \tag{2.2}$$

Za usmjerene mreže pak, moramo definirati dva stupnja čvora; ulazni k_i^{in} koji označava broj poveznica koje usmjeravaju na *i*-ti čvor (”ponor”) i izlazni k_i^{out} koji označava broj poveznica koje izlaze iz *i*-tog čvora (”izvor”). Ukupan stupanj čvora k_i možemo prikazati kao zbroj ulaznog i izlaznog stupnja:

$$k_i = k_i^{in} + k_i^{out}\,. \tag{2.3}$$

Želimo li saznati ukupan broj poveznica u mreži, moramo prosumirati sve izlazne ili sve ulazne čvorove :

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}\,. \tag{2.4}$$

Primjećujemo kako u gornjoj jednadžbi nemamo faktor $\frac{1}{2}$ budući da ulazne i izlazne čvorove brojimo zasebno pa se dvostruko prebrojavanje ne dogodi. Sume su jednake budući da svaka poveznica ima i ”izvor” i ”ponor”. Također, srednje stupnjeve izlaznih i ulaznih čvorova možemo izračunati kao :

$$\overline{k_i^{in}} = \frac{1}{N}\sum_{i=1}^{N} k_i^{in} = \overline{k_i^{out}} = \frac{1}{N}\sum_{i=1}^{N} k_i^{out} = \frac{L}{N}\,. \tag{2.5}$$

U ovom radu od interesa su samo neusmjerene mreže, tako da ćemo se od sada bazirati samo na njih. Pretpostavljat ćemo, npr. da su u nekoj socijalnoj strukturi poznanstva uzajamna, da su interakcije bioloških molekula dvostrane itd.

Sljedeći parametar od interesa je distribucija stupnja čvorova p_k, koja predstavlja vjerojatnost da nasumično izvučeni čvor kompleksne mreže ima stupanj *k*. Vjerojatnost mora biti normirana na 1 :

$$\sum_{k} p_k = 1\,. \tag{2.6}$$

Općenito, poznajemo li za svaki čvor njegov stupanj, distribuciju p_k možemo do-

biti prebrojavanjem čvorova s istim stupnjem :

$$p_k = \frac{N_k}{N}\,, \tag{2.7}$$

gdje N_k predstavlja broj čvorova sa stupnjem k. S druge pak strane, poznajemo li distribuciju p_k, broj čvorova za dani stupanj možemo dobiti kao :

$$N_k = p_k N\,. \tag{2.8}$$

Distribucija p_k odigrala je jednu od ključnih uloga u znanosti o kompleksnim mrežama, poglavito otkrićem *mreža bez skale* [5]. Poznavanjem nje možemo izračunati neke od temeljnih značajki stupnja vezanja, a samim time i kompleksne mreže. Tako, npr. srednji stupanj čvora $\overline{k}$ možemo izračunati izrazom:

$$\overline{k} = \sum_k k p_k\,. \tag{2.9}$$

Naravno, dobar opis opis neke distribucije podrazumijeva i računanje viših momenata (standardna devijacija, asimetrija...) za što nam također služi raspodjela p_k.

2.2 Matrica susjednosti

Kako bismo lakše popisali poveznice u sustavu koristimo se *matricom susjednosti* A_{ij} iz koje možemo lako isčitati koliki je stupanj pojedinog čvora, ali isto tako koji su čvorovi međusobno povezani. Elementi matrice su definirani kao :

- $A_{ij} = 1$ ako su *i*-ti i *j*-ti čvor povezani;
- $A_{ij} = 0$ ako *i*-ti i *j*-ti čvor nisu povezani.

Za neusmjerene mreže, koje su u ovom radu područje našeg interesa, vrijedi uzajamna interakcija dvaju čvorova, pa je tako $A_{ij} = A_{ji}$, tj. naša matrica je simetrična. Na slici 2.2 dan je primjer jednostavne neusmjerene mreže s $N = 4$ čvora. Kompleksna mreža sa N čvorova imat će pripadnu NxN matricu (N redaka i N stupaca) gdje redovi općenito označavaju čvor iz kojeg poveznica izlazi, a stupci čvorove na koje poveznice upućuju. Za neusmjerene mreže, simetričnost matrice slijedi iz činjenice da su svakoj poveznici između dva čvora oba čvora i ”izvor” i ”ponor”.

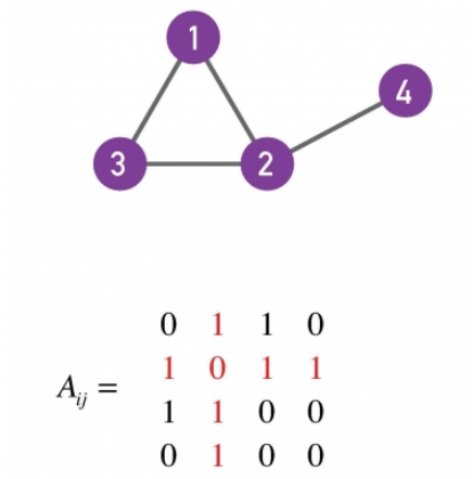
Uz pomoć matrice susjednosti možemo dobiti stupanj vezanja i-tog čvora. Za neusmjerene mreže, možemo odabrati hoćemo li sumirati i-ti redak ili i-ti stupac :

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{j=1}^{N} A_{ji} \tag{2.10}$$

Općenito, za usmjerene mreže ovo nije slučaj, budući da matrica susjednosti za njih nije simetrična. Za kraj, spomenimo kako je broj elemenata različitih od 0 u matrici dvostruko veći od broja poveznica L :

$$2L = \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} \, . \tag{2.11}$$

Matrica susjednosti uvelike olakšava generiranje sintetičkih mreža i rad na istima kada zadatak zahtjeva detaljno poznavanje strukture kompleksne mreže. Mi smo je uvelike koristili pri računalnom generiranju situacija od interesa i prilikom numeričkog obrađivanja rezultata značajnih za ovaj rad. Njihov kompaktan zapis i jednostavnost pristupanja pojedinom elementu svakako ubrzavaju i samu izvedbu računalnog koda.



Slika 2.2: Jednostavna neusmjerena mreža zajedno sa pripadnom matricom susjednosti A_{ij}. Preuzeto sa dozvolom autora [9]

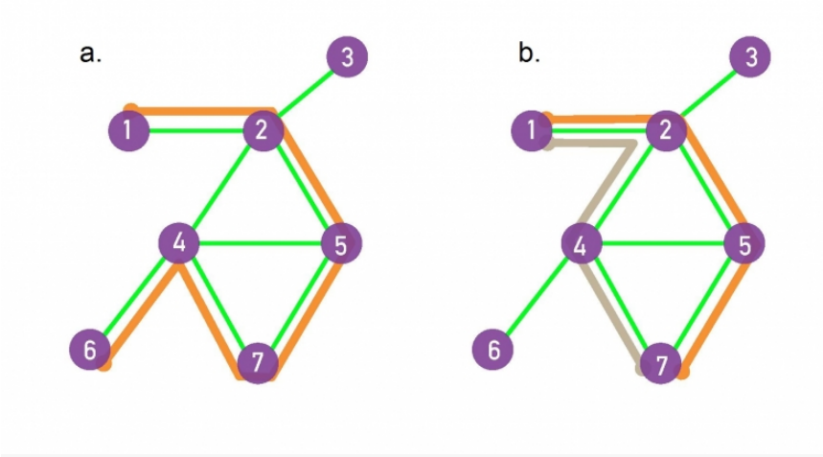
2.3 Mrežni putevi i povezanost

U fizikalnim, biološkim i kemijskim kompleksnim sustavima fizička udaljenost je od krucijalnog značaja. Uzmemo li za primjer kristalnu rešetku nekog kemijskog spoja,

međudjelovanje atoma uvelike ovisi o udaljenosti na kojoj se atomi nalaze jedan u odnosu na drugog.

Postavlja se pitanje: kako bismo definirali udaljenost u kompleksnim mrežama? Naime, u nekim kompleksnim sustavima, kakav je npr. Facebook, dva korisnika mogu biti na različitim dijelovima zemaljske kugle, a da su opet Facebook *prijatelji*, tj da postoji direktna poveznica između ta dva korisnika. S druge pak strane, osobe koje su susjedi u istoj zgradi uopće ne moraju biti *prijatelji* na Facebooku, dakle ne postoji direktna poveznica između njih dvoje. Ovo nas upućuje kako je u nekim mrežama fizička udaljenost od malog značaja.

Put između dva čvora *i* i *j* u kompleksnoj mreži je definiran povezanim čvorovima u kompleksnoj mreži kroz koje moramo proći kako bismo stigli od *i*-tog do *j*-tog čvora. Udaljenost definiramo kao broj poveznica koje spajaju dva krajnja čvora od interesa. Na Slici 2.3. možemo vidjeti kako unutar iste kompleksne mreže možemo imati više puteva koji povezuju dva čvora. Tako npr, vidimo da od čvora 1 do čvora 6 na Slici 2.3. možemo stići putem (1-2-5-7-4-6), čija je udaljenost $n = 5$ ali isto tako i putevima (1-2-4-6), udaljenosti $n = 3$ i (1-2-5-4-6), udaljenosti $n = 4$.[1]



Slika 2.3: Shematski prikaz različitih puteva unutar jednostavnih kompleksnih mreža. Preuzeto sa dozvolom autora [9]

Često je od praktičnog interesa razmatrati najkraći put između dva čvora u mreži, d_{ij}. U principu, d_{ij} može biti izračunat direktno iz matrice susjednosti. Tako, npr, ukoliko postoji direktna poveznica između *i* i *j*, najkraći put će imati udaljenost $d_{ij} = 1$, budući da postoji matrični element $A_{ij} = 1$. Ukoliko je $A_{ij} = 0$, između čvorova ne postoji put duljine $d_{ij} = 1$.

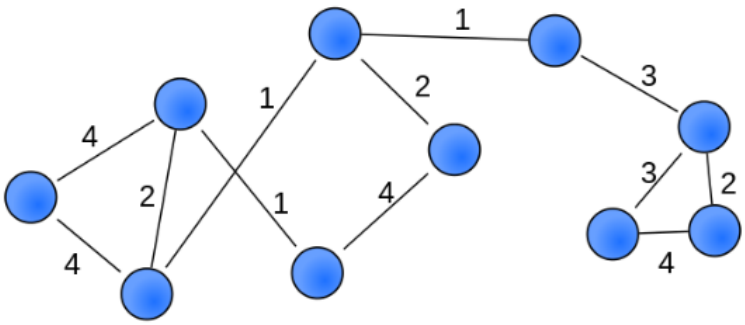
[1] Ovdje smo, radi jednostavnosti, zanemarili puteve koji dva ili više puta prolaze kroz isto čvorište. Tako, npr. put (1-2-3-2-4-6) ne razmatramo.

Za najkraći put duljine $d_{ij} = 2$ mora postojati par matričnih elemenata A_{ik} i A_{jk} takvih da je $A_{ik}A_{jk} = 1$. Ukoliko je njihov umnožak 0, tada put duljine 2 ne postoji i moramo tražiti dalje. Općenito, broj puteva duljine d između dva čvora u mreži možemo naći preko formule :

$$N_{ij}^{(d)} = A_{ij}^{d} \, . \tag{2.12}$$

Ovakva metoda pretraživanja je jednostavna i pogodna na malenim, pokaznim mrežama, međutim do problema nailazimo želimo li naći najkraći put u realnim mrežama, koje veličinom uvelike nadmašuju shematske. Za tu svrhu potrebni su nam algoritmi pretrage.

Općenito, različite poveznice u realnim mrežama nemaju isti kapacitet (Slika 2.4.), pa ih razlikujemo težinama w_{ij} koje mogu označavati njihovu snagu, intenzitet, prostornu udaljenost itd. [10] [11]. Ovakve mreže nazivamo *težinske mreže*.



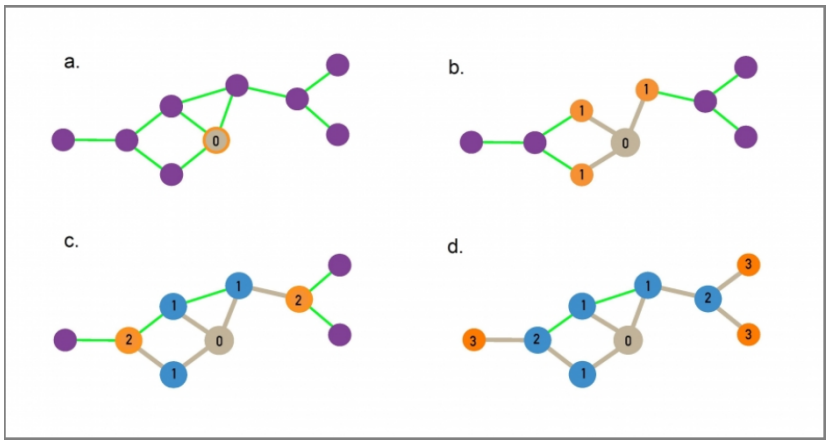
Slika 2.4: Primjer *težinske mreže*. Brojevi iznad poveznica su težinski faktori w_{ij}. Preuzeto s dozvolom autora [9]

Želimo li saznati najkraći put između dvaju čvorova na velikim, realnim i težinskim mrežama, pogodan je *Dijkstra* algoritam, nazvan po računalnom znanstveniku Edsgeru W. Dijkstri koji ga je 1956. otkrio i opisao [12]. U puno realnih slučajeva mrežne poveznice moramo ponderirati njihovim težinama, pa je u većini problema *Dijkstra* algoritam od koristi, kao npr. u VLSI *routing* problemu [13]. Za detaljniji opis *Dijkstra* algoritma i primjena vidjeti [14].

Međutim, u ovom radu koristimo se specifičnim slučajem, aproksimacijom pri kojoj su sve težine u mreži jednake $w_{ij} = 1$. U ovom slučaju koristimo specifični slučaj *Dijkstra* algoritma, *BFS* algoritam (engl. *breadth first search*) (Slika 2.5), koji ćemo i detaljnije opisati.

Koraci u *BFS* algoritmu su sljedeći :

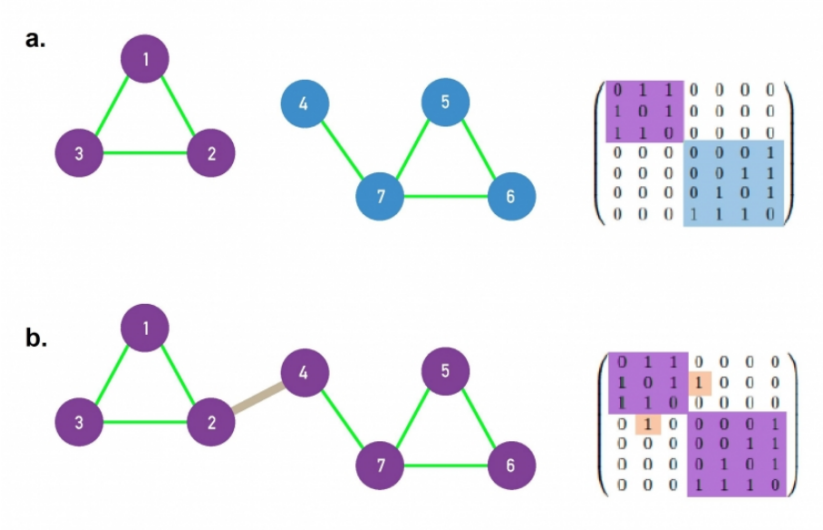
- Krenimo od *i*-tog čvora, početka našeg puta i označimo ga s $n = 0$.
- Prepoznajmo njegove susjede, označimo ih s $n = 1$ i stavimo ih u *neposjećeni skup*.
- Uzmimo prvog susjeda označenog s $n = 1$, pronađimo njegove susjede i maknimo ga iz *neposjećenog skupa*. Susjede označavamo s $n = 2$ i stavljamo ih u *neposjećeni skupa*.
- Ponavljamo treći korak sve dok ne pronađemo *j*-ti čvor ili dok ne iscrpimo *neposjećeni set*.
- Udaljenost čvorova je oznaka n za *j*-ti čvor. Ukoliko čvor nema oznake, ne postoji put između dva čvora i pišemo $d_{ij} = \infty$



Slika 2.5: Shematski prikaz BFS algoritma. Preuzeto sa dozvolom autora [9]

Udaljenost čvorova d_{ij} pomaže nam pri definiciji pojma *povezivosti*. Kompleksnu mrežu u kojoj možemo pronaći barem jedan par čvorova koji nisu povezani ($d_{ij} = \infty$) zovemo *nepovezanom*, a njezine podmreže zovemo *komponentama*. Komponente su podskupovi čvorova u kojima su svi čvorovi međusobno povezani. U suprotnom, ako je za svaki *i* i *j*, $d_{ij} \neq \infty$, mreža je *povezana*. Slika 2.6. daje nam dobru skicu nepovezane i povezane kompleksne mreže, zajedno sa pripadnim matricama susjednosti. Na slici vidimo kako umetanjem poveznice (*mosta*) između čvorova 2 i 4 možemo iz nepovezane strukture prijeći u povezanu. Općenito, *most* je poveznica koju kad izrežemo iz mreže, mreža postaje nepovezana.

Općenito, matricu poveznica nepovezane mreže možemo napisati preko blok dijagonalne matrice (Slika 2.6.a), takve da su svi članovi različiti od 0 grupirani u blokove duž dijagonale matrice, dok su svi članovi van blokova 0. Blokovi u ovom slučaju predstavljaju upravo *komponente* mreže.



Slika 2.6: Shematski prikaz nepovezane (a.) i povezane (b.) kompleksne mreže zajedno sa pripadnim matricama poveznica. Preuzeto sa dozvolom autora [9].

Općenito, u velikim mrežama, istraživanje povezanosti može postati problematičnim u okviru linearne algebre. Tada je pogodno koristiti gore opisani *BFS* algoritam, alat koji je i u ovom radu korišten prilikom istraživanja "*random-snowball* sukoba".

Kako bi kompletirali osnovne pojmove u ovom uvodu o kompleksnim mrežama, spomenimo i *lokalni koeficijent grupiranja*. Istraživanja na realnim sociološkim mrežama pokazuju kako njihovi čvorovi imaju tendenciju snažnog povezivanja u *komponente* (*klastere*) sa visokom gustoćom poveznica [3] [15]. Za čvor *i* sa pripadnim stupnjem k_i, lokalni koeficijent grupiranja opisujemo sa :

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{2.13}$$

gdje je L_i broj poveznica između susjeda zadanog čvora. Primjećujemo kako C_i varira između 0 i 1, gdje $C_i = 0$ označava stanje u kojem ne postoji nijedna poveznica između čvorova susjednih *i*-tom, a stanje $C_i = 1$ označava da su svi čvorovi susjedni *i*-tom međusobno povezani. Ukratko, C_i opisuje gustoću vezanja promatrane *grupe članova*.

