Usporedba jezičnih mreža pravnih i književnih tekstova

Miličić, Tanja

Sveučilište u Rijeci – Filozofski fakultet u Rijeci

Preddiplomski dvopredmetni studij informatike i povijesti umjetnosti

Tanja Miličić

Usporedba jezičnih mreža pravnih i književnih tekstova

Završni rad

Mentor: Doc. dr. sc. Ana Meštrović

Rijeka, srpanj, 2015.

Sadržaj

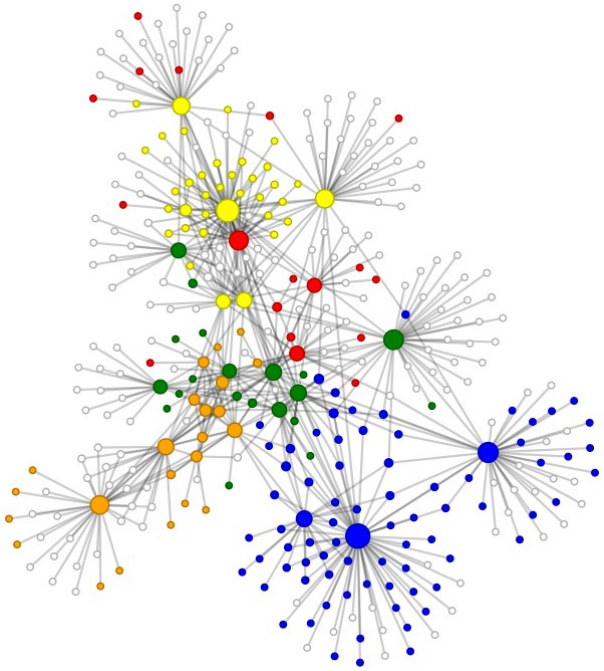
Sažetak

Predmet istraživanja u ovom radu je usporedba jezičnih mreža pravnih i književnih tekstova pisanih na engleskom jeziku. Mreže su generirane kao usmjereni grafovi s vezama koje sadrže težine. Unutar mreže riječi predstavljaju čvorove koji su povezani ukoliko su riječi susjedne u rečenici. Dok težine njihovih veza označavaju koliko se puta pojedini par riječi ponavlja. Ispitane su mjere mreža na globalnoj, lokalnoj i središnjoj razini. Na globalnoj i središnjoj razini promatrane su mjere poput prosječne snage, prosječnog stupnja, gustoće, asortativnost, broja zajednica i dr., dok su na lokalnoj razini uzeti u obzir snaga, stupanj, koeficijent grupiranja i prosječna snaga pojedinog čvora. Cilj rada bio je ispitati i prikazati kako se dvije različite kategorije teksta odražavaju na vrijednosti mjera kompleksnih mreža. Dobiveni rezultati pokazuju kako se gledajući mreže na globalnoj razini ne mogu vidjeti značajne razlike. No usporedbom mreža na lokalnoj razini, konkretno na prosječnoj snazi čvora (selektivnosti), primijećene su razlike između dvije kategorije.

Ključne riječi: jezične mreže, kompleksne mreže, usmjereni graf, mjere mreža, NetworkX

1. Uvod

Mreže ili grafovi se javljaju kao dominantna struktura u različitim domenama uključujući biološke sisteme, sociološke fenomene, prometne i tehnološke infrastrukture. U svojoj najjednostavnijoj formi mreže se sastoje od skupa vrhova (čvorova) međusobno povezanih pomoću bridova (veza). Takva pojednostavljena reprezentacija reducira sustave na apstraktne strukture bilježeći samo osnovne podatke. Ona se može proširiti dodatnim informacijama pa tako na primjer veze mogu biti usmjerene ili neusmjerene te mogu sadržavati podatke o težini veze između dva susjedna vrha. Analiza takvih mreža omogućuje nam da lakše razumijemo kompleksne sisteme, njihove komponente i interakcije u svrhu identificiranja strukture i funkcije koja proizlazi iz cijelog skupa elemenata od kojih se sastoje. Tijekom istraživanja mreža koje opisuju realne sustave došlo se do spoznaje kako one imaju određena univerzalna svojstva kao što su pojavljivanje vrhova visokog stupnja (engl. *hubs*), visoki koeficijent grupiranja, efekt malog svijeta, grupiranje u zajednice i mnoga druga [1]. Kompleksne mreže su upravo iz tog razloga, ali najviše zbog dostupnosti velike količine podataka i računala visokih performansi, tek nedavno postale fokusom istraživanja.



Slika 1.1: Kompleksna mreža.

Jedan od kompleksnih sustava s kojim se susrećemo gotovo u svakom trenutku jest jezik kojega također možemo prikazati pomoću mreža odnosno grafova. Riječi su primjer jednostavnih elemenata koji zajedno mogu tvoriti kompleksnu strukturu poput novela ili romana. Ako razmotrimo jednu cjelovitu knjigu ili bilo koji tekst kao usmjereni graf u kojem su riječi vrhovi međusobno povezani u slučaju da se radi o susjednim riječima te čiji bridovi sadrže težine ovisno o tome koliko se puta pojedini par riječi ponavlja, onda možemo u potpunosti sagledati cijelu strukturu te mreže. Da takve mreže susjednih riječi pokazuju slična svojstva kao i ostale koje opisuju realne sustave dokazali su Ramon Ferrrer i Cancho te Ricard V. Solé [2]. Analizu statističkih svojstava između običnog teksta i teksta u kojemu su riječi pomiješane prikazali su A. P. Masucci i G. J. Rodgers te su tim eksperimentom predstavili prosječnu snagu kao mjeru pomoću koje su pokazali razliku u strukturi, a nazvali su je selektivnost [3]. Nadalje, Costa *et al.* [4] usporedili su kako se pojedine mjere mreža ponašaju usporedbom originalnih tekstova s njihovim pojednostavljenim verzijama te su opisali metode za procjenu kompleksnosti tekstova baziranima na mjerama kompleksnih mreža. Istraživana je i usporedba jezičnih mreža književnih tekstova i blogova na globalnoj i lokalnoj razini [5] gdje je predloženo da se mjerom selektivnosti mogu vidjeti razlike u kategorijama teksta.

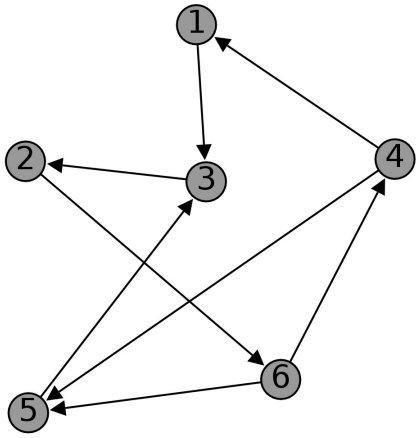
Spomenuti primjeri su samo neki od brojnih istraživanja jezičnih mreža, a rezultati postignuti u njima potiču na daljnje proučavanje mjera mreža te njihovu relaciju s obzirom na različite strukture teksta. To je ujedno i motivacija za ovaj rad u kojemu je cilj prikazati kako se dvije potpuno različite kategorije teksta (pravni i književni tekstovi) odražavaju na vrijednosti mjera kompleksnih mreža. U prvom dijelu rada ukratko su objašnjeni osnovni pojmovi te matematičko prikazivanje mreža. Drugi dio opisuje mjere mreža koje su korištene prilikom usporedbe tekstova na globalnoj, lokalnoj i središnjoj razini. Zatim je u sljedećem dijelu objašnjena implementacija određenih mjera te naposljetku metodologija i rezultati usporedbe mreža.

