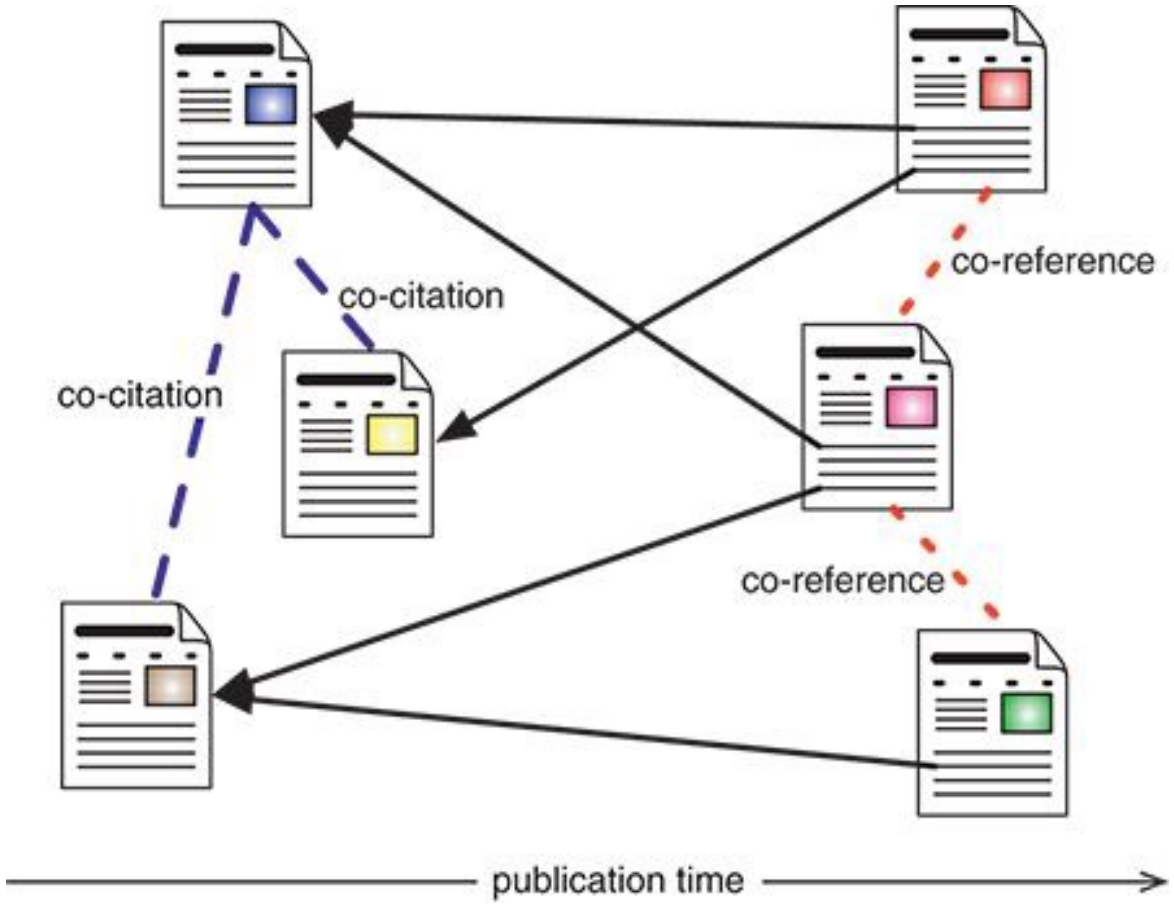
Kompleksne mreže

4. predavanje

Usmjerene mreže

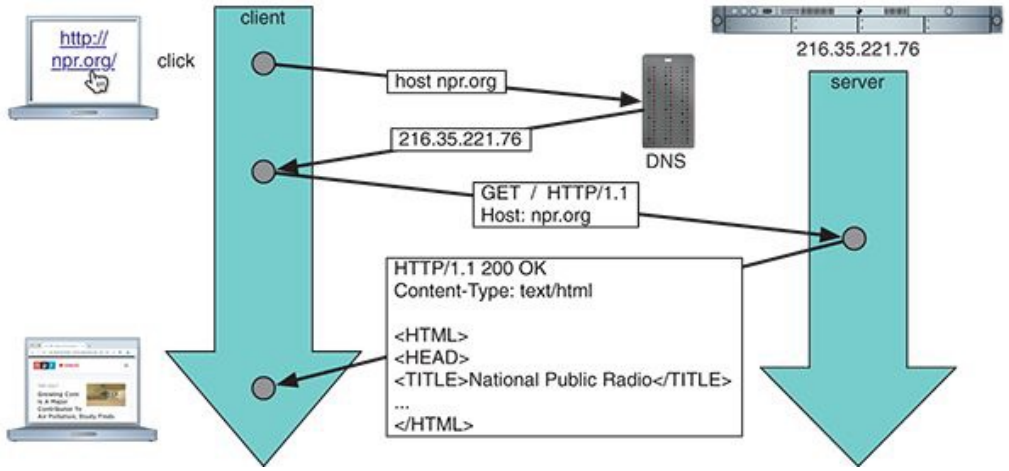
- Socijalne mreže – prijateljstvo simetrično (teoretski ☺)
- Promet (avionski, željeznički, cestovni) – dvosmjeran
- Usmjerene mreže:
  - Twitter
