Dinamički procesi na kompleksnim mrežama

Petric Maretić, Hermina

Master's thesis / Diplomski rad

2014

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:651545

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Hermina Petric Maretić

DINAMIČKI PROCESI NA KOMPLEKSNIM MREŽAMA

Diplomski rad

Voditelj rada:
prof. dr. sc. Sanja Singer

Zagreb, rujan 2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Sadržaj **iii**

Uvod **2**

1 Prikaz i analiza mreža **3**
1.1 Osnovni pojmovi teorije grafova 3
1.2 Neke veličine bitne za analizu mreža 4
1.3 Općenita svojstva mreža 9

2 Teorijski modeli mreža **11**
2.1 Modeli statičkih mreža 11
2.2 Modeli difuzije 15

3 Model linearnog utjecaja **19**
3.1 Formalan opis modela 20
3.2 Procjenjivanje parametara modela 20
3.3 Difuzija informacija u mreži vijesti 22

4 Analiza japanske automobilske industrije **25**
4.1 Analiza najvažnijih kompanija 27
4.2 Mreže utjecaja nekih kompanija 31

Bibliografija **37**

Uvod

Pojavom velikih količina informacija i napretkom društva prirodno se pojavila potreba za povezivanjem podataka i analizom informacija. Osnovni koncept, ali i intuitivno prva asocijacija za povezivanje su mreže. Pojavom mreža složenih struktura uzrokovanih kompleksnošću podataka koje je potrebno povezati i vezama čija protočnost mora podržavati širok aspekt struktura komunikacije, pojavila se i potreba za strukturiranjem i analizom mreža.

Što je zapravo kompleksna mreža i kako je prepoznati? Bitno je uočiti razliku između kompleksnog i kompliciranog sustava – avion je vrlo kompliciran, ali nije kompleksan. Iako ne postoji prihvaćena definicija kompleksnog sustava, možemo ih prepoznati po nekim obilježjima. Pogledajmo primjer Interneta. Za njega ne postoji globalni plan za strukturu mreže, već se ona razvija samostalno. S druge strane, avion ima predefiniran plan gradnje. Internet se mijenja "iznutra", odnosno svatko može dodati čvor i tako promijeniti mrežu koja s vremenom evoluira. Usto, kompleksne mreže često imaju sposobnost prilagođavanja, pa mogu i odolijevati raznim napadima. Jasno je da za avion s nekoliko kvarova ne možemo reći da funkcionira na zadovoljavajućoj razini, dok u slučaju kvara nekoliko rutera, Internet i dalje dobro funkcionira na globalnoj razini.

Dinamički procesi na mrežama prirodan su način za modeliranje brojnih pojava u prirodnim, ali i u društvenim i tehnološkim krugovima. U ovom ćemo radu predstaviti dinamičku analizu mreža na dvije razine – kroz difuziju informacija u kompleksnoj mreži te kroz analiziranje promjena na mrežama u različitim vremenskih trenutcima i analizu osjetljivosti na izbacivanje nekih vrhova. Rad je strukturiran tako da prvo dajemo teorijsku pozadinu kompleksnih mreža i definiramo neke modele, a zatim to sve dodatno objašnjavamo kroz primjenu.

Prvo se poglavlje bavi osnovnim pojmovima potrebnim za razumijevanje prikaza mreže te nekim veličinama i svojstvima bitnim za analizu mreže.

Drugo poglavlje predstavlja neke teorijske modele mreža, kao i osnovne modele difuzije na mrežama.

Treće poglavlje detaljno opisuje model linearnog utjecaja i prikazuje primjenu

istog na difuziju informacija u mreži vijesti.

Četvrto poglavlje bavi se analizom razvoja i suradnje u japanskoj automobilskoj industriji kroz godine.

Poglavlje 1

Prikaz i analiza mreža

Teorija koja opisuje i analizira kompleksne mreže svoje temelje ima u teoriji grafova i statističkoj analizi. Kompleksnu mrežu opisujemo pomoću grafa G, tako da podatke prikazujemo čvorovima, a veze između njih bridovima.

Ovisno o svojstvima mreže, taj graf može biti usmjeren ili neusmjeren, težinski ili bez težina. Stoga u ovom poglavlju iznosimo osnovne pojmove teorije grafova, kao i neke specijalne veličine bitne za analizu mreža.

1.1 Osnovni pojmovi teorije grafova

Definicija 1.1.1. ***(Neusmjereni) Graf*** *je uređeni par $G = (V, E)$, gdje je $V = V(G)$ neprazan skup čije elemente nazivamo vrhovima, a $E = E(G)$ je familija dvočlanih podskupova od V koje nazivamo bridovima.*

Pritom vrhove u i v koji su pridruženi bridu e nazivamo krajevima brida e, a brid čiji se krajevi podudaraju nazivamo ***petljom****.*

Definicija 1.1.2. ***Usmjereni graf*** *je uređeni par $G = (V, E)$, gdje je $V = V(G)$ neprazan skup čije elemente nazivamo vrhovima, a E=E(G) je skup uređenih parova elemenata iz V, čije elemente nazivamo usmjerenim bridovima.*

Definicija 1.1.3. ***Multiskup*** *M na skupu S je uređeni par (S, m), gdje je m funkcija, $m : S \to \mathbb{N}_0$.*

Definicija 1.1.4. ***Neusmjereni (Usmjereni) multigraf*** *je neusmjereni (usmjereni) graf čiji bridovi čine multiskup.*

Napomena 1.1.5. *Ako želimo naglasiti da se ne radi o multigrafu, graf nazivamo* ***jednostavnim grafom****.*

Definicija 1.1.6. ***Težinski graf*** *G je graf kojem je pridružena funkcija $f : E \to \mathbb{R}$, koja svakom bridu grafa pridružuje njegovu težinu. Tu funkciju nazivamo* ***težinskom funkcijom****.*

Napomena 1.1.7. *Graf koji nije težinski možemo nazivati još i* ***grafom bez težina****.*

Definicija 1.1.8. *Za par vrhova u i v kažemo da su* ***susjedni*** *ako postoji brid e kojemu su oni krajevi. Pritom kažemo da je brid e* ***incidentan*** *s vrhovima u i v te u slučaju usmjerenog grafa pišemo $e = (u, v)$, a u slučaju neusmjerenog $e = \{u, v\}$. U općenitom slučaju koristi se oznaka $e = uv$.*

Definicija 1.1.9. ***Stupanj vrha*** *v, u oznaci $\deg(v)$, u neusmjerenom (multi)grafu G je broj bridova grafa G incidentnih s v, pri čemu se svaka petlja računa kao dva brida.*

Izolirani vrh *ima stupanj 0.* ***Ulazni stupanj vrha*** *v u usmjerenom (multi)-grafu G je broj bridova grafa G koji završavaju u vrhu v, odnosno broj bridova oblika $e = (x, v)$. Analogno se definira* ***izlazni stupanj vrha****.*

Definicija 1.1.10. ***Šetnja*** *u (usmjerenom) (multi)grafu je niz $(v_0, e_1, v_1, e_2, v_2, \ldots, e_k, v_k)$, gdje je e_i brid $v_{i-1}v_i$, za $i = 1, \ldots, k$.* ***Put*** *je šetnja u kojoj su svi vrhovi različiti (osim eventualno početnog i završnog).* ***Duljina puta*** *je broj bridova u nizu za bestežinske (multi)grafove, odnosno suma vrijednosti težinske funkcije svih bridova u nizu za težinske (multi)grafove.*

Kažemo da su dva vrha ***povezana*** *ako postoji put između njih.*

Kažemo da je graf ***povezan*** *ako su svi njegovi vrhovi povezani.*

Definicija 1.1.11. ***Komponenta povezanosti*** *(usmjerenog) (multi)grafa je maksimalni povezani podgraf grafa.*

Definicija 1.1.12. *Neka je $G = (V, E)$ (usmjereni) (multi)graf i $n = |V|$ broj vrhova. Neka je $V = \{v_1, v_2, \ldots, v_n\}$.* ***Matrica susjedstva*** *grafa G je matrica $A \in M^{n\times n}$ koja na mjestu (i, j) ima vrijednost jednaku broju bridova incidentnih vrhovima v_i i v_j.*

Napomena 1.1.13. *Primijetimo da u slučaju jednostavnog grafa matrica A poprima vrijednosti u skupu $\{0, 1\}$.*

1.2 Neke veličine bitne za analizu mreža

U analizi mreže vrlo je bitno moći odrediti važnost određenog vrha te istaknuti one vrhove koji su najutjecajniji. U tu svrhu definiramo nekoliko veličina, tzv. centralnosti, među kojima su najpopularnije centralnost stupnja (engl. degree centrality),

centralnost blizine (engl. closeness centrality), centralnost svojstvenog vektora (engl. eigenvector centrality) i centralnost između vrhova (engl. betweenness centrality).

U neusmjerenom grafu **centralnost stupnja** vrha i jednaka je stupnju tog vrha, $k_i = \deg(v)$.

U usmjerenom grafu definiramo **ulaznu centralnost stupnja** $k_{in,i}$, **izlaznu centralnost stupnja** $k_{out,i}$ i **ukupnu centralnost stupnja** $k_i = k_{in,i} + k_{out,i}$.

Napomena 1.2.1. *Često se kao centralnost stupnja koristi i omjer stupnja vrha s najvećim mogućim stupnjem vrha,*

$$k_i = \frac{\deg(v)}{n-1}. \tag{1.1}$$

Ova veličina na najjednostavniji način opisuje koliko je dobro vrh i povezan s ostalim vrhovima u grafu.

Napomena 1.2.2. *Ako nije drugačije naglašeno, u daljnjem tekstu sve su definicije izražene za neusmjerene grafove bez težina. Definicije se proširuju na težinske i/ili usmjerene grafove na način analogan gornjim primjerima.*

Definicija 1.2.3. *Neka je $G = (V, E)$ graf i označimo s $l_{i,j}$ duljinu najkraćeg puta od vrha i do vrha j. Tada*

$$d_G = \max_{i,j} l_{i,j} \tag{1.2}$$

*zovemo **promjer grafa** G.*

Ako je još $n = |V|$ broj vrhova grafa G, tada

$$\langle l \rangle = \frac{1}{n(n-1)} \sum_{i,j} l_{i,j} \tag{1.3}$$

*označava **prosječnu duljinu najkraćeg puta**.*

Napomena 1.2.4. *Ako ne postoji put između vrha i i vrha j, formalno definiramo duljinu najkraćeg puta kao $l_{i,j} = \infty$. Ipak, često želimo izračunati svojstva i u nepovezanim grafovima, pa za tu udaljenost koristimo veliku konstantu ili broj vrhova u grafu.*

Definicija 1.2.5. *Neka je $G = (V, E)$ graf te $i \in V$ neki njegov vrh. Za vrh $j \in V$ označimo s $l_{i,j}$ duljinu najkraćeg puta od vrha i do vrha j. Tada sa*

$$g_i = \frac{1}{\sum_{j \neq i} l_{i,j}} \tag{1.4}$$

*definiramo **centralnost blizine**.*

Velika centralnost blizine upućuje na vrh čiji je prosječni najkraći put do drugih čvorova malen, odnosno baš na vrh koji je blizu ostalim čvorovima.