Sada kada smo predstavili bazične koncepte vezane uz kompleksne mreže, spremni smo predstaviti dvije klase mreža : *nasumične* (*engl. random*) i mreže *bez skale* (engl.

scale-free network) zajedno sa svojim najpoznatijim modelima.

3 Erdős - Rényi kompleksna mreža

Glavna zadaća mrežne znanosti je modeliranje ponašanja realnih kompleksnih sustava. Velik dio stvarnih kompleksnh sustava nema jasan obrazac interakcije među komponentama, a samim time mrežna reprezentacija raspodjele poveznica među čvorovima izgleda nasumično. Ovo je osnovna motivacija za modeliranje nasumičnih mreža.

Anatol Rapoport (1911.-2017.) bio je prvi koji je počeo proučavati nasumične mreže. Motiviran problematikom matematičke biologije, točnije povezanosti neurona, objavio je, zajedno s Rayom Solomonoffom, ključni rad [1] u kojem je pokazao da, povećavamo li srednji stupanj čvora $\overline{k}$, dolazi do faznog prijelaza od nepovezanog grafa do grafa s divovskom komponentom (engl. *giant component*).

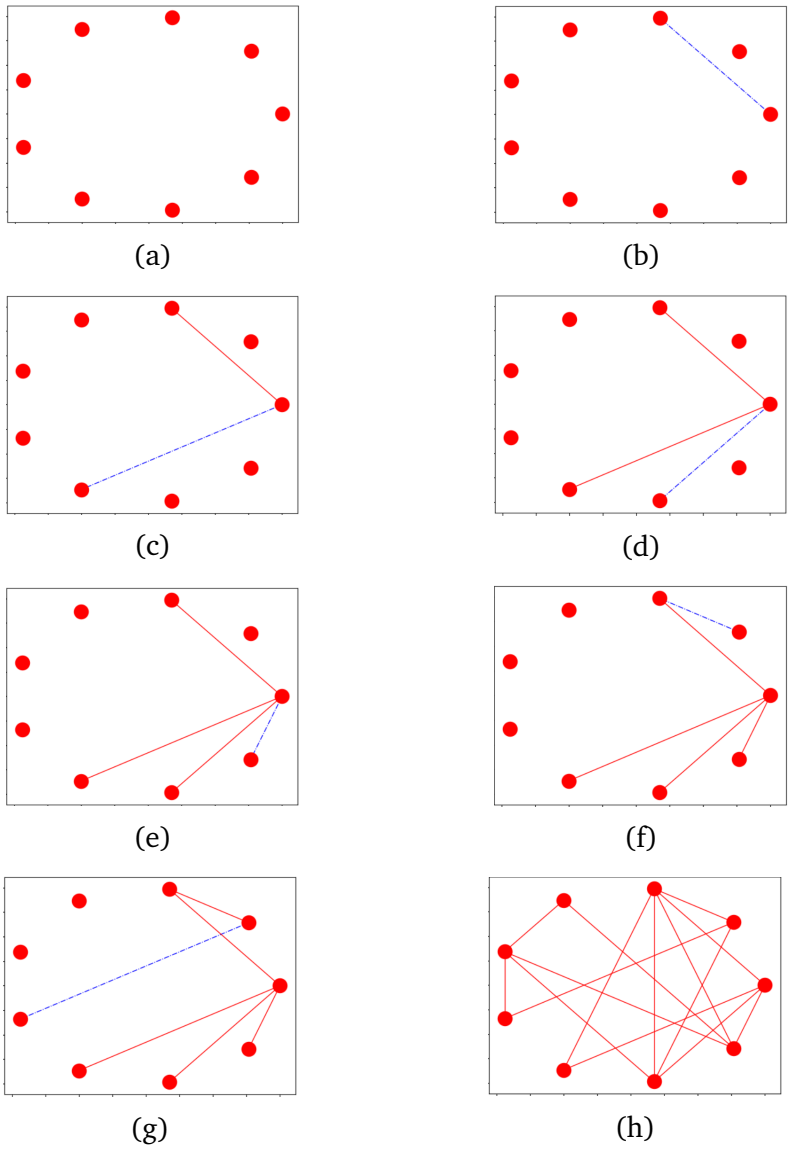
Znanost nasumičnih mreža svoj uzlet dobiva radovima dvojice mađarskih matematičara, Paula Erdősa i Alfreda Rényia, od kojih je najznačajniji i najpoznatiji *O evoluciji nasumičnih grafova* [2], objavljenom 1960., gdje je dvojac uporabom teorije vjerojatnosti i teorije grafova iznio neke glavne značajke nasumičnih mreža. Njima u čast, najpoznatiji model nasumičnih mreža nazivamo *Erdős - Rényi* mrežom (dalje u tekstu kao E-R mreža).

Polazišna osnova nasumičnog E-R modela, je da se on sastoji od N označenih čvorova između kojih sa vjerojatnošću p nasumično stavljamo poveznice. Pokazat ćemo kako uz pomoć ovih pretpostavki generiramo E-R mrežu. Potom ćemo detaljnije izvesti kako izgleda distribucija stupnja čvorova za ovakav model i kako evoluiraju nasumične mreže. U tom kontekstu upoznat ćemo se s pojmom *divovske komponente* i proučiti koji režimi mreža postoje te koji su od tih režima najbliži realnim mrežama. Za kraj, predstavit ćemo pojam *mrežne skale* i vidjeti što nam govori o raspodjeli stupnjeva čvora.

3.1 Generiranje mreže

E-R mrežu generiramo zadavanjem vjerojatnosti povezivanja dva čvora p i ukupnim brojem čvorova N. Kako bi uspjeli slikovito prikazati proces, uzeli smo $N = 9$ čvorova

i vjerojatnost $p = 0.5$ (Slika 3.1.). Algoritam prvo generira mrežu od N nepovezanih čvorova. Potom se proizvoljno odabere početni čvor (na slici krajnji desni čvor, po sredini slike) od kojeg se počinju dodavati poveznice (Slika 3.1. b)- e)).



Slika 3.1: Generiranje Erdős - Rényi kompleksne mreže. Slike a) - e) prikazuju dodavanje poveznica prvom čvoru, slike f) i g) nastavak procesa za drugi čvor, a slika h) konačan izgled mreže. Generirano u programu *Python 2.7*.

Poveznice se dodaju tako da se za svaki čvor generira nasumični broj između 0 i 1. Ako je generirani broj manji od p, dodaje se poveznica koja spaja taj čvor i početni. Na slici 1. vidimo kako su našem početnom čvoru dodani njegov 2., 6., 7., i 8. susjed (promatramo u smjeru obrnutom od kazaljke na satu), dok su 1., 3., 4. i 5. susjed "preskočeni". Proces se nastavlja, pa tako sada idući čvor u smjeru obrnutom od kazaljke na satu (Slika 1. f), g)). Važno je napomenuti kako se za svaki idući "početni" čvor prestaje "provjeravati" sa prethodnim "početnim" čvorovima, budući

da su oni već prije "provjereni". Slika 3.1. h) prikazuje konačan izgled kompleksne mreže.

Generirana mreža je pojednostavljena, premalena i sa prevelikim p da bi mogla opisivati fenomenologiju stvarnih mreža, no predstavlja dobar uvid u to kako je možemo proizvesti računalno. Dalje u radu upotrebljavali smo puno veće mreže i manje vjerojatnosti.

3.2 Distribucija stupnja čvora

U nasumičnoj mreži stupanj čvora nije uniforman, već varira od čvora do čvora. Tako, neki čvorovi imaju veliki broj poveznica, dok ih neki imaju jako malo ili ih ponekad nemaju uopće. Važan čimbenik koji opisuje ovo svojstvo je distribucija stupnja čvora p_k. Ona predstavlja vjerojatnost da pojedini čvor ima točno k poveznica s prema drugim čvorovima u mreži.

Kako bismo dobili izraz za distribuciju p_k potrebno je znati sljedeće :

- Vjerojatnost da *i*-ti čvor ima k poveznica je p^k;
- Vjerojatnost da *i*-ti čvor nije spojen sa preostalih $(N-1-k)$ čvorova je $(1-p)^{N-1-k}$;
- k poveznica među $(N-1)$ čvorova možemo raspodijeliti na $\binom{N-1}{k}$ načina.

Distribuciju stupnja čvora p_k dobivamo kao produkt gornja tri člana. Prepoznajemo da se radi o binomnoj distribuciji :

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \,. \tag{3.1}$$

Dakle, oblik ove distribucije ovisi o veličini mreže N i vjerojatnosti p. Znamo li samo ta dva parametra, za nasumičnu mrežu možemo izračunati srednji stupanj čvora $\overline{k}$:

$$\overline{k} = \sum_{k=1}^{N} k p_k = Np \,, \tag{3.2}$$

a preko srednje vrijednosti kvadrata stupnja čvora $\overline{k^2}$:

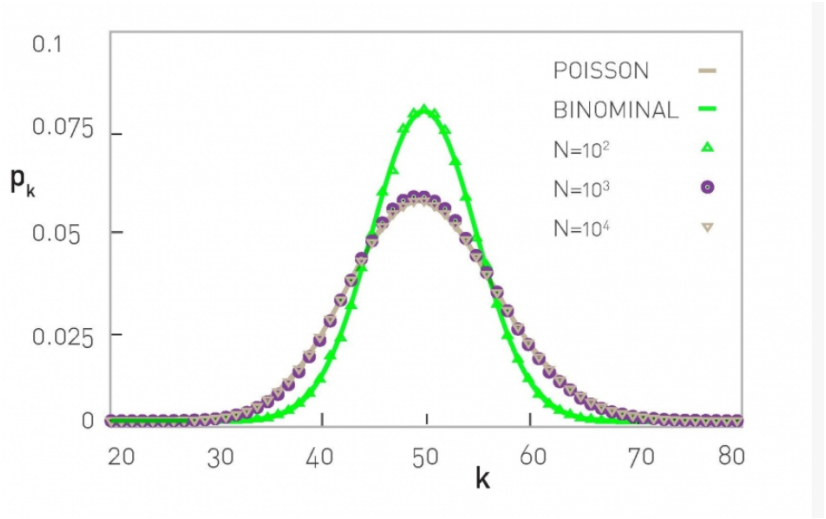
$$\overline{k^2} = \sum_{k=1}^{N} k^2 p_k = p(1-p)N + p^2 N^2 \,, \tag{3.3}$$

možemo izračunati varijancu σ_k, odnosno odstupanje raspodjele stupnja čvora od srednje vrijednosti $\overline{k}$:

$$\sigma_k = (\overline{k}^2 - \overline{k^2})^{\frac{1}{2}} = [p(1-p)N]^{\frac{1}{2}} . \tag{3.4}$$

Za mreže kojima je $N \gg \overline{k}$, što je svojstvo većine realnih mreža, binomnu distribuciju možemo aproksimirati Poissonovom (Slika 3.2) :

$$p_k = \mathrm{e}^{-\overline{k}} \frac{\overline{k}^k}{k!} . \tag{3.5}$$



Slika 3.2: Usporedba Poissonove i binomne distribucije za različite veličine kompleksnih mreža i isti srednji stupanj čvora $\overline{k} = 50$. Vidimo da je poklapanje izvrsno za sve mreže koje su dva reda (ili više) veličine veće od srednjeg stupnja čvora $\overline{k}$. Za one veličinom usporedive sa srednjim stupnjem (na slici $N = 100$), moramo koristiti točan oblik binomne distribucije. Preuzeto s dozvolom autora [9].

Pri korištenju Poissonove distribucije važno je napomenuti nekoliko stvari :

- Povećavamo li vjerojatnost vezanja čvorova p, mreža postaje gušća, a vrh raspodjele se pomiče udesno.

- Poissonova distribucija je samo aproksimacija binomne. Binomna distribucija je ta koja je u potpunosti precizna u svim slučajevima, no činjenica da za većinu realnih mreža vrijedi uvjet $N \gg \overline{k}$, distribucija dobro opisuje velik broj nasumičnih fenomena.

- Poissonova distribucija ne ovisi eksplicitno o veličini mreže N. Dakle, jednadžba (3.5) predviđa da mreže različite veličine N i iste srednje vrijednosti

stupnja čvora $\overline{k}$ ne možemo razlikovati.

Poissonova distribucija omogućuje nam da na elegantan način uvedemo pojam *skale* mreže. Naime, za Poissonovu distribuciju sa srednjom vrijednošću stupnja čvora $\overline{k}$, varijanca te distribucije je:

$$\sigma_k = \sqrt{\overline{k}}\,. \tag{3.6}$$

Na primjer, kompleksna mreža kojoj je $\overline{k} = 100$, ima varijancu $\sigma_k = 10$. Konkretno, izvučemo li nasumično iz takve mreže jedan čvor, s velikom vjerojatnošću očekujemo njegov stupanj u intervalu $k = 100 \pm 10$. Vidimo, dakle, da se stupnjevi čvorova u nasumičnim mrežama vrlo malo razlikuju, tj. nalaze se u uskom intervalu oko $\overline{k}$, što definira $\overline{k}$ kao dobru skalu za nasumičnu mrežu. Malo je vjerojatno da izvučeni čvor bude van tog intervala zadanog varijancom, tj. čvorovi malih stupnjeva i čvorovi velikih stupnjeva (engl. *hubovi*) su u nasumičnim mrežama vrlo rijetki. Kasnije ćemo vidjeti da isto svojstvo ne vrijedi za mreže bez skale.

3.3 Fazni prijelazi u Erdős-Rényi mreži

Kako bismo objasnili što točno predstavlja fazni prijelaz u E-R mreži, uvedimo prvo pojam *najveće komponente*. Ona predstavlja komponentu (klaster) mreže koja sadrži najveći broj čvorova, a označavamo ga s N_G. Budući da smo prije predstavili $\overline{k}$ kao najvažniji, skalirajući parametar nasumične mreže, zanima nas kako veličina najveće komponente N_G ovisi o $\overline{k}$. Rubne slučajeve je lako kvalitativno analizirati :

- Za $\overline{k}$ = 0 (tj. vjerojatnost povezivanja $p = 0$) svi su čvorovi izolirani pa je $N_G = 1$.
- Za $\overline{k} = N - 1$ (tj. vjerojatnost povezivanja $p = 1$) svi čvorovi su povezani sa svim ostalima i dio su najveće komponente $N_G = N$, odnosno $\frac{N_G}{N} = 1$.

Pitanje koje se postavlja je kako se mijenja veličina najveće komponente za vrijednosti srednjeg stupnja čvora između krajnjih vrijednosti. Intuitivnu pretpostavku o glatkom rastu vrijednosti $\frac{N_G}{N}$ od 0 do 1 srušili su Erdős i Rényi u svom radu iz 1960 [2]. Tamo su pokazali kako omjer $\frac{N_G}{N}$ ima vrijednost 0 za male $\overline{k}$, sve dok vrijednost $\overline{k}$ ne dosegne svoju kritičnu vrijednost $\overline{k}_c = 1$. Dakle, prosječno je dovoljno

imati 1 poveznicu po čvoru u mreži, želimo li da nam se pojavi najveća komponenta mreže, koju još zovemo i *gigantska komponenta*. Kritični uvjet pojave gigantske komponente možemo zapisati pomoću kritične vjerojatnosti p_c :

$$p_c \approx \frac{1}{N} \tag{3.7}$$

Ona nam govori da je povećanjem mreže dovoljna manja vjerojatnost p_c kako bismo dobili gigantsku komponentu. Pojava gigantske komponente je samo jedna od faza koje se pojavljuju u modelu. Razlikujemo 4 topološki različite faze (Slika 3.3.) :

- Podkritični režim ($0 < \overline{k} < 1$). Za $\overline{k} = 0$ mreža je nepovezana. Povećavanje $\overline{k}$ ekvivalentno je dodavanju $p\frac{N(N-1)}{2}$ poveznica u graf. Budući da je $\overline{k}$ i dalje jako malen, opažamo samo sitne komponente. Ove komponente su veličinom usporedive, tako da najčešće ne možemo pronaći "pobjednika" koji bi predstavljao gigantsku komponentu.

- Kritična točka ($\overline{k} = 1$). Razdvaja fazu bez gigantske komponente i fazu sa gigantskom komponentom.

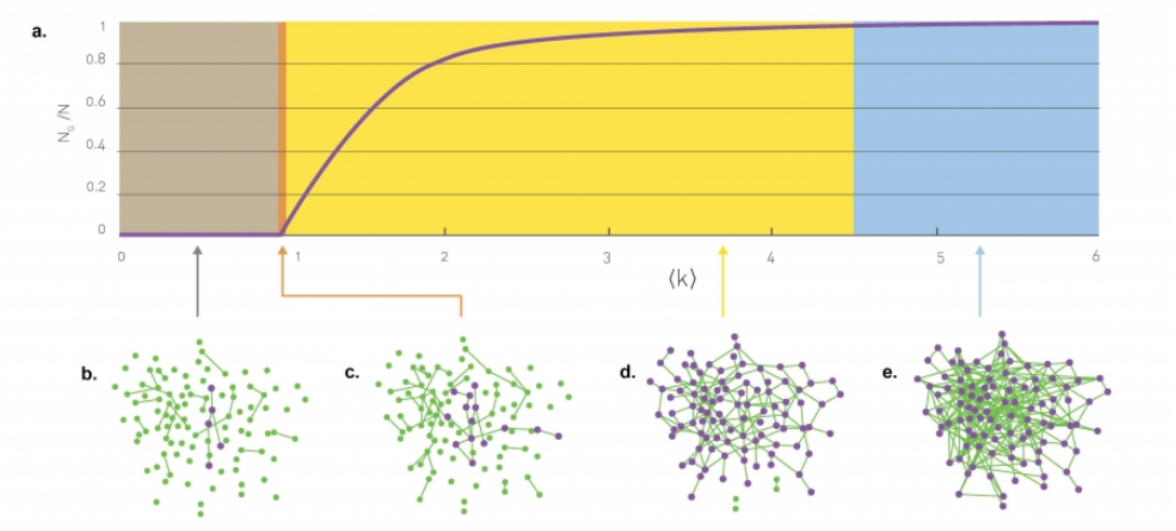
- Superkritična faza ($\overline{k} > 1$). Faza koja najbolje opisuje realne nasumične mreže. U blizini kritične točke, veličina gigantske komponente je opisana linearnom ovisnošću :

  $$\frac{N_G}{N} \sim \overline{k} - 1 \tag{3.8}$$

  Za velike $\overline{k}$, ovisnost N_G o $\overline{k}$ nije linearna. U superkritičnom režimu zajedno postoje gigantska komponenta koja sadrži petlje i malene komponente koje nazivamo *stablima*.