  - Komunikacijske i informacijske mreže
    - Email
    - Wikipedia

Mreža citata



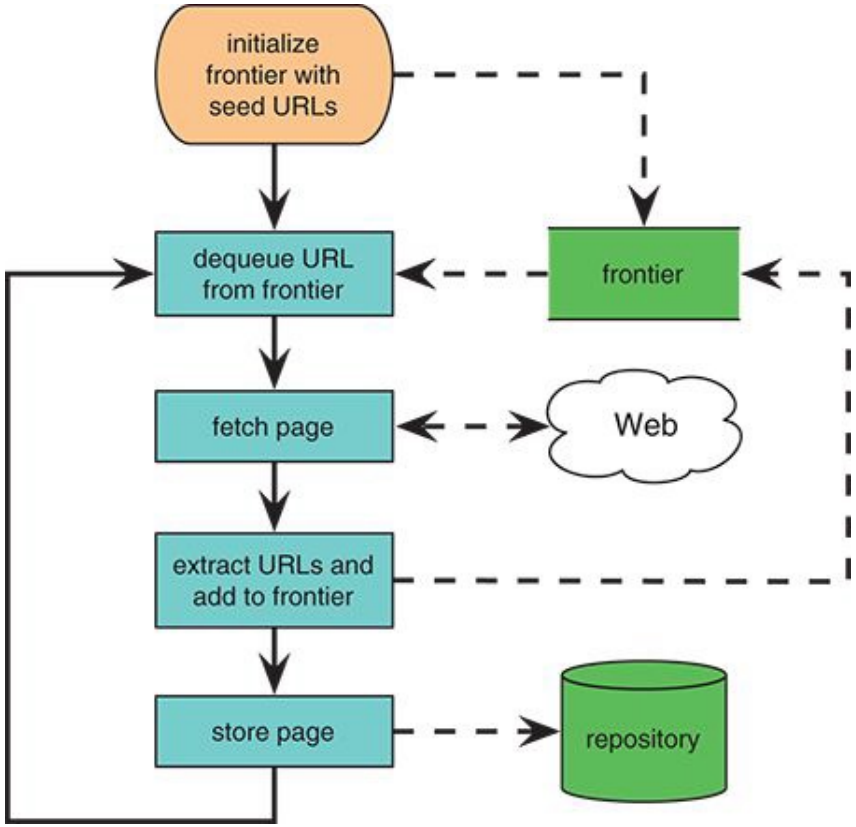
- Citat – usmjerena veza jednog rada na drugi
- Dodatne neusmjerene veze
  - Citirani zajedno (co-citation)
  - Citiraju isti rad (co-reference)

HTTP (Hyper Text Transfer Protocol)