2. Definicija i prikaz mreža

Mreža ili graf sastoji se od skupa *N* vrhova i skupa *K* bridova koji ih povezuju. Spomenuti elementi nazivaju se različitim terminima ovisno o znanstvenom području unutar kojega se koriste, pa su tako u fizici i računalnim znanostima uobičajeni termini čvor i veza umjesto vrha i brida. S obzirom na usmjerenost veze mogu biti usmjerene ili neusmjerene, a mogu imati i težine proporcionalne intenzitetu ili kapacitetu veza između različitih čvorova. Veze se, ukoliko se radi o usmjerenima, vizualno prikazuju sa strelicama na strani čvora prema kojemu pokazuju. Postoji nekoliko različitih načina matematičkih prikaza mreža ili grafova. Jedan od njih jest matrica susjedstva A s elementima A_{ij} koju za graf $G=(V,\ E)$ sa *N* čvorova i *K* veza $(|V|=N,|E|=K)$ prikazujemo kao $|V|\times|V|$ matricu tako da vrijedi:

$$A_{ij}=\begin{cases}1 & ako(i,j)\in E,\ \ \forall i,j\in 1,\ldots,|V| \\ 0 & \text{inače}\end{cases}$$

Dakle matrica susjedstva pokazuje da li postoji veza između dva čvora ili ne. Mnoge mreže, pa tako i jezične mreže koje se analiziraju u radu, imaju veze s težinama. Takve mreže se mogu prikazati tako da se elementima matrice susjedstva daju vrijednosti težina koje odgovaraju danoj vezi. Tada se obično elementi matrice označuju kao w_{ij} [1]. Na slici 2.1. prikazan je usmjereni graf sa šest čvorova i osam veza sa svim težinama vrijednosti 1, a slika 2.2. prikazuje njegovu pripadajuću matricu susjedstva.



Slika 2.1: Usmjereni graf.

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

Slika 2.2: Matrica susjedstva.

3. Mjere mreža

Analizom raznih mjera na globalnoj, lokalnoj i središnjoj razini možemo otkriti topološke karakteristike promatrane mreže. No s obzirom da postoji široki raspon mjera često je teško odabrati one koje će biti prikladne za prikaz određenog problema te je stoga vrlo važno poznavanje najreprezentativnijih mjera kao i njihova svojstva te interpretacije. Prilikom analize jezičnih mreža u ovom radu mreže su generirane kao usmjereni grafovi s težinama pa su s obzirom na to u nastavku rada objašnjene mjere koje vrijede upravo za takvu vrstu grafa.

3.1. Lokalne mjere

Na lokalnoj razini nas zanima iznos mjere za svaki pojedini čvor te se dobiveni rezultati obično vizualno prikazuju pomoću grafikona. U nastavku su opisane neke od najčešćih mjera korištenih pri analizi mreža kao što su stupanj, centralnost stupnja, snaga i koeficijent grupiranja.

3.1.1.Stupanj i snaga čvora

Stupanj čvora jedna je od osnovnih mjera, a predstavlja broj veza s kojima je čvor povezan. Kako kod usmjerenog grafa čvorovi mogu imati veze koje vode od sebe prema ostalim čvorovima i obrnuto, veze koje od drugih čvorova vode prema njima, razlikujemo ulazne (engl. *in-degree*) i izlazne (engl. *out-degree*) stupnjeve koje možemo izračunati pomoću sljedećih formula:

$$k_i^{\text{in}} = \sum_{j=1}^{n} A_{ij}, \qquad k_j^{out} = \sum_{i=1}^{n} A_{ij}$$

Ukoliko u matrici susjedstva umjesto vrijednosti 1 i 0, koje nam govore postoji li veza između dva promatrana čvora ili ne, zapišemo težinu veze (i, j) možemo definirati sljedeću mjeru koja predstavlja snagu čvora:

$$s_i^{\text{out/in}} = \sum_j w_{\text{ij/ji}}$$

Dakle snaga čvora jest zbroj svih težina na vezama njegovih susjeda, a kao takvu su ju definirali Barrat *et al.* u [6]. Nadalje omjerom prethodno opisanih mjera dobivamo prosječnu snagu čvora:

$$e_i^{\text{out/in}} = \frac{s_i^{\text{out/in}}}{k_i^{\text{out/in}}}$$

Prosječnu snagu čvora A. P. Masucci i G. J. Rodgers [3] nazivaju mjerom selektivnosti, a pokazali su kako se njome mogu vidjeti razlike u strukturama tijekom usporedbe mreža. U posljednje vrijeme znanstvenici su pokazali sve veću zainteresiranost za otkrivanjem kompleksnijih relacija između čvorova. Jedna od karakteristika koja dopušta bogatiji prikaz mreže je težina njezinih veza. Neovisno o kojem kompleksnom sustavu se radi uvijek postoje određena obilježja koja povezuju njezine objekte. Tako na primjer, u okviru poslovne mreže, težine mogu predstavljati broj projekata na kojima akteri surađuju. Kod mreže citata težine mogu imati vrijednost ovisno o tome koliko se puta određeni članak spominje na drugim mjestima i sl.

3.1.2. Mjere centralnosti

Često pitanje koje se postavlja tijekom analize mreža jest o tome koji su čvorovi najvažniji. Ovo pitanje se veže uz koncept centralnosti, a kako možemo na razne načine definirati važnost čvora postoji i više odgovarajućih mjera centralnosti. Možda je jedna od najjednostavnijih ona vezana za stupanj čvora, odnosno za broj susjeda s kojima je čvor povezan. Kod usmjerenih grafova razlikujemo ulazni (engl. i*n-degree*) i izlazni stupanj čvora (engl. o*ut-degree*). Iako je centralnost stupnja čvora jednostavna mjera ona nam često može otkriti neke važne činjenice. Primjerice ako promatramo mrežu citata članaka možemo lako otkriti koji je članak najutjecajniji tako da tražimo onaj s najvećim ulaznim stupnjem [1]. Ukoliko se radi o mreži u kojoj veze sadrže određene težine centralnost stupnja čvora može se predstaviti kao zbroj težina veza s kojima je povezan. No tu dolazi do problema jer se time ne uzima u obzir broj veza pa tako čvor može imati centralnost stupnja 5 no mi ne možemo znati da li se radi o 5 veza s težinom 1, jednoj vezi s težinom 5 ili o nekoj drugoj kombinaciji.

Upravo su se ovime, između ostalog, bavili Opsahl *et al.* u [7] gdje su pokušali spojiti snagu i stupanj čvora koristeći parametar α koji određuje važnost broja veza u usporedbi s njihovim težinama te su definirali sljedeću mjeru:

$$C_{\text{D-out/in}}^{w\,\alpha} = k_i^{\text{out/in}} \times \left(\frac{s_i^{\text{out/in}}}{k_i^{\text{out/in}}} \right)^{\alpha}$$

u kojoj se parametar α slobodno namješta s obzirom na željeni ishod. Ukoliko se parametar postavi između 0 i 1 tada će čvorovi s više veza imati veću vrijednost stupnja centralnosti, a ako se parametar postavi iznad 1 onda će veću vrijednost imati čvorovi s manjim brojem veza.