- Povezana faza ($\overline{k} > \ln(N)$). Za dovoljne velike vjerojatnosti p i posljedično velike $\overline{k}$, male komponente odvojene od gigantske prestaju postojati i svi čvorovi postaju dio gigantske komponente, tj $\frac{N_G}{N} = 1$. Ovaj fazni prijelaz događa se na $\overline{k} = \ln(N)$.

  Važno je napomenuti da se blizu točke faznog prijelaza i dalje može koristiti Poissonova raspodjela, budući da je za velike mreže $\frac{\ln N}{N} \approx 0$, ($N >> \overline{k}$).



Slika 3.3: Ovisnost veličine gigantske komponente o srednjem stupnju čvorova $\overline{k}$ (a) i pripadajući shematski prikazi podkritične (b), kritične (c), superkritične (d) i povezane faze (e). Preuzeto sa dozvolom autora [9].

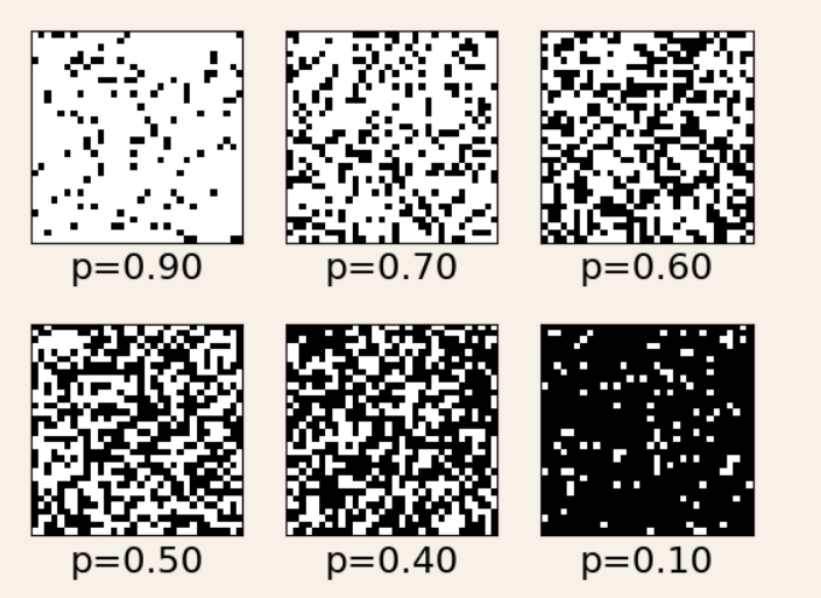
Nakon svega navedenog možemo zaključiti da nastanak mreže nije gladak proces. Sitne komponente i izolirani čvorovi postaju dijelom gigantske komponente kroz fazni prijelaz.

Od interesa u ovom radu je analogija faza kompleksnih mreža s fenomenom *perkolacija*. Perkolacije predstavljaju jedan od najjednostavnijih modela neuređenih sustava [16].

Kao najjednostavniji primjer zamislimo kvadratnu mrežu veličine NxN gdje čvorovi mreže mogu predstavljati različita fizikalna svojstva. Čvorovi su zauzeti s vjerojatnošću p ili nezauzeti s vjerojatnošću $1 - p$. Zauzeće, odnosno nezauzeće čvorišta može predstavljati, npr. vodljivi i izolatorski dio mreže.

Struja može protjecati samo ako duž cijele mreže postoji put prvih susjeda vodiča (gigantska komponenta). Ukoliko je vjerojatnost zauzeća (vodljivosti) pojedinog čvora $p = 0$, svi su čvorovi izolatori i struja ne može protjecati. Daljnjim povećavanjem vjerojatnosti stvaraju se sitne nakupine koje se na nekoj kritičnoj koncentraciji p_c pretvaraju u gigantsku komponentu, strukturu koja prožima cijelu mrežu s jednog kraja na drugi. Kažemo da je struji omogućeno da *perkolira*. Dakako, za $p = 1$, svi čvorovi su vodiči. Još jedan primjer perkolacija može biti sistem čije se uređenje mijenja iz paramagnetskog u feromagnetsko za određenu kritičnu koncentraciju magneta p_c.

Važno je napomenuti kako je perkolacija geometrijski fenomen, dok su fazni prijelazi na kompleksnim mrežama topološki fenomen. Preciznije, prvi susjedi u mreži na



Slika 3.4: Perkolacije na 32 x 32 kvadratnoj mreži. Bijeli pikseli prikazuju zauzeće sa vjerojatnošću p. Preuzeto s dozvolom autora [16].

kojoj promatramo perkolaciju su upravo čvorovi geometrijski najbliži promatranom čvoru, dok u općenitoj kompleksnoj mreži to ne mora biti slučaj. Međutim, ništa nam ne brani da čvorove neke općenite mreže smjestimo u točno određene geometrijske koordinate i promatramo kada će neko svojstvo pridijeljeno mreži perkolirati kroz nju.

Tema perkolacija na kompleksnim mrežama od sve je većeg interesa za znanstvenike [17], a mi ćemo se ovoj temi vratiti kad budemo pričali o ”*snowball-random* sukobu” na kvadratnoj mreži.

3.4 Mreže bez skale

Za kraj ovog poglavlja donosimo usporedbu nasumičnih mreža i mreža bez skale (*scale-free network*).

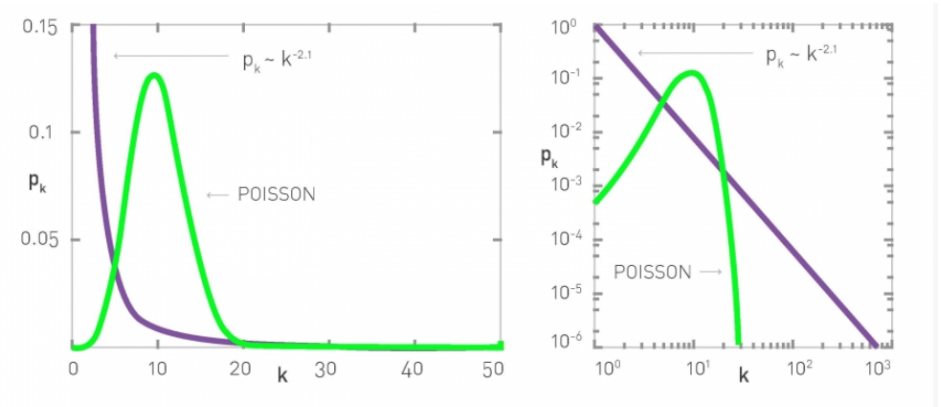
Ako pretpostavimo da je World Wide Web nasumična mreža, tada očekujemo da će stupnjevi njegovih čvorova (Web dokumenata) pratiti Poissonovu razdiobu. Međutim, pokazano je [18] da je to vrlo slab opis distribucije stupnjeva koja opisuje WWW.

Štoviše, WWW, genetske mreže i mnoge druge realne kompleksne sustave opisuju mreže koje nemaju unutarnju skalu (*mreža bez skale*, engl. *scale-free network*).

Razdioba vjerojatnosti za njihov stupanj vezanja prikazan je zakonom potencija:

$$p_k = Ck^{-\gamma} \tag{3.9}$$

gdje konstantu C dobivamo iz uvjeta normiranja (2.6). Na slici (3.5.) dana je usporedba zakona potencije, koji predstavlja *scale - free* mrežu i Poissonove razdiobe koja opisuje nasumičnu mrežu.



Slika 3.5: Poissonova distribucija i distribucija zakona potencija za $\gamma = 2.1$ i prosječnim stupnjem čvorova $\overline{k} = 10$. Slika desno je u logaritamskoj skali, pogodnoj za proučavanje razdiobe za stupnjeve nekoliko redova veličine veće od srednjeg. Preuzeto s dozvolom autora [9].

Gornja slika nam pruža uvid u nekoliko različitih slučajeva:

- Za jako malene stupnjeve k distribucija zakona potencija je iznad Poissonove raspodjele. U skladu s očekivanjim, kod Poissonove raspodjele je vrlo malena vjerojatnost da nađemo čvorove puno manje od srednjeg stupnja $\overline{k}$, dok je za distribuciju zakonom potencija vrlo velik broj čvorova s malenim stupnjem.

- Za stupnjeve oko $\overline{k}$, Poissonova distribucija je iznad zakona potencija, reflektirajući činjenicu da su u nasumičnim mrežama gotovo svi čvorovi usporedivog stupnja.

- Za vrlo visoke stupnjeve čvorova k, opet prevladava zakon potencija. Vjerojatnost da pronađemo čvor s visokim stupnjem je puno redova veličine veća u zakonu potencija. Ovakve čvorove nazivamo *hubovima* i oni su jedna od glavnih karakteristika *mreža bez skale*.

Matematički formalnije, možemo pogledati što se događa s *n*-tim momentom raspodjele zakonom potencija. Može se pokazati [9] kako je *n*-ti moment raspodjele dan

izrazom:

$$\overline{k^n} = C\frac{k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1}}{n-\gamma+1}\,, \tag{3.10}$$

gdje su k_{max} i k_{min} maksimalni, odnosno minimalni stupanj čvora u mreži. Za realne mreže vrijednost potencije γ iznosi između 2 i 3 [19]. Uvrstimo li, npr. $\gamma = 2.1$ u gornju jednadžbu, primjećujemo kako izraz divergira za sve momente $n > 1$. Dakle, prvi moment, koji predstavlja srednji stupanj čvora $\overline{k}$ postoji, no već drugi moment, kojeg koristimo kod definicije varijance, divergira ($\overline{k^2} \to \infty$).

Dakle, kod *mreža bez skale*, nemamo dobro definirano odstupanje od srednje vrijednosti, a samim time srednja vrijednost $\overline{k}$ gubi smisao kao prediktivni faktor i opis skale. Oprimjereno, izvlačimo li nasumično čvor iz *mreže bez skale* kojoj znamo srednji stupanj $\overline{k}$, ne očekujemo izvući čvor čiji je stupanj usporediv sa srednjim.

Zaključno, *mreže bez skale* imaju vrlo velik broj čvorova sa stupnjem vezanja puno manjim od srednjeg, za razliku od nasumičnih mreža čija je distribucija stupnjeva takva da svi čvorovi imaju usporediv stupanj. Također, u *mreža bez skale* mreži se pojavljuju *hubovi*, čvorovi sa vrlo visokim stupnjem k, efekt koji priroda nasumičnih mreža ne dozvoljava.

4 Barabási - Albert kompleksna mreža

U prijašnjem poglavlju upoznali smo se s činjenicom da mnogi realni kompleksni sustavi ne mogu biti predstavljeni nasumičnom kompleksnom mrežom. Pretpostavka Erdősa i Rényia kako mreže koje izgledaju nasumično i jesu nasumične opovrgnuta je mapiranjem distribucije stupnja čvora k za mnoge realne sustave, gdje je pokazano da distribucija ne prati Poissonovu krivulju, već zakon potencija.

U ovom poglavlju zanima nas koji je mehanizam u pozadini zakona potencija. Zašto u stvarnim mrežama postoji nezanemariva vjerojatnost pronalaženja čvora sa jako velikim stupnjem (*hub*)? Odgovor na ova pitanja snažno je vezan uz činjenicu da se u realnim mrežama čvorovi ne povezuju nasumično, već *preferencijalno*. Budući da čvorovi pri povezivanju u mreži radije "biraju" članove mreže s većim stupnjem, narušen je uvjet nasumičnosti.

Preferencijalno vezanje u sustavima svoje je povoje dobilo prije gotovo 100 godina. Mađarski matematičar Gyorgi Polya još je 1923. u radu s Eggenbergerom predstavio model urne [20]. U ovom modelu razmatrani su objekti od interesa (atomi, ljudi, automobili) predstavljena kao kuglice bijele i crne boje stavljene u neku posudu. Kuglice se nasumično izvlače i ako se izvuče bijela kuglica, u posudu se dodaje još jedna bijela kuglica i obrnuto. George Udmy Yule objasnio je 1925. kako se modelom preferencijalnog vezanja može objasniti potencijska ovisnost broja vrsta po rodu biljaka (poznati *Yule proces*). Francuski inženjer Robert Gibrat u svojoj je knjizi 1931. [21] pokazao kako kapital tvrtki ne raste neovisno o veličini tvrtke. Kapital većih tvrtki povećava se brže, efekt preferencijalnog vezanja kojeg nazivamo *proporcionalnim rastom*.

Baš je proporcionalni rast bio kamen temeljac kojim su se Barabási i Albert 1999. [5] poslužili u objašnjavanju nastanka *mreža bez skale*. Njihov model (u nastavku teksta B-A model) kompleksnih sustava počiva na dvije jednostavne premise:

- Kompleksna mreža nastaje dodavanjem novih čvorova u graf. Broj čvorova pri nastajanju mreže nije fiksiran.
- Čvorovi se povezuju tako da se svaki novi čvor dodan u graf sa većom vjerojatnošću povezuje s već postojećim čvorovima VISOKOG stupnja čvora.

U prvom dijelu ovog poglavlja preciznije ćemo objasniti mehanizam preferenci-

jalnog vezanja uz kratko objašnjenje računalnog algoritma koji generira B-A kompleksnu mrežu. Zatim ćemo pokazati kako se mijenja stupanj promatranog čvora pri dodavanju novog čvora u mrežu i pokazati zašto se pojavljuju čvorovi izrazito visokih stupnjeva (*hubovi*). Za kraj, dajemo izvod distribucije stupnja čvora kojim pokazujemo kako iz osnovnih pretpostavki modela slijedi svojstvo *mreže bez skale*

4.1 Generiranje mreže

Osnovna pretpostavka nasumičnog mrežnog modela je fiksni broj čvorova N među kojima raspodjeljujemo poveznice s vjerojatnošću p. U realnim sustavima je, pak, pokazano da sustav raste dodavanjem novih komponenti (čvorovi) koji se s većom vjerojatnošću vežu za komponente u mreži čiji je stupanj čvora veći.

Razmotrimo, na primjer, WWW. Još davne 1991, kad je osnovan, imao je samo jednu stranicu. Danas ih ima više od 10^9 [9]. Trivijalno je primijetiti kako nije svih 10^9 dokumenata stvoreno odjednom, već su postepeno dodavani. Dakako, proces stvaranja novih web stranica traje i danas.

U kontekstu WWW-a lako je razumjeti koncept preferencijalnog vezanja. Naime, svima nam je poznata tek mala frakcija web stranica, najpoznatije od kojih su Facebook, Google itd. Njihova poznatost leži u činjenici da je broj njihovih korisnika, nemjerljivo veći od od ostatka WWW-a. Posljedica toga je da na brojnim drugim web stranicama imamo *linkove* koje usmjeravaju baš na gore navedene popularne stranice. Posljedično, poznatost Facebooka, Googlea i inih stranica povećava vjerojatnost da se novi korisnik WWW-a, bilo u vidu usputnog korisnika ili nove web stranice, poveže baš s njima.

Ovaj efekt, u kojem ”bogati postaju bogatiji” opisuje razvoj mnogih sociooloških i prirodnih sustava. Dakle, generiranje ovakve ovakvih sustava (prirodnih ili umjetnih), preko poznatog B-A modela, slijedi dva osnovna pravila :

- Mrežu gradimo postupnim dodavanjem novih čvorova.
- Vjerojatnost povezivanja novog čvora na postojeći čvor *i* proporcionalna je TRENUTNOM stupnju koji već postojeći čvor u mreži ima :

$$\Pi_i = \frac{k_i}{\sum_j k_j}\,, \tag{4.1}$$

gdje je Π_i vjerojatnost da se tek dodani čvor poveže s *i*-tim čvorom u mreži. Suma u nazivniku s desne strane jednadžbe (4.1) predstavlja normiranje vjerojatnosti.

Proporcionalno vezanje je, fundamentalno gledano, probabilistička metoda, jer se novododani čvor može povezati s bilo kojim drugim čvorom u mreži, bio taj čvor *hub* ili čvor sa samo jednom poveznicom. Međutim, jednadžba (4.1) nam daje kvalitativnu naznaku ponašajnog obrasca B-A modela; vjerojatnost da se novi čvor spoji na *hub* je veća!

Kako bismo računalno generirali B-A kompleksnu mrežu, zadajemo početni broj čvorova N_0 i nasumično raspodijelimo poveznice među njima (Slika 4.1 (a)).

Potom dodajemo $N - N_0$ čvorova, jednog po jednog ih povezujući ih sa m već postojećih čvorova u mreži prateći vjerojatnosnu distribuciju Π_i.

Algoritam generiranja mreže radi na principu kumulativnih vjerojatnosti. Uzmimo, radi jednostavnosti, primjer gdje novododani čvor "bira" između samo dva postojeća, vjerojatnosti 0.3 i 0.7. Tada, u *Pythonu* radimo listu kumulativnih vjerojatnosti koja je za ovaj jednostavan primjer [0.0, 0.3, 1.0]. Program generira nasumični broj između 0 i 1, koji, ako je u intervalu [0.3, 1.0] bira poveznicu sa vjerojatnošću 0.7, a ako je u intervalu [0.0, 0.3] bira poveznicu sa vjerojatnošću 0.3. Naglasimo još jednom kako se prilikom dodavanja svakog novog čvora vjerojatnosna razdioba Π_i, a samim time i kumulativna vjerojatnost, nanovo kalibrira.