- Zahtjevi
  - Zaglavlje, prazna linija, tijelo poruke
  - GET
    - Zahtjev za stranicom
    - Samo zaglavlje i prazna linija
  - POST
    - Tijelo poruke sadrži dodatne parametre sadržaja
    - Npr. slanje unosa u formu
  - Obavezan *hostname* (jedan server više webova)
- Odgovor
  - Zaglavlje, prazna linija, tijelo poruke
  - Zaglavlje
    - Tip servera, vrijeme, broj vraćenih okteta…
    - Kod odgovora (200-uspjeh, 404 nije pronađen)
  - Tijelo – HTML stranica

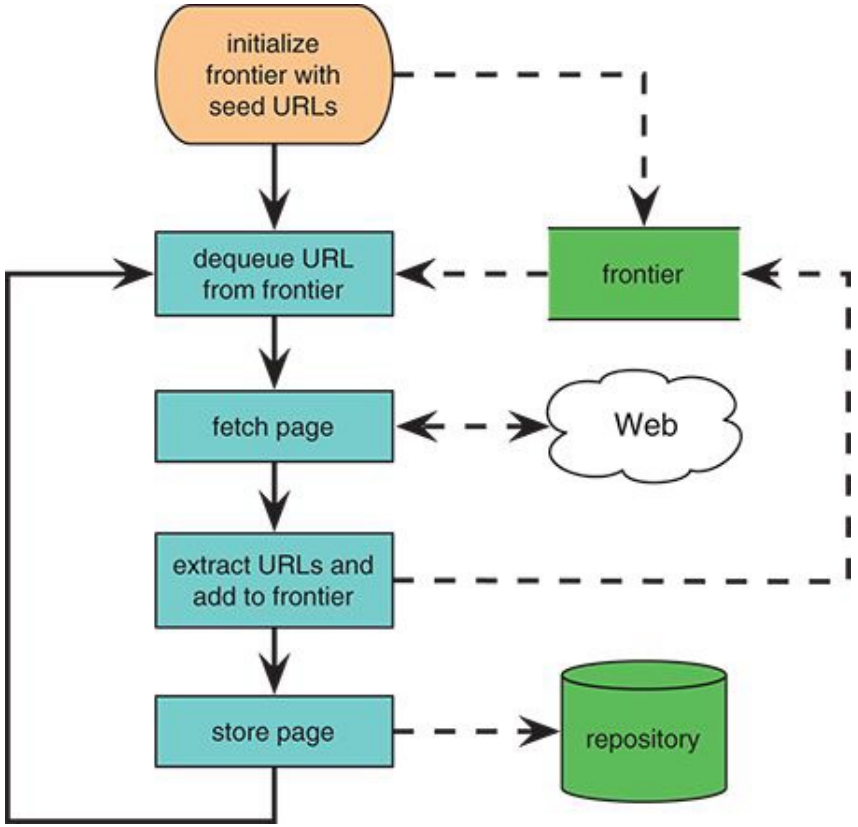
Web crawler



- Programi za automatsko skidanje web stranica
- Najčešća primjena – pretraživači (npr. Google)
  - Indeksiranje sadržaja – mapiranje sadržaja (ključnih riječi i fraza te stranica na kojima se nalaze)
  - Rangiranje
  - Najsofisticiraniji i najosjetljiviji dio pretraživača (Brin i Page)