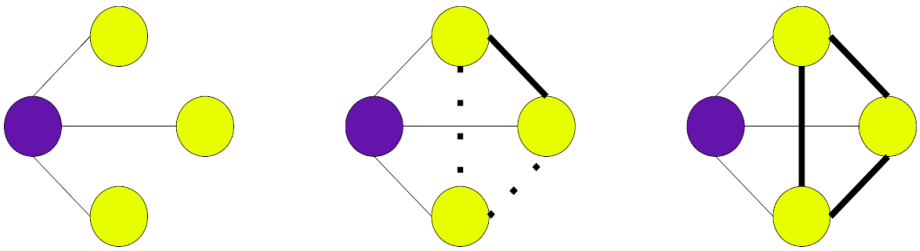
Još neki primjeri centralnosti čvora jesu oni koji se odnose na identifikaciju i dužinu najkraćih puteva unutar mreže. Tako razlikujemo centralnost blizine (engl. *closeness*) i centralnost međupoloženosti (engl. *betweenness*). Centralnost blizine čvora oslanja se na dužinu puteva do svih ostalih čvorova u mreži te pokazuje preko kojeg se čvora najbolje može prenositi informacija. Dok centralnost međupoloženosti pokazuje vjerojatnost da se čvor nalazi na najkraćem putu između bilo koja druga dva čvora. Mjere udaljenosti opisane su u poglavlju 3.2.2.

3.1.3. Koeficijent grupiranja

Koeficijent grupiranja mjera je koja se može gledati na lokalnoj i na globalnoj razini mreže. Ona pokazuje vjerojatnost da su dva susjeda određenog čvora međusobno povezana, odnosno mjeri gustoću trokuta unutar mreže. Lokalni koeficijent grupiranja možemo definirati sljedećom formulom:

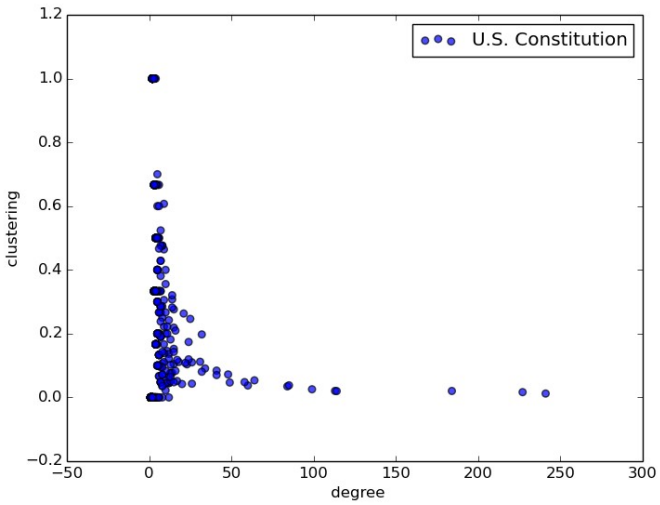
$$C_i = \frac{(\textit{broj povezanih parova susjeda čvora } i)}{(\textit{broj susjeda čvora } i)}$$

Dok se prosječni koeficijent grupiranja računa kao zbroj svih lokalnih vrijednosti podijeljeno s brojem čvorova.



Slika 3.1: Lokalni koeficijent grupiranja čvora sa vrijednostima: C=0, C=1/3, C=3/3

U mrežama koje prikazuju realne sustave koeficijent grupiranja pokazuje netrivijalno ponašanje u odnosu na stupanj čvora. Naime, čvorovi s niskim stupnjem obično su dio relativno dobro povezane zajednice tj. imaju visok koeficijent grupiranja za razliku od onih s visokim stupnjem koji obično povezuju čvorove koji nisu direktno povezani [6]. Slika 3.2 prikazuje upravo ovo svojstvo na primjeru jedne jezične mreže.



Slika 3.2: Grafikon u kojemu x os predstavlja stupanj, a y os koeficijent grupiranja čvora.

3.2. Globalne mjere

Na globalnoj razini promatramo mrežu kao cjelinu te nas zanimaju prosječne vrijednosti pojedinih mjera. Pokazalo se je kako kompleksne mreže imaju određena univerzalna svojstva na globalnoj raznini poput male udaljenosti između čvorova, grupiranje čvorova u zajednice i dr. U sljedećem podpoglavlju opisana je mjera gustoće mreže, slabe i jake komponente te asortativnost. Zatim su prikazane mjere vezane za udaljenost čvorova.

3.2.1. Povezanost mreže

Povezanost mreže, odnosno čvorova unutar mreže, možemo promatrati na više načina. Jedna od osnovnih mjera koja pokazuju u kolikoj su mjeri čvorovi dobro povezani jest gustoća mreže. Gustoća mreže je omjer postojućih veza i maksimalnog broja mogućih veza u mreži.

$$D=\frac{K}{N(N-1)}$$

Ta vrijednost kreće se od 0 do 1, što je ona veća to je veća i gustoća mreže odnosno njezini čvorovi su bolje međusobno povezani. Govoreći o povezanosti čvorova dolazimo do pojma komponenti. Naime u mreži mogu postojati čvorovi koji su potpuno odvojeni od drugih ili se nalaze unutar gušće povezane zajednice te tako tvore jednu posebnu komponentu. Za graf kod kojega postoje potpuno odvojene komponente kaže se da je nepovezan (engl. *disconnected*), dok se za graf kod kojega postoji put od svakog čvora do svih ostalih kaže da je povezan (engl. c*onnected*). Kod usmjerenih grafova možemo govoriti o slabo i jako povezanim komponentama. Ukoliko postoji najveći podskup čvorova u kojemu postoji veza u oba smjera između svakog para u podskupu kažemo da se radi o jako povezanoj komponenti. Za razliku od slabo povezanih komponenti između kojih postoji veza u samo jednom smjeru [1].

Unutar kompleksnih mreža možemo razlikovati čvorove prema određenim vrijednostima te ih svrstati u drugačije tipove ovisno o tome koju vrijednost odaberemo. Mjera koja pokazuje tendenciju da se čvorovi povezuju s drugim čvorovima sličnih karakteristika naziva se asortativnost [8].

3.2.2. Mjere udaljnosti

Kada govorimo o mjerama udaljenosti zanima nas kolika je dužina puta između čvorova. Put se može definirati za neusmjerene i usmjerene grafove, a predstavlja skevencu čvorova u kojoj je svaki uzastopni čvor povezan. Kod neusmjerenih grafova postoji put do svih čvorova u mreži, tj. neovisno o tome odakle krenuli uvijek ćemo moći doći do svih ostalih čvorova. To ne vrijedi i za usmjerene grafove s obzirom da ne mora uvijek postojati veza preko koje bismo mogli doći do nekog određenog čvora. Dužina puta u mreži je broj veza potrebnih za prolazak od čvora A do čvora B. Ukoliko se radi o grafu čije veze imaju težine tada se obično dužina puta računa kao zbroj težina svih puteva potrebnih za prolazak od odredišta do cilja.