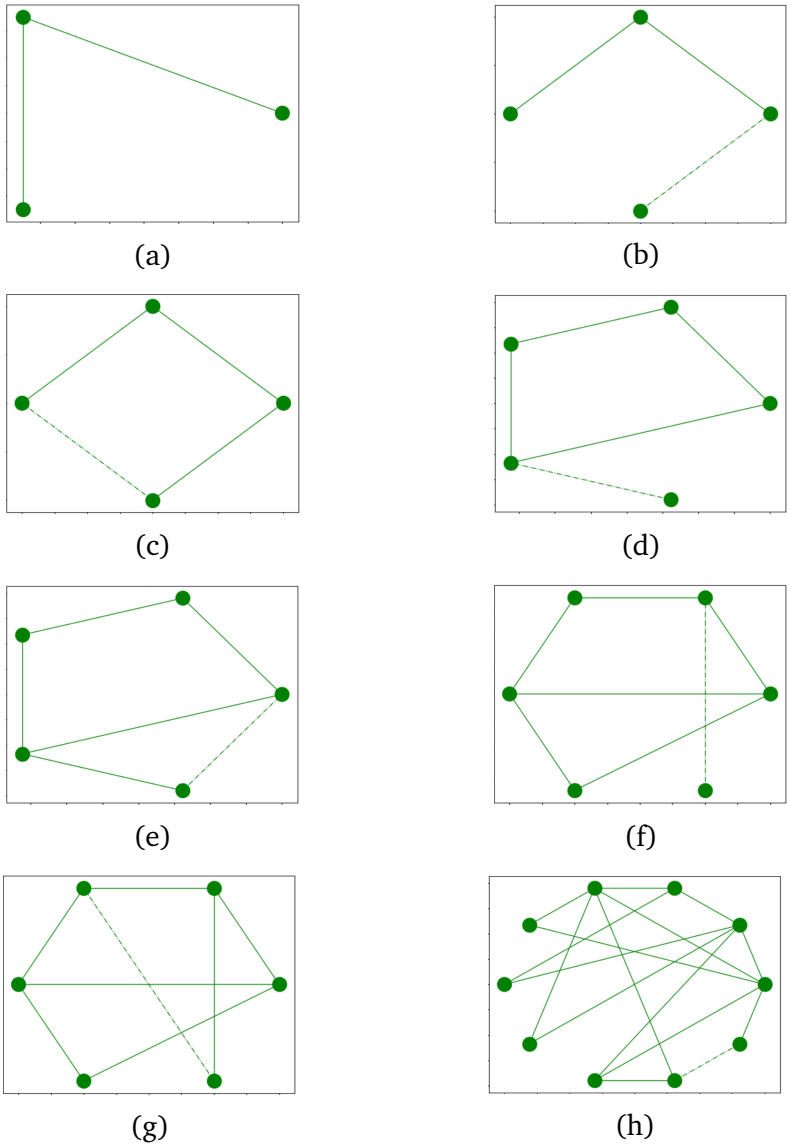
Naravno, Slika 4.1 nipošto ne služi kao etalon B-A mreže iz koje bi isčitavali svojstva. Ona je definitivno premalena kako bi opisala stvarne značajke koje B-A model nosi, značajke o kojima ćemo više reći u nastavku teksta.

4.2 Dinamika stupnja čvora

Sad kad smo se upoznali s osnovnim principima i realizacijom nastanka B-A mreže, vratimo se na detaljniju raspravu o mrežnim svojstvima.

Za razliku od nasumičnih mreža, kod kojih su stupnjevi čvorova unaprijed "zadani" raspodjelom koja ovisi o dva parametra (N i p) i samim time stacionarni, kod mreža koje *rastu* u vremenu, možemo proučavati kako se mijenja stupanj pojedinog čvora kroz vremenske korake.

U B-A modelu već postojećem čvoru u mreži može narasti stupanj dodavanjem



Slika 4.1: Generiranje Barabási-Albert kompleksne mreže za početnu konfiguraciju od $N_0 = 3$ čvora (a). Novi čvorovi u mreži preferencijalno se povezuju sa čvorovima većeg stupnja (b)-(g). Slika (h) predstavlja konačan izgled mreže za $N = 10$ i $m = 2$. Generirano u programskom jeziku *Python 2.7*.

novog čvora, kojeg povezujemo s m postojećih čvorova u mreži. Odaberemo li neki čvor i u mreži, možemo proučavati kako se njegov stupanj mijenja kroz dodavanje novog čvora. Za tu potrebu, tretirajmo k_i kao kontinuiranu varijablu, što je sasvim razumna aproksimacija za velike mreže. Stopa kojom raste stupanj *i*-tog čvora dodavanjem novog čvora u mrežu je:

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = m\Pi_i = m\frac{k_i}{\sum_{j=1}^{N-1} k_j}\,, \tag{4.2}$$

gdje nam faktor m sugerira činjenicu da *i*-ti čvor ima m šansi da bude izabran.

Suma u nazivniku ne broji čvor koji je tek dodan, tako da tu istu sumu možemo pisati kao :

$$\sum_{j=1}^{N-1} k_j = 2mt - m\,. \tag{4.3}$$

Uvrštavanjem izraza (4.3), jednadžba (4.2) postaje:

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = \frac{k_i}{2t-1}\,. \tag{4.4}$$

Za jako velika vremena t, jedinicu u nazivniku možemo zanemariti. S ovim na umu integracija po vremenu t od početnog vremena t_i (vrijeme kada smo dodali promatrani čvor) do nekog krajnjeg vremena t daje:

$$\ln\left(\frac{k_i(t)}{k_i(t_i)}\right) = \frac{1}{2}\ln\left(\frac{t}{t_i}\right). \tag{4.5}$$

Sređivanjem, uz činjenicu da je $k_i(t_i) = m$, dobivamo:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta}, \tag{4.6}$$

gdje je koeficijent $\beta = \frac{1}{2}$. Iz (4.6) možemo izvući nekoliko zaključaka :

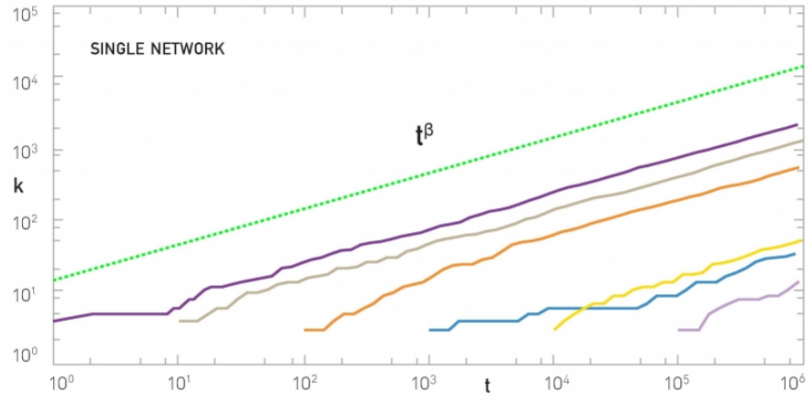
- Svakom se čvoru u mreži stupanj povećava uz gore navedeni zakon potencija. Zakon, kao i dinamički eksponent β, isti je za sve čvorove u mreži.

- Rast stupnja čvora prati "slabiju" ovisnost od jednostavne linearne, o čemu nam svjedoči eksponent $\beta < 1$. Ovo je posljedica narastajućeg suparništva među čvorovima uzrokovanog dodavanjem novog čvora u svakom koraku. Svaki idući čvor bira iz većeg "bazena" mogućnosti.

- Ranije dodani čvorovi imaju veći stupanj vezanja u datom trenutku t. Njihovo vlastito vrijeme dolaska t_i je manje nego za kasnije dodane čvorove tako da je rezultat jednadžbe (4.6) za njih veći.

- Brzina kojom se *i*-tom čvoru dodaju poveznice dana je sa $\frac{\mathrm{d}k_i}{\mathrm{d}t} = \frac{m}{2}\frac{1}{\sqrt{t_i t}}$. što nam govori da kako vrijeme t prolazi, čvor dobiva sve manje i manje poveznica, uzrok čega je rastuća kompeticija između sve većeg broja čvorova u mreži.

Važno je naglasiti kako vrijeme t u kontekstu ovog izvoda ne treba shvatiti kao strogo definiranu, univerzalnu fizikalnu veličinu i pritom joj *apriori* pridijeliti mjernu



Slika 4.2: Rast stupnja čvorova za različita vremena pristizanja t_i u B-A modelu. Crtkana linija predstavlja analitičko predviđanje dano sa (4.6). Preuzeto s dozvolom autora [9]

jedinicu. Vrijeme kao dinamički parametar stupnja čvora, koji evoluira na gore opisan način može predstavljati različite fizikalne skale u ovisnosti o sustavu kojeg mreža predstavlja. Tako se, npr. nove stranice na WWW dodaju u potpuno različitom vremenskom razmaku nego što se, npr. javljaju nove genske upute unutar ljudske DNA. Dakle, vrijeme koje je ovdje bitno je *događajno* (engl. *event time*) i definirano promjenom u topologiji mreže.

Zaključno, Barabási - Albert mreža opisuje ponašanje realnih sistema u kojima broj čvorova raste u vremenu. Dinamičkom analizom utvrdili smo kako ovakvom evolucijom profitiraju "stariji" čvorovi, tj. oni koji su prije postali dio mreže, što ih s vremenom pretvara u *hubove*.

4.3 Distribucija stupnja čvora

Sad kad smo na kvantitativan način opisali dobivanje *hubova* u realnim mrežama, preostaje nam pokazati kako iz opisanih uvjeta mrežne dinamike dolazi do raspodjele stupnja čvora zakonom potencija.

Neka nam je $N(k,t)$ broj čvorova sa stupnjem k u nekom vremenskom trenutku t. Tada trenutnu distribuciju stupnja čvora možemo pisati kao $p_k(t) = \frac{N(k,t)}{N(t)}$ gdje je $N(t)$ ukupni broj čvorova u datom trenutku. Budući da nam je vrijeme definirano kao korak dodavanja novog čvora u mrežu, a dodajemo ih jedan po jedan, možemo

u datom trenutku pisati $N(t) = t$. Naše preferencijalno vezanje je:

$$\Pi_k = \frac{k}{2mt} \tag{4.7}$$

gdje faktor 2 u nazivniku izraza reflektira činjenicu da svaka nova poveznica dodana u graf povezuje dva čvora. Budući da proučavamo broj čvorova stupnja k, zanimaju nas slučajevi u kojima se broj čvorova s datim stupnjem mijenja. To su slučajevi kada se:

- Novi čvor povezuje sa čvorom stupnja k. Tada tom čvoru stupanj raste na $k+1$, a broj čvorova $N(k,t)$ opada.
- Novi čvor povezuje sa čvorom stupnja $k-1$. Čvoru raste stupanj na k, a broj čvorova $N(k,t)$ se povećava.

Pogledajmo sada koliko se ukupno poveznica spoji na čvor stupnja k dolaskom novog čvora. Prvo, treba nam vjerojatnost da će se poveznica spojiti na čvor stupnja k. Taj podatak je zapisan kao preferencijalno vezanje Π_k. Drugo što nam treba je broj čvorova sa stupnjem k, kojeg trivijalno dobijemo kao $Np_k(t)$. Treće, moramo uzeti u obzir broj poveznica m dodanih u svakom koraku. Što je m veći, veća je vjerojatnost da se novopridošli čvor spoji na čvor stupnja k. Konačno, množenjem gornjih članova dobivamo očekivani broj poveznica spojenih na čvor stupnja k:

$$\frac{k}{2mt} \text{ x } Np_k(t) \text{ x } m = \frac{k}{2}p_k(t)\,. \tag{4.8}$$

Dobiveni rezultat možemo primijeniti na oba gore opisana slučaja, pa za broj čvorova stupnja k nakon dodavanja novog čvora imamo:

$$(N+1)p_k(t+1) = Np_k(t) + \frac{k-1}{2}p_{k-1}(t) - \frac{k}{2}p_k(t)\,. \tag{4.9}$$

Ovaj izraz vrijedi za sve $k > m$. Budući da nam u mreži nedostaju čvorovi sa $k < m$, pišemo zasebno jednadžbu za čvorove čiji je stupanj točno jednak m, što je polazišna osnova za rješavanje rekurzije:

$$(N+1)p_m(t+1) = Np_m(t) + 1 - \frac{m}{2}p_m(t)\,. \tag{4.10}$$

Iskoristimo sada činjenicu da tražimo stacionarnu distribuciju stupnja čvora, za koju vrijedi relacija $p_k(\infty) = p_k$. Sada možemo pojednostaviti lijeve strane jednadžbi (4.9) i (4.10) kao:

$$(N+1)p_k(t+1) - Np_k(t) = Np_k(\infty) + p_k(\infty) - Np_k(\infty) = p_k\,, \tag{4.11}$$

$$(N+1)p_m(t+1) - Np_m(t) = p_m\,. \tag{4.12}$$

Ovime, jednadžbe (4.9) i (4.10) poprimaju svoju rekurzivnu formu:

$$p_k = \frac{k-1}{k+2}p_{k-1}\,, \tag{4.13}$$

$$p_m = \frac{2}{m+2}\,. \tag{4.14}$$

Rekurziju počinjemo od najmanjemog mogućeg stupnja čvora $k = m$ i uz pomoć (4.13) računamo vjerojatnosti za više stupnjeve. Prvih nekoliko su:

$$p_{m+1} = \frac{2m}{(m+2)(m+3)}\,, \tag{4.15}$$

$$p_{m+2} = \frac{2m(m+1)}{(m+2)(m+3)(m+4)}\,, \tag{4.16}$$

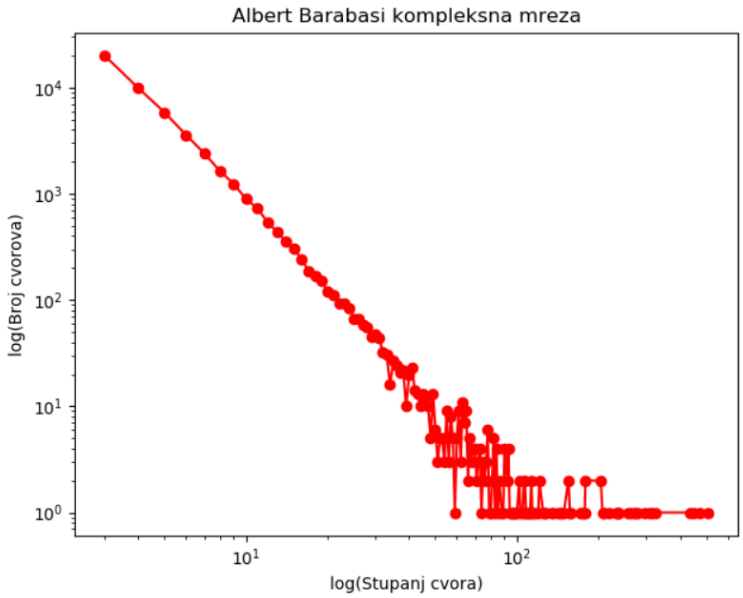
$$p_{m+3} = \frac{2m(m+1)}{(m+3)(m+4)(m+5)}\,. \tag{4.17}$$

Primjećujemo obrazac ponašanja raspodjele (dokaz valjanosti izvoda u [22]) i konačno pišemo za B-A mrežu:

$$p_k = \frac{2m(m+1)}{(k+1)(k+2)(k+3)}\,, \tag{4.18}$$

što za velike stupnjeve čvora k postaje $p_k \approx k^{-3}$, što je upravo potencijska ovisnost, značajka *mreža bez skale* mreža koje smo obradili u prethodnom poglavlju. Na slici (4.3) vidimo raspodjelu stupnja čvora koju smo sami generirali.

Zaključno, pokazali smo kako distribucija stupnja čvora u B-A modelu prati zakon potencije sa eksponentom $\gamma = -3$. Eksponent za velike mreže ne ovisi o m i m_0, što se reflektira u činjenici kako za realne mreže različitog porijekla i povijesti dobivamo



Slika 4.3: Log-log distribucija stupnja čvora k za mrežu od $N = 50000$ čvorova dodavanih sa $m = 3$ poveznice. Generirano u programskom jeziku *Python 2.7*.

istu distribuciju.

Sada, kada poznajemo osnovne pojmove i svojstva vezane uz mreže i njihova dva reprezentatvna modela, okrećemo se problemu uzorkovanja na kompleksnim mrežama. Kao što ćemo vidjeti, efikasnost uzorkovanja metodom snježne grude uvelike ovisi kako su stupnjevi čvorova unutar mreže raspodijeljeni.

5 Uzorkovanje

Nerijetko kada želimo ispitati svojstva nekog velikog kompleksnog sustava, nismo u mogućnosti ispitati svaku komponentu zasebno. Tada posežemo za metodom uzorkovanja, tj. selekcijom podskupa komponenti ispitivanjem kojih možemo donijeti zaključke o karakteristikama cijelog sustava. Dvije eklatantne prednosti uzorkovanja naspram obuhvaćanja cjelokupnog sustava su jeftinija izvedba i brže skupljanje podataka.

Svaka opservacija mjeri jedno ili više svojstava komponenti kompleksnog sustava. U anketnom uzorkovanju često se primjenjuju *težinske* korekcije, naročito kada je riječ o *stratificiranom* uzorkovanju [23]. Podaci prikupljeni uzorkovanjem analiziraju se alatima statistike i teorije vjerojatnosti.

5.1 Općenito o uzorkovanju

Kako bismo što kvalitetnije predstavili uzorkovanje kao metodu ispitivanja svojstava sustava, nužno je predstaviti ključne pojmove:

- Ciljna populacija (engl. *target population*) - skupina komponenata sustava na koje se istraživanje odnosi i na koju poopćavamo rezultate dobivene uzorkovanjem.

Definicija populacije u nekim je slučajevima vrlo trivijalna. Na primjer, želimo li za neku tvornicu cipela ispitati koliki je udio cipela proizvedenih s greškom (odlijepljeni potplati, dimenzionalne nepravilnosti itd.), naša ciljna populacija očito obuhvaća sve cipele proizvedene u tvornici.

S druge pak strane, postoje slučajevi u kojima populacija nije toliko ”opipljiva”; Joseph Jagger, engleski industrijalac proučavao je u 19. st. ruletna kola u Monte Carlu tražeći koja od njih nisu ispravna. U ovom slučaju, ciljna populacija predstavljena je ukupnim ponašanjem koje kolo ima (tj. vjerojatnosnom raspodjelom ishoda na beskonačno mnogo igara), dok su uzorci predstavljeni rezultatima na konačnom broju igara. Slična razmatranja možemo primijeniti i na fizikalne sustave, npr. ispitivanje magnetske susceptibilnosti u paramagnetima.

- Uzorak - dio ciljne populacije na kojem se ispituju svojstva od interesa i temeljem kojega se zaključuje o svojstvima populacije.

Postoje razni mehanizmi kojima možemo skupljati uzorke iz populacije, a dijele se na dvije velike skupine : *probabilistički* i *neprobabilistički* uzorci [24].

