Analiza Socijalnih Mreza

Izvestaj projektnog zadatka

Aleksa Pavlovic 18/3347

Elektrotehnicki fakultet

Univerzitet u Beogradu

Januar 2018

Sadrzaj

1. Uvod

U ovom dokumentu predstavljen je problem iskazan projektnim zadatkom na predmetu Analiza Socijalnih Mreza, pristup resavanja pomenutog problema, kao i diskusija dobijenih rezultata.

2. Opis problema

U okviru projekta, potrebno je analizirati podatke dobijene sa sajta IMDB, koje sadrze informacije o top 1000 filmova iz perioda od 2006. do 2016. godine. Ove podatke treba modelovati kao socijalnu mrezu glumaca, filmova i zanrova, na osnovu cijih metrika pokusavamo dobiti dodatne informacije i zakljucke o datim podacima.

Ovi podaci, datim u vidu CSV fajla, imaju sledece vrednosti:

- *Rank* - rang filma na top listi
- *Title* - naziv filma
- *Genre* - zanrovi kojima film pripada
- *Description* - opis radnje filma
- *Director* - reziser filma
- *Actors* - glavna postava glumaca u filmu
- *Year* - godina kada je film napravljen
- *Runtime (Minutes)* - trajanje filma u minutima
- *Rating* - rejting filma na skali od 1 do 10
- *Votes* - broj korisnickih glasova koje je film dobio
- *Revenue (Millions)* - zarada filma izrazena u milionima dolara
- *Metascore* - prosek ocena kriticara koje je film dobio

3. Koriscene tehnologije

Pri resavanje navedenog problema upotrebljene su sledece tehnologije:

- *Python 3.6*
  - *pandas* - biblioteka za manipulaciju i obradu podataka
  - *networkx* - biblioteka za rad sa mrezama
  - *matplotlib* - biblioteka za vizualizaciju numerickih podataka
- *Gephi 0.9.2*
  Alat za vizualizaciju i manipulaciju grafovskih struktura

4. Pretprocesiranje podataka

Pre pocetka analize, potrebno je obraditi dobijene podatke kako bi oni bili pogodniji za dalju analizu.

U sklopu ovog resenja, nakon citanja i ubacivanja podataka iz dobijenog CSV fajla u *pandas*-ov *DataFrame*, kolone *Genre* i *Actors* su transformisane iz stringa zanrova i glumaca odvojenih zarezom u liste ovih vrednosti.

Takodje su primecene rupe u kolonama *Revenue* i *Metascore*, ali to nije od znacaja za ovu analizu, jer se ove kolone nece koristiti.

5. Reprezentacija podataka grafom

Za potrebe zadatka, treba napraviti tri grafovske strukture:

- *Graf glumaca*
  Neusmereni tezinski graf kod koga su cvorovi glumci, a grane izmedju dva glumca predstavljaju indirektnu vezu koja govori o tome koliko puta su ova dva glumca zajedno igrala u nekom filmu.

- *Graf zanrova*
  Neusmereni tezinski graf kod koga su cvorovi zanrovi, a veza dva zanra govori o tome koliko puta je neki film potpao pod oba zanra.
- *Graf filmova*
  Usmereni tezinski graf kod koga su cvorovi filmovi, a grana od filma A do filma B postoji ukoliko postoje glumci koji su igrali u oba filma, a film B je izasao nakon filma A.

Za rad sa ovim strukturama, kao sto je vec pomenuto, koristi se biblioteka *networkx*, a da bi pomocu nje mogli oformiti graf, najpre se podaci iz *pandas DataFrame*-a transformisu u respektivnu matricu susednosti (po jedna za svaki tip grafa).

Pomenuta biblioteka omogucava i eksportovanje grafova u *.graphml* format, koji se onda moze otvoriti u programu *Gephi*, gde se dalje graf lakse vizualizuje i analizira.

6. Rezultati analize i odgovori na postavljena istrazivacka pitanja

U ovoj sekciji su dati odgovori na istrazivacka pitanja postavljenih u okviru projekta.

6.1 Ko su glumci koji su glumili sa najviše drugih glumaca?

	Actor	Num. of collaborations
0	Mark Wahlberg	42
1	Hugh Jackman	41
2	Christian Bale	37
3	Brad Pitt	37
4	Jake Gyllenhaal	33
5	Anne Hathaway	33
6	Tom Hardy	33
7	Michael Fassbender	33
8	Channing Tatum	33
9	Scarlett Johansson	32

6.2 Koliki je prosečan broj glumaca sa kojima je jedan glumac igrao?

Prosecan broj glumaca sa kojima je jedan glumac igrao, sto ujedno iznosi i prosecnu centralnost po stepenu u ovoj mrezi, iznosi **5.8**.