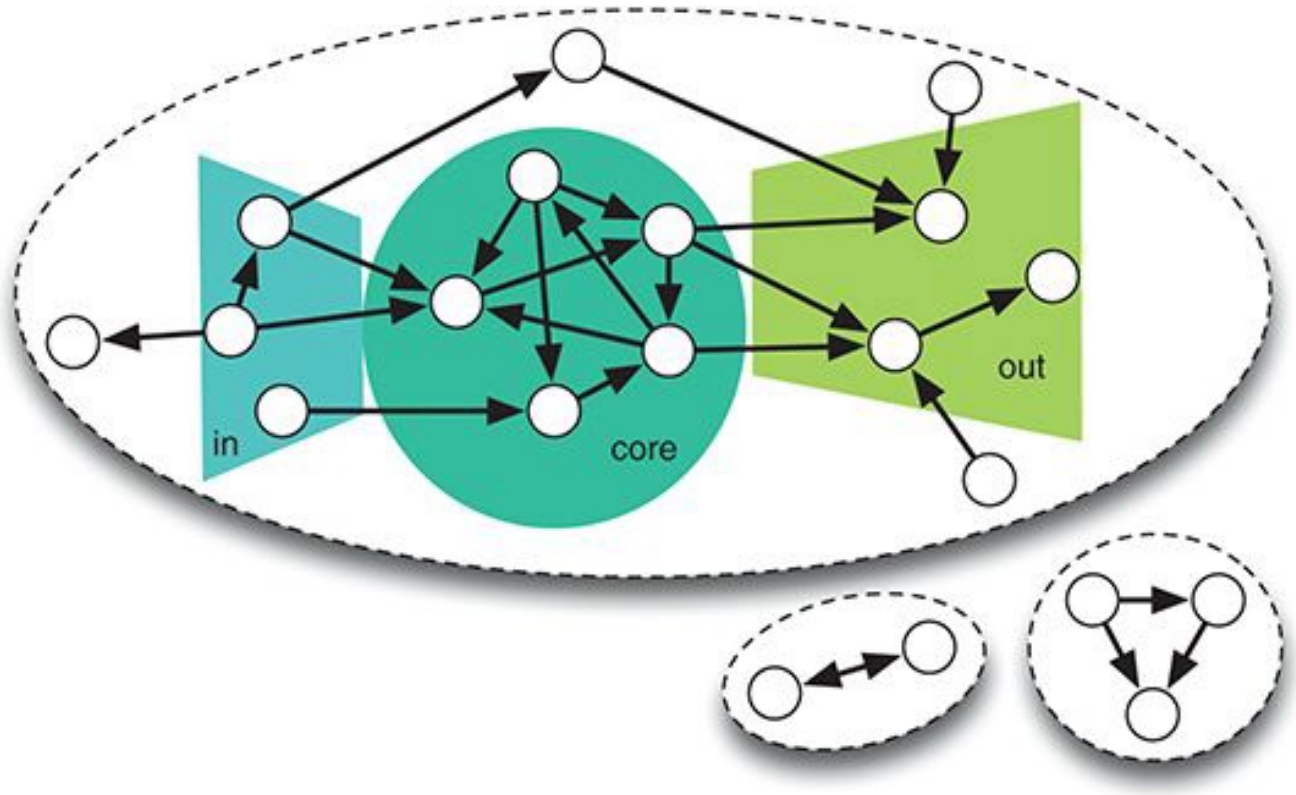
- Druge primjene:
  - Poslovna inteligencija
  - Digitalne biblioteke
  - Webometric alati – utjecaj institucija
  - Maliciozno korištenje (skupljanje email adresa za slanje spam poruka)
  - Znanstvene svrhe (npr. Struktura Weba)