Probabilistički uzorak podrazumijeva svaki uzorak uzet iz ciljne populacije čije jedinke imaju poznatu vjerojatnost različitu od 0 da budu uzete u uzorak. Ove karakteristike osiguravaju *nepristranost* uzorka čiji se svaki element može težinski ponderirati u skladu s apriori vjerojatnošću izvlačenja.

Opišimo navedeno jednostavnim primjerom. Zamislimo neki omanji gradić u kojem želimo saznati ukupan dohodak svih žitelja. Budući da je gradić malen, u stanju smo brzo identificirati sve žitelje po kućanstvu i nasumično odabrati jednu osobu iz kućanstva kojoj ćemo mjeriti visinu dohotka. Nasumični odabir vršimo tako da svakoj osobi u kućanstvu dodijelimo broj od 1 do K (gdje je K broj ukućana) te potom nasumično izvlačimo broj. Ljudi koji žive sami sigurno će biti izabrani, tako da njihove dohodke jednostavno dodajemo u uzorak. No, u kućanstvu koje broji $K > 1$ ukućana svaki ukućan može biti izabran s vjerojatnošću $\frac{1}{K}$ pa njihove dohodke dodajemo K puta u procjenu (*težinsko ponderiranje*).

U gore navedenom primjeru vjerojatnosti biranja u uzorak nisu jednake, ali su poznate, što je nužan uvjet da uzorak bude probabilistički. Mi ćemo se u našem istraživanju fokusirati na probabilističku uzorkovanje gdje je vjerojatnost uzimanja u uzorak jednaka za svaku jedinku. Ovakav tip uzorkovanja naziva se *jednostavni nasumični uzorak* (engl. *simple random sample*). Još neke poznate vrste probabilističkih uzoraka su *sistematsko* uzorkovanje, *klastersko* uzorkovanje, *stratificirano* uzorkovanje itd. Sve ove vrste uzorkovanja imaju dvije stvari zajedničke :

1. Svi elementi imaju poznatu vjerojatnost različitu od 0 da budu uzeti u uzorak

2. Svaki element u uzorku izabran je nasumično

Neprobabilistički uzorci podrazumijevaju sve uzorke uzete iz ciljne populacije u kojoj postoje jedinke koje ne mogu biti uzete u uzorak (nulta vjerojatnost) ili populacije u kojoj ne znamo vjerojatnosnu raspodjelu uzimanja u uzorak.

Zamislimo da u gornjem primjeru po datom kućanstvu intervjuiramo osobu koja nam otvori vrata. Tada je sasvim jasno da više ne postoji nepristranost, budući da u datom trenutku uopće ne moraju sve osobe biti u kući. Štoviše, osobe koje su u kući dok vršimo anketiranje vjerojatnije su, npr. nezaposlene, tako da uzorak kojim

pokušavamo saznati prosječna primanja nije reprezentativan - on će vrlo vjerojatno podcijeniti vrijednost ukupnog dohodka.

Dakle, budući da vjerojatnost izbora nekog elementa populacije u uzorak nije poznata ili je jednaka 0, ne možemo tvrditi da je on reprezentativan, što uvelike smanjuje mogućnost uopćavanja ispitanog svojstva na cijelu populaciju. Svejedno, u dosta slučajeva pribjegava se korištenju ne-probabilističkih metoda budući da su one jednostavnije i generalno ekonomičnije - izdaci za njih su znatno manji i provedivi su u kraćem vremenskom roku.

Najpoznatije vrste neprobabilističkih uzoraka su kvotni uzorci, prigodni uzorci, namjerni uzorci i uzorci snježne grude (engl. *snowball*) koji su nama od interesa u ovom radu [25].

U narednom dijelu teksta, predstavit ćemo dvije metode uzoraka, reprezentativne za probabilističku i neprobabilističku skupinu: jednostavno nasumično uzorkovanje kao probabilističkog "predstavnika" i metodu snježne grude kao neprobabilističkog.

5.2 Jednostavno nasumično uzorkovanje

Jednostavno nasumično uzorkovanje (engl. *simple random sampling*) je ono u kojem svi elementi ciljne populacije imaju jednaku vjerojatnost p izbora u uzorak [25].

Uzmimo kao primjer slučaj u kojem želimo ispitati potrošačke navike građana grada Zagreba u uzorku od $M = 100$ nasumično odabranih ljudi. Radi jednostavnosti, uzmimo da Zagreb broji $N = 1000000$ ljudi. Tada bi vjerojatnost uzimanja u uzorak za svakog stanovnika bila $p_r = M/N = 0.00001$.

Teorijski gledano, jednostavni slučajni uzorak možemo uzimati s povratom ili bez povrata elementa. Radi jednostavnosti, uzmimo da u ciljnoj populaciji postoje samo svojstva "0" i "1", pridijeljena svakom članu u populaciji. Neka nam je poznat broj elemenata K sa svojstvom "1" u populaciji. Tada je vjerojatnosna distribucija izvlačenja k elemenata sa svojstvom "1" u uzorku od M članova hipergeometrijska:

$$p(k) = \frac{\binom{K}{k}\binom{N-K}{M-k}}{\binom{N}{M}}, \tag{5.1}$$

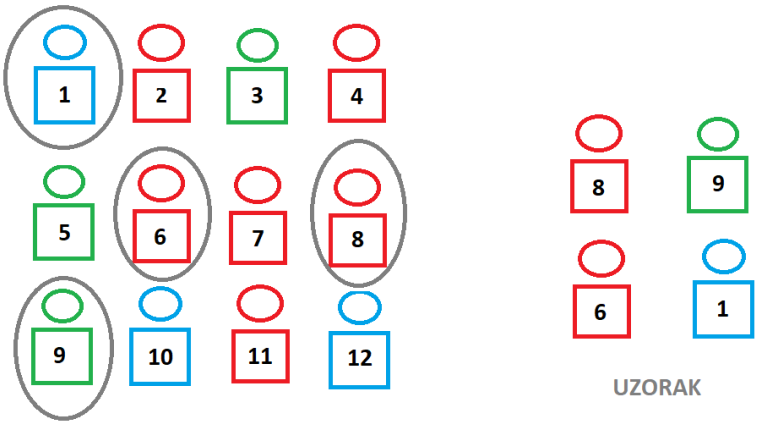
što u slučaju kada je veličina uzorka znatno manja od veličine ciljne populacije

$M \ll N$ možemo aproksimirati binomnom raspodjelom :

$$p(k) = \binom{M}{k} p_1^k (1-p_1)^{M-k}, \tag{5.2}$$

gdje je p_1 udio elemenata u populaciji sa svojstvom "1" ($p_1 = \frac{K}{N}$). Kako bismo koristili metodu slučajnog uzorka, moramo poznavati sve elemente u populaciji. Najčešći način biranja elemenata u slučajni uzorak prati sljedeće korake :

1. Svim elementima u populaciji pridijelimo jedinstveni broj.
2. Računalno generiramo broj u intervalu između najmanjeg i najvećeg broja pridijeljenog u populaciji te element označen tim brojem stavljamo u uzorak.
3. Ukoliko računalo isti broj generira dva ili više puta, ponavljamo postupak (ne dodajemo isti element dvaput).
4. Postupak ponavljamo dok ne napunimo uzorak.



Slika 5.1: Shematski prikaz nasumičnog uzorkovanja (*random sampling*). Elementima ciljne populacije pridijeljeni su brojevi od 1 do 12. Izvučena su $M = 4$ nasumična broja (1,6,8,9) te su elementi s tim brojnim oznakama uzeti u uzorak. Različite boje elemenata reprezentiraju različita svojstva od interesa u populaciji

Prednost nasumičnog uzorkovanja je u jednostavnoj metodi kojom generiramo uzorak i u činjenici da ovakvo skupljanje uzorka omogućava standardne statističke izračune grešaka, prema kojima su, uostalom, baždareni brojni programi za računalnu obradu podataka.

No, ovakvo uzorkovanje ima brojne, često nepremostive nedostatke. Na primjer, želimo li ispitati neko svojstvo na ciljnoj populaciji čitavog grada, poznavanje i dostupnost svakog stanovnika je vrlo teško izvediva. Logistički, postupak je vrlo skup

i spor, jer traži mnoštvo anketara koji stalno putuju kako bi ispitali stanovnike po zemljopisno raspršenim mjestima.

5.3 Uzorkovanje metodom snježne grude

Metoda snježne grude (engl. *snowball method*) je, u sociologiji i statistici [26] definirana kao ne-probabilističko uzorkovanje u kojem već postojeći elementi unutar uzorka ”regrutiraju” nove elemente u uzorak kroz mrežu poznanstava. Ova se metoda uzorkovanja koristi kod ”skrivenih” populacija, kakve su na primjer ovisnici o heroinu [27], žene uključene u prostituciju [28] i sukobljeni u ratu [29] kojima istraživački ne možemo jednostavno pristupiti.

Kako bi bolje dočarali ovu metodu, oprimjerimo je jednim fiktivnim, ali sasvim mogućim istraživanjem. Zamislimo da želimo saznati kolika je stopa zaposlenosti unutar neke ekstremne navijačke skupine. Ukoliko i sami nismo njen dio, štoviše njen utjecajan dio, tada nam definitivno nije lako pristupiti širokom broju ljudi unutar skupine. Tada možemo uložiti napor i vrijeme da lociramo i upoznamo jednog pripadnika te navijačke skupine, koji postaje *izvorište* našeg uzorka. Nakon što *izvorište* ispuni anketu, njegova je zadaća pronaći nove pripadnike skupine koji će popuniti anketu. Tada ti pripadnici postaju nova *izvorišta* i proces se nastavlja dok se ne skupi uzorak. Ovakav uzorak, dakle, raste poput snježne grude, pa je po toj analogiji i dobio ime.

Problemi u potonjem istraživanju proizlazi iz pristranosti skupljenog uzorka. Naime, neki od scenarija koji se mogu dogoditi su sljedeći :

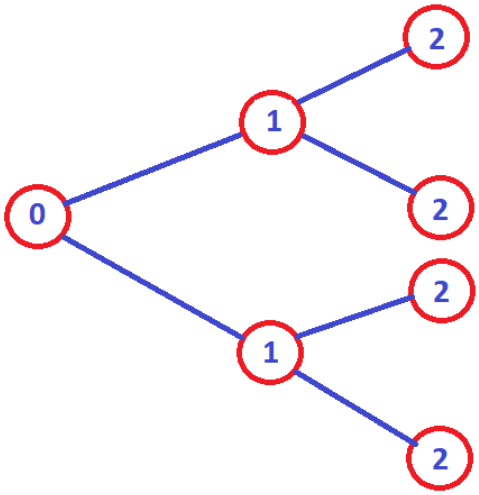
- Ništa nam ne garantira da ćemo skupiti reprezentativan uzorak. Naše *izvorište* će vrlo vjerojatno biti bolji prijatelj unutar skupine sa onim navijačima koji imaju sličan životni stil.

- Istraživanje će puno vjerojatnije stići do osoba koje imaju više prijatelja unutar skupine (činjenica koja će biti relevantna u razmatranjima uzorkovanja na kompleksnim mrežama). Njihova umreženost uvelike smanjuje mogućnost da budu nezaposleni.

- Ne možemo znati hoće li naše ankete biti popunjene van ciljne populacije. U realnoj stiuaciji, čak i kad sklopimo povjerenje sa našim *izvorištem*, nije nam za-

jamčeno da će nova *izvorišta* osjećati jednaku razinu odgovornosti prema nama. Tako se može dogoditi da našu anketu popune ljudi koji su dobri prijatelji navijača, a nisu dio same skupine.

- Proces može ”puknuti” prije nego se napuni uzorak. To će se dogoditi u slučaju kada nitko u nekom koraku ne prihvati anketu.



Slika 5.2: Shema *snowball* uzorkovanja. *Izvorište*, označeno sa 0, daje anketu poznanicima (1), koji je šalju svojim poznanicima (2) itd. Ovakav proces nastavlja se dok se ne skupi željeni broj elemenata u uzorku ili dok ih se ne skupi što je više moguće.

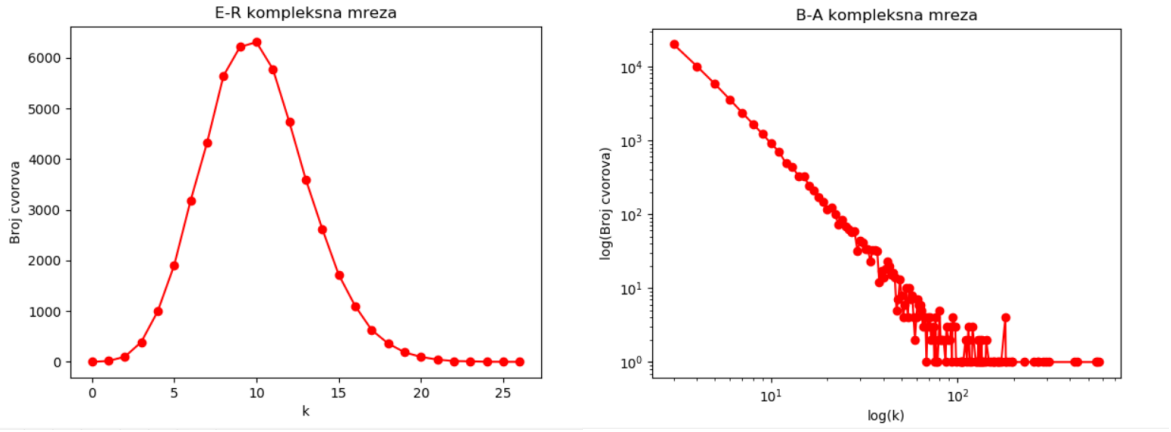
Sada, kada smo ukratko predstavili kako u stvarnosti izgledaju dva tipa uzorkovanja, prelazimo na sam istraživački rad. U idućem poglavlju opisat ćemo kako smo modelirali dva tipa uzorkovanja na kompleksnim mrežama. Odsad pa nadalje, koristit ćemo praktičniju terminologiju, gdje ćemo jednostavno nasumično uzorkovanje zvati *random*, a uzorkovanje metodom snježne grude *snowball*.

5.4 Uzorkovanje na kompleksnim mrežama

Uzorkovanje na ciljnoj populaciji predstavljenoj kompleksnom mrežom relativno je nova i uzbudljiva domena istraživačkog interesa [30]. U ovakvom pristupu čvorovi kompleksne mreže predstavljaju elemente ciljne populacije koju možemo uzeti u uzorak.

Nama je od interesa u ovom radu bilo ispitati koliko se razlikuju *random* i *snowball* uzorkovanje na kompleksnim mrežama, točnije na dva najpoznatija modela (B-A i E-R).

Odmah vrijedi napomenuti kako je istraživanje bilo strogo modelsko, uz mnoga pojednostavljenja i razumne pretpostavke. Krenimo od samih modela kompleksnih mreža koje smo generirali (slika 5.3). Oba modela su iste veličine, $N = 50000$, gdje smo pazili da pri generiranju E-R mreže postignemo natkritični režim, koji je najčešći slučaj u realnim sustavima.



(a) (b)

Slika 5.3: Raspodjele stupnjeva čvorova za kompleksne mreže na kojima ćemo vršiti uzorkovanje. E-R mrežu (a) generirali smo u natkritičnom režimu, sa $N = 50000$ čvorova i srednjim stupnjem čvorova $\overline{k} = 10$. B-A mreža (b) ima također $N = 50000$ čvorova, koji se dodavanjem u mrežu spajaju s $m = 3$ čvora. Generirano u *Pythonu 2.7*.

Jednom kad generiramo mrežu, prebacujemo se na raspodjelu svojstava na mreži, tj. proizvoljno odabiremo karakteristike svakog čvora koje ćemo ispitivati uzorkovanjem. Odlučili smo se za najjednostavniji slučaj, u kojem imamo samo dva svojstva ; svojstvo ”1” i svojstvo ”0”. Ova dva svojstva mogu predstavljati razne stvari od interesa u istraživanju : izbornu opredijeljenost u dominantno dvostranačkom sustavu, zaposlenost, odnos feromagnet-paramagnet, itd. Ova dva svojstva raspodijelili smo na nekoliko načina koji spadaju u dvije velike skupine:

- Nasumična (*random*) raspodjela je ona u kojoj zadajemo točan udio p_1^r svojstva 1 u uzorku. Svaki čvor u mreži ima svoj jedinstveni broj, od 1 do 50000. Nasumično generiramo $50000 \, \mathrm{x} \, p_1^r$ brojeva i čvorovima označenim pripadnim brojevima dodjeljujemo svojstvo ”1”, pazeći naravno, ukoliko je generiran isti broj više puta, da postupak ponovimo. Ovakva raspodjela svojstava ne ovisi o povezanosti čvorova u mreži.

- Raspodjele ovisne o stupnju čvora u mreži su one u kojima posjedovanje svoj-

stva ”1” za neki čvor ovisi o stupnju tog čvora k. Svojstvo ”1” u mreži smo raspodijelili po tri zakona;

1. Zakon opće potencije : $p_1(k) = \mathrm{A}\left(\frac{k}{k_{\max}}\right)^{\zeta}$
2. Logaritamska raspodjela : $p_1(k) = \frac{\log(\mathrm{A}k)}{\log(\mathrm{A}k_{\max})}$
3. Raspodjela tangensom hiperbolnim : $p_1(k) = \tanh(\frac{k}{\mathrm{A}})$

U gornjim raspodjelama $k_{\max}$, najveći stupanj čvora pronađen u mreži služi za normalizaciju vjerojatnosti, dok koeficijent A i potenciju ζ zadajemo sami. Računalno, za svaki čvor kojem je pridijeljena vjerojatnost $p_1(k)$ generiramo nasumični broj između 0 i 1 pa ako je broj manji od $p_1(k)$, čvoru dodijeljujemo svojstvo ”1”. Ukupan udio p_1 svojstva ”1” u mreži po završetku generiranja razdiobe svojstava nalazimo tako da pobrojimo sve čvorove sa svojstvom ”1” i podijelimo taj broj sa N. Od interesa će nam biti konfiguracije koje ćemo baždariti tako da udio p_1 bude oko 0.5 (odnosno 50 %).