Prosječna duljina puta (L) jedna je od čestih mjera koje se koristi, a predstavlja prosječnu vrijednost svih najkraćih puteva u mreži. Najkraća duljina puta d_{ij} je najmanji zbroj fizičke udaljenosti između dvaju čvorova od svih mogućih puteva od i do j. Najveća udaljenost između svih mogućih najkraćih puteva naziva se dijametar mreže.

$$L=\frac{1}{N(N-1)}\sum_{i\neq j} d_{ij}$$

S obzirom na to da se računajući ovu mjeru u sumu uključuju samo povezani parovi čvorova može doći do "iskrivljenja" mreže ukoliko se ona sastoji od velikog broja nepovezanih čvorova. U takvom slučaju će prosječna duljina puta imati malu vrijednost što se inače očekuje za mrežu s visokim stupnjem povezanosti [9]. Latora i Marchiori predstavili su mjeru učinkovitosti (engl. *efficiency*) pomoću koje su pokušali prikazati učinkovitost kojom se informacije prenose unutar mreže [10]. U radu su definirali mjeru E za analizu globalne i lokalne učinkovitosti. Globalna učinkovitost grafa G definirana je pomoću sljedeće formule:

$$E_{glob}(G)=\frac{1}{N(N-1)}\sum_{i\neq j\in G}\frac{1}{d_{ij}}$$

Dok je lokalna učinkovitost prosječna učinkovitost lokalnih podgrafova koji sadrže susjede čvora i.

$$E_{loc}=\frac{1}{N}\sum_{i\in G} E(G_i)$$

3.3. Mjere mreža na središnjoj razini

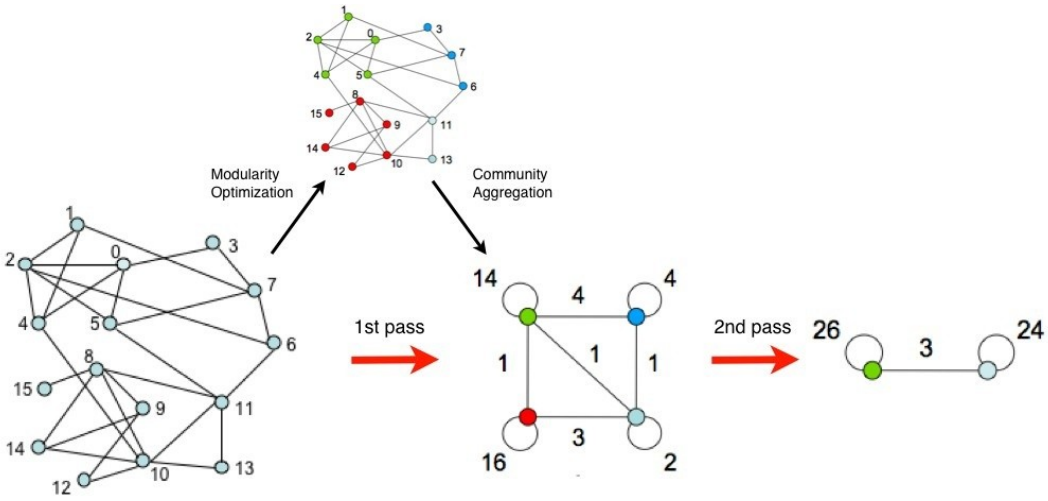
3.3.1. Zajednice u mrežama

Svojstvo koje se čini da je zajedničko većini mreža jest grupiranje u zajednice odnosno, podjela čvorova u grupe unutar kojih su veze između čvorova guste ali su izvan njih rijetke. Otkrivanje zajednica je od velikog značenja s obzirom da nam one mogu otkriti nepoznate module poput tema u informacijskim mrežama ili cyber-zajednica u socijalnim mrežama. Postoji nekoliko tipova algoritama za otkrivanje zajednica, pa tako jedni (engl. *divisive algorithms*) otkrivaju veze unutar zajednica i uklanjaju ih iz mreže, drugi (engl. *agglomerative algorithms*) rekurzivno spajaju slične čvorove/zajednice, i treći koji su bazirani na metodi optimizacije odnosno poboljšavanju objektivne funkcije. Kvaliteta dobivene particije se obično računa pomoću tzv. mjere modularnosti, skalarne vrijednosti u rasponu od -1 do 1 koja mjeri gustoću veza unutar zajednica nasuprot veza između zajednica. U slučaju mreža s težinama modularnost se definira kao:

$$Q=\frac{1}{2m}\sum_{i,j}\left[A_{ij}-\frac{k_i k_j}{2m}\right]\delta\left(c_i,c_j\right)$$

gdje je k_i zbroj težina veza povezanih sa čvorom *i*, c_i je zajednica kojoj pripada čvor *i*, funkcija $\delta(u,v)$ je 1 ako $u=v$ i 0 inače, dok je $m=\frac{1}{2}\sum_{i,j}A_{ij}$

Za potrebe ovog rada korišten je Louvainov algoritam opisan u [11], a koji koristi heurističku metodu pronalaženja zajednica baziranu na optimizaciji modularnosti. Algoritam se sastoji od dvije glavne faze koje se naizmjenično ponavljaju, a prikazane su na slici 3.3.

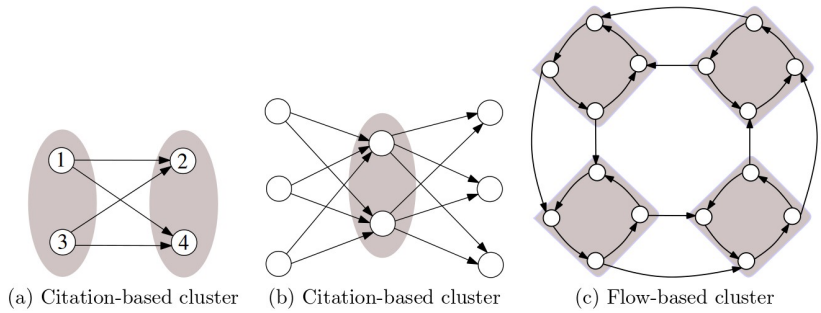


Slika 3.3: Faze Louvainovog algoritma.

Inicijalo se svakom čvoru dodjeli posebna zajednica tako da na početku imamo onoliko zajednica koliko ima čvorova. Zatim se za svaki čvor *i* uzmu u obzir njegovi susjedi *j* za koje se izračunava iznos modularnosti u slučaju da čvor *i* preselimo u zajednicu čvora *j*. Promatrani čvor se premješta u zajednicu za koju je taj iznos maksimalan te ujedno i pozitivan, ukoliko takva vrijednost ne postoji čvor će ostati u trenutnoj zajednici. Ova faza završava kada se dostigne lokalni maksimum modularnosti. Druga faza se sastoji od generiranja novog grafa u kojemu su čvorovi sada zajednice dobivene u prvoj fazi. Težine veza između novih čvorova dobiju se zbrojem težina veza između čvorova u odgovarajuće dvije zajednice dok veze između čvorova unutar istih zajednica dovode do *self-loops* za tu zajednicu. S obzirom da takvim pristupom dolazimo do mreža sa sve manjim brojem čvorova većina vremena za izračun potrebna je u prvoj fazi te se ono smanjuje sa svakom novom iteracijom.