Web crawler



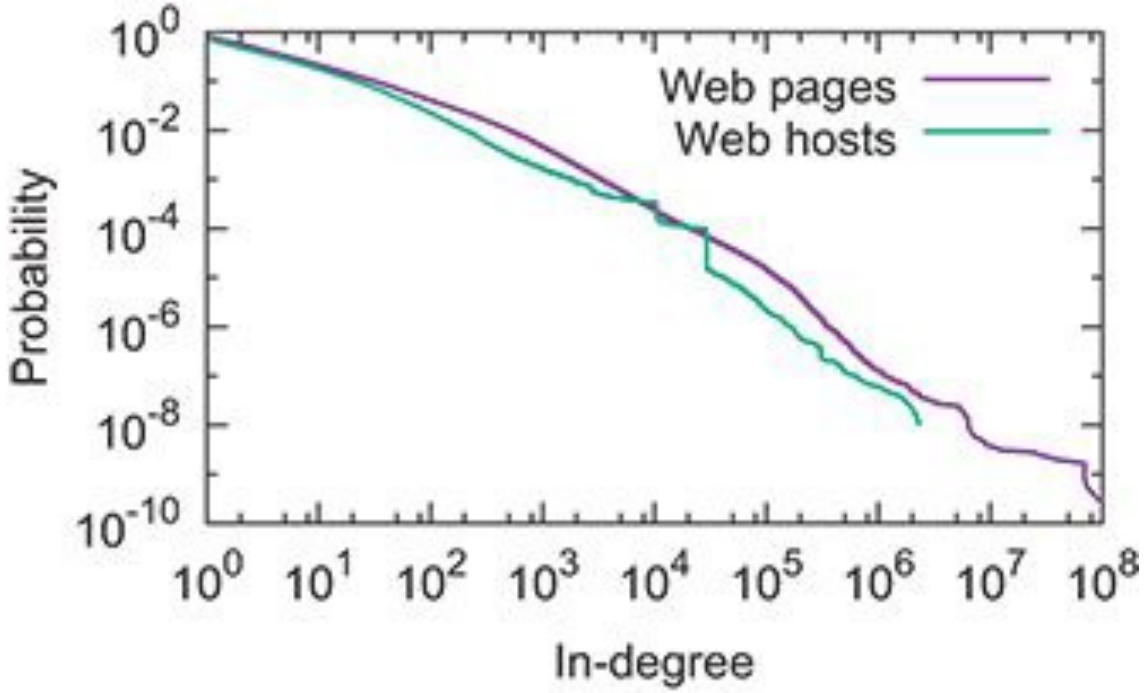
- BFS pristup
- Heuristike za prioritiziranje veza
- Što idemo dublje manja šansa za pronalazak dobrih stranica
- Optimizacija i paralelizacija

Struktura Weba – Web graf



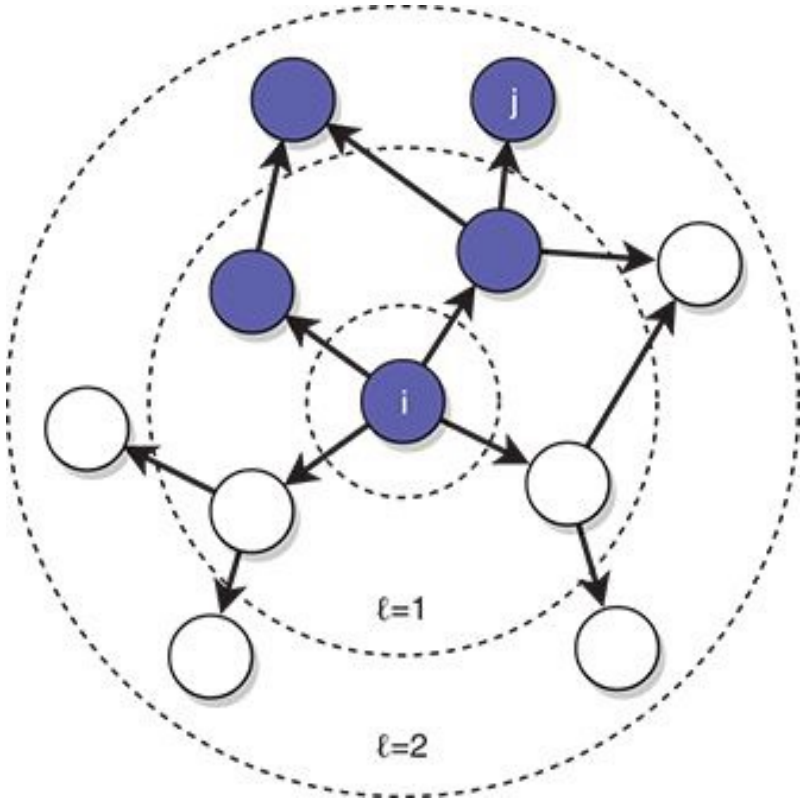
- Dobiven *crawlerima*
- Postoje različite slabo povezane komponente
- Najveća komponenta 90% svih stranica