6.3 Ko su najproduktivniji glumici i u kojim žanrovima su najviše igrali?

	Actor	Num. of movies	Most frequent genre
0	Mark Wahlberg	15	Drama
1	Hugh Jackman	14	Drama
2	Brad Pitt	13	Drama
3	Christian Bale	13	Drama
4	Robert Downey Jr.	12	Action
5	Channing Tatum	12	Comedy
6	Anne Hathaway	12	Drama
7	Tom Hardy	12	Drama
8	Scarlett Johansson	12	Drama
9	Johnny Depp	12	Fantasy

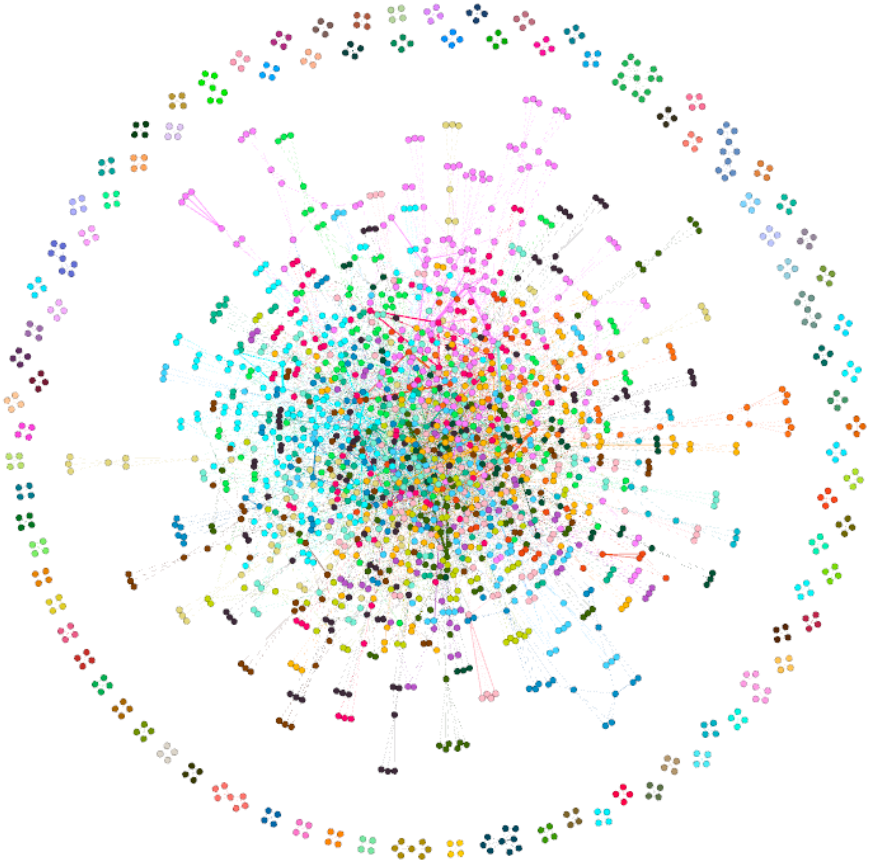
6.4 Koje zajednice glumaca se mogu uočiti prilikom analize mreže?

Na sledecoj slici prikazan je graf glumaca gde su cvorovi grupisani po modularnosti. Ovde se ne primecuju neke znacajnije komune glumaca, sem na obodu, gde se nalaze glumci koji su glumili u samo jednom filmu, pa samim tim cine kliku.