Dosad nabrojene veličine dobro opisuju povezanost vrha s ostatkom grafa, ali mogu zanemariti vrhove koji nemaju veliki stupanj, a povezuju razna, inače slabo povezana područja. Centralnost između vrhova opisuje važnost vrha u ulozi mosta između različitih dijelova mreže.

Definicija 1.2.6. *Neka je σ_{hj} broj najkraćih puteva od h do j, a $\sigma_{hj}(i)$ broj najkraćih puteva od h do j koji prolaze kroz i. Tada definiramo*

$$L_i = \sum_{h \neq j \neq i} \sigma_{hj}(i) \tag{1.5}$$

naglašenu centralnost *(engl. stress centrality) i*

$$b_i = \sum_{h \neq j \neq i} \frac{\sigma_{hj}(i)}{\sigma_{hj}} \tag{1.6}$$

centralnost između vrhova*.*

Kad bismo promatrali utjecajnost nekog elementa u mreži, vrlo važan faktor bio bi povezanost s drugim utjecajnim elementima mreže. Matematički to možemo zapisati pomoću sustava jednadžbi

$$x_v = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t, \tag{1.7}$$

gdje je $A = (a_{v,t})$ matrica susjedstva, t i v vrhovi grafa $G = (V, E)$, a $\lambda > 0$ neka konstanta. Zapišemo li taj sustav matrično, vidimo da se zapravo radi o svojstvenom problemu:

$$Ax = \lambda x. \tag{1.8}$$

Uz dodatni uvjet da sve vrijednosti u x budu pozitivne, može se pokazati da traženim svojstvima rezultira samo svojstveni vektor uz najveću svojstvenu vrijednost [9].

Definicija 1.2.7. *Neka je $G = (V, E)$ graf i $A = (a_{v,t})$ matrica susjedstva. Tada je* ***centralnost svojstvenog vektora*** *jednaka svojstvenom vektoru pridruženom najvećoj svojstvenoj vrijednosti λ matrice A.*

Napomena 1.2.8. *Zanimljivo je da je Google, u današnje doba vjerojatno najpopularniji pretraživač "www" mreže, donedavno za računanje rezultata pretraživanja koristio algoritam koji je u svojoj osnovi upravo računanje centralnosti svojstvenog vektora [4].*

Primjer 1.2.9. *Prethodnu ideju možemo promatrati i malo općenitije. Pogledajmo primjer mreže znanja u nekoj organizaciji ili "www" mreže. Informacija o tome koji su ljudi dobar izvor informacija i koji nas ljudi mogu uputiti na dobre izvore informacija vrlo su bitne. Slično, kada koristimo internet, vrlo su nam bitni autoriteti, odnosno stranice koje sadrže ono što nas zanima (poput Wikipedije, raznih portala...), ali su nam jednako tako važne, a možda čak i važnije, stranice koje nas znaju uputiti na pravo mjesto (poput tražilice ili glavne stranice fakulteta s popisom poveznica na sve predmete). Stoga odvojeno promatramo članove koji imaju puno znanja (autoritete) i članove koji nam znaju reći koga trebamo pitati (čvorišta - engl. hub). Naravno, ti se skupovi mogu preklapati, ali preklapanje nije nužno.*

Sad smo došli do slične relacije kao u slučaju centralnosti svojstvenog vektora. Dobar autoritet je onaj na koji pokazuje puno dobrih čvorišta, a dobro čvorište je ono koje pokazuje na puno dobrih autoriteta. To možemo zapisati ovako:

$$x_v = \frac{1}{\lambda_1} \sum_{t \in G} a_{v,t} y_t, \quad y_v = \frac{1}{\lambda_2} \sum_{t \in G} a_{t,v} x_t, \tag{1.9}$$

gdje x_v označava svojstvo autoriteta vrha v, a y_v svojstvo čvorišta vrha v, $A = (a_{v,t})$ matrica susjedstva, t i v vrhovi grafa $G = (V, E)$, a $\lambda_1, \lambda_2 > 0$ neke konstanta. Kao i u slučaju centralnosti svojstvenog vektora, jednadžbe možemo zapisati matrično:

$$Ax = \lambda_1 y \tag{1.10}$$

$$Ay = \lambda_2 x. \tag{1.11}$$

Kombiniranjem jednadžbi dobivamo dva svojstvena problema:

$$A^t A x = \mu x, \quad \mu = \lambda_1 \lambda_2 \tag{1.12}$$

i

$$A A^t x = \nu x, \quad \nu = \lambda_2 \lambda_1. \tag{1.13}$$

To motivira sljedeću definiciju:

Definicija 1.2.10. *Neka je $G = (V, E)$ graf i $A = (a_{v,t})$ matrica susjedstva. Tada je* ***centralnost autoriteta*** *jednaka svojstvenom vektoru pridruženom najvećoj svojstvenoj vrijednosti μ matrice $A^t A$, a* ***centralnost čvorišta*** *(engl. hub centrality) jednaka svojstvenom vektoru pridruženom najvećoj svojstvenoj vrijednosti η matrice AA^t.*

Napomena 1.2.11. *Budući da su $A^t A$ i AA^t simetrične i pozitivno semidefinitne, singularne vrijednosti matrice A su korijeni svojstvenih vrijednosti matrica AA^t i*

A^tA. Također, lijevi singularni vektori od A jednaki su svojstvenim vektorima od AA^t, a desni svojstvenim vektorima od A^tA. Time se svojstveni problem iz definicije 1.2.10 pretvara u singularni što nam, zbog točnosti i brzine SVD-a, daje efikasan način za izračunavanje gornjih svojstava. [5]

U kompleksnim mrežama jedna od vrlo bitnih stvari je sklonost formiranju jako povezanih grupa, tzv. klastera.

Definicija 1.2.12. *Neka je t_i broj veza među susjedima vrha i te neka je k_i stupanj vrha i. Tada* ***koeficijent klasteriranja*** *(engl. clustering cofficient) definiramo kao*

$$C(i) = \frac{t_i}{k_i(k_i - 1)/2} \tag{1.14}$$

za $k_i > 1$. Za $k_i \leq 1$ definiramo $C(i) \equiv 0$. Na razini cijele mreže definiramo ***prosječni koeficijent klasteriranja*** *(engl. average clustering coefficient) kao*

$$\langle C \rangle = \frac{1}{N} \sum_i C(i). \tag{1.15}$$

Definicija 1.2.13. *Definiramo funkciju distribucije $\mathrm{P}(k)$ kao vjerojatnost da je nasumično odabran vrh stupnja k. Ta se distribucija zove* ***distribucija stupnja*** *(engl. degree distribution). U slučaju usmjerenog grafa, definiramo dvije distribucije $\mathrm{P}(k_{in})$ i $\mathrm{P}(k_{out})$.*

Srednji stupanj *mreže je*

$$\langle k \rangle = \frac{1}{N} \sum_i k_i = \sum_k k\,\mathrm{P}(k) = \frac{2E}{N}. \tag{1.16}$$

Analogno definiramo srednji ulazni i izlazni stupanj (primijetimo da moraju biti jednaki) za usmjerene grafove.

*Tada je n-**ti moment distribucije***

$$\langle k^n \rangle = \sum_k k^n\,\mathrm{P}(k). \tag{1.17}$$

Na analogan način definiramo i **distribuciju centralnosti između vrhova**, te **prosječnu centralnost između vrhova**.

1.3 Općenita svojstva mreža

Sociološki pokus iz 1967. inspirirao je znanstvenike na promatranje još jednog svojstva u kompleksnim mrežama. Nasumično odabranim ljudima u SAD-u podijeljena su pisma adresirana na neke druge, također nasumično odabrane, ljude u SAD-u. Cilj je bio dostaviti pismo do primatelja, ali uz pravilo da se pismo smije prenositi samo preko poznanika. Eksperiment je pokazao da je potreban vrlo malen broj poznanika da bi pismo dostiglo cilj, u prosjeku njih 6. Ipak, treba napomenuti da je eksperiment uspio samo u približno 20% slučajeva, dok u približno 80% slučajeva pismo nikad nije stiglo na destinaciju [7]. Ovo svojstvo u kompleksnim mrežama nazvat ćemo svojstvom malog svijeta (engl. small-world property). Mreža koja ima ovo svojstvo mora imati neki mali broj kao prosječnu duljinu najkraćeg puta. Ipak, logično je da on smije rasti s veličinom mreže.

Definicija 1.3.1. *Neka je $\langle l \rangle$ prosječna duljina najkraćeg puta. Za kompleksnu mrežu kažemo da ima* ***svojstvo malog svijeta*** *ako*

$$\langle l \rangle \backsim O(\log(N)). \tag{1.18}$$

Oblik funkcija distribucija koje opisuju svojstva mreže dijeli mreže na dvije velike klase. U klasu **statistički homogenih mreža** spadaju sve mreže čije distribucije stupnja i distribucije centralnosti između vrhova imaju "lagane repove" poput normalne ili Poissonove distribucije. **Statistički heterogene mreže** karakterizirane su iskrivljenim (engl. skewed) distribucijama s "teškim repovima", najčešće distribucijama zakona potencija (engl. power law distribution) [1].

U homogenim mrežama sva se svojstva ponašaju onako kako se na prvi pogled i očekuje od mreže. Tako se većina statističkih svojstava promatranih na razini vrhova grupira oko očekivanih vrijednosti tih svojstava. Funkcija distribucije stupnja postiže maksimum u očekivanju i "brzo pada", što nam govori da će najveći broj vrhova imati stupanj jednak očekivanom stupnju, odnosno mod će biti vrlo blizak očekivanju. S druge strane, u heterogenim mrežama točka očekivanja nije istaknuta točka u funkciji distribucije.

Također, pri otkrivanju homogenih mreža koristi se činjenica da u njima varijanca nije velika.

Definicija 1.3.2. ***Koeficijent heterogenosti*** *definira se kao*

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle}, \tag{1.19}$$

gdje $\langle k \rangle$ predstavlja prvi moment (očekivanje) distribucije stupnja, a $\langle k^2 \rangle$ drugi moment distribucije stupnja.

Mreža se smatra homogenom ako vrijedi $\kappa \backsim \langle k \rangle$, a heterogenom u slučaju $\kappa \to \infty$. U praksi ćemo mrežu smatrati heterogenom ako vrijedi $\kappa \gg \langle k \rangle$.

Poglavlje 2

Teorijski modeli mreža

Prvi korak u boljem razumijevanju strukture mreže i predviđanju ponašanja mreže je aproksimiranje i objašnjavanje mreže nekim teorijskim modelom. U nastavku ćemo zato navesti nekoliko najpoznatijih i najkorištenijih statičkih modela mreže i objasniti osnovne modele difuzije informacija na takvim mrežama kako bismo napravili temelj za daljnje razumijevanje teorije.

Napomena 2.0.3. *U ovom poglavlju radi jednostavnosti promatrat ćemo neusmjerene grafove bez težina.*

2.1 Modeli statičkih mreža

Među najjednostavnijim modelima mreža koji uključuju stohastičnost definitivno je **Erdős–Renyijev** model. Model predlaže dvije vrlo slične metode stvaranja mreže:

1. Neka je n broj vrhova i $|E|$ broj bridova grafa koji želimo modelirati. Graf $G_{n,|E|}$ konstruiramo tako da za svaki od $|E|$ bridova na nasumičan način odaberemo krajeve iz skupa od n vrhova.