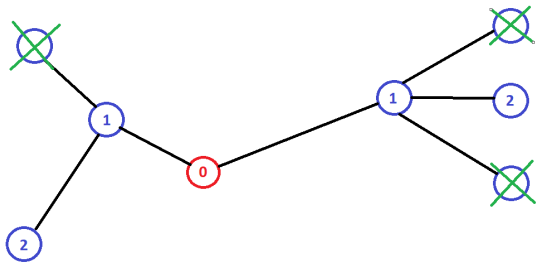
Jednom kada smo generirali mrežu i na njoj raspodijelili svojstva po nekoj od gore navedenih raspodjela, spremni smo uzorkovati. Uzorkovanje se vrši *random* i *snowball* metodom, a za veličinu uzoraka biramo $M = 100$. Kako bismo dobili dobru statistiku procjena koju daju uzorci, radimo ansambl od $Z = 10000$ uzoraka za zadanu raspodjelu na mreži. U svakom uzorku u ansamblu računamo udio svojstva ”1”, a raspodjele udjela u ansamblu prikazujemo grafički.

Random uzorkovanje provodimo već opisanim principom generiranja nasumičnih brojeva dok ne popunimo uzorak.

Snowball uzorkovanje modeliramo sljedećim koracima:

1. U 1. koraku nasumično izabiremo prvi čvor u uzorku, koji će biti naše *izvorište*.
2. Nadalje, promatramo prve susjede *izvorišta*, tj. čvorove izravno povezane s *izvorištem*. Njih nalazimo zapisane u matrici susjednosti A_{ij}. Svaki prvi susjed ima jednaku vjerojatnost uzimanja u uzorak p_s, koju nazivamo *snowball parametrom*.
3. Prema zadanom parametru p_s, program odabire određen broj ”prvih susjeda” koji će biti uzeti u uzorak i postati nova *izvorišta*.
4. Proces ponavljamo dok ne napunimo uzorak.

5. Ukoliko se dogodi da u nekom od koraka ne bude izabran nijedan prvi susjed trenutnih *izvorišta*, nasumično izabiremo novo *izvorište* i ponavljamo proces dok ne napunimo uzorak.



Slika 5.4: Shema *snowball* uzorkovanja. *Izvorište*, označeno s 0 i crvenom bojom, skuplja u uzorak prve susjede (1) sa vjerojatnošću p_s. Oba prva susjeda izvorišta prihvaćaju ući u uzorak, no neki njihovi prvi susjedi, prekriženi zelenom bojom, nisu "regrutirani" u uzorak.

Važno je ukazati na nekoliko stvari kod *snowball* modela uzorkovanja. Prvo, za svaki od uzoraka u ansamblu nasumično biramo *izvorište*, čime u ovoj metodi podižemo reprezentativnost. Drugo, u realnim, društvenim sustavima često nismo u mogućnosti izabirati nova *izvorišta*, pa se zadovoljimo sa maksimalnim brojem ispitanika što možemo skupiti. Ovdje to nije slučaj, budući da nam je cilj statističkim metodama usporediti *snowball* i *random* uzorke jednake veličine. Također, nama je ovdje u principu dostupna čitava mreža, budući da se radi o topološkim, a ne geometrijskim objektima[2] u kojima nijedan čvor ne ostaje nepovezan.

Prilikom statističke obrade podataka služili smo se programom *QtiPlot*, koji ima već ugrađene algoritme za obrade skupova podataka. *Random* i *snowball* ansamble za dane raspodjele svojstava prikazivat ćemo na zajedničkim grafovima radi jasnoće, a procjene koje daju uzorci pisat ćemo kao :

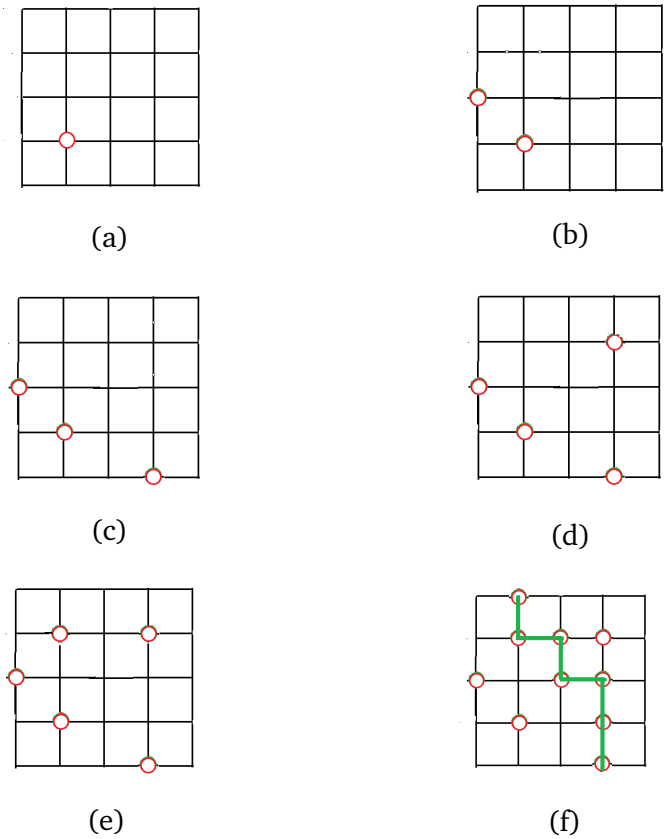
$$p_1 = \overline{p_1} \pm \sigma_{p_1} \tag{5.3}$$

gdje je $\overline{p_1}$ srednja vrijednost udjela svojstva "1" za dano uzorkovanje, a σ_{p_1} standardno odstupanje od srednje vrijednosti.

[2]Prostorne dimenzije mreže su zanemarene. U stvarnom slučaju, to bi odgovaralo situaciji u kojoj nam nije bitno gdje živi ispitanik, jednako nam je dostupan ako živi u susjedstvu kao i jako daleko od nas.

5.5 *Snowball-random "sukob"*

Kao posljednji dio ovog istraživanja, predstavit ćemo *snowball-random* "sukob", svojevrsno natjecanje između dva načina uzorkovanja koje smo osmislili kao usporedbu u kojoj uzorak nije definiran BROJČANO, kao u prethodnim dijelovima teksta, već GEOMETRIJSKI.



Slika 5.5: Potezi *random* igrača na mreži veličine $N = 16$. Čvorovi mreže su jednostavno presjecišta linija koje čine kvadratnu mrežu. Prvih nekoliko poteza prikazano je na slikama (a)-(e), gdje crveni kružići označavaju čvorove izabrane u uzorak, a na slici (f) prikazan je ispunjen geometrijski uzorak, tj. put prvih susjeda s jednog kraja mreže na drugi.

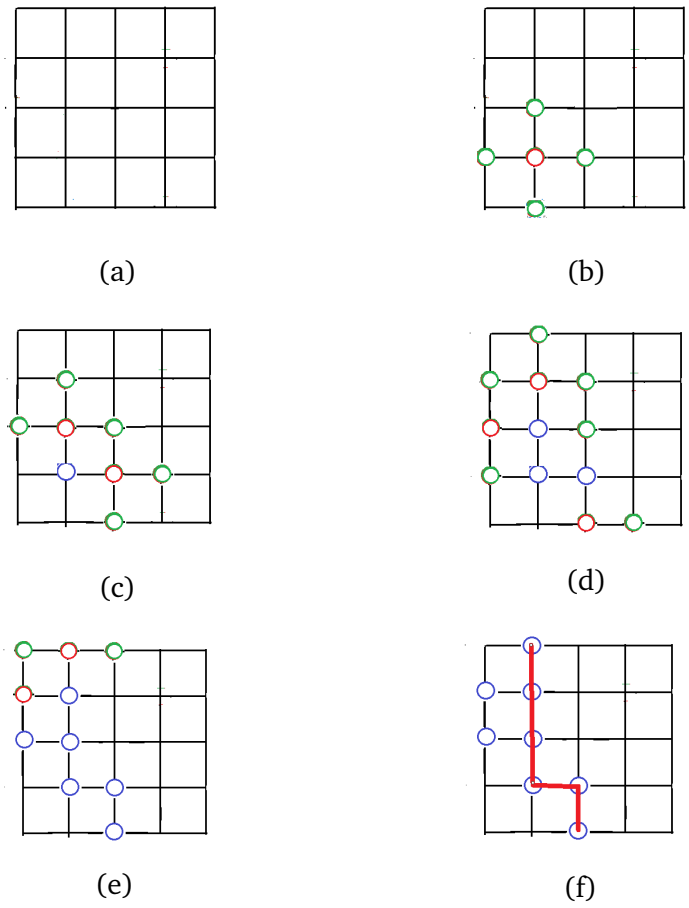
"Sukob" se odvija na kvadratnoj mreži dimenzija $\sqrt{N} \times \sqrt{N}$ gdje je N broj čvorova u kvadratnoj mreži. Prije nego objasnimo "sukob", valja predstaviti osnovne značajke kvadratne mreže. Gotovo svaki čvor ima jednak stupanj, $k = 4$ jer je povezan samo sa svojim prvim susjedima (ovdje su prvi susjede geometrijski najbliži čvorovi). Prvi susjedi nekog čvora na koordinati (x,y) su (x,y+1), (x,y-1), (x+1,y) i (x-1,y). Naravno, prilikom igranja "utakmice" uzeli smo u obzir da čvorovi na rubovima mreže nemaju $k = 4$, već $k = 2$ (u kutevima) i $k = 3$ na stranicama.

Svaki "sukob" sastoji se od "poteza", koje random i snowball igrač naizmjence

”odigravaju” na svojim vlastitim mrežama.

Random ”potez” se sastoji u nasumičnom biranju čvora u mreži, kojeg na taj način bilježimo u svojevrstan uzorak. Slika 5.5 služi kao ilustrativan primjer kako izgleda biranje.

Snowball ”potez” (Slika 5.6) u početku izabire *izvorište*. Potom u novom potezu bira prve susjede početnog čvora u uzorak sa snowball vjerojatnošću p_s (prva generacija). U idućem potezu bira između prvih susjeda prve generacije (druga generacija) itd. Ukoliko nijedan susjed n-te generacije nije izabran u uzorak, snowball igrač NEMA PRAVO IZABRATI NOVO *IZVORIŠTE* U ISTOM POTEZU, nego tek u idućem! (efektivno, nijedan čvor u potezu nije izabran).



Slika 5.6: Potezi *snowball* igrača na mreži veličine $N = 16$. Crveni kružići označavaju čvorove uzete u uzorak u trenutnom potezu, zeleni označavaju njihove prve susjede koji će u idućem potezu biti razmotreni za uzorak, a plavi su čvorovi uzeti u uzorak u prijašnjim potezima (a)-(d). Slika pod (e) prikazuje ”pobjedu” *snowball* igrača, popunjen geometrijski uvjetovan uzorak duž y-osi.

Po gore opisanim pravilima, snowball i nasumični igrač "grade uzorak". Uzorak je izgrađen onda kada se duž y-osi u kvadratnoj mreži stvori put prvih susjeda (perkolacija!). Puteve duž x-osi ovdje ne razmatramo. Dakle, "pobjednik" je onaj igrač koji prvi izgradi ovakav uzorak. Postojanje puta prvih susjeda provjeravamo BFS (*breadth first search*) algoritmom opisanim u prvom poglavlju.

Fizikalno, ovakva igra može nam pomoći opisivati korelacije, svojstva i ponašanja u različitim sustavima. Na primjer, čvorove u kvadratnoj mreži možemo shvatiti kao prekidače koje u svakom potezu "spuštamo" i na taj način omogućujemo da struja proteče u nekom preferiranom smjeru (u ovom slučaju po površini duž osi y). U tom slučaju, na pozicijama y=0 i y=$\sqrt{N}$ nalaze se izolatori pa duž osi x struja ne može proteći.

Cilj ovog istraživanja je saznati kako "pobjeda" *snowball* igrača ovisi o *snowball* parametru p_s i veličini mreže N. U istraživanju smo proučili "sukobe" na mrežama veličine $N = 100$, $N \approx 1000$ i $N = 10000$ čvorova. Za "sukobe" na svakoj od tih mreža varirali smo parametar p_s i odigravali "sezone" od $Z = 1000$ "sukoba", kako bismo saznali u kolikom postotku utakmica p_p pobjeđuje *snowball* igrač.

Također, od interesa nam je bilo koliko resursa troši *snowball* igrač kad pobjeđuje, tj. koliki je omjer broja izabranih čvorova *snowball* i *random* igrača. Naime, ukoliko *snowball* igrač za svoje pobjede troši puno više resursa, tada takva metoda građenja geometrijskog uzorka i nije baš najisplativija. Dakle, zanimat će nas režimi u kojima *snowball* igrač pobjeđuje a da pritom "odabere" manje čvorova od *random* igrača.

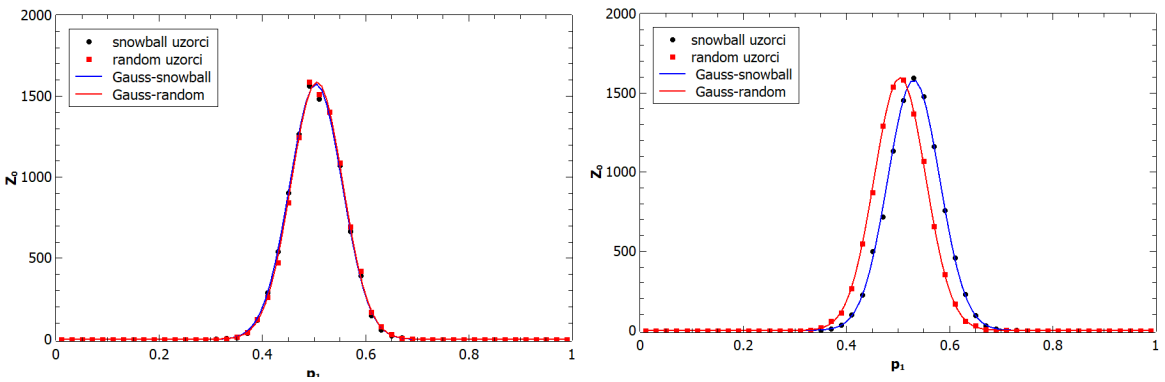
6 Rezultati računalnih modela

U posljednjem dijelu ovog rada predstavit ćemo računalno dobivene kvantitativne usporedbe *random* i *snowball* uzorkovanja na dvjema kompleksnim mrežama za navedene raspodjele te rezultate *snowball-random* "sukoba" u ovisnosti o *snowball* parametru p_s. Svi uzorci obrađivani su u programu *QtiPlot*, a svi rezultati bit će prikazani grafički.

E-R mrežu generirali smo u natkritičnom režimu, s $N = 50000$ čvorova i srednjim stupnjem čvorova $\overline{k} = 10$. Na mreži smo prvo napravili nasumičnu raspodjelu svojstva "1" i raspodjelu zakonom opće potencije (slika 6.1.). Potom smo uzeli $Z = 10000$ uzoraka veličine $M = 100$ *random* i *snowball* metodom. Raspodjele po udjelima p_1 prikazali smo grafički i na svaku prilagodili Gaussian :

$$Z_0(p_1) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}((p_1-\overline{p_1})/\sigma)^2}, \tag{6.1}$$

gdje je Z_0 broj uzoraka za dati udio p_1, $\overline{p_1}$ je srednja vrijednost, a σ standardno odstupanje.



(a) (b)

Slika 6.1: Grafička usporedba *random* i *snowball* uzorkovanja na E-R mreži sa $N = 50000$ čvorova i srednjim stupnjem čvora $\overline{k} = 10$. Pod (a) su prikazana uzorkovanja za nasumičnu razdiobu svojstva "1" sa $p_r = 0.5$. Na x-osi prikazani su udjeli svojstva "1", a na y-osi broj uzoraka za određeni udio Z_0. *Random* i *snowball* uzorkovanje daju iste procjene $p_1^r = p_1^s = 0.50 \pm 0.05$. Pod (b) su prikazana uzorkovanja za zakon opće potencije i $p_1 = 0.49904$. *Snowball* uzorak u ovom slučaju precjenjuje pravu vrijednost pa je tako $p_1^r = 0.50 \pm 0.05$, a $p_1^s = 0.53 \pm 0.05$.

S obzirom na navedeno, rezultate uzorkovanja smo pisali u obliku $p_1^{r,s} = \overline{p_1}^{r,s} \pm \sigma^{r,s}$.

Budući da su nam od interesa u ovom radu bile raspodjele u kojima su svojstva

”0” i ”1” u mreži zastupljena jednako, parametre za pojedine raspodjele svojstava smo odabrali tako da dobijemo $p_1 \approx 0.5$.

Kod nasumične raspodjele, postavili smo parametar $p_r = 0.5$ (točno 50% svojstva ”1” u mreži). Rezultati uzorkovanja prikazani su na Slici 6.1. (a). U skladu s očekivanjima, *random* i *snowball* uzorkovanje daju potpuno isti rezultat ; $p_1^r = p_1^s = 0.50 \pm 0.05$.

Naime, uzrok tome je činjenica da pri nasumičnoj razdiobi ne uzimamo u obzir stupanj čvora k, već svojstva raspodjeljujemo neovisno o njemu. Stoga se naš model *snowball* uzorkovanja ne razlikuje od *random* uzorkovanja, budući da nema nikakvih topoloških preferenci pri raspodjeljivanju svojstava. Isto rezoniranje i rezultati vrijede i za B-A mrežu!

Na Slici 6.1.(b) prikazano je uzorkovanje za raspodjelu svojstava zakonom opće potencije $p_1(k) = \mathrm{A}\left(\frac{k}{k_{\max}}\right)^{\zeta}$. Uz $\zeta = 0.5$ i $A = 0.85$ na čitavoj mreži se dobije udio svojstva ”1” $p_1 = 0.49904$.

Kod ovakve raspodjele svojstava primjećujemo da *snowball* uzorkovanje precjenjuje vrijednost p_1, pa tako dobivamo $p_1^r = 0.50 \pm 0.05$ i $p_1^s = 0.53 \pm 0.05$.

Ovakav rezultat ponukao nas je da provjerimo događa li se precjenjivanje i za neke druge raspodjele svojstava ovisne o stupnju čvora k, pa smo izabrali logaritamsku raspodjelu $p_1(k) = \frac{\log(\mathrm{A}k)}{\log(\mathrm{A}k_{\max})}$ i raspodjelu tangensom hiperbolnim $p_1(k) = \tanh(\frac{k}{\mathrm{A}})$ (Slika 6.2.).