Graf glumaca obojen po klasi modularnosti

6.5 Da li se glumci u mreži grupišu na osnovu filmskog žanra u kome najčešće glume?

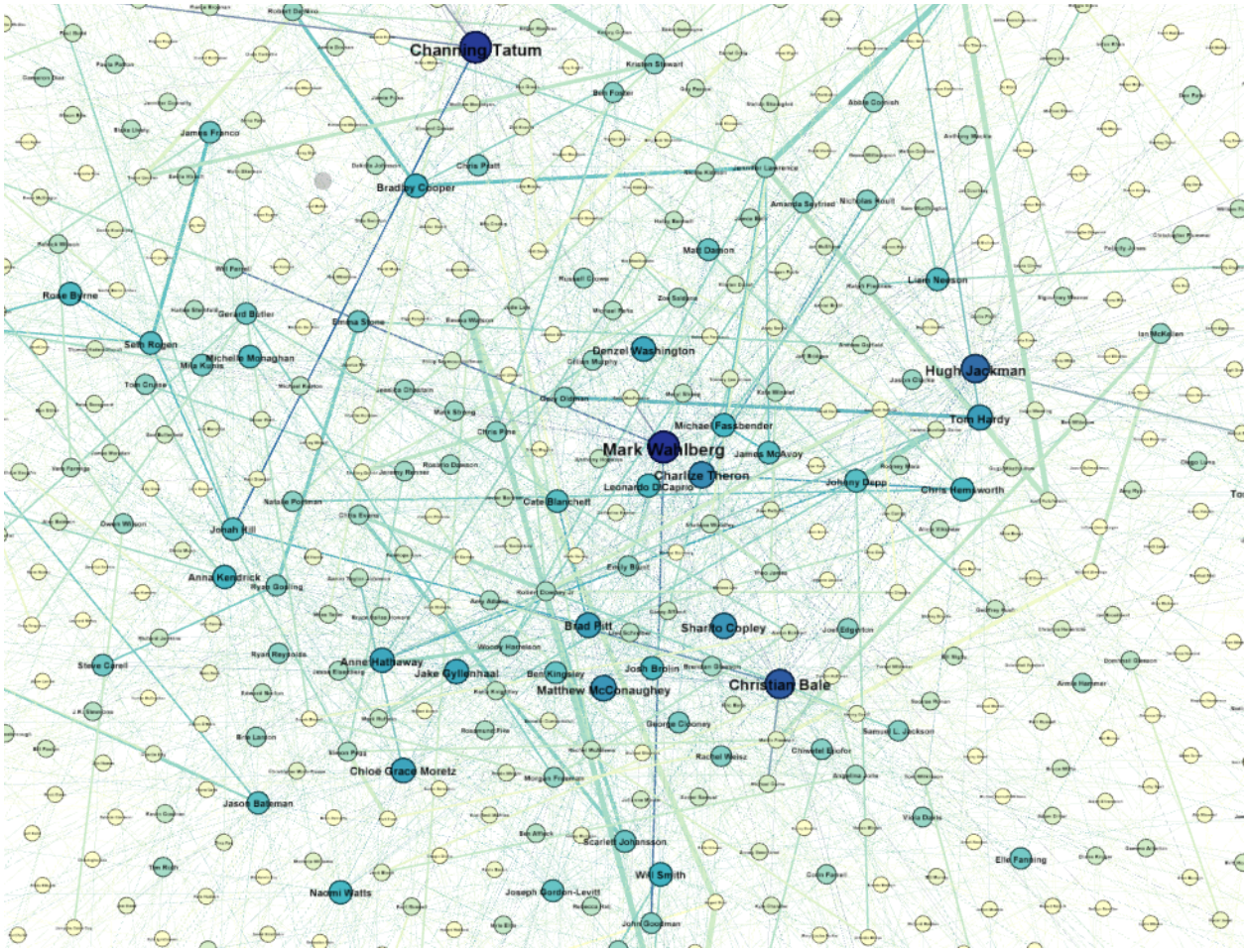
Sledeca slika predstavlja graf glumaca u kome boju cvora odredjuje najcesci zanr glumca koji taj cvor oznacava. Primecuje se da su oni blago grupisani po najcescim zanrovima, i to u najvecoj meri na obodu grafa, gde se nalaze glumci koji su, kao sto je u prethodnom odgovoru pomenuto, glumili u samo jednom filmu, pa sa svojim kolegama grade kliku.



Graf glumaca obojen po zanru u koji spada najveci broj njihovih filmova

6.6 Ko su glumci koji povezuju različite zajednice glumaca?

Glumci koji povezuju najvise drugih glumaca imaju najvecu relacionu centralnost. Stoga, iscrtavamo graf glumaca tako sto bojimo i skaliramo cvorove po ovoj metrici, kako bismo dobili bolji uvid u postojanje ovakvih tipova glumaca. Sa slike se moze videti da s neki od glumaca koji povezuju najveci broj svojih kolega Mark Wahlberg, Channing Tatum, Christian Bale, Hugh Jackman itd.



6.7 Kolika je gustina mreže?

Gustina mreze glumaca iznosi **0.003**.