2. Neka je n broj vrhova i $p \in [0, 1]$ vjerojatnost povezivanja dvaju vrhova. Tada graf $G_{n,p}$ konstruiramo tako da svaki od $\frac{n(n-1)}{2}$ bridova generiramo s vjerojatnošću p, odnosno ne generiramo s vjerojatnošću $1 - p$.

Prva metoda opisuje izvornu formulaciju Erdős–Renyijevog grafa iz 1959. godine, a druga jednostavnu preinaku modela koju je iste godine predložio Gilbert. Zbog jednostavnosti, u ovom radu koristit ćemo Gilbertovu metodu.

Broj bridova u grafu generiranom na ovaj način u prosjeku iznosi

$$\langle |E| \rangle = \frac{n(n-1)}{2} p. \tag{2.1}$$

S obzirom na to da svaki brid pridonosi stupnju dvaju vrhova, tada prosječni stupanj brida u grafu iznosi

$$\langle k \rangle = \frac{2\langle |E| \rangle}{n} = (n-1)p \simeq np. \tag{2.2}$$

Dok mreže u stvarnom životu, unatoč rastu, najčešće imaju konačan i ograničen srednji stupanj, vidi se da taj stupanj u ovom modelu divergira kako n raste. Iz tog se razloga najčešće koristi normalizirana vjerojatnost

$$p(N) = \frac{\langle k \rangle}{N}. \tag{2.3}$$

Promotrimo distribuciju stupnja grafa $G_{n,p}$. Vjerojatnost da vrh v ima stupanj k jednaka je vjerojatnosti da je povezan s k različitih vrhova i nije povezan s $(n-k-1)$ preostalih vrhova. Primijetimo još i da je generiranje jednog brida neovisno o generiranju nekog drugog brida. Sada je jasno da je distribucija stupnja opisana binomnom distribucijom:

$$\mathrm{P}(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \tag{2.4}$$

Poznato je da binomna distribucija konvergira prema Poissonovoj kad $n \to \infty$ i kad je np fiksan. Stoga, za dovoljno veliki n i dovoljno mali p, Poissonovu distribuciju s parametrom $\lambda = np$ možemo koristiti kao aproksimaciju za binomnu distribuciju s parametrima n i p. Primijetimo da je u modelu $pn = \langle k \rangle$ konstantno, pa za dovoljno velike n koristimo aproksimaciju Poissonovom distribucijom:

$$\mathrm{P}(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \tag{2.5}$$

Promotrimo sad neka osnovna svojstva ovog modela:

- Vjerojatnost da su proizvoljna dva susjeda nekog vrha v povezana jednaka je p. Stoga je koeficijent klasteriranja jednak

  $$\langle C \rangle = \frac{np}{n} = p = \frac{\langle k \rangle}{n}. \tag{2.6}$$

  Koeficijent se smanjuje širenjem grafa, pa zaključujemo da model nema ugrađen mehanizam grupiranja.

- Može se pokazati da prosječni najkraći put ima logaritamsku ovisnost o veličini grafa

$$\langle l \rangle \simeq \frac{\log n}{\log \langle k \rangle}. \tag{2.7}$$

  Iz toga slijedi da u modelu nalazimo svojstvo malog svijeta.

- Distribucija stupnja može se opisati Poissonovom distribucijom koja spada među distribucije "lakog repa", pa zaključujemo da model opisuje homogenu mrežu.

Sljedeći model koji ćemo promatrati čine **generalizirani slučajni grafovi**. Model se također ponaša na slučajan način, ali s razlikom da možemo nametnuti proizvoljnu (moguću) distribuciju stupnja. Konstrukcija grafa teče tako da stupnjeve vrhova odredimo iz distribucije, pa u skladu s tim dodajemo bridove na slučajan način. Svojstva modela:

- Pokazuje se da koeficijent klasteriranja može poprimiti puno veće vrijednosti nego u Erdős–Renyijevom modelu, ali kako $n \to \infty$, koeficijent opet iščezava.

- Može se pokazati da prosječni najkraći put i dalje ima logaritamsku ovisnost o veličini grafa, pa je svojstvo malog svijeta i dalje prisutno. Dapače, pokazuje se da sama prisutnost nasumičnosti u izgradnji mreže inducira pojavu svojstva malog svijeta [2].

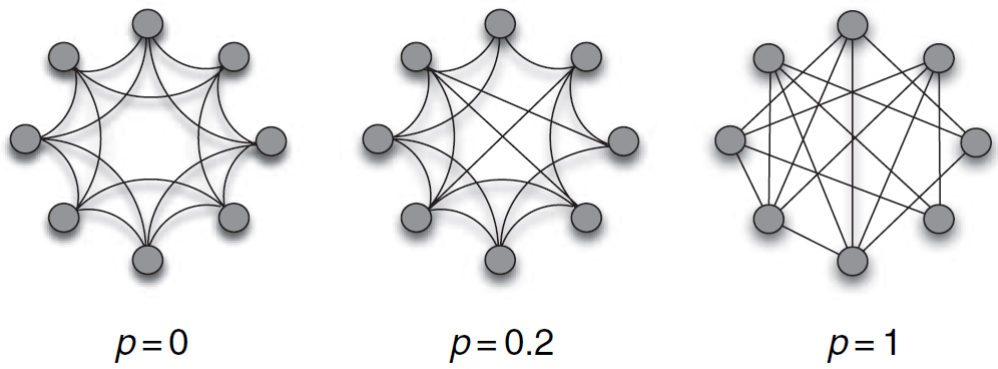
- Heterogenost mreže ovisi o distribuciji stupnja.

Mreže nastale od podataka u stvarnom svijetu često imaju vrlo jaki efekt grupiranja. Stoga se nameće potreba za traženjem modela u kojem će koeficijent klasteriranja to i prikazati. Jedan takav model je **Watts–Strogatzov** model. Neka je n broj vrhova, $p \in [0, 1]$ i $\langle k \rangle = 2m$ željeni prosječni stupanj grafa. Graf konstruiramo tako da svaki od n vrhova simetrično povežemo s $2m$ najbližih susjeda. Tada sa svakim od m desnih susjeda ostajemo povezani s vjerojatnošću $1 - p$, odnosno nasumično biramo neki drugi vrh s vjerojatnošću p. Na taj smo način napravili prečace u grafu, dok smo istovremeno zadržali grupiranost s početka. Primijetimo da i kad $p \to 1$, graf nije statistički ekvivalentan slučajnom zbog činjenice da svaki vrh zadržava minimalno m bridova. Slika 2.1 prikazuje grafove iz Watts–Strogatzovog modela za $n = 8$ i $m = 2$.

Svojstva modela:

- Za $p = 0$ koeficijent klasteriranja vrlo je velik jer su susjedi svakog vrha međusobno jako dobro povezani. Za $p \to 1$ koeficijent klasteriranja iščezava zbog nasumičnosti.

- Jasno je da je za $p = 0$ prosječni najkraći put jako dug, $\langle l \rangle \sim n$, ali već za male vrijednosti p pojava prečaca drastično smanjuje put.
- Distribucija stupnja može se analitički izračunati i pokazuje se da je "lakog repa" te da mreža ima svojstva homogene mreže [3].



Slika 2.1: Watts–Strogatzov model za različite vrijednosti parametra p

Napomena 2.1.1. *Zanimljivo je da postoji velik interval za izbor parametra p za koji će ova mreža imati svojstvo malog svijeta i velik koeficijent klasteriranja [10].*

Svi dosad navedeni modeli bili su relativno strogo određeni. **Eksponencijalni slučajni grafovi** predstavljaju klasu modela koji se danas u svrhu modeliranja mreža najčešće koriste. Neka je n broj vrhova grafa i neka je X proizvoljna matrica susjedstva grafa s n vrhova. Promotrimo vjerojatnost da iz skupa svih matrica susjedstva grafova s n vrhova, u oznaci A_n, izaberemo baš matricu X. Tu vjerojatnost definiramo kao

$$\mathrm{P}(X) = \frac{\exp[\sum_i \theta_i z_i(X)]}{\sum_{Y \in A_n} \exp[\sum_i \theta_i z_i(Y)]}, \tag{2.8}$$

gdje su θ_i parametri modela, a $z_i(X)$ neke statističke opservacije. Prilagodljivost modela očituje se u opcijama pri određivanju statistika $z_i(X)$. Te veličine mogu predstavljati jednostavna svojstva grafa poput srednjeg stupnja $\langle l \rangle$, kao i komplicirana svojstva poput distribucije stupnja i/ili distribucije pripadnih atributa.

2.2 Modeli difuzije

Osnovni modeli difuzije u mreži nastali su pri proučavanju širenja epidemija različitih zaraza. Danas se slični modeli koriste za opisivanje difuzije virusa, znanja, inovacija i mnogih drugih pojava u mreži. Difuzija je tema koja se aktivno promatra već više od 50 godina, a u posljednjim je godinama doživjela veliki porast popularnosti izumom računala koja mogu napraviti raznovrsne simulacije. Naime, time je modeliranje difuzije dobilo primjene u tehnologiji i marketingu, gdje se koriste tzv. virusne marketinške metode.

Osnovni modeli epidemija pretpostavljaju da se populacija može podijeliti u nekoliko odvojenih klasa, ovisno o tome u kojem su stanju bolesti. Gotovo u svim modelima prisutne su klasa podložnih jedinki S (engl. susceptibles - oni koji se mogu zaraziti), zaraznih jedinki I (engl. infectious - oni koji su već zaraženi i mogu širiti zarazu), a vrlo često i klasa oporavljenih jedinki R (engl. recovered - oni koji su bili, ali više nisu zaraženi). Ovisno o problemu, često se koriste i neke dodatne klase, poput klase ljudi koji su prirodno imuni na bolest ili onih koji su zaraženi, ali još nisu zarazni.

Najjedostavniji modeli imaju pretpostavku da su sve jedinke jednake i da komuniciraju na homogen način. Promotrimo širenje epidemije u populaciji od N jedinki. Označimo s $X^{[m]}(t)$ broj jedinki u klasi $[m]$ u trenutku t. Tada je $N = \sum_m X^{[m]}(t)$.

Procese koji se događaju na mreži možemo podijeliti u dvije kategorije. Prva opisuje procese spontanog prelaska iz jednog stanja u drugo, poput prelaska iz zaraženog u oporavljeno stanje ($I \to R$). Spontani prelazak iz stanja $[m]$ u stanje $[h]$ možemo opisati na ovaj način:

$$X^{[m]} \to X^{[m]} - 1 \tag{2.9}$$

$$X^{[h]} \to X^{[h]} + 1. \tag{2.10}$$

Promjena broja jedinki u klasi $X^{[m]}$ izazvana spontanim prelaskom može se opisati kao:

$$\partial_t X^{[m]} = \sum_h \nu_h^m a_h X^{[h]}, \tag{2.11}$$

gdje je a_h stopa tranzicije iz klase $[h]$, a $\nu_h^m = 1, -1$ ili 0 je promjena broja jedinki $X[m]$ koja ovisi o tome prelazi li jedinka u klasu $[m]$, iz klase $[m]$ ili nijedno od toga.