Otkrivanje zajednica unutar mreža je vrlo zahtijevan zadatak za koji postoje brojne metode i različiti algoritmi. Prethodno opisani algoritam namijenjen je za neusmjerene grafove tako da se usmjereni graf najprije treba pretvoriti u neusmjereni kako bi se na njemu mogla primijeniti ova metoda. To ne čini veliku razliku s obzirom na to da algoritam identificira zajednice prema gustoći veza između čvorova. Kod usmjerenih grafova zajednice se osim prema gustoći veza mogu otkriti i prema određenim uzorcima koje nam otkrivaju smjerovi njihovih veza. Tako su F. D. Malliaros i M. Vazirgainnis u [12] sakupili tadašnja

saznanja i ideje o otkrivanju zajednica kod usmjerenih grafova pa su se, između ostaloga, osvrnuli i na metodu koja se bazira na uzorcima. Konkretno, to bi značilo da se čvorovi mogu grupirati u zajednice prema sličnim uzorcima povezanosti te ne moraju nužno biti susjedi.



Slika 3.4: Primjeri različitih uzoraka za detektiranje zajednica.

4. Implementacija

Mjere opisane u prethodnom poglavlju implementirane su u programskom jeziku Pythonu uz korištenje biblioteke NetworkX [13]. Biblioteka je namijenjena za stvaranje, manipulaciju i proučavanje strukture kompleksnih mreža te sadrži brojne gotove funkcije koje su korištene i za potrebe ovog rada. U nastavku su prikazane implementacije funkcija za izračun prosječne snage, koeficijenta grupiranja, globalne i lokalne učinkovitosti te manji dio kôda louvainovog algoritma za otkrivanje zajednica unutar grafa. Sve navedene funkcije, osim koeficijenta grupiranja, nisu bile prethodno implementirane u NetworkX-u.

Prosječna snaga čvora vrlo je jednostavna mjera no ipak, kako će se to kasnije pokazati, može biti od velikog značenja za ispitivanje strukture mreža. To je omjer snage i stupnja čvora, a gotove funkcije unutar NetworkX biblioteke čine ovo vrlo jednostavnim za izračun. Prikazan je kôd za ulaznu prosječnu snagu:

```
def in_selectivity_dict(network_edgelist):
  graph = nx.read_weighted_edgelist(
    network_edgelist, create_using=nx.DiGraph()
  )
  selectivity_dict = {}
  for node in graph.nodes_iter():
    s = graph.in_degree(node, weight='weight')
    k = graph.in_degree(node)
    if k > 0:
      s_in = s / k
      selectivity_dict[node] = s_in
    else:
      selectivity_dict[node] = 0
  return selectivity_dict
```

Koeficijent grupiranja je kompleksnija mjera koju se obično računa isto za usmjerene i neusmjerene grafove, tj. smjerovi unutar veza su zanemareni. Postoje podijeljena mišljenja o tome da li je takva metoda ispravna ili ne. Brojni znanstvenici istraživali su koeficijent grupiranja za usmjerene grafove gdje su određeni uzorci koji postoje unutar veza od velikoga značenja. Također se za izračun mogu koristiti i težine veza. Iako funkcija za izračun ove mjere postoji unutar NetworkX-a ovdje je, za bolje razumijevanje i daljnje modificiranje, prikazana jedna od mogućih metoda koja prvo pretvara graf u neusmjereni te nakon toga računa zatvorene trojke pojedinog čvora.

```
def clustering_dict(network_edgelist):
  graph = nx.read_weighted_edgelist(
    network_edgelist, create_using=nx.Graph()
  )
  clustering_dict = {}
  for node in graph.nodes_iter():
    ei = 0
    temp = [n for n in nx.all_neighbors(graph, node)]
    temp = set(temp)
    for i in temp:
      for j in temp:
        if graph.has_edge(i, j) and i != j:
          ei += 1
    k = graph.degree(node)
    try:
      clustering = float(ei) / (k * (k - 1))
      clustering_dict[node] = clustering
    except  ZeroDivisionError:
      clustering_dict[node] = 0
      continue
  return clustering_dict
```

Globalna i lokalna učinkovitost implementirane su tako da se može izabrati da li se žele uzeti u obzir težine na vezama ili ne. Za veze s težinama koristi se funkcija koja izračunava duljine najkraćih puteva pomoću Dijkstra algoritma te ih zapisuje kao najmanji zbroj težina potrebnih da se prođe od izvorišta do cilja.

```
def global_efficiency(g, w='unweighted'):
  graph = g
  n = len(graph)
  sum_dij = 0
  if w == 'weighted':
    for node in graph.nodes():
      shortest_paths = nx.single_source_dijkstra_path_length(graph, node)
      sum_dij += sum(1 / d_ij for d_ij in shortest_paths.values() if d_ij != 0)
  elif w == 'unweighted':
    for node in graph.nodes():
      shortest_paths = nx.single_source_shortest_path_length(graph, node)
      sum_dij += sum(1 / d_ij for d_ij in shortest_paths.values() if d_ij != 0)
  try:
    e = 1. / (n * (n - 1)) * sum_dij
  except ZeroDivisionError:
    e = 0
  return e

def local_efficiency(g):
  graph = g
  sum_EGi = 0
  for node in graph:
    neighbors = graph.neighbors(node)
    Gi = graph.subgraph(neighbors)
    E_Gi = global_efficiency(Gi)
    sum_EGi += E_Gi
  return 1. / len(graph) * sum_EGi
```

Nadalje za podjelu čvorova na zajednice implementiran je algoritam opisan u [11], a ovdje je prikazan samo jedan mali dio kôda s obzirom na njegovu veličinu. Implementacija se sastoji od dvije klase, jedna za stvaranje i manipulaciju zajednicama i druga za upravljanje glavnim funkcijama algoritma. Glavna funkcija, best_partition(), naizmjenično poziva metode za inicijalizaciju zajednice, optimizaciju modularnosti te metodu za stvaranje novog grafa iz dobivenih zajednica. Nakon svake iteracije izračunava se modularnost, odnosno kvaliteta generiranih zajednica, a funkcija se zaustavlja kada više ne dolazi do novih promjena i modularnost se ne može povećati.

```
class Louvain:
 def __init__(self, network):
   self.graph = nx.read_weighted_edgelist(network)
   self.new_graph = nx.Graph()
   self.community = Community()
   self.partition = {}
   self.modularity = 0.

 def best_partition(self):
   stop = False
   self.community.initialize(self.graph)
   self.modularity_optimization(self.graph)
   self.partition = self.community.community_dict.copy()
   self.community_aggregation(self.graph)
   self.community.initialize(self.new_graph)
   while not stop:
     mod = self.calculate_modularity()
     self.modularity_optimization(self.new_graph)
     self._renumber_partition()
     new_mod = self.calculate_modularity()
     if mod == new_mod:
       self.modularity = new_mod
       stop = True
     self.community_aggregation(self.new_graph)
     self.community.initialize(self.new_graph)
```

5. Eksperiment

U ovom poglavlju prikazani su rezultati i metodologija ispitivanja mjera kompleksnih mreža dvaju različitih kategorija tekstova. Za analizu mjera i implementaciju određenih algoritama korištena je biblioteka otvorenoga kôda NetworkX napisana u programskom jeziku Pythonu. Za crtanje grafikona korištena je biblioteka matplotlib dok je kao određena pomoć pri radu poslužio Gephi [14], softver za analizu i vizualizaciju mreža.