6.8 U kojoj meri je mreža povezana i centralizovana?

Povezanost mreze iznosi **1.54**, sto predstavlja prosecan broj cvorova koje treba ukloniti da bi neka dva cvora postala nepovezana.

Procenti poklapanja metrika centralnosti sa grafom tipa zvezde su:

- *centralnost po stepenu* - **1.83%**
- *relaciona centralnost* - **2.54%**
- *centralnost po bliskosti* - **22.94%**
- *centralnost po svojstvenom vektoru* - **24.41%**

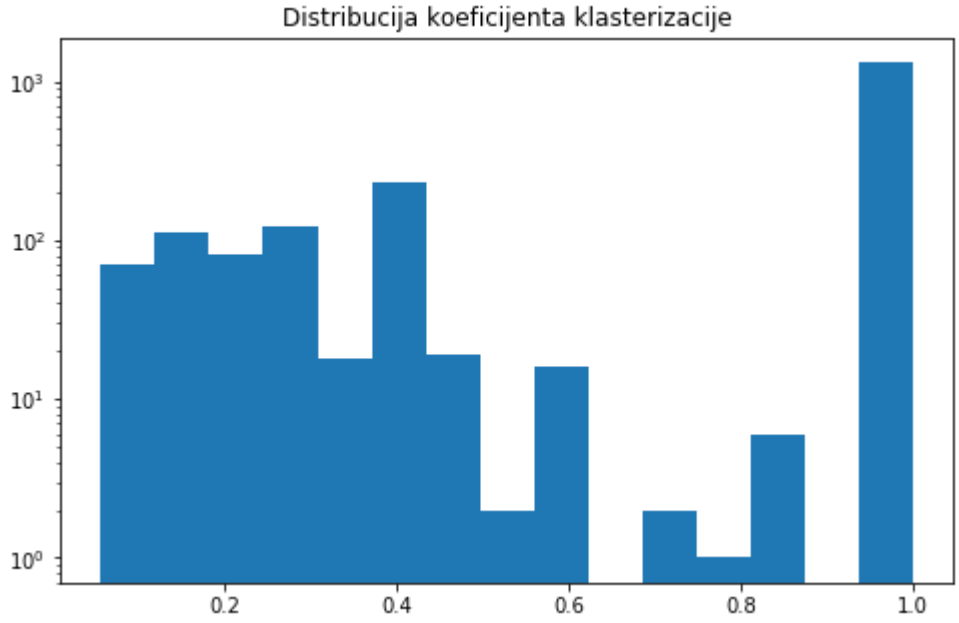
Cinjenica da su sve ove vrednosti relativno male nam ukazuje na to da u mrezi ne postoje cvorovi koji su u boljoj poziciji u odnosu na ostale.

6.9 Kolike su prosečne distance u okviru mreže i dijametar mreže?

Prosecna distanca u grafu glumaca je **4.28**, a njegov dijametar iznosi **9**.

6.10 Koliki je koeficijent klasterizacije mreže i njenih čvorova?

Prosecan koeficijent klasterizacije mreze je **0.76**, a distribucija po cvorovima je data sledecim histogramom:

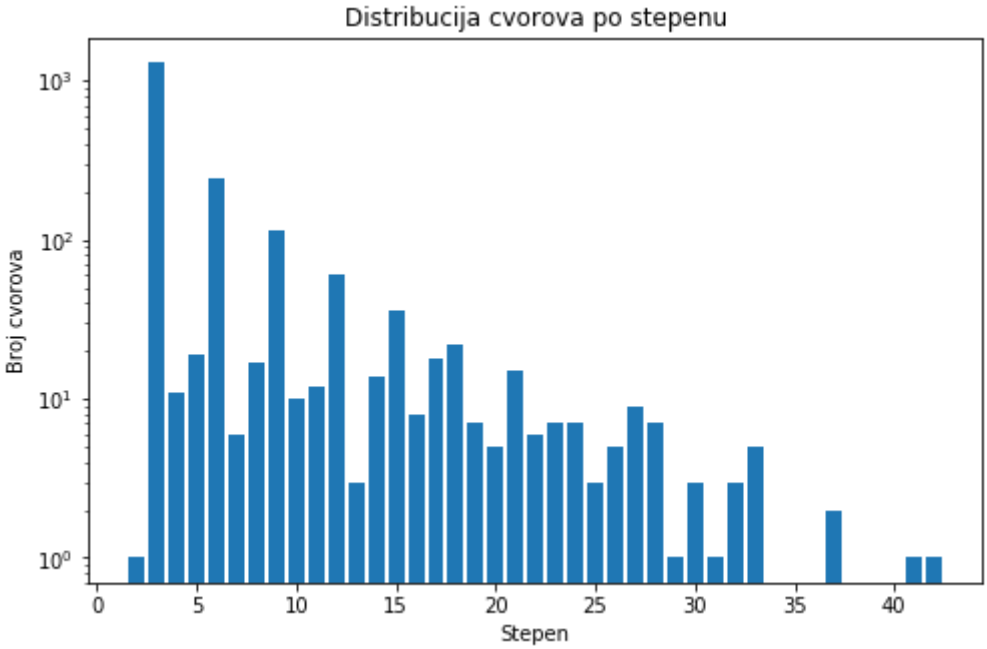


Primecujemo da veliki broj cvorova imaju koeficijent klasterizacije 1.0, sto nam ponovo ukazuje na postojanje velikog broja klika u mrezi (glumci koji su igrali samo u jednom filmu).

6.11 Kakva je distribucija čvorova po stepenu i da li prati neku zakonomernost?

Distribucija cvorova u mrezi glumaca prati *power law* distribuciju koju ispoljavaju

scale-free mreze, gde veliki broj cvorova ima mali stepen, dok je nekolicina cvorova izrazito visokog stepena.



6.12 Da li mreža iskazuje osobine malog sveta?

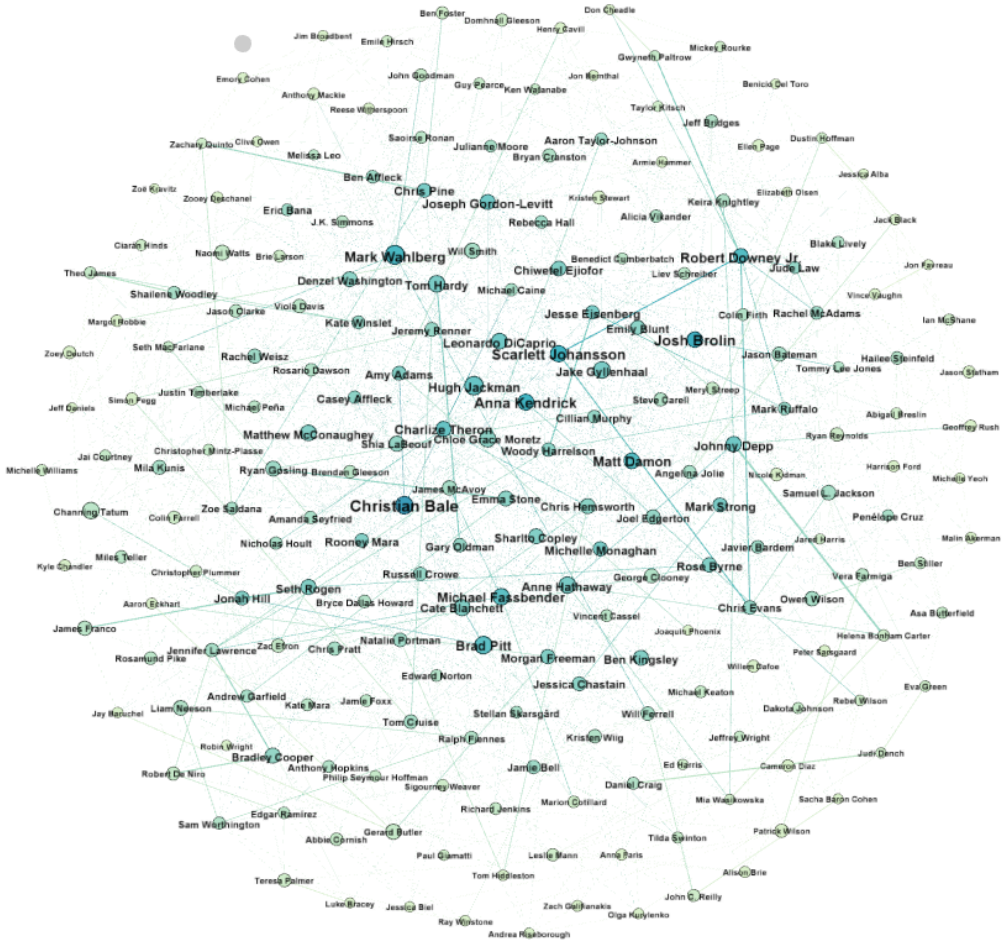
Da, zbog toga sto je stepen klasterizacije prilicno visok (0.76), dok je prosecna distanca izmedju cvorova mala (4.28).