Druga kategorija opisuje binarne interakcije između jedinki koje rezultiraju promjenom stanja barem jedne od njih. Primjer za takav proces je zaraza podložne jedinke u interakciji sa zaraženom jedinkom, a opisuje se kao:

$$S + I \to 2I. \tag{2.12}$$

Promjena broja jedinki u klasi $X^{[m]}$ izazvana takvim procesima može se opisati kao:

$$\partial_t X^{[m]} = \sum_{h,g} \nu^m_{h,g} a_{h,g} \frac{X^{[h]} X^{[g]}}{N}, \tag{2.13}$$

gdje je $a_{h,g}$ stopa tranzicije, a $\nu^m_{h,g} = 1, -1$ ili 0 promjena broja jedinki $X^{[m]}$. Jednadžba 2.13 odnosi se na homogenu aproksimaciju u kojoj je za svaku jedinku iz klase $[h]$ vjerojatnost da bude u interakciji s jedinkom iz klase $[g]$ jednaka $\frac{X^{[g]}}{N}$.

Kombiniranjem jednadžbi 2.11 i 2.13 možemo prikazati promjenu broja jedinki u klasi [m] za cijeli generalizirani model:

$$\partial_t X^{[m]} = \sum_{h,g} \nu^m_{h,g} a_{h,g} \frac{X^{[h]} X^{[g]}}{N} + \sum_h \nu^m_h a_h X^{[h]}, \tag{2.14}$$

gdje, uz pretpostavku da je ukupan broj jedinki konstantan, jednadžbe moraju zadovoljavati:

$$\sum_m \partial_t X^{[m]} = 0. \tag{2.15}$$

Opisani generalizirani model najčešće se koristi u nekoliko specifičnih oblika. Najjednostavniji je svakako model **SI** (engl. susceptible - infected). U tom modelu jedinke mogu biti u samo dva stanja - podložne zarazi ili zaražene. Jednom kad se jedinka zarazi, ona se ne može oporaviti već ostaje zaražena i zarazna. U tom modelu, vjerojatnost da zaražena jedinka prenese zarazu svojem podložnom susjedu u infinitezimalno malom vremenskom intervalu dt jednaka je βdt, gdje je β stopa širenja patogena. Stoga je za podložnu jedinku sa n zaraženih susjeda vjerojatnost dobivanja zaraze u intervalu dt jednaka $1 - (1 - \beta dt)^n$. Pojednostavljeno gledano, ako je $i(t) = I(t)/N$ udio zaraženih jedinki, svaka podložna jedinka s k susjeda u prosjeku će imati $n = ki$ zaraženih susjeda i, uz činjenicu da je $\beta dt \ll 1$, iz binomne formule slijedi da vjerojatnost dobivanja zaraze možemo aproksimirati kao $1 - (1 - \beta dt)^{ki} \simeq \beta kidt$.

Problem možemo još malo pojednostaviti pretpostavkom da svaki vrh ima jednak broj bridova, odnosno $k \simeq \langle k \rangle$. Tada širenje zaraze u cijeloj mreži pomoću SI modela možemo opisati kao:

$$\frac{di(t)}{dt} = \beta \langle k \rangle i(t)[1 - i(t)], \tag{2.16}$$

gdje je $1 - i(t) = s(t) = S(t)/N$ udio podložnih jedinki.

Jedno puno korišteno, a jednostavno proširenje modela SI je model **SIS** (engl. susceptible-infected-susceptible). Jedinke opet mogu biti samo u dva stanja (podložne ili zaražene), ali zaražene jedinke se mogu oporaviti i ponovno postati podložne s

vjerojatnošću μdt, gdje je μ stopa oporavka. Širenje zaraze u cijeloj mreži pomoću modela SIS možemo opisati kao:

$$\frac{di(t)}{dt} = -\mu i(t) + \beta\langle k\rangle i(t)[1 - i(t)]. \tag{2.17}$$

Možda najpoznatiji model je svakako **SIR** (engl. susceptible-infected-removed) model, koji predviđa da će jedinke nakon zaraženog stanja doći u stanje u kojemu više nisu interesantne za širenje zaraze, jer nisu zarazne, a ni podložne zarazi. To se stanje često odnosi na slučaj smrti ili stečenog trajnog imuniteta (kao što je slučaj s npr. vodenim kozicama). Zaražena će jedinka sa stopom μ izaći iz klase zaraženih i u ući u klasu izbačenih jedinki R, čija je gustoća $r(t) = R(t)/N$. Širenje zaraze u ovom modelu opisujemo pomoću sustava jednadžbi:

$$\frac{ds(t)}{dt} = -\beta\langle k\rangle i(t)[1 - r(t) - i(t)] \tag{2.18}$$

$$\frac{di(t)}{dt} = -\mu i(t) + \beta\langle k\rangle i(t)[1 - r(t) - i(t)] \tag{2.19}$$

$$\frac{dr(t)}{dt} = \mu i(t), \tag{2.20}$$

gdje prve dvije jednadžbe na sličan način kao i dosad opisuju promjenu udjela podložnih, odnosno zaraženih jedinki, dok treća opisuje promjenu broja izbačenih jedinki pomoću stope oporavka μ.

Širenje informacija i tračeva osnovni su primjeri društvenog procesa zaraze. Društvenim procesom smatramo svaki proces širenja znanja, ideja ili navika, a koji najčešće ovisi o sklonosti pojedinca da prihvati zarazu. Jednostavan prijevod iz epidemiološkog u društveni vokabular bio bi sljedeći: Podložna jedinka je osoba I (engl. ignorant) koja još nije naučila ili saznala novu informaciju, zarazna jedinka S (engl. spreader) je ona koja širi informacije i tračeve, a oporavljena jedinka osoba koja je svjesna informacije, ali je više ne širi dalje (nezainteresirana). Primijetimo da je došlo do zamjene oznaka I i S.

Društveni se proces zaraze ipak razlikuje u nekoliko bitnih detalja. Širenje informacija je svjestan i namjeran proces, dok širenje zaraze u principu nije. Također, često je poželjno doći do novih informacija, pa prelazak iz podložnog u zaraženo stanje prestaje biti pasivan proces. Ove promjene, iako su bitne u razumijevanju procesa, ne moraju nužno uvesti promjene u model širenja zaraze. S druge strane, činjenica da u društvenim procesima prelazak iz zaraženog u oporavljeno stanje najčešće nije spontan, uvest će promjene u model. Naime, u društvenim procesima osoba obično prestane širiti informacije kad naiđe na ljude koji tu informaciju već znaju.

Kao i u širenju epidemije, prelazak iz neinformiranog (podložnog) stanja u stanje širenja vijesti (zarazno) događa se sa stopom λ kad neinformirana osoba dođe u kontakt s nekim tko širi informaciju. Prijelaz iz zaraznog stanja u nezainteresirano događa se sa stopom α kad zaražena osoba dođe u kontakt s nekom drugom zaraženom ili nezainteresiranom osobom. Te prijelaze možemo prikazati na sljedeći način:

$$\begin{cases} I + S \xrightarrow{\lambda} 2S \\ S + S \xrightarrow{\alpha} S + R \\ S + R \xrightarrow{\alpha} 2R \end{cases} \tag{2.21}$$

Problem ćemo malo pojednostaviti pretpostavkom da svaka osoba u svakom vremenskom koraku komunicira s $\langle k \rangle$ ljudi. Sada širenje informacija možemo opisati pomoću sljedećih jednadžbi:

$$\frac{di(t)}{dt} = -\lambda \langle k \rangle i(t)s(t) \tag{2.22}$$

$$\frac{ds(t)}{dt} = \lambda \langle k \rangle i(t)s(t) - \alpha \langle k \rangle s(t)[s(t) + r(t)] \tag{2.23}$$

$$\frac{dr(t)}{dt} = \alpha \langle k \rangle s(t)[s(t) + r(t)]. \tag{2.24}$$

Navedeni modeli mogu se na različite načine transformirati kako bi se promatrao neki specifičan problem ili kako bismo promatrali difuziju u heterogenoj mreži. Ipak, u ovom radu te verzije modela nećemo opisivati, jer su pojednostavljeni modeli dovoljni za shvaćanje osnovnih ideja modeliranja difuzije na mrežama. Zainteresiranog čitatelja upućujemo na [8].

Poglavlje 3

Model linearnog utjecaja

Kao što smo vidjeli u prethodnom poglavlju, difuzija informacija vrlo je popularno područje istraživanja u posljednjih nekoliko godina, ali modeliranje difuzije pokazuje se vrlo kompleksnim zadatkom. Prvi problem javlja se već pri prikupljanju podataka. Naime, vrlo je teško pratiti difuziju informacija na velikim količinama podataka i pratiti njihovo širenje kroz različite mreže. Također, kad bismo i uspjeli izgraditi cijelu relevantnu mrežu i odrediti širenje informacija kroz svaku zasebnu društvenu mrežu, kao i između njih, modeliranje difuzije i dalje je vrlo kompleksan problem za nas, ali i za današnja računala.

Svi dosad opisani modeli pretpostavljaju da su nam dostupni podaci koji opisuju cijelu mrežu, da se informacije mogu širiti samo bridovima mreže koju imamo i da je struktura mreže dovoljna za objašnjenje uočenog ponašanja. Ipak, vrlo je često mreža po kojoj se difuzija događa implicitna ili čak nepoznata. Najčešće znamo samo koji se čvor mreže "zarazio", ali ne i tko ga je zarazio. Zaista, mreža koja proučava širenje virusa primijetit će koji su ljudi u kojem trenutku zaraženi, ali gotovo je nemoguće znati točan put zaraze, odnosno od koga je tko zaražen. Slično, promatramo li širenje informacija na nekoj društvenoj mreži, lako je vidjeti kad je netko napisao novu informaciju, ali vrlo je teško saznati gdje je tu informaciju dobio.

U ovom ćemo poglavlju opisati model linearnog utjecaja (engl. linear influence model) koji služi modeliranju difuzije informacija u implicitnim mrežama, a poseban naglasak stavlja na utjecaj svakog čvora posebno na stopu difuzije u cijeloj mreži.

Model linearnog utjecaja kreće od pretpostavke da je broj novozaraženih čvorova ovisan o tome koji su čvorovi i kada zaraženi u prošlosti. Svaki čvor ima pridruženu funkciju utjecaja (engl. influence function), a broj novozaraženih čvorova u trenutku t modelira se kao funkcija svih utjecaja čvorova koji su zaraženi prije trenutka t. Na primjeru širenja vijesti portalima, to bi značilo da je broj portala koji će objaviti neku informaciju ovisan o tome koji su portali već objavili tu informaciju u nedavnoj

prošlosti. Kad portal u objavi neku vijest, izazvao je lavinu na način da će $I_u(1)$ portala objaviti tu vijest u sljedećoj vremenskoj jedinici, pa $I_u(2)$ u sljedećoj, itd.

3.1 Formalan opis modela

Neka je W skup vrhova koji sudjeluju u procesu difuzije. Kako se informacija širi, smatramo da je vrh "zaražen" ako je spomenuo neku informaciju. Promatrat ćemo širenje samo kroz trenutke t_u kad je neki čvor u spomenuo informaciju, bez proučavanja kako je dobio tu informaciju. Definiramo volumen, $V(t)$, kao broj čvorova koji spominju informaciju u trenutku t. Pretpostavljamo da svaki čvor ima svoju nenegativnu funkciju utjecaja $I_u(t)$. Vezu između volumena $V(t)$ i funkcija utjecaja svih čvorova u koji su spomenuli informaciju u vremenima $t_u(t_u < t)$ modeliramo na sljedeći način:

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u), \tag{3.1}$$

gdje $A(t)$ predstavlja skup svih čvorova u koji su se aktivirali prije trenutka $t(t_u < t)$.