Sa slike vidimo da i za ove raspodjele svojstava postoji precjenjivanje od strane *snowball* uzorkovanja. Kod logaritamske raspodjele uz $A = 0.3$ dobili smo $p_1 = 0.51899$. *Random* uzorkovanje standardno daje dobru procjenu ($p_1^r = 0.52 \pm 0.05$) dok *snowball* metoda daje $p_1^s = 0.58 \pm 0.05$. Primjećujemo da je precjenjivanje kod ovakve raspodjele svojstava nešto veće nego kod zakona opće potencije.

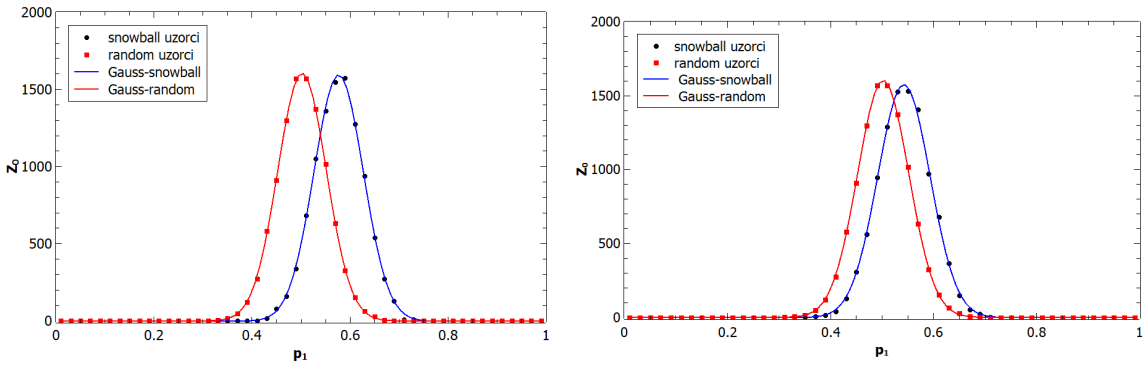
Konačno, za raspodjelu tangensom hiperbolnim, uz $A = 18$ generiran je udio $p_1 = 0.49748$. *Random* i *snowball* metoda daju iste procjene kao i za raspodjelu zakonom opće potencije, tj. $p_1^r = 0.50 \pm 0.05$ i $p_1^s = 0.53 \pm 0.05$.

Ovakvi rezultati zahtijevaju malo detaljniji osvrt. Na jako velikoj E-R mreži, gotovi svi čvorovi u mreži imaju jednak stupanj k i samim time bi vjerojatnost pridijeljivanja svojstva ”1” bila jednaka za gotovo sve čvorove[3]. Dakle, očekivali bi da uzorkovanje

[3]Naravno, postoje čvorovi koji nemaju jednak stupanj, ali u ogromnim mrežama odstupanja od srednjih vrijednosti su malena i rijetka.

ne ovisi o stupnju čvora, a samim time i da naši *random* i *snowball* modeli daju isti rezultat.

Međutim, naša mreža ima "samo" $N = 50000$ čvorova, tako da ipak postoji nezanemariv broj čvorova čiji stupanj k odstupa od $\overline{k}$. Samim time, postoji značajan udio čvorova sa različitim vjerojatnostima pridijeljivanja svojstva "1", a uzorkovanje koje ovisi o k (*snowball*) i ono koje ne ovisi (*random*) će se razlikovati.



(a) (b)

Slika 6.2: Grafička usporedba *random* i *snowball* uzorkovanja na E-R mreži sa $N = 50000$ čvorova i srednjim stupnjem čvora $\overline{k} = 10$. Pod (a) su prikazana uzorkovanja za logaritamsku razdiobu svojstva "1" sa $p_1 = 0.51999$. Na x-osi prikazani su udjeli svojstva "1", a na y-osi broj uzoraka za određeni udio Z_0. *Snowball* metoda daje $p_1^s = 0.58 \pm 0.05$, dok *random* uzorkovanje standardno daje dobru procjenu $p_1^r = 0.52 \pm 0.05$. Pod (b) su prikazana uzorkovanja za raspodjelu tangensom hiperbolnim i $p_1 = 0.49748$. *Snowball* uzorak i u ovom slučaju precjenjuje pravu vrijednost pa je tako $p_1^r = 0.50 \pm 0.05$, a $p_1^s = 0.53 \pm 0.05$.

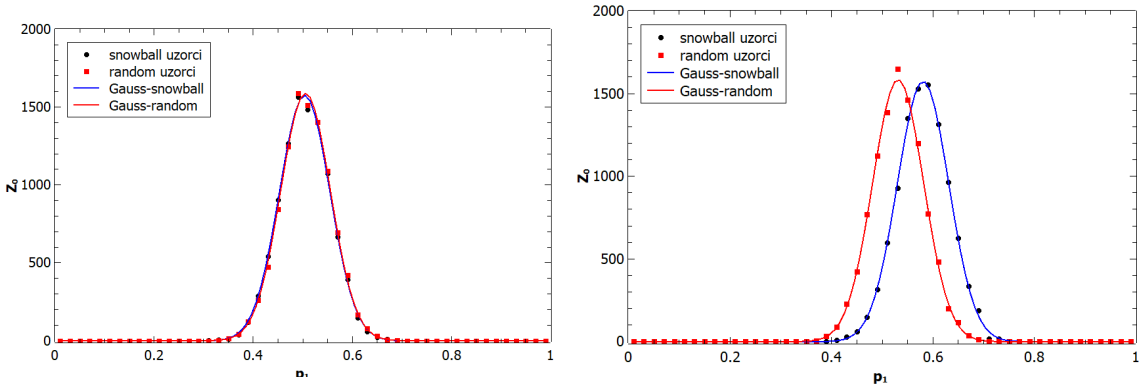
To razlikovanje je fino demonstrirano na gornjim slikama kao pomicanje vrha *Gaussijana* udesno za *snowball* uzorkovanje. Ostaje nam pokušati objasniti zašto *snowball* uzorkovanje precjenjuje stvarnu vrijednost p_1 u kompleksnoj mreži.

Ponudit ćemo kvalitativno objašnjenje. Naime, budući da u mreži postoji nezanemariv broj čvorova sa većim stupnjem od srednjeg stupnja čvora $\overline{k}$, postoji značajna vjerojatnost da *izvorišta* našeg *snowball* uzorkovanja budu spojena sa čvorovima višeg stupnja, a samim time da u uzorak koji je relativno malen ($M = 100$) budu uzeti čvorovi koji vjerojatnije imaju svojstvo "1" (budući da smo izabirali funkcije koje monotono rastu sa k). Stoga, čvorovi višeg stupnja će u uzorku biti zastupljeniji no što su u čitavoj kompleksnoj mreži.

Greška u uzorkovanju je poprilično velika ($\sigma = 0.05$, odnosno 5%). To dolazi zbog činjenice da su naši uzorci vrlo maleni i da ne mogu procjenjivati red veličine manji od 0.01. No, cilj ovog rada je ionako dati samo okvirnu skicu i nagovještaj

ponašanja *snowball* i *random* uzorkovanja. U budućim istraživanjima, dakako, s malo više računalnih mogućnosti, cilj je raditi puno veći ansambl sa puno većim uzorcima kako bismo dobili što precizniju procjenu.

Ista argumentacija za greške vrijedi i kod B-A modela. Tu smo kompleksnu mrežu generirali sa $N = 50000$ čvorova dodavajući ih u mrežu sa $m = 3$ poveznice. Uzorkovanja smo proveli sa istim razdiobama svojstava. Na slici 6.3. vidimo nasumičnu razdiobu i razdiobu po zakonu opće potencije.



(a) (b)

Slika 6.3: Grafička usporedba *random* i *snowball* uzorkovanja na B-A mreži sa $N = 50000$ čvorova i $m = 3$ dodane poveznice. Pod (a) su prikazana uzorkovanja za nasumičnu razdiobu svojstva "1" sa $p_r = 0.5$. Na x-osi prikazani su udjeli svojstva "1", a na y-osi broj uzoraka za određeni udio Z_0. *Snowball* metoda i *random* metoda daju iste rezultate $p_1^s = p_1^r = 0.50 \pm 0.05$. Pod (b) su prikazana uzorkovanja za razdiobu zakonom opće potencije i $p_1 = 0.52716$. *Snowball* uzorak precjenjuje pravu vrijednost pa je tako $p_1^r = 0.53 \pm 0.05$, a $p_1^s = 0.58 \pm 0.05$.

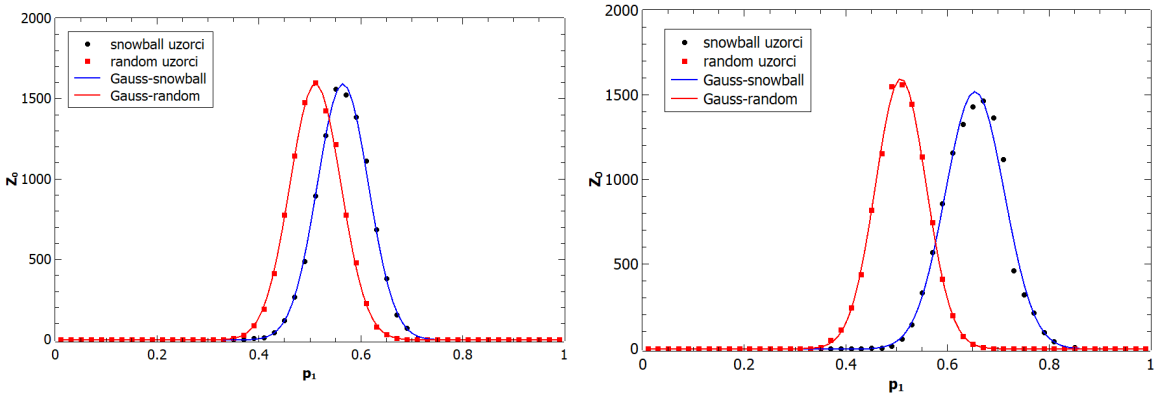
Uzorkovanja u slučaju nasumične razdiobe, kao i kod E-R modela, daju istu procjenu, tj. $p_1^r = p_1^s = 0.50 \pm 0.05$.

U slučaju razdiobe zakonom opće potencije, uz A=1 i $\zeta = 0.14$ dobivamo na čitavoj mreži $p_1 = 0.52741$. Kao i kod E-R kompleksne mreže i ovdje *snowball* metoda uzorkovanja precjenjuje *random* uzorkovanje.

Logaritamska razdioba daje, uz A=20 udio na mreži $p_1 = 0.5057$. Snowball uzorkovanjem dobivamo $p_1^s = 0.56 \pm 0.05$, a *random* metoda već standardno daje dobru procjenu. Razdioba tangensom hiperbolnim, pak, uz A=8.5, daje $p_1 = 0.50674$. *Snowball* uzorkovanje kod ovakve raspodjele svojstava na B-A mreži daje već osjetnije precjenjivanje, uz $p_1^s = 0.65 \pm 0.05$ i $p_1^r = 0.51 \pm 0.05$.

Valja naglasiti kako smo precjenjivanje kod B-A kompleksne mreže očekivali. Intrinzično svojstvo ovog modela je pojavljivanje *hubova*, čvorova s jako velikim stup-

njem k. *Hubovi* će uvijek imati osjetnu veću vjerojatnost posjedovanja svojstva "1" za raspodjele koje monotono rastu s k. Stoga, naša *izvorišta* pri skupljanju uzorka će često biti spojena baš s *hubovima*, što će dovesti do nesrazmjernog uzimanja *hubova* u uzorke i samim time precjenjivana udjela svojstava u uzorku.



(a) (b)

Slika 6.4: Grafička usporedba *random* i *snowball* uzorkovanja na B-A mreži sa $N = 50000$ čvorova i $m = 3$ dodane poveznice. Pod (a) su prikazana uzorkovanja za logaritamsku razdiobu svojstva "1" sa $p_1 = 0.5057$. Na x-osi prikazani su udjeli svojstva "1", a na y-osi broj uzoraka za određeni udio Z_0. *Snowball* metoda i *random* metoda daju različite rezultate $p_1^s = 0.56 \pm 0.05$ i $p_1^r = 0.51 \pm 0.05$. Pod (b) su prikazana uzorkovanja za razdiobu tangensom hiperbolnim i $p_1 = 0.50674$. *Snowball* uzorak precjenjuje pravu vrijednost pa je tako $p_1^r = 0.65 \pm 0.05$, a $p_1^r = 0.51 \pm 0.05$.

Kod B-A modela očekujemo da će se precjenjivanje događati bez obzira na to koliko povećavali mrežu, a mogući problem od interesa u daljnjem istraživanju mogao bi biti ovisnost precjenjivanja o veličini mreže. Želimo li pak reprezentativniji snowball uzorak, tada je od interesa povećati sam uzorak pošto ćemo na taj način uzimati sve manje *hubova* u uzorak i samim time dobiti reprezentativniji prikaz čvorova i njihovih stupnjeva u mreži. Dakako, u nekim realnim situacijama, od interesa nam je potrošiti što manje resursa, tako da je problem od daljnjeg interesa i onaj balansiranja dovoljno dobre procjene i "cijene" koju za tu procjenu trebamo platiti.

Valja primjetiti i kako na ovih nekoliko primjera vidimo da je precjenjivanje na B-A mreži veće nego na E-R mreži, što je u skladu s očekivanjima. Također, vidimo da za različite raspodjele svojstava imamo nešto drukčije precjenjivanje, pa bi bilo vrlo interesantno proučiti kako se ponašaju uzorkovanja za različite razdiobe te, po mogućnosti, razviti i teorijski model koji bi nam pomogao u dubljem razumijevanju simulacijskih rezultata.

Također, daljnje istraživanje zahtijevat će precizniju kvantifikaciju statističke značajnosti precijenjivanja koje daje *metoda snježne grude*. Ako kažemo da je neko precijenjivanje statistički značajno, onda smo zapravo ustvrdili da precijenjivanje, koja je nađeno, bez obzira na veličinu precijenjivanja, nije slučajno, već da vrlo vjerojatno postoji i među populacijom. Zanima nas, naime, kako se odnosi razlika aritmetičkih sredina dva tipa uzorkovanja sa standardnom greškom te razlike (*t*-vrijednost). Radi jednostavnosti, uzmimo kako su naša dva tipa uzorkovanja nezavisna. Tada *t*-vrijednost možemo zapisati kao [24]:

$$t = \frac{\overline{p_1^s} - \overline{p_1^r}}{s_{\overline{p_1^s} - \overline{p_1^r}}}, \tag{6.2}$$

gdje je nazivnik *standardna greška* razlike aritmetičkih sredina koja se za nezavisna uzorkovanja računa kao:

$$s_{\overline{p_1^s} - \overline{p_1^r}} = \sqrt{\frac{(\sigma^s)^2}{Z_s} + \frac{(\sigma^r)^2}{Z_r}}. \tag{6.3}$$

Brojnici u izrazu pod korijenom predstavljaju standardne devijacije kod uzorkovanja *metodom snježne grude* i *nasumičnog* uzorkovanja, dok su u nazivnicima veličine ansambla uzoraka, koje su u našem radu jednake ($Z_s = Z_r = Z = 10000$). Izračunatoj *t*-vrijednosti možemo pridružiti tabličnu vjerojatnost [24] koja nam kaže kolika je šansa da je naše precijenjivanje (ili podcijenjivanje) slučajno. Uvriježena granična vrijednost statističke značajnosti [24] je $t_c = 1.96$, odnosno 5%. Dakle, za sve *t*-vrijednosti veće od navedene, vjerojatnost slučajnog precijenjivanja manja je od 5%, tj. kažemo kako je naše precijenjivanje statistički značajno.

Promotrimo radi primjera slučaj u kojem su razlike aritmetičkih sredina najmanje: raspodjela zakonom opće potencije na E-R mreži. Standardne devijacije uzorkovanja su iste ($\sigma^s = \sigma^r = 0.05$). Uz pomoć (6.2) i (6.3) dobivamo $t = 42.4$, što je znatno veće od granične vrijednosti t_c, pa samim time možemo reći kako je naše precijenjivanje statistički značajno.

Pri svim drugim raspodjelama svojstava ovisnih o k razlike aritmetičkih sredina su veće ili jednake gore opisanom slučaju. Budući da su standardne devijacije, isto kao i veličine ansambla, svugdje iste (ista standardna greška razlike), slijedi kako bi u svim preostalim slučajevima dobivali *t*-vrijednost veću ili jednaku kao u primjeru, što sva naša precijenjivanja čini statistički značajnima.

Ovo je pojednostavljeni primjer, budući da su naša dva uzorkovanja međusobno zavisna (provode se na ISTOJ populaciji). Tada u izraz (6.3) moramo uključiti i koeficijent korelacije q [24]:

$$s_{\overline{p_1^s}-\overline{p_1^r}} = \sqrt{\frac{(\sigma^s)^2}{Z_s} + \frac{(\sigma^r)^2}{Z_r} - 2q\frac{\sigma_s}{\sqrt{Z_s}}\frac{\sigma_r}{\sqrt{Z_r}}}\,. \tag{6.4}$$

Računanje koeficijenta korelacije nije dio ovog diplomskog rada, već daljnjeg istraživanja, no radi kvalitativnog opisa dovoljno je znati kako se radi o vrijednosti u rasponu od -1 (snažna negativna korelacija) do 1 (snažna pozitivna korelacija)[4]. Tada, promotrimo li izraz (6.4), možemo zaključiti kako postojanje pozitivne korelacije između dva tipa uzorkovanja pridonosi smanjenju standardne greške razlike, a samim time i većoj *t*-vrijednosti, odnosno većoj statističkoj značajnosti precijenjivanja. Za negativnu korelaciju, primijenimo na gornjem primjeru krajnju vrijednost; snažnu negativnu korelaciju $q = -1$. Tada računski dobivamo $t = 30$, što je i dalje osjetno više od t_c.