6.13 Kolika je prosečna udaljenost, a kolika maksimalna udaljenost nekog glumca od Kevina Bejkona (Kevin Bacon)?

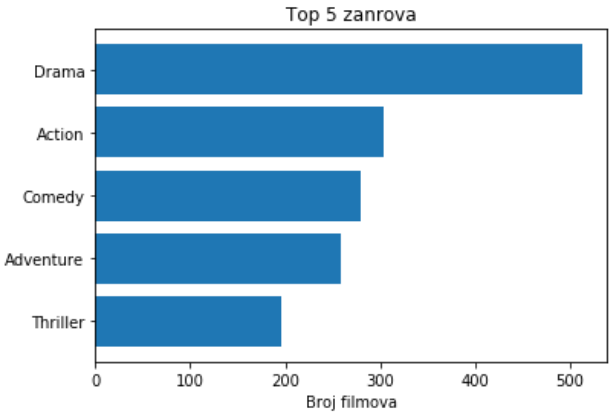
Ako posmatramo komponentu u kojoj se Kevin Bejkon nalazi, prosecna udaljenost cvorova od Kevin Bejkona je **3.77**, dok maksimalna iznosi **7**.

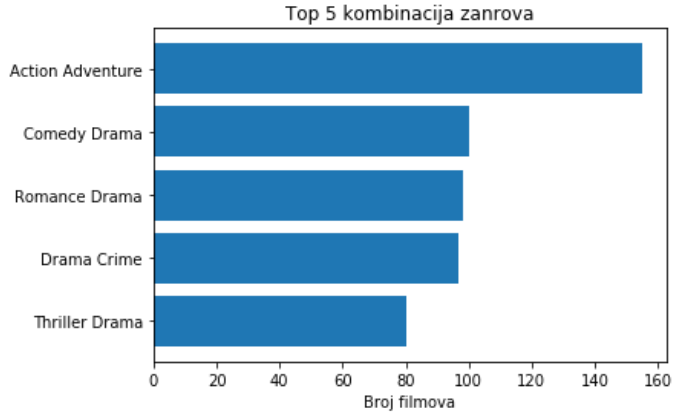
6.14 Koji glumci predstavljaju jezgro mreže?

Sledeca slika predstavlja mrezu glumaca filtrirane po k-core vrednosti za k=6. Ovo znaci da se u filtriranoj mrezi nalaze samo oni glumci koji imaju minimum 6 grana do glumaca koji takodje dele ovu osobinu. Stoga, ovi glumci cine najbolje povezani deo mreze.