Postoje dva pristupa modeliranju funkcije utjecaja:

- Parametarski pristup govori da funkcija $I_u(l)$ slijedi neku parametarsku formu, najčešće eksponencijalnu $I_u(l) = c_u e^{-\lambda_u l}$ ili oblik potencije $I_u(l) = c_u l^{-\alpha_u}$ s parametrima koji ovise o čvoru u.

- Neparametarski pristup dijeli vrijeme u diskretne intervale (npr. sate) i predstavlja funkciju utjecaja $I_u(l)$ kao nenegativni vektor duljine L, gdje l-ta vrijednost vektora predstavlja vrijednost $I_u(l)$, a duljina L označava pretpostavku da L vremenskih jedinica nakon što je objavio vijest, utjecaj čvora na širenje te vijesti nestaje.

Iako je jednostavniji, parametarski pristup ima veliku manu zbog toga što pretpostavlja da se utjecaj svih čvorova ponaša na isti način. Stoga ćemo u nastavku opisivati model koji koristi neparametarski pristup.

3.2 Procjenjivanje parametara modela

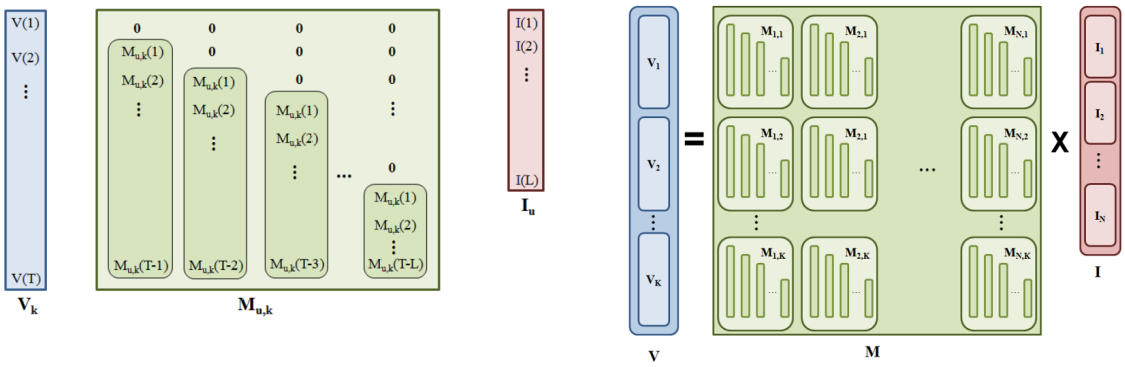
Neka podaci koje imamo opisuju skup od N čvorova i K različitih zaraza (informacija) te način na koji su se zaraze širile čvorovima kroz vrijeme. Definiramo funkciju $M_{u,k}(t)$ kao funkciju indikator, $M_{u,k}(t) = 1$ ako je čvor u zaražen sa k u trenutku t, a 0 inače.

Podsjećamo da je volumen $V_k(t)$ zaraze k u trenutku t definiran kao broj čvorova koji su zaraženi zarazom k u trenutku t. S druge strane, sada modeliramo volumen $V_k(t)$ kao sumu utjecaja čvorova u koji su zaraženi prije trenutka t:

$$V_k(t+1) = \sum_{u=1}^{N} \sum_{l=0}^{L-1} M_{u,k}(t-l) I_u(l+1). \tag{3.2}$$

Vanjska suma prolazi po svim čvorovima, a unutarnja po vremenskim trenucima u kojima funkcija utjecaja nije trivijalna. Funkcija indikator osigurava da ćemo na ovaj način zbrojiti točno trenutne vrijednosti funkcija utjecaja svih čvorova zaraženih u prošlosti.

S obzirom na to da je broj čvorova koji šire informaciju potencijalno vrlo velik, metoda najčešće ne procjenjuje utjecaje svih tih čvorova, već se bazira na utjecajima nekoliko (stotinjak) aktivnijih čvorova i njima pokušava aproksimirati cijeli volumen informacije $V(t)$. Također, model se najčešće koristi u slučajevima kada imamo veći broj mogućih zaraza nego čvorova, kao u primjeru širenja informacija kroz novinske kuće.



Slika 3.1: Matrična jednadžba za jedan vrh i jednu zarazu s blokom $M_{u,k}$ s lijeve strane te matrična jednadžba za više vrhova i zaraza s cijelom blok matricom M s desne strane. Preuzeto iz [11]

Kako bismo procijenili vrijednosti $I_u(l)$ formirat ćemo matričnu jednadžbu na sljedeći način:
Volumen zaraze $V_k(t)$ možemo shvatiti kao vektor duljine T. Sada definiramo vektor volumena V duljine $K \cdot T$ kao konkatenaciju vektora V_k za sve zaraze $k = 1, ..., K$. Zatim na sličan način definiramo vektor utjecaja I duljine $N \cdot L$ kao konkatenaciju vektora $I_u, u = 1, ..., N$. Također, definiramo binarnu matricu indikator M dimenzije

$K \cdot T \times N \cdot L$. Radi se o blok-matrici u kojoj svaki $M_{u,k}$ predstavlja jedan par čvor-zaraza. Ako je čvor u zaražen sa k u trenutku t, u matrici $M_{u,k}$ nalazit će se dijagonalna pruga jedinica, $M_{u,k}[t+l, l+1] = 1$, za $l = 0, ..., \min(L-1, T-t)$. Struktura jednog bloka $M_{u,k}$ grafički se može prikazati kao na lijevoj strani slike 3.1, a struktura cijele blok-matrice M može se vidjeti na desnoj strani slike 3.1.

S obzirom na to da je sustav predefiniran te da samo dio čvorova pokušava opisati ponašanje cijele mreže, vjerojatno ne postoji egzaktno rješenje. Stoga se traženje funkcija utjecaja, odnosno vektora I svodi na optimizacijski problem minimiziranja kvadratne greške:

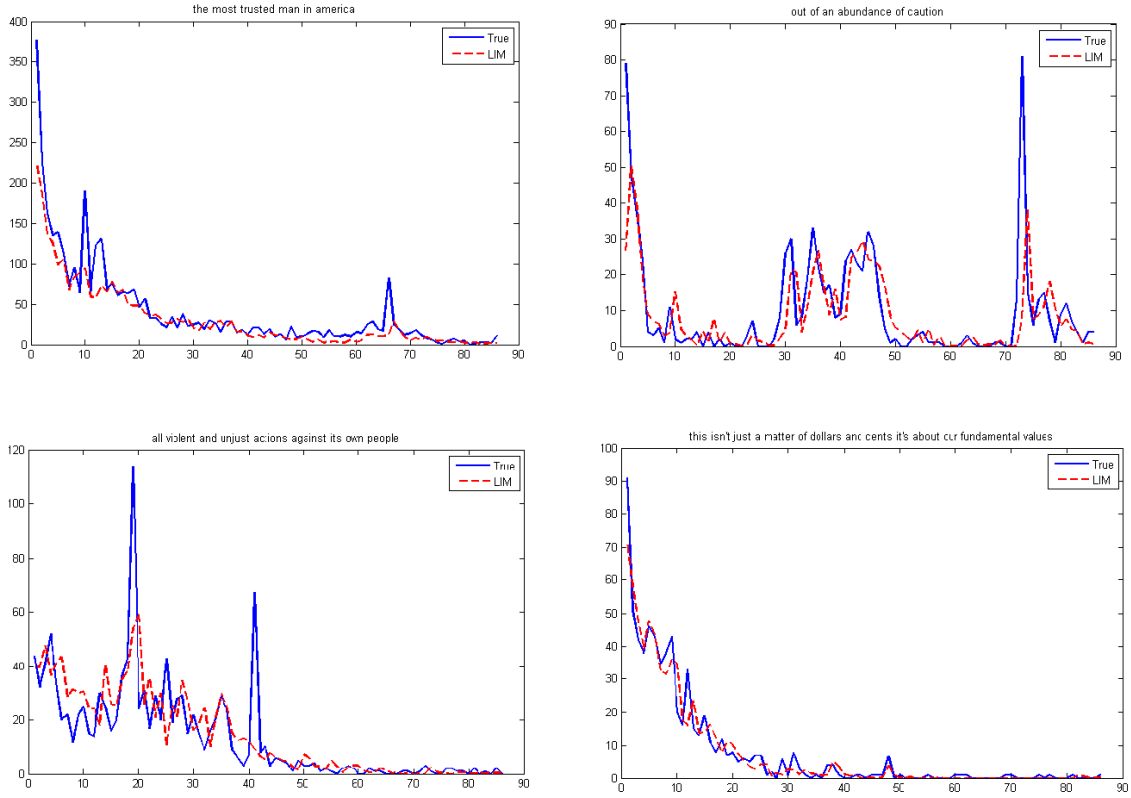
$$\|V - M \cdot I\|_2^2 \to \min, \quad I \geq 0, \tag{3.3}$$

gdje $\| \cdot \|_2^2$ predstavlja kvadrat euklidske norme.
Gornji problem poznat je i kao nenegativni problem najmanjih kvadrata (engl. non-negative least squares (NNLS) problem) i može se riješiti i za velike brojeve čvorova i zaraza [11].

3.3 Difuzija informacija u mreži vijesti

Kao primjer primjene LIM modela, proučit ćemo difuziju informacija u mreži vijesti. Vijesti u novinarskoj mreži prikupljane su između 1. rujna 2008. godine i 31. kolovoza 2009. godine na velikom broju novinskih kuća i blogova. Iz njih su izdvojene bitne rečenice i fraze, a kako bi se osiguralo da je cijeli životni tok neke fraze uhvaćen, zadržane su samo one rečenice koje se prvi put pojavljuju nakon 5. rujna 2008. godine. Iz vrlo velikog broja rečenica izabrano je 1000 fraza s najvećom frekvencijom pojavljivanja unutar 5 dana oko vrhunca njihove popularnosti. Podaci su preuzeti s [6].

Slika 3.2 prikazuje četiri grafa, koji pripadaju sljedećim frazama, respektivno: "The most trusted man in America", "Out of an abundance of caution", "All violent and injust actions against its own people" i "This isn't just a matter of dollars and cents, it's about our fundamental values". Svaki graf opisuje popularnost rečenice kroz vrijeme i prikazuje usporedbu stvarne popularnosti s onom predviđenom LIM modelom. Kao što se moglo i očekivati, vidimo da je model bolje predvidio popularnost fraza koje imaju nešto jednoličniji pad popularnosti. Ipak, predviđanje LIM modelom u svim je slučajevima dalo vrlo dobre rezultate.



Slika 3.2: Popularnost nekih rečenica i model linearnog utjecaja

Poglavlje 4

Analiza japanske automobilske industrije

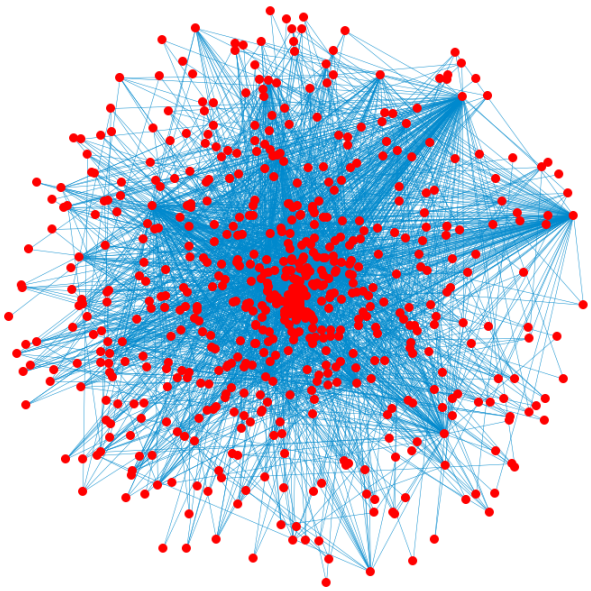
U ovom ćemo poglavlju analizirati suradnju kompanija u japanskoj automobilskoj industriji. Pokušat ćemo opisati razvoj industrije te se posebno fokusirati na najutjecajnije tvrtke. Također, promotrit ćemo efekt koji na mrežu ima izbacivanje najvažnije tvrtke.

Podaci o suradnji kompanija u japanskoj automobilskoj industriji prikupljeni su za tri vremenska stanja: 1983. godinu, 1993. godinu te 2001. godinu. Kompleksna mreža koja ih opisuje izgrađena je na sljedeći način: Vrhovi mreže označavaju kompanije, dok usmjereni bridovi označavaju vezu dobavljač-kupac. Navedena mreža bit će reprezentirana kao usmjereni graf bez težina.

Tablica 4.1: Glavne vrijednosti po godinama

	1983	**1993**	**2001**
Broj vrhova	356	679	627
Broj bridova	1480	2437	2171
Gustoća	0.012	0.005	0.006
Broj komponenti	1	1	1
Promjer grafa	5	5	6
Povezanost	1	1	1
Prosječna duljina najkraćeg puta	1.535	1.681	1.697
Prosječni koeficijent klasteriranja	0.099	0.107	0.11

Tablica 4.1 prikazuje glavna svojstva mreže dobavljača u japanskoj automobilskoj industriji kroz godine. Možemo primijetiti da mreža za svaku od navedenih godina sadrži samo jednu komponentu. To nam govori da nema potpuno izdvojenih grupa



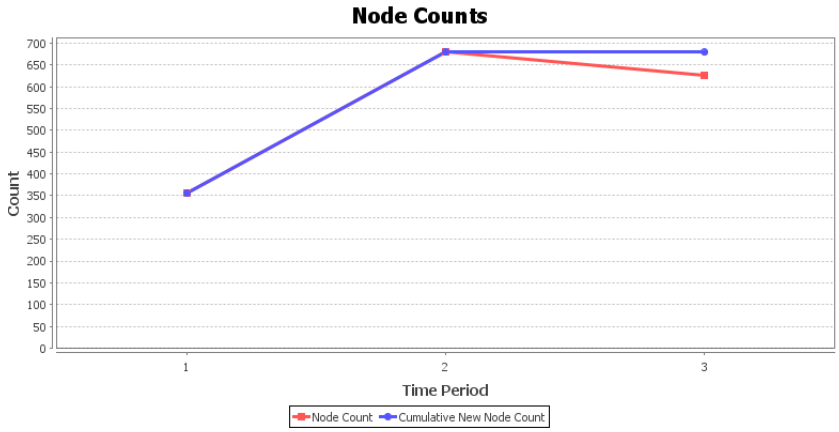
Slika 4.1: Mreža automobilske industrije iz 2001. godine

koje ne surađuju s ostalima. Iako je gustoća grafa prilično malena, promjer grafa pokazuje nam da nije teško povezati dvije različite kompanije. Zaista, kako bismo povezali bilo koje dvije kompanije, u najgorem nam je slučaju potrebno 5, odnosno 6 koraka. Također, prosječna duljina najkraćeg puta poprima vrlo malene vrijednosti, između 1,5 i 1,7. Jasno je da su te vrijednosti za svaku godinu manje od $\log_2 n$, pa bismo mogli zaključiti da mreža dobavljača u automobilskoj industriji ima svojstvo malog svijeta. Ipak, treba biti oprezan jer nam je za takav zaključak potrebno poznavanje asimptotskog ponašanja mreže, a ono nam u ovom slučaju nije dostupno. Valja primijetiti i kako je koeficijent klasteriranja svake godine prilično malen, što znači da u industriji u principu ne postoje ekskluzivne grupe suradnje s malo komunikacije prema ostalim kompanijama.

Napomena 4.0.1. *Ova se svojstva vrlo često promatraju ne samo za cijelu mrežu, već i za svaku komponentu posebno. Naravno, kako su naše mreže povezane, ne možemo govoriti o odvojenoj analizi po komponentama.*

Velika promjena broja vrhova između 1983. godine i 1993. godine može govoriti dvije stvari. Čini se jasno da je u tim godinama došlo da značajnog razvoja industrije i ulaska mnogih novih kompanija u istu. Ipak, treba imati na umu da podaci iz prve mreže govore o industriji 1983. godine, pa možda nisu potpuni, već opisuju samo

najutjecajnije kompanije. Velika promjena gustoće mreže navodi nas na to da je prvi zaključak ipak, barem djelomično, točan. Naime, dok je u slučaju nepotpunih informacija veća gustoća prilično teško objašnjiva, logično je da je industrija u kojoj sudjeluje manji broj kompanija bolje povezana od one u kojoj ih sudjeluje velik broj. To nam potvrđuju i povijesni podaci, koji govore o značajnom porastu proizvodnje između 1970. i 1990. godine[1].



Slika 4.2: Usporedba broja kompanija i kumulativnog broja kompanija po godinama

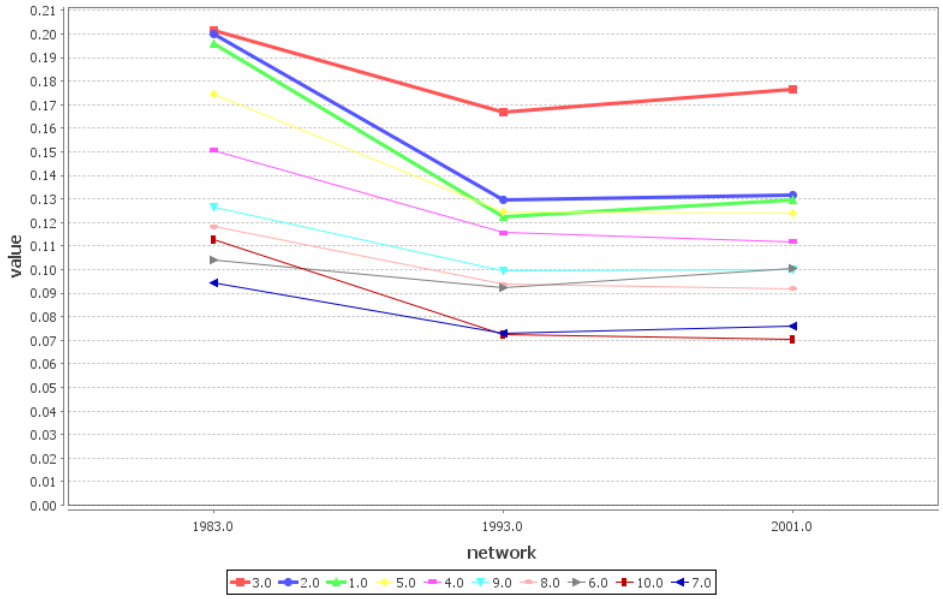
Na slici 4.2 možemo vidjeti da gotovo nijedna kompanija nije nestala iz mreže između 1983. i 1993. godine, a uključilo se mnogo novih. U tom razdoblju japanska industrija orijentirala se na izvoz i proboj na nova tržišta. U 70-im godinama prošlog stoljeća Mitsubishi i Honda postale su prve japanske automobilske tvrtke koje su izvozile na američko tržište, dok je 80-ih taj trend nastavljen masovnim probojem ostalih kompanija. Nagli porast potražnje doveo je do otvaranja velikog broja novih tvrtki te velikog porasta proizvodnje. Između 1993. i 2001. nije nastala gotovo nijedna nova kompanija, ali ih je nekoliko nestalo. Ta se pojava može objasniti i završetkom doba proboja na nova tržišta i stabilizacijom sektora.

4.1 Analiza najvažnijih kompanija

S obzirom na to da se radi o stvarnim poslovnim podacima kompanija u japanskoj automobilskoj industriji, imena svih kompanija zamijenjena su brojevima radi zadržavanja anonimnosti.

Analiza važnosti kompanija za svako svojstvo prikazana je na 10 najvažnijih čvorova za dotično svojstvo.

[1] `http://en.wikipedia.org/wiki/Automotive_industry_in_Japan`

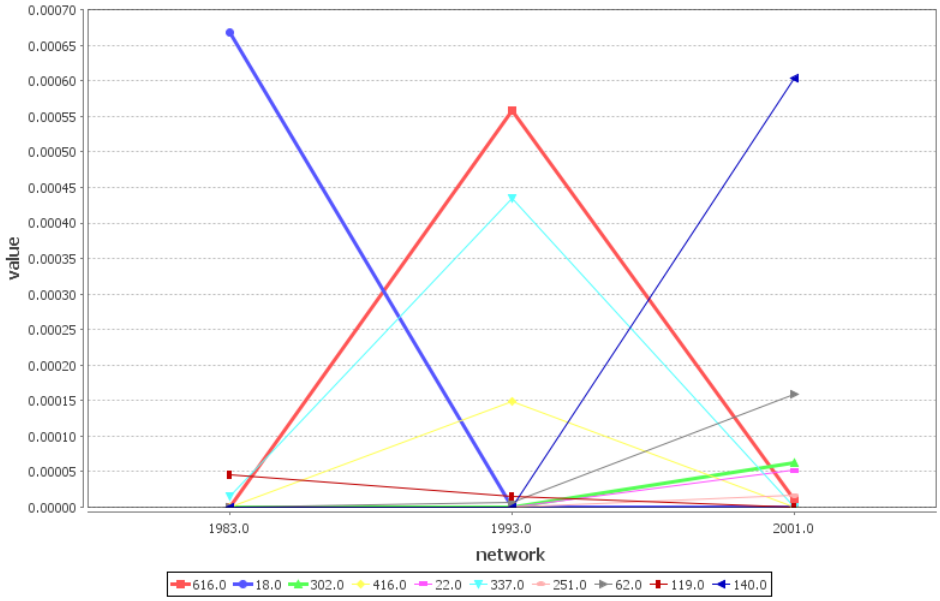


Slika 4.3: Centralnost stupnja

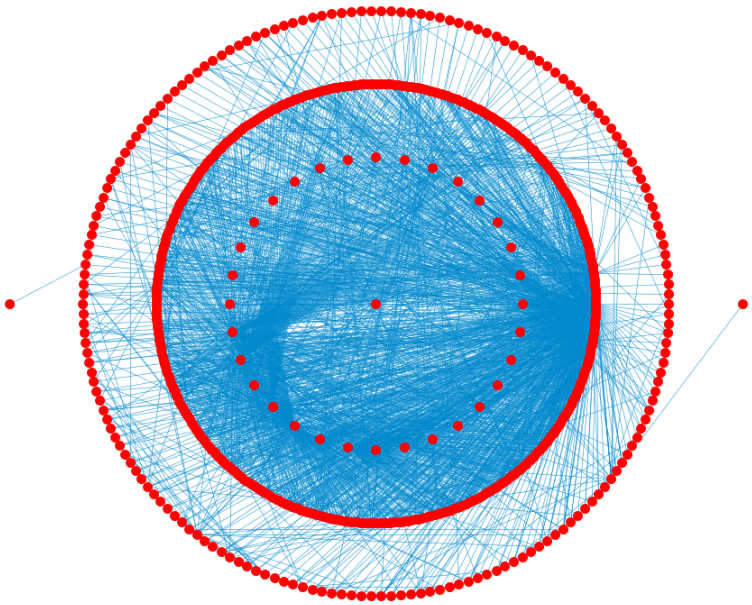
Slika 4.3 prikazuje vrijednost centralnosti stupnja po godinama za deset najvažnijih kompanija. Tih deset kompanija najviše surađuje s ostalima. Prilično velika centralnost u 1983. godini, čak 20 % za neke vrhove, može se objasniti činjenicom da je broj kompanija bio manji, pa su vjerojatno bile međusobno povezanije. Vidimo da centralnost za svih 10 čvorova pada između 1983. i 1993. godine. Možemo pretpostaviti da je razlog tome dolazak velikog broja novih kompanija, koje su uzrokovale disperziju suradnje. Iako su prava imena čvorova skrivena, valja napomenuti da je tvrtka pod šifrom 3 surađivala s najvećim brojem ostalih tvrtki, i to kroz sve tri vremenske faze. Također, iako ne možemo govoriti o podacima vezanim uz konkretne tvrtke, javno je dostupna činjenica da su Toyota, Honda, Daihatsu, Nissan, Suzuki, Mazda, Mitsubishi, Subaru, Isuzu, Kawasaki, Yamaha i Mitsuoka najmoćnije kompanije u japanskom automobilskom sektoru tog doba. Stoga možemo pretpostaviti da je većina kompanija koje opisuje slika 4.3 na tom popisu.

Centralnost između vrhova ističe čvorove koji funkcioniraju kao mostovi u komunikaciji između vrhova i smatraju se potencijalno utjecajnima, jer stoje između mnogo grupa i mogu odlučiti koje informacije prenijeti dalje, a koje ne. U našem primjeru radi se o kompanijama koje su jaka poveznica među ostalim kompanijama u automobilskoj industriji.

Na slici 4.4 možemo primijetiti da su vrijednosti centralnosti između vrhova vrlo malene, pa možemo reći da u ovoj mreži nema čvorova koji su jako utjecajni na taj



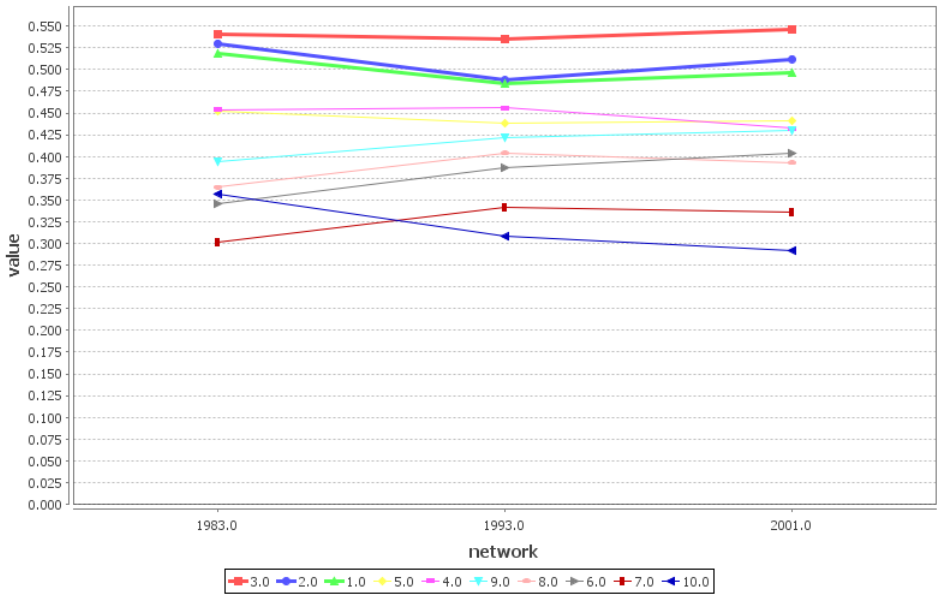
Slika 4.4: Centralnost između vrhova



Slika 4.5: Kružni prikaz mreže s naglaskom na centralnost između vrhova

način. Također, malene vrijednosti objašnjavaju naizgled kaotičan graf na slici 4.4, jer se u ovakvoj situaciji već sitnim promjenama na mreži, koje se prirodno dogode razvojem industrije kroz godine, stanje centralnosti može drastično promijeniti. Tako je 1983. godine tvrtka broj 18 bila najjači most u industrijskom mreži, a do 1993. to je svojstvo u potpunosti izgubila i prepustila ga tvrtci broj 616. 2001. godine najveću centralnost između vrhova imala je tvrtka broj 140.

Ipak, kako bismo objasnili važnost vrhova s visokom centralnošću između vrhova, nudimo prikaz koji to svojstvo naglašava. Slika 4.5 je kružni prikaz mreže iz 2001. godine. U centru se nalazi vrh s najvećim stupnjem centralnosti između vrhova (140), a na vanjskoj kružnici vrhovi s najmanjim stupnjem centralnosti. Ovdje se vidi uloga središnjih vrhova kao mostova.



Slika 4.6: Centralnost autoriteta

U neusmjerenim se grafovima centralnost svojstvenog vektora promatra kao vrlo bitna veličina. Naime, iz definicije je jasno da će ona zapravo biti jednaka i centralnosti autoriteta i centralnosti čvorišta. U usmjerenim će se grafovima te centralnosti ipak značajno razlikovati i davati puno jasniju informaciju od centralnosti svojstvenog vektora.

Slike 4.6 i 4.7 prikazuju vrijednosti centralnosti autoriteta, odnosno čvorišta za najvažnije vrhove. Odmah možemo primijetiti kako su vrijednosti vrlo različite. Kao što smo vidjeli u poglavlju 1.2, najveća čvorišta su oni vrhovi koji pokazuju na dobre autoritete. U mreži automobilske industrije čvorišta će biti kompanije koje čine

Slika 4.7: Centralnost čvorišta

popularne dobavljače među kompanijama koje su popularni kupci. Kako su najveći autoriteti oni vrhovi na koje pokazuje puno jakih čvorišta, popularni autoriteti bit će upravo one kompanije koji naručuju od popularnih dobavljača. Iako se čini pomalo cirkularno, kao i u motivaciji pripadnih definicija, ove veličine na najprirodniji način izdvajaju kompanije koje su dobri kupci (kao autoritete) i one koje su dobri dobavljači (kao čvorišta).

Popis popularnih klijenata nije se kroz godine previše mijenjao, pa se tako pokazuje da je tvrtka 3 najveći autoritet kroz sva tri razdoblja. S druge strane, izmjenjivanje dobavljača na ljestvici najpopularnijih značajno je veće, pa je 1983. najbolja bila tvrtka 108, 1993. tvrtka 92, a 2001. godine tvrtka 58 zauzela je poziciju napopularnijeg dobavljača.

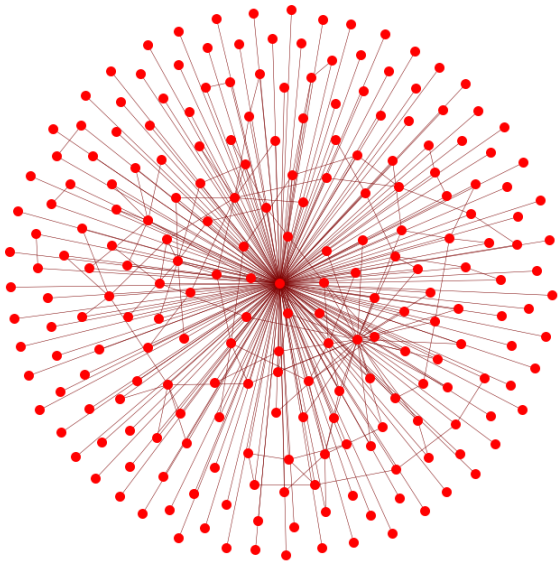
4.2 Mreže utjecaja nekih kompanija

Vidjeli smo da kompanija koja se nalazi pod šifrom 3 ima vrlo velike vrijednosti nekoliko centralnosti, što nam govori da na različite načine ima jak utjecaj na industriju.

U ovom ćemo potpoglavlju dati grafički prikaz tog utjecaja te proučiti što bi se dogodilo kad bi ta kompanija naglo nestala iz mreže.

Da bismo taj utjecaj stavili u pravilan kontekst, razmotrit ćemo i efekt micanja nasumično odabranog čvora iz mreže.

Na slici 4.8 vidimo mrežu utjecaja tvrtke pod šifrom 3. Kao što se moglo i očekivati, s obzirom na činjenicu da je to vrh s najvećom centralnošću stupnja i autoriteta, mrežu utjecaja čini prilično velik dio početne mreže.

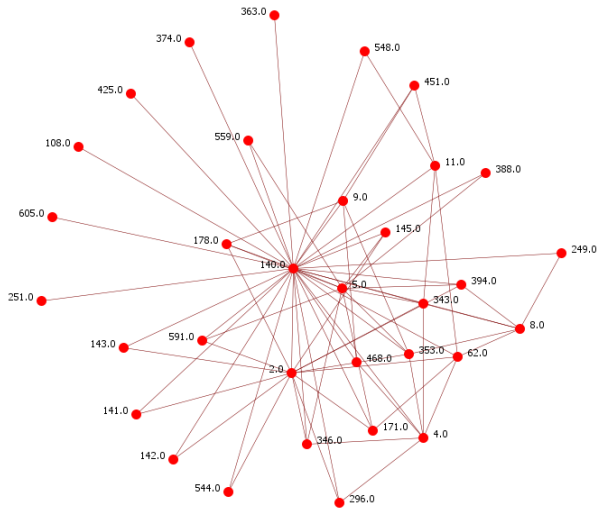


Slika 4.8: Mreža utjecaja kompanije pod šifrom 3

Kao primjer još jedne mreže utjecaja, uzeli smo onu koja pripada kompaniji 140. Taj je vrh u 2001. godini imao najveću centralnost između vrhova. Već je iz slike 4.9 jasno da je utjecaj tog vrha mnogo manji od utjecaja vrha 3. Ipak, valja napomenuti kako je i ovo utjecajan čvor. Naime, prosječan čvor u početnoj mreži ima 3.76 susjeda te bi njegova mreža utjecaja bila prilično siromašna.

Tablica 4.2 prikazuje promjene u mreži nastale micanjem jednog nasumičnog vrha iz mreže. Kako bi rezultat bio vjerodostojan, pokus vađenja nasumično odabranog vrha ponovljen je 1000 puta, a tablica prikazuje usrednjenje rezultata svih 1000 pokusa.

Vidimo da micanje jednog nasumično izabranog vrha na mreži ne stvara drastične promjene. Kao što smo mogli i očekivati, u tablici 4.2 vidimo da se broj komponenti malo povećao, a povezanost malo smanjila. Naime, uvijek postoji mogućnost da nasumično odaberemo baš onaj čvor koji povezuje neke, inače odvojene, dijelove. S obzirom na to da se mreža sastojala samo od jedne komponente, broj komponenti može se samo povećati u nekim slučajevima, i pritom smanjiti povezanost. Ipak, broj takvih slučajeva vrlo je malen.



Slika 4.9: Mreža utjecaja kompanije pod šifrom 140

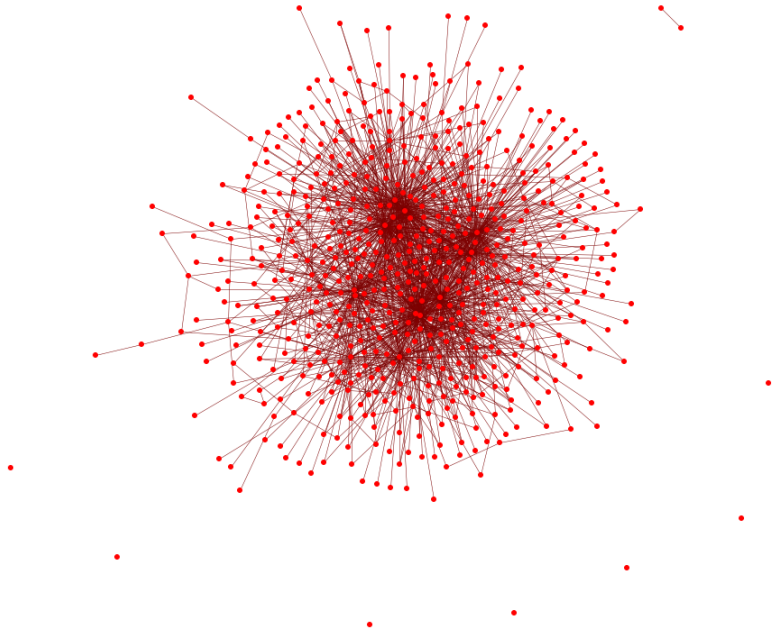
Tablica 4.2: Promjene u mreži nastale micanjem nasumično odabranog vrha

	Prije	**Poslije**	**Promjena**
Gustoća	0.006	0.006	+0.03%
Broj komponenti	1	1.14	+0.14%
Broj izoliranih vrhova	0	0.122	∞
Povezanost	1	$1 - 2 \times 10^{-4}$	$-2 \times 10^{-4}\%$
Prosječna duljina najkraćeg puta	1.697	1.696	-0.06%
Prosječni koeficijent klasteriranja	0.11	0.11	-0.09%

Slično, možemo primijetiti da je u 1000 izvedenih pokusa postojalo 122 izoliranih vrhova. Niti to nije velika promjena, uzmemo li u obzir da to znači da je u prosjeku svaki deseti put, ili rjeđe, nastalo nekoliko izoliranih vrhova.

Proječna duljina najkraćeg puta također se malo smanjila, što je logično iz najjednostavnijeg razloga – smanjimo li veličinu grafa, svi će nam preostali čvorovi biti bliži. Valja napomenuti da ovdje koristimo prosječnu duljinu najkraćeg puta u širem smislu – računamo duljine samo do onih čvorova do kojih možemo doći.

Vidjeli smo da micanje nasumično izabranog čvora nije prouzročilo drastične promjene. S druge strane, micanje bitnog čvora hoće. Slika 4.10 prikazuje mrežu automobilske industrije bez čvora 3. Već odavde vidimo da se mreža razdvojila na nekoliko



Slika 4.10: Mreža automobilske industrije bez vrha 3

Tablica 4.3: Promjene u mreži nastale micanjem vrha 3

	Prije	**Poslije**	**Postotak promjene**
Gustoća	0.006	0.005	-9.89%
Broj komponenti	1	9	+800%
Fragmentacija	0	0.029	∞
Promjer	6	6	-
Broj izoliranih vrhova	0	7	∞
Povezanost	1	0.971	-0.03%
Prosječna duljina najkraćeg puta	1.697	1.707	+0.6%
Prosječni koeficijent klasteriranja	0.11	0.094	-14.87%

komponenti te da ima nekoliko izoliranih vrhova.

Rezultati prikazani u tablici 4.3 daju nam bolji uvid u nastalu situaciju. Veliko smanjenje gustoće bilo je i očekivano, s obzirom na to da smo iz mreže izbacili vrh s najvećim brojem susjeda. Možemo vidjeti da je nastalo čak 8 novih komponenti, od čega ih 7 čini izolirane vrhove. To su vrlo velike brojke ako uzmemo u obzir činjenicu da smo iz velike povezane mreže izbacili samo jedan vrh. Novonastale komponente vrlo su malene, pa fragmentacija nije velika.

Izbačeni vrh je bitan te je logično da se prosječna duljina najkraćeg puta poveća. Ipak, zbog slabo-klasterirane strukture mreže, to povećanje nije preveliko. Kako je imao mnogo veza prema drugim vrhovima, ali i zato što je bio snažan autoritet, pa je bio povezan s jakim čvorištima, taj je vrh povećavao koeficijent klasteriranja mnogim susjedima. Zbog toga nije čudno da je prosječni koeficijent klasteriranja drastično pao kad smo izbacili taj vrh.

Bibliografija

[1] Albert László Barabási i Réka Albert, *Emergence of scaling in random networks*, science **286** (1999), br. 5439, 509–512.

[2] Alain Barrat, Marc Barthelemy i Alessandro Vespignani, *Dynamical processes on complex networks*, sv. 1, Cambridge University Press Cambridge, 2008.

[3] Alain Barrat i Martin Weigt, *On the properties of small-world network models*, The European Physical Journal B-Condensed Matter and Complex Systems **13** (2000), br. 3, 547–560.

[4] Sergey Brin i Lawrence Page, *The anatomy of a large-scale hypertextual Web search engine*, Computer networks and ISDN systems **30** (1998), br. 1, 107–117.

[5] James W Demmel, *Applied numerical linear algebra*, Siam, 1997.

[6] Stanley Milgram, *The small world problem*, Psychology today **2** (1967), br. 1, 60–67.

[7] Mark EJ Newman, *The structure and function of complex networks*, SIAM review **45** (2003), br. 2, 167–256.

[8] Leo Spizzirri, *Justification and Application of Eigenvector Centrality.*

[9] Duncan J Watts i Steven H Strogatz, *Collective dynamics of 'small-world'networks*, nature **393** (1998), br. 6684, 440–442.

[10] Jaewon Yang i Jure Leskovec, *Modeling information diffusion in implicit networks*, Data Mining (ICDM), 2010 IEEE 10th International Conference on, IEEE, 2010, str. 599–608.

[11] Jaewon Yang i Jure Leskovec, 2010, `http://snap.stanford.edu/lim/#data`, posjećena 30. kolovoza 2014.

Sažetak

Ovaj se rad sastoji od dvije velike smislene cjeline. U prvoj je cjelini opisan model linearnog utjecaja koji služi za analizu i predikciju difuzije informacija u mreži. Za taj je model dana motivacija, zatim formalan opis modela, a na kraju i primjena modela na problem difuzije informacija u mreži vijesti objavljenih u razdoblju od rujna 2008. godine do rujna 2009. godine. U svrhu boljeg razumijevanja modela, dani su i primjeri nekih drugih, osnovnih načina modeliranja difuzije u kompleksnim mrežama.

Za potrebe rada u prvom su poglavlju dani osnovni pojmovi koji služe za prikaz i analizu mreža. Objašnjena je teorijska pozadina izgradnje mreža, kao i mnoge veličine i svojstva mreža koji se koriste u analizi.

U drugoj je cjelini napravljena analiza suradnje u japanskoj automobilskoj industriji na temelju podataka iz 1983., 1993. i 2001. godine. Za svaku je od tih godina napravljena mreža suradnje te su one međusobno uspoređivane na globalnoj razini, uz promatranje promjene osnovnih svojstava mreže. Također, napravljena je usporedba na razini vrhova koja nam je predočila koje su kompanije bile najutjecajnije i kako se mijenjao njihov utjecaj kroz vrijeme. Zatim je napravljena analiza važnosti tvrtke koja se pokazala najutjecajnijom, kao i usporedba štete na mreži nastale izbacivanjem tog vrha i nekog nasumično odabranog vrha. Prikazana je mreža utjecaja te kompanije, kao i cijela mreža suradnje bez te kompanije.

Summary

This thesis consists of two parts. The first part describes the linear influence model which is used to analyze and predict the diffusion of information in the network. First, we provide motivation for this model, followed by a formal description of the model and finally we apply the model to the problem of diffusion of information in a network containing news published in the period from September 2008 to September 2009. In order to provide better understanding of the model, we give examples of other fundamental ways of modeling the diffusion in complex networks.

For the purposes of this thesis, the first section provides basic concepts used to display and analyze networks. We explain the theoretical background of network construction, as well as many network's properties that are used in its analysis.

The second part is an analysis of the cooperation in the Japanese automotive industry, based on data from the years 1983, 1993 and 2001. For each of those years a cooperation network is built and they are compared to each other on a global scale, by observing changes in the basic properties of the network. In addition, we make a comparison at the node level, presenting which companies were influential in that period and how their influence changed over time. We provide an importance analysis for the company that has proven to be the most influential, and we compare the network damage caused by ejection of that company and of a randomly selected one. We show the influence network of that company, as well as the entire cooperation network without that company.

Životopis

Hermina Petric Maretić rođena je 30. svibnja 1991. godine u Zagrebu, gdje pohađa Osnovnu školu Ivana Filipovića i Osnovnu školu Silvija Strahimira Kranjčevića. Istovremeno pohađa i Osnovnu glazbenu školu Pavla Markovca, gdje maturira na odjelu gitare. 2009. godine završava Klasičnu gimnaziju u Zagrebu i u istom gradu, na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta, upisuje preddiplomski studij inženjerske matematike. Za vrijeme studija, kao bivši natjecatelj u debati, počinje voditi debatne sastanke u Klasičnoj gimnaziji.

Po završetku preddiplomskog studija Matematike, 2012. godine, dobiva priznanje za izniman uspjeh na preddiplomskom studiju. Te godine stažira kao programer u Ericssonu Nikoli Tesli i volontira u Ljetnoj tvornici znanosti u Samoboru, gdje drži radionicu iz kriptografije. Iste godine upisuje Diplomski studij Matematike i Računarstva. 2013. godine ljeto provodi na EPFL-u (Ecole polytechnique fédérale de Lausanne) kao stažist u Laboratoriju za računalnu biologiju i bioinformatiku te upisuje paralelni studije Matematičke statistike na PMF-u u Zagrebu. Iste godine počinje suradnju s prof. dr. sc. Mariom Štorgom, koja rezultira rektorovom nagradom za rad "Dinamička analiza mreža - evolucija istraživačkog područja temeljem radova objavljenih u sklopu DESIGN konferencije 2002-2014".

U toku studija bila je demonstrator iz kolegija Linearna algebra 1 i 2, te Mjera i integral. Pred kraj diplomskog studija Matematike i Računarstva, od Matematičkog odsjeka dobiva priznanje, a od Prirodoslovno-matematičkog fakulteta nagradu za izniman uspjeh na studiju. Iste godine sudjeluje na eStudentovom natjecanju Mozgalo u timu s Antom Čabrajom i Anamarijom Fofonjka, gdje osvajaju drugu nagradu.