Za završni dio ovog rada ostavili smo *random - snowball* sukob, opisan u prijašnjem poglavlju. Proučavali smo kako ovisi postotak *snowball* p_p o parametru p_s U "sezonama" od $Z = 1000$ "sukoba". Mjerenja spomenute ovisnosti prilagodili smo logističkom funkcijom (slika 6.5.) za tri veličine mreža ($N = 100$, $N \approx 1000$ i $N = 10000$) :

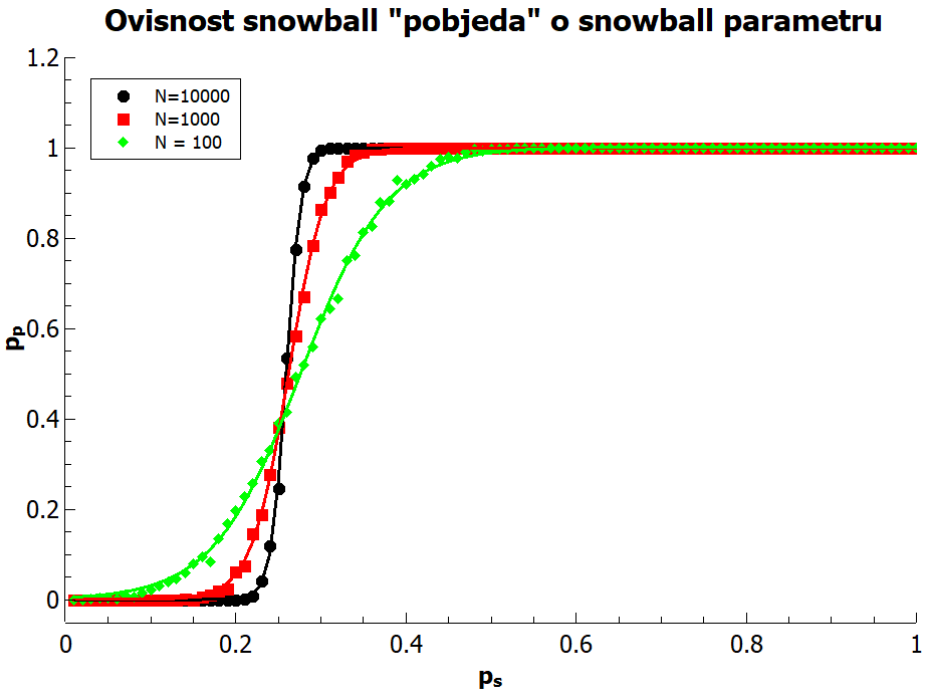
$$p_s(p_p) = \frac{L}{1 + \mathrm{e}^{-t(p_p - x_c)}} \tag{6.5}$$

gdje L označava maksimalnu vrijednost funkcije ("zasićenje"), t je nagib, a x_c je "središnja točka" funkcije. Konkretno, u našem slučaju, za sve tri veličine mreža $L = 1$, što je vidljivo iz grafa na slici 6.5. To znači da će u sva tri slučaja postojati neki p_s nakon kojeg će *snowball* "igrač" pobijediti u SVIM "sukobima" u sezoni.

Parametar t govori nam koliko brzo će funkcija narasti sa vrijednosti 0 na vrijednost L. U našem slučaju, kvalitativno nam predočava koliko je širok "pojas" *snowball* parametra p_s u kojem postoji kompeticija, tj. u kojem se neće dogoditi da bilo jedan bilo drugi igrač pobijedi u SVAKOM "sukobu" u sezoni ($p_p = 0,1$)

Na Slici 6.5. vidimo kako je za veće mreže, "pojas" u kojem dolazi do kompeticije sve uži, tj t sve više raste. Konkretno, za mrežu $N = 100$ dobili smo $t = 19.6 \pm 0.2$, za

[4] za $q = 0$ nema korelacije među uzorkovanjima



Slika 6.5: Prilagodba logističke funkcije $p_s(p_p) = \frac{L}{1+e^{-t(p_p - x_c)}}$. Na osi x prikazan je *snowball* parametar, dok je na osi y postotak pobjeda *snowball* igrača u sezoni od $Z = 1000$ sukoba. Što je mreža veća, t (nagib) je veći, tj. funkcija postiže maksimalnu vrijednost L puno brže.

$N = 1000$ imamo $t = 45.6 \pm 0.3$, a za $N = 10000$, nagib je $t = 114.2 \pm 0.08$.

Na kraju, imamo parametar x_c koji nam kaže kritičnu vrijednost p_s nakon koje *snowball* "igrač" pobjeđuje u više "sukoba" u sezoni od *random* "igrača" ($p_p > 0.5$). Ovaj parametar se smanjuje povećanjem mreže, pa je tako za $N = 100$ njegova vrijednost $x_c = 0.2761 \pm 0.0005$, za $N = 1000$ imamo $x_c = 0.2619 \pm 0.0002$ i konačno za $N = 10000$ vrijednost je $x_c = 0.2590 \pm 0.0007$. Od daljnjeg interesa u istraživanju bit će limes parametra x_c kada $N \to \infty$.

Još nam je od interesa u ovom problemu bio režim p_{s_0} u kojem *snowball* "igrač" pobjeđuje u većini "sukoba" u "sezoni" ($p_p > 0.5$) a da pritom troši manje "resursa" od *random* "igrača", tj. da pobijedi u sukobu a da je pritom uzeo manje čvorova u uzorak. Donja granica ovakvog režima je za sve veličine mreža x_c. Dobivamo :

- Za $N = 100$; $p_{s_0} \in [0.2761, 0, 3012]$
- Za $N \approx 1000$; $p_{s_0} \in [0.2619, 0, 3078]$
- Za $N = 10000$; $p_{s_0} \in [0.2590, 0, 3113]$

Vidimo da je "pojas" u kojem nam se "isplati" graditi naš geometrijski uzorak

jako uzak. Daljnja istraživanja građenja geometrijskih uzoraka provodit ćemo na složenijim geometrijama od jednostavne kvadratne.

7 Zaključak

U ovom radu usporedili smo dva tipa uzorkovanja: uzorkovanje *metodom snježne grude* i *nasumično* uzorkovanje. Uzorkovanja su provedena na dva najpoznatija mrežna modela, Erdős-Rényi (E-R) i Barabási-Albert (B-A), na kojima smo raspodijelili svojstva "0" i "1". Svojstva su raspodijeljena nasumično i u ovisnosti o stupnju čvora k.

Parametre pri raspodjeli svojstava smo birali tako da dobijemo udio svojstva $p_1 \approx 0.5$. Rezultati koje smo dobili uzorkovanjem pri nasumičnoj raspodjeli svojstava pokazuju kako između *nasumičnog* uzorkovanja i uzorkovanja *metodom snježne grude* ne postoji razlika (na oba mrežna modela), što je i očekivano s obzirom na činjenicu da nasumična raspodjela nema ovisnost o stupnju čvora.

Kod raspodjela ovisnih o stupnju čvora k, *metoda snježne grude* na oba mrežna modela precjenjuje udio svojstva "1", s tim da je na B-A mreži precjenjivanje veće. Također, precjenjivanje je različito za različite raspodjele ovisne o stupnju čvora k. Rezultati ovog rada otvaraju širok spektar daljnjih istraživačkih mogućnosti: sustavnije razmatranje uzorkovanja pri različitim parametrima kompleksnih mreža; matematičko modeliranje uzorkovanja *metodom snježne grude*; razmatranje dodatnih raspodjela svojstava u ovisnosti o k i dr.

U posljednjem dijelu rada povezali smo modele uzorkovanja sa fizikalnim fenomenom *perkolacija* preko *random-snowball* "sukoba" i pokazali pri kojim uvjetima *snowball* "igrač" pobjeđuje u postizanju perkolacije na kvadratnoj mreži. Otvorili smo mogućnost daljnjeg istraživanja ovog fenomena i to na mrežama koje predstavljaju različite fizikalne sustave i njihova svojstva.

Ovo istraživanje moglo bi biti od daljnje koristi u analizi, kako društvenih, tako i fizikalnih kompleksnih sustava. Od iznimnog fizikalnog značaja je istraživanje svojstava u fizici čvrstog stanja, od već spomenutih perkolacija, preko Bose-Einstenovog kondenzata [9] sve do svojstava kristalnih sustava. Od temeljnog će značaja biti dizajniranje *metode snježne grude* u fizikalnom ispitivanju takvih sustava.

Dodaci

Dodatak A Erdős-Rényi mreža

A.1 Generiranje mreže

```
import random
import networkx as nx
import matplotlib.pyplot as plt

def erdos_renyi(G, p):

        nodes = list(G.nodes())
        brojNodes = len(nodes)
        for i in range(0, brojNodes):
                node1 = nodes[i]
                print(i)
                for j in range(i, brojNodes):
                        node2 = nodes[j]
                        if node1 != node2:
                                r = random.random()
                                if r <= p:
                                        G.add_edge(node1, node2)
                                        ne = [(node1, node2)]
                                        display_graph(G, '', ne)
                                else:
                                        display_graph(G, '', '')
```

A.2 Shema dodavanja poveznica

```
def display_graph(G, cvor, ne):
        pos = nx.circular_layout(G)
        if cvor == '' and ne == '':
                novi_cvor = []
```

```
            ostali_cvorovi =G.nodes()
            novi_rub = []
            ostali_rubovi = G.edges()

        elif cvor== '':
            novi_cvor = []
            ostali_cvorovi = G.nodes()
            novi_rub = ne
            ostali_rubovi = list(set(G.edges()) - set(novi_rub) -
                                 set([(b,a) for (a,b) in novi_rub]))

        else:
            novi_cvor = [cvor]
            ostali_cvorovi = list( set(G.nodes())-set(novi_cvor) )
            novi_rub = ne
            ostali_rubovi = list( set(G.nodes())-set(novi_rub) -
                                 set([(b,a) for (a,b) in novi_rub]) )

    nx.draw_networkx_nodes(G, pos, nodelist = novi_cvor,
                           node_color='g')
    nx.draw_networkx_nodes(G, pos, nodelist = ostali_cvorovi,
                           node_color='r')
    nx.draw_networkx_edges(G, pos, edgelist = novi_rub,
                         edge_color='b', style='dashdot')
    nx.draw_networkx_edges(G, pos, edgelist = ostali_rubovi,
                           edge_color='r')


        plt.show()
```

A.3 Distribucija stupnja čvora

```
def plot_deg_dist(G):
    all_degrees = [j for (i,j) in G.degree()]
```

```
        unique_degrees = list(set(all_degrees))
        unique_degrees.sort()

        count_of_degrees = []
        for i in unique_degrees:
                c = all_degrees.count(i)
                count_of_degrees.append(c)

        plt.plot(unique_degrees, count_of_degrees, 'ro-')
        plt.xlabel('Stupanj')
        plt.ylabel('Broj cvorova')
        plt.title('Distribucija E-R')
        plt.show()
```

Dodatak B Barabási-Albert mreža

B.1 Generiranje mreže

```
import random
import networkx as nx
import matplotlib.pyplot as plt

def dodaj_cvor(G, nd, n0, m):
for cvor in range(n0,nd):
        if cvor%100 == 0:
                print(cvor)

        vjerojatnost = []
        stupnjevi = [j for (i,j) in nx.degree(G)]
        norma = sum(stupnjevi)

        for i in stupnjevi:
                vjerojatnost.append (float(i)/norma)
```

```
        G.add_node(cvor)

        p0 = 0
        kumulativna = []
        for i in vjerojatnost:
                kumulativna.append(p0+i)
                p0 += i

        dodane_poveznice = []
        broj_dodanih = 0

        while broj_dodanih < m:
            r = random.random()

            for indeks, vrijednost in enumerate(kumulativna):
                        if r<vrijednost:

                                if not(indeks in dodane_poveznice):
                                        G.add_edge(cvor, indeks)
                                        dodane_poveznice.append(indeks)
                                        broj_dodanih += 1
                                        display_graph(G,'',[(cvor,
                                                indeks)])
                                break
```

B.2 Shema dodavanja poveznica

```
def display_graph(G, cvor, ne):
        pos = nx.circular_layout(G)
        if cvor == '' and ne == '':
                novi_cvor = []
                ostali_cvorovi =G.nodes()
                novi_rub = []
```

```
            ostali_rubovi = G.edges()

        elif cvor== '':
            novi_cvor = []
            ostali_cvorovi = G.nodes()
            novi_rub = ne
            ostali_rubovi = list(set(G.edges()) - set(novi_rub) -
                                 set([(b,a) for (a,b) in novi_rub]))

        else:
            novi_cvor = [cvor]
            ostali_cvorovi = list( set(G.nodes())-set(novi_cvor) )
            novi_rub = ne
            ostali_rubovi = list( set(G.nodes())-set(novi_rub) -
                                 set([(b,a) for (a,b) in novi_rub]) )

    nx.draw_networkx_nodes(G, pos, nodelist = novi_cvor,
                           node_color='g')
    nx.draw_networkx_nodes(G, pos, nodelist = ostali_cvorovi,
                           node_color='r')
    nx.draw_networkx_edges(G, pos, edgelist = novi_rub,
                         edge_color='b', style='dashdot')
    nx.draw_networkx_edges(G, pos, edgelist = ostali_rubovi,
                           edge_color='r')


        plt.show()
```

B.3 Distribucija stupnja čvora

```
def plot_deg_dist(G):
    all_degrees = [j for (i,j) in G.degree()]

        unique_degrees = list(set(all_degrees))
        unique_degrees.sort()
```

```
    count_of_degrees = []
    for i in unique_degrees:
            c = all_degrees.count(i)
            count_of_degrees.append(c)

    plt.plot(unique_degrees, count_of_degrees, 'ro-')
    plt.xlabel('Stupanj')
    plt.ylabel('Broj cvorova')
    plt.title('Distribucija B-A')
    plt.show()
```

Dodatak C Uzorkovanja

C.1 Nasumično uzorkovanje

```
def randomSample(nd, M, B):
    uzorci = np.zeros([M])

    j=random.sample(range(nd),M)

    for i in range (0,M):
            p=j[i]
            uzorci[i]+=B[p]

    return sum(uzorci)/M
```

C.2 Uzorkovanje metodom snježne grude

```
def snowball(p10, nd, M, poveznice, B):

    snowTrenutni = [random.randint(0,nd-1)]
    zabranjeni = []
    uzorci = []
    cvorovi = np.arange(nd)
```

```
p1 = p10
while len(uzorci) < M:

        if len(snowTrenutni) < 1:
                kandidat = random.randint(0,nd-1)
                while kandidat in zabranjeni:
                        kandidat = random.randint(0,nd-1)
                snowTrenutni = [kandidat]
                p1 = p10
                print "novi"

        snowNovi = []
        for cvor in snowTrenutni:
                uzorci.append(B[cvor])
                zabranjeni.append(cvor)
                if len(uzorci) == M:
                        break

                povezani = poveznice[cvor, :]
                povezaniCvorovi = cvorovi[povezani==1]

                for cvorTMP in povezaniCvorovi:
                        if cvorTMP in zabranjeni:
                                pass
                        else:
                                k = random.random()
                                if k<p1:
                                    snowNovi.append(cvorTMP)

        snowTrenutni = snowNovi

return sum(uzorci)/M
```

Literatura

[1] Ray Solomonoff and Anatol Rapoport. Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117, June 1951.

[2] P. Erdős and A. Rényi. On the Evolution of Random Graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[3] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, June 1998.

[4] Alain Barrat and Martin Weigt. On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, 13:547–560, 2000.

[5] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999.

[6] S. N. Dorogovtsev, J. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(21):4633–4636, November 2000.

[7] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, April 2001.

[8] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378, July 2000.

[9] Albert-László Barabási and Márton Pósfai. *Network Science*. Cambridge University Press, July 2016. Google-Books-ID: iLtGDQAAQBAJ.

[10] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, March 2004.

[11] Steve Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer-Verlag, New York, 2011.

[12] E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numer. Math.*, 1(1):269–271, December 1959.

[13] Sven Peyer, Dieter Rautenbach, and Jens Vygen. A generalization of Dijkstra's shortest path algorithm with applications to VLSI routing. *Journal of Discrete Algorithms*, 7(4):377–390, December 2009.

[14] Dijkstra's Algorithm: Definition, Applications & Examples - Video & Lesson Transcript. 2019-07-12, `http://study.com/academy/lesson/dijkstra-s-algorithm-definition-applications-examples.html`.

[15] Paul W. Holland and Samuel Leinhardt. Transitivity in Structural Models of Small Groups. *Comparative Group Studies*, 2(2):107–124, May 1971.

[16] Armin Bunde and Shlomo Havlin, editors. *Fractals and Disordered Systems*. Springer-Verlag, Berlin Heidelberg, 2 edition, 1996.

[17] Deokjae Lee, Y. S. Cho, K.-I. Goh, D.-S. Lee, and B. Kahng. Recent advances of percolation theory in complex networks. *Journal of the Korean Physical Society*, 73(2):152–164, July 2018. arXiv: 1808.00905.

[18] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the World-Wide Web. *Nature*, 401(6749):130, September 1999.

[19] Barabási Albert-László. Barabási Albert-László - Books.

[20] F. Eggenberger and G. Pólya. Über die Statistik verketteter Vorgänge. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 3(4):279–289, 1923.

[21] Robert Gibrat. *Les Inégalités économiques*. Recueil Sirey, Paris, 1931. OCLC: 250098599.

[22] Bela Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001.

[23] Hattori A Lance P. Sampling and Evaluation – A Guide to Sampling for Program Impact Evaluation — MEASURE Evaluation. 2019-07-24, `https://www.measureevaluation.org/resources/publications/ms-16-112`.

[24] Boris Petz, Vladimir Kolesarić, and Dragutin Ivanec. Petzova statistika. Osnovne statističke metode za nematematičare. *Naklada Slap* 2007. str. 111-155. ISBN: 978-953-191-058-3.

[25] Goran Milas. Istrazivacke metode u psihologiji i drugim drustvenim znanostima. *Naklada Slap*. 2005, ISBN: 978-953-191-283-1.

[26] Leo A. Goodman. Snowball Sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, March 1961.

[27] Charles Kaplan, Dirk Korf, and Claire Sterk. Temporal and Social Contexts of Heroin-Using Populations An Illustration of the Snowball Sampling Technique. *The Journal of Nervous and Mental Disease*, 175(9):566–574, September 1987.

[28] Jennifer L. Syvertsen, Angela M. Robertson, Daniela Abramovitz, M. Gudelia Rangel, Gustavo Martinez, Thomas L. Patterson, Monica D. Ulibarri, Alicia Vera, Nabila El-Bassel, Steffanie A. Strathdee, and Proyecto Parejas. Study protocol for the recruitment of female sex workers and their non-commercial partners into couple-based HIV research. *BMC public health*, 12:136, February 2012.

[29] Tamar Arieli. Israeli-Palestinian border enterprises revisited. *Journal of Borderlands Studies*, 24(2):1–14, June 2009.

[30] Alireza Rezvanian, Mohammad Rahmati, and Mohammad Reza Meybodi. Sampling from complex networks using distributed learning automata. *Physica A: Statistical Mechanics and its Applications*, 396:224–234, February 2014.