6.15 Koji filmski žanrovi su najpopularniji? U kojim kombinacijama se najčešće javljaju?

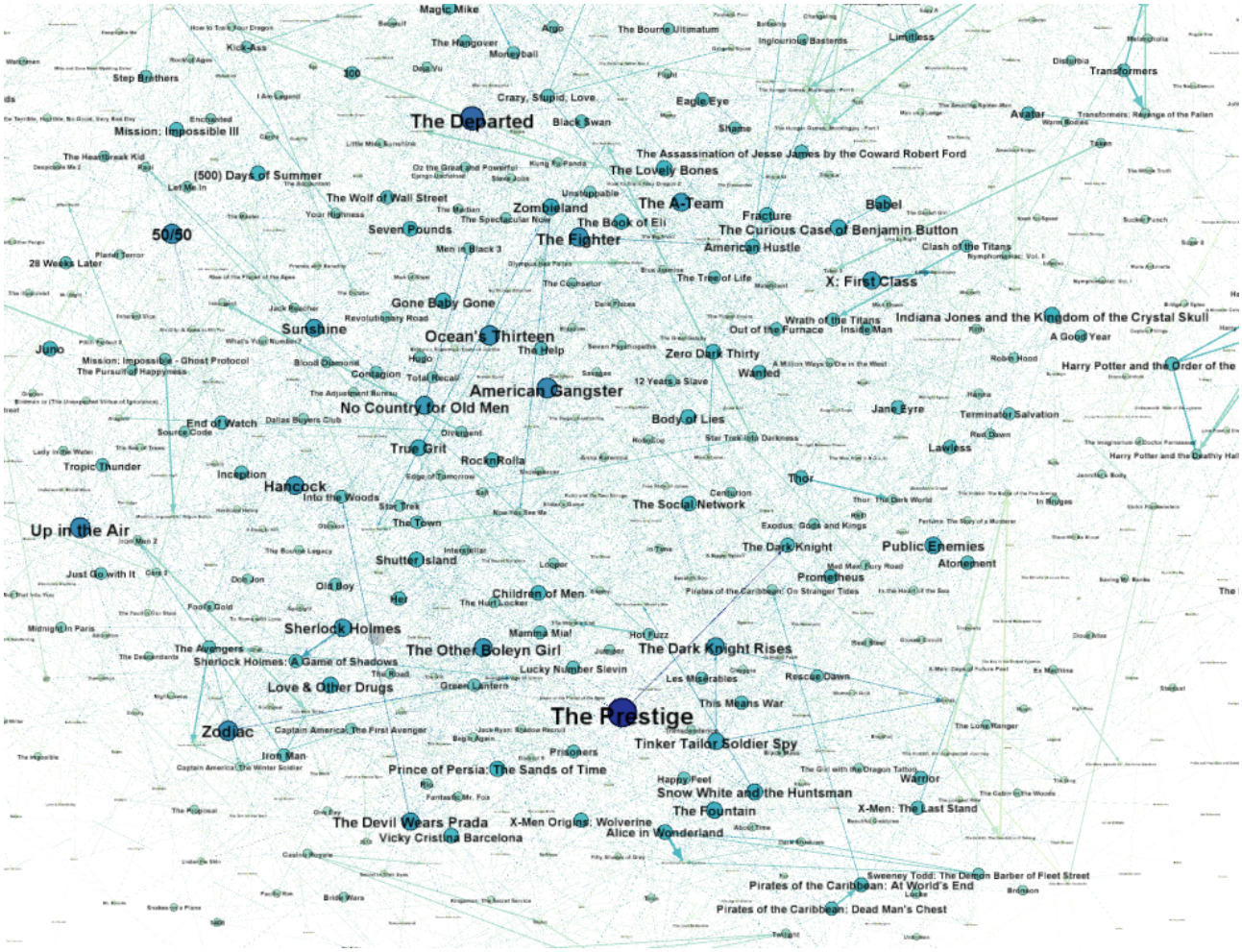




6.16 Koji filmovi su najviše uticali da njihovi glumci igraju u kasnijim filmovima?

Na osnovu usmerenog grafa filmova, gde film A ima granu do filma B ukoliko je neko od glumaca glumio u oba filma, pri cemu je film B izasao nakon filma A, kreiramo vizualizaciju gde velicina i boja cvora koji predstavlja film zavisi od njegovog izlaznog stepena. Ako je ovaj parametar veliki, to znaci da su se glumci iz tog filma pojavljivali u vecem broju drugih popularnih filmova, te se moze zakljuciti da je prvobitni film donekle zasluzan za uspeh svojih glumaca.