5.1. Podaci

Korpus korišten prilikom ispitivanja sadrži četiri kraće novele te četiri pravna teksta na engleskom jeziku. Književni tekstovi su Metamorfosis, The Bet, Gooserberries i Through the Looking-Glass, a preuzeti su sa stranice Project Gutenberg [15] . Pravni tekstovi preuzeti su sa stranice Legal Information Institute [16], a radi se o Uniform Commercial Code, Staff Regulations, U.S. Constitution, Code of Federal Regulations. Ono što je potaknulo da se izaberu spomenute kategorije tekstova jest njihova očita razlika u strukturi, a time se je htjelo pokazati kako se te razlike odražavaju na mjere kompleksnih mreža. Pravni tekstovi sadrže značajan broj ponavljanja određenih fraza i kratica za razliku od književnih tekstova koji su napisani u “opuštenijem” obliku te se obično pazi da se koristi što širi kapacitet različitih riječi. Spomenuti tekstovi imaju varijabilne dužine te drugačiji broj različitih riječi pa su s obzirom na to uspoređivani parovi tekstova dvaju kategorija čiji je broj čvorova, odnosno broj različitih riječi približne veličine.

	Književni tekstovi	Pravni tekstovi
Broj riječi	59812	92425
Broj različitih riječi	7596	7451

Table 5.1: Statistika o broju riječi u svakom od korpusa.

Prvi korak tijekom stvaranja mreža jest “čišćenje” teksta što podrazumijeva uklanjanje specijalnih znakova te interpunkcija koji ne označavaju kraj rečenice ili koji nisu dio kratica i decimalnih brojeva. Sve mreže, sveukupno njih osam, generirane su kao usmjereni grafovi s težinama. Čvorovi su riječi koje su međusobno povezane ukoliko se nalaze jedna uz drugu unutar rečenice. Težine na vezama pokazuju koliko se puta određeni par riječi ponavlja unutar teksta. Nakon stvaranja mreža generirane su liste veza (engl. *edgelists*) koje su se dalje koristile prilikom izračuna mjera, a u kojima jedan red u listi sadrži izvorišni i odredišni čvor te težinu veze između njih.

5.2. Rezultati

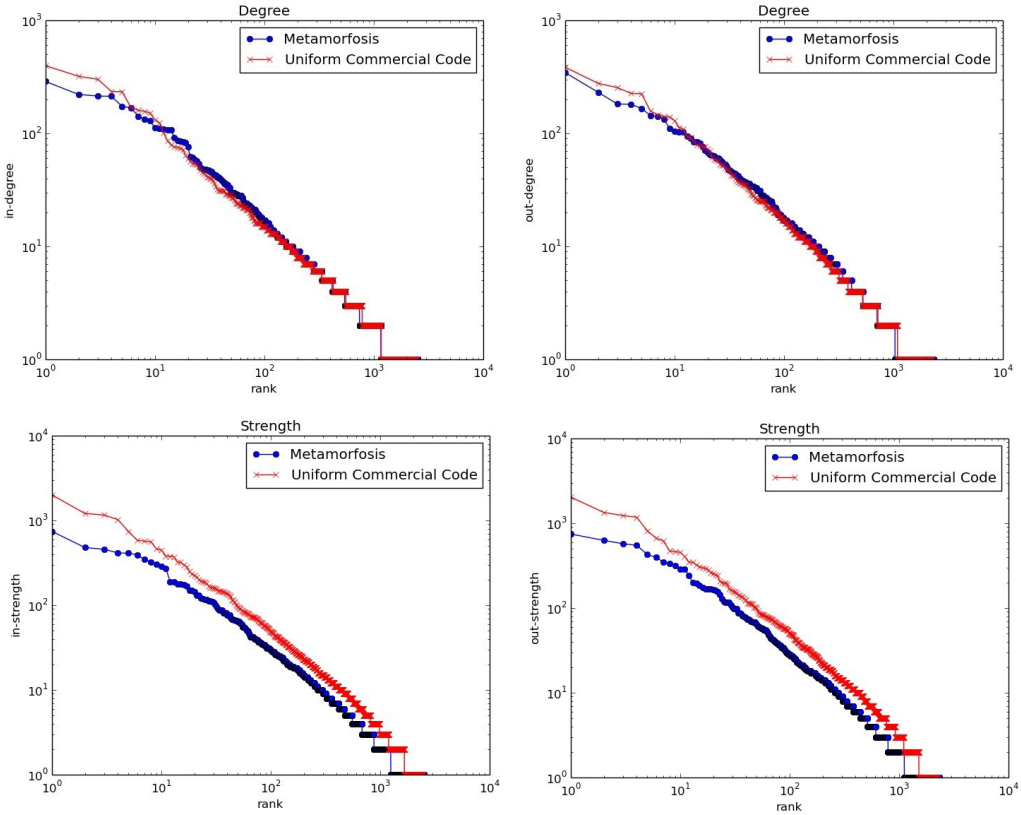
U ovome dijelu prikazani su rezultati mjera mreža opisanih u trećem poglavlju. Tablica 5.12 prikazuje mjere na globalnoj i središnjoj razini dok su pojedine lokalne mjere prikazane pomoću grafikona.

Mjere	**Tekstovi**							
	Metamorfosis	**Uniform Commercial Code**	**The Bet**	**Staff Regulations**	**Gooserberries**	**U.S. Constitution**	**Through the Looking-Glass**	**Code of Federal Regulations**
Broj riječi	22375	40915	2866	4473	4017	4375	30554	42662
Broj čvorova (N)	2599	2586	876	907	1087	843	3034	3115
Broj veza (K)	11274	11282	1927	2131	2603	2237	12812	12810
Prosječan stupanj (k)	4.34	4.54	2.20	2.35	2.39	2.65	4.22	4.11
Prosječna snaga (s)	15.19	28.32	5.64	8.61	6.16	9.02	15.96	23.43
Prosječan koeficijent grupiranja	0.326	0.392	0.169	0.211	0.194	0.285	0.275	0.299
Prosječna duljina najkraćeg puta (L)	2.96	2.87	3.10	3.22	2.83	2.99	2.80	2.73
Dijametar (D)	9	8	12	10	10	10	11	9
Prosječna duljina najkraćeg puta (L) s težinama	3.62	4.58	3.62	5.03	3.32	3.95	3.44	4.28
Dijametar s težinama	97	65	22	75	18	67	63	106
Globalna učinkovitost	0.224	0.193	0.170	0.172	0.170	0.187	0.202	0.173
Lokalna učinkovitost	0.048	0.045	0.010	0.015	0.015	0.020	0.037	0.038
Gustoća	0.0017	0.0018	0.0025	0.0026	0.0022	0.0032	0.0014	0.0013
Jako povezane komponenete	385	315	222	176	311	176	631	639
Slabo povezane komponente	1	7	5	4	7	2	3	14
Asortativnost	-0.29	-0.31	-0.25	-0.28	-0.29	-0.31	-0.28	-0.28
Modularnost	0.310	0.291	0.450	0.431	0.419	0.394	0.327	0.330
Broj zajednica	20	27	18	17	20	15	24	38