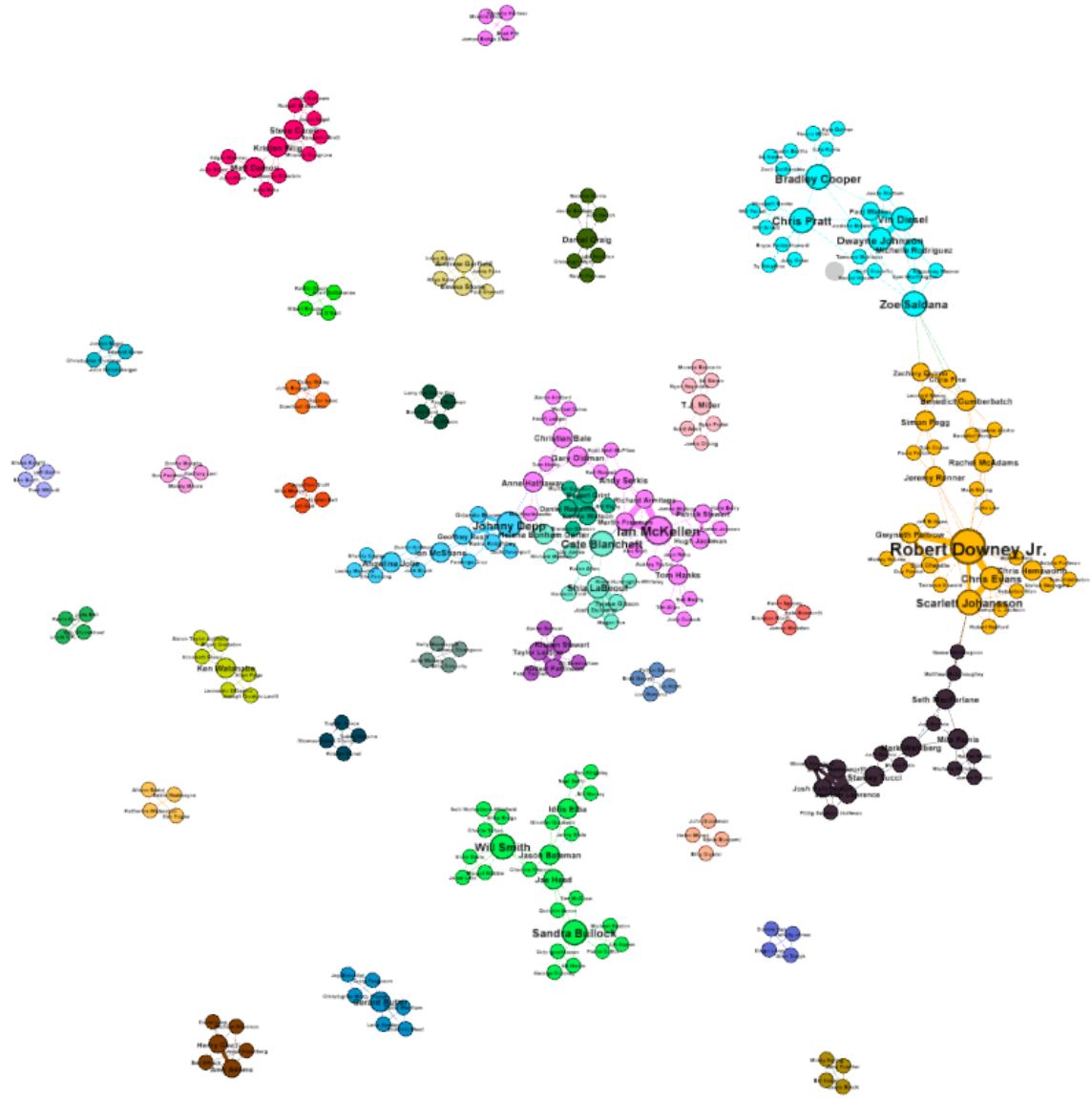
Sa slike mozemo primetiti da su neki od najuticajnijih ovakvih filmova The Prestige, The Departed, The Fighter, 50/50 itd.



6.17 Kako se svojstva mreže menjaju ukoliko se pre formiranja filmovi filtriraju po zaradi?

Ako napravimo novi graf u kome ce se nalaziti samo top 100 filmova sortiranih po zaradi, mozemo primetiti nekoliko promena u parametrima:

- srednji stepen cvorova mreze se smanjuje na **3.79**
- gustina mreze se povecala na **0.014**
- dijametar mreze se povecava na **12**
  (prosecna distanca ostaje slicna kao i pre sa 4.15)
- prosecni koef. klasterizacije se povecava na **0.86** i primecujemo dosta izrazenije klastere (primetni klasteri glumaca iz popularnih filmskih saga: Batman, Avengers, Harry Potter, Fast&Furious itd.)



6.18 Koji režiser je režirao najveći broj filmova?

Ridley Scott, sa 8 reziranih filmova.

6.19 Da li režiseri imaju omiljene glumce koje često angažuju u svojim filmovima?

U sledecoj tabeli su izdvojeni reziseri koji su rezirali najmanje 3 filma i koji su kastovali nekog glumca u bar 80% svojih filmova. Interesantno je primetiti tri rezisera u ovoj tabeli koji i sami glume u svim svojim filmovima.

	Director	Num. of movies	Most freq. casted actor
0	Lars von Trier	4	Charlotte Gainsbourg (100.0%)
1	Dennis Dugan	4	Adam Sandler (100.0%)
2	Seth MacFarlane	3	Seth MacFarlane (100.0%)
3	Ben Stiller	3	Ben Stiller (100.0%)
4	Neill Blomkamp	3	Sharlto Copley (100.0%)
5	Ethan Coen	3	Josh Brolin (100.0%)
6	Sylvester Stallone	3	Sylvester Stallone (100.0%)

6.20 Koje godine je filmska produkcija bila najveća?