Table 5.2: Usporedba mjera jezičnih mreža književnih i pravnih tekstova

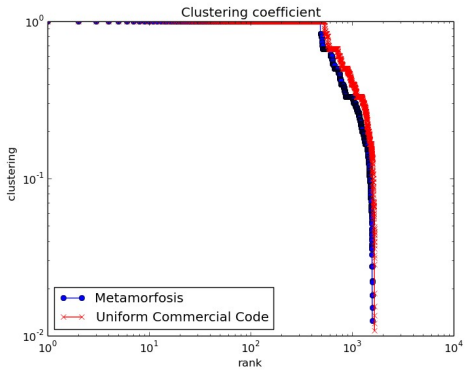
Uspoređujući dobivene rezultate može se primijetiti kako su vrijednosti gotovo svih mjera poprilično bliske u obje kategorije. Značajnije razlike se na primjer vide usporedbom broja riječi i broja čvorova gdje pravni tekstovi imaju veći broj riječi iako su iz njih generirane mreže koje su približno velike kao i one od njihovih parova književnih tekstova. To se javlja zbog toga što se kod pravnih tekstova veliki broj riječi često ponavlja. Iz tog se razloga očekuje i da oni obično imaju veću prosječnu snagu čvora što se i vidi usporedbom dobivenih rezultata. Prosječan koeficijent grupiranja kao i prosječna duljina najkraćeg puta, dijametar te gustoća mreža su također relativno približnih vrijednost. Prosječna duljina najkraćeg puta i dijametar izračunati su prvo tako što su težine na vezama zanemarene, a drugi puta su one uzete u obzir. Ako usporedimo dijametar sa i bez težina vidimo kako se je on znatno povećao za razliku od prosječne duljine najkraćeg puta koja nije narasla u velikoj mjeri kada su se uključile težine veza. Naravno, treba uzeti u obzir da se računanjem prosječnih vrijednosti gubi velika količina podataka te da je možda bolje usporediti pojedine objekte unutar mreže. Globalne i lokalne učinkovitosti pokazuju niske vrijednosti što bi značilo da uklanjanje jednog čvora može utjecati na povezanost ostatka mreže. Malo veće razlike vide se usporedbom jako i slabo povezanih komponenti te u broju zajednica no također se ne može reći kako za određenu vrstu teksta vrijedi neko univerzalno pravilo. Asortativnost je kod svih tekstova u minusu što bi značilo da unutar mreža čvorovi nemaju tendenciju vezati se za druge čvorove slične njima samima, tj. da su čvorovi s visokim stupnjem uglavnom vezani za one s niskim stupnjem.

U nastavku su prikazane lokalne mjere za svaki čvor kao što su stupanj, snaga, prosječna snaga čvora te koeficijent grupiranja. Za lakšu usporedbu su korišteni grafikoni koji prikazuju za pojedine čvorove njihove vrijednosti mjera mreža sortirane prema rastućem redu. Slika 5.1 prikazuje ulazne i izlazne vrijednosti stupnja i snage čvora za jedan književni (Metamorfosis) i pravni tekst (Uniform Commercial Code). Usporedbe ostalih parova tekstova pokazuju poprilično slične rezultate.



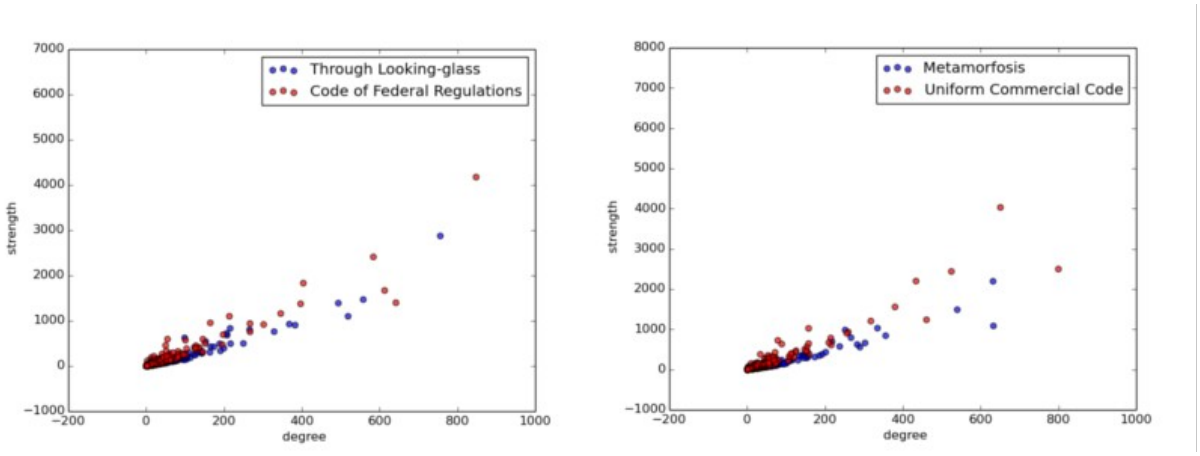
Slika 5.1: Ulazni i izlazni stupanjevi i snage pojedinih čvorova.

Promatrajući rezultate vidimo kako se čvorovi unutar mreža ne razlikuju u velikoj mjeri po stupnju no određena razlika se vidi kod snage. To se je moglo i pretpostaviti budući da su prosječne snage mreža pravnih tekstova veće, a što se može iščitati iz prethodno prikazane tablice. Nadalje slika 5.2 prikazuje vrijednosti lokalnog koeficijenta grupiranja za iste tekstove. No niti ovdje se ne vide značajnije devijacije.

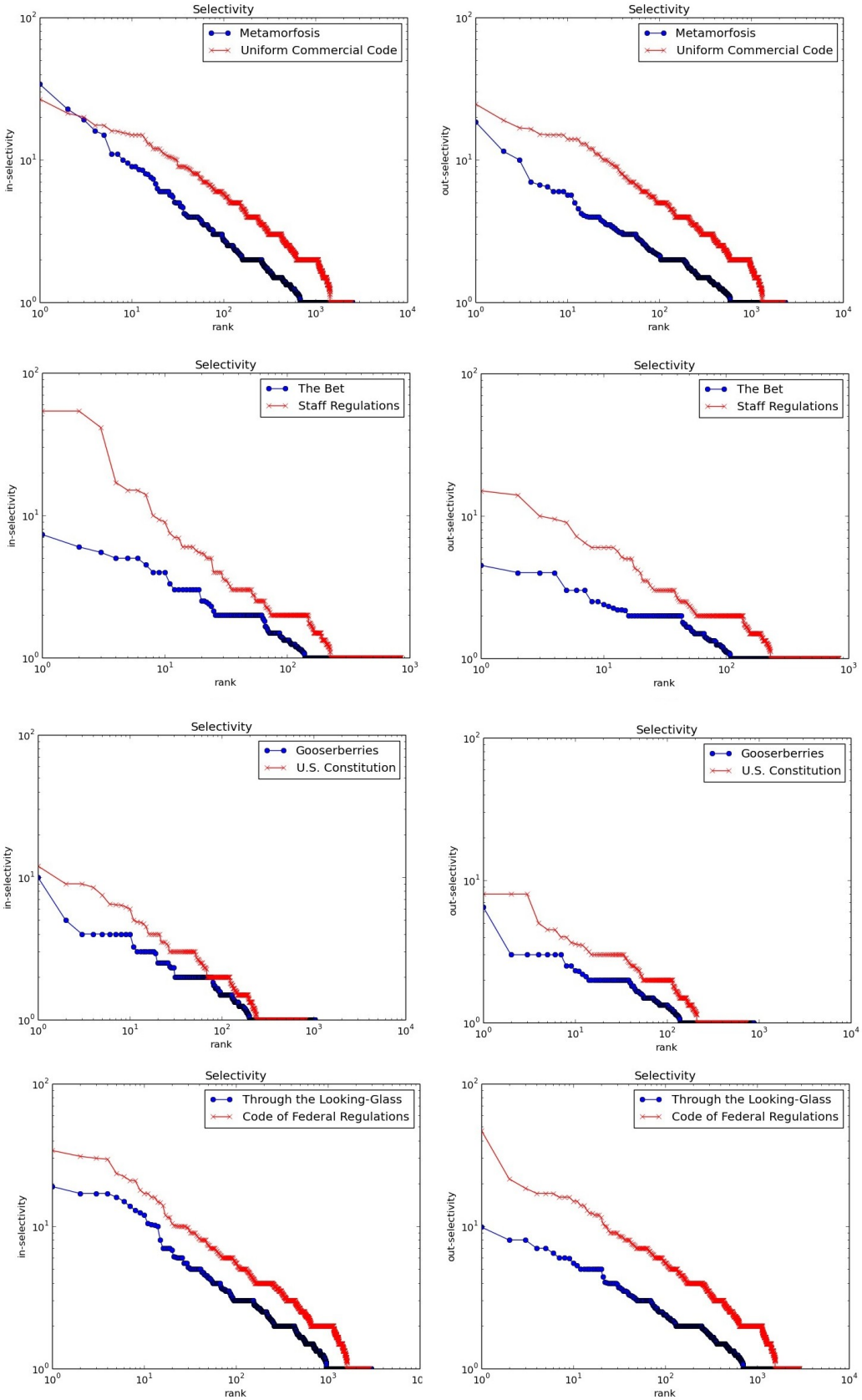


Slika 5.2: Lokalni koeficijent grupiranja pojedinih čvorova.

Do vidljivih razlika dolazimo usporedbom prosječne snage čvora ili, kako bi to rekli A. P. Masucci i G. J. Rodgers, selektivnosti. Ovakvi rezultati su se i mogli očekivati budući da je prosječna snaga omjer snage i stupnja čvora, a ako promotrimo sliku 5.1 možemo vidjeti kako je stupanj kod čvorova pravnog teksta manji dok je snaga veća. Na slici 5.3 to je još uočljivije. Ovdje su prikazani čvorovi tako da su na x osi rangirani prema stupnju, a na y osi prema snazi. Iz toga se može vidjeti kako pravni tekstovi u prosjeku za približan stupanj imaju veću snagu. Samim time je i prosječna snaga čvora kod pravnih tekstova veća u odnosu na književne tekstove. Slika 5.4 prikazuje usporedbu ulaznih i izlaznih prosječnih snaga za pojedine čvorove četiriju različitih parova teksta.



Slika 5.3: Grafikon u kojemu x os predstavlja stupanj, a y os snagu čvora.



Slika 5.4: Prosječne snage pojedinih čvorova.

6. Zaključak

U radu su prikazana osnovna svojstva kompleksnih mreža te njihovih mjera koje su ispitane na dvije različite kategorije teksta. Ono što je potaknulo na odabir baš ovih kompleksnih sustava jest velika razlika u njihovim strukturama, ali i činjenica da se slični slučajevi nisu istraživali u velikoj mjeri. Prethodno su u [5] ispitane mjere jezičnih mreža na hrvatskom jeziku generiranih iz književnih tekstova i onih preuzetih s raznih portala, a pokazuju gotovo iste rezultate. To svakako potiče na daljnje proučavanje tekstova u obliku kompleksnih mreža u svrhu njihove klasifikacije ili, u nekoj mjeri, ocjenjivanja kvalitete sadržaja. Naravno treba uzeti u obzir da je prirodni jezik vrlo kompleksan te da ga nije uvijek dovoljno proučavati na statističkoj razini, odnosno da bi se u ispitivanje trebala uključiti i računalna analiza prirodnog jezika za što kvalitetnije rezultate.

Primjer uzet u ovom radu je trivijalan s obzirom na to da se radi samo o dvije kategorije teksta no pokazuje u kojem bi se smjeru bilo najbolje kretati. Dobiveni rezultati pokazuju kako se na globalnoj razni ne vide velike razlike između kategorija osim u prosječnoj snazi i broju riječi gdje je u oba slučaja vrijednost veća kod pravnih tekstova. Gledajući prosječnu snagu na lokalnoj razini, odnosno na razini pojedinog čvora, primjećuju se razlike u strukturi. Isti slučaj javio se je i u spomenutom radu gdje su uspoređeni književni tekstovi i tekstovi iz portala. Dakle čini se kako su stupanj i snaga čvora korisne mjere za otkrivanje različitih struktura mreža. Za daljnje proučavanje vrijedilo bi uključiti još različitih kategorija tekstova te, uz znanje o kompleksnim mrežama, uključiti i prednosti računalne analize prirodnih jezika.

Popis priloga

Popis slika

Popis tablica

Popis literature

[1] M. Newman, *Networks: An Introduction*, 1 edition. Oxford ; New York: Oxford University Press, 2010.

[2] R. F. i Cancho and R. V. Solé, “The small world of human language,” *Proc. R. Soc. Lond. B Biol. Sci.*, vol. 268, no. 1482, pp. 2261–2265, Nov. 2001.

[3] A. P. Masucci and G. J. Rodgers, “Differences between normal and shuffled texts: structural properties of weighted networks,” *ArXiv08022798 Phys.*, Feb. 2008.

[4] D. R. Amancio, S. M. Aluisio, O. N. Oliveira Jr., and L. da F. Costa, “Complex networks analysis of language complexity,” *EPL Europhys. Lett.*, vol. 100, no. 5, p. 58002, Dec. 2012.

[5] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, “Comparison of the language networks from literature and blogs,” *ArXiv14052702 Phys.*, May 2014.

[6] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, “The architecture of complex weighted networks,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 11, pp. 3747–3752, 2004.

[7] F. A. Tore Opsahl, “Node centrality in weighted networks: Generalizing degree and shortest paths,” *Soc. Netw. - SOC Netw.*, vol. 32, no. 3, pp. 245–251, 2010.

[8] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski, “Edge direction and the structure of networks,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 24, pp. 10815–10820, Jun. 2010.

[9] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, “Characterization of complex networks: A survey of measurements,” *Adv. Phys.*, vol. 56, no. 1, pp. 167–242, Jan. 2007.

[10] V. Latora and M. Marchiori, “Efficient Behavior of Small-World Networks,” *Phys. Rev. Lett.*, vol. 87, no. 19, Oct. 2001.

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008.

[12] F. D. Malliaros and M. Vazirgiannis, “Clustering and Community Detection in Directed Networks: A Survey,” *Phys. Rep.*, vol. 533, no. 4, pp. 95–142, Dec. 2013.

[13] D. A. Schult and P. Swart, “Exploring network structure, dynamics, and function using NetworkX,” *Proc. 7th Python Sci. Conf. SciPy 2008*, vol. 2008, pp. 11–16, 2008.

[14] M. Bastian, S. Heymann, and M. Jacomy, others, “Gephi: an open source software for exploring and manipulating networks.,” *ICWSM*, vol. 8, pp. 361–362, 2009.

[15] “Project Gutenberg,” *Project Gutenberg*. [Online]. Available: http://www.gutenberg.org/. [Accessed: 12-Jul-2015].

[16] “LII / Legal Information Institute.” [Online]. Available: https://www.law.cornell.edu/. [Accessed: 12-Jul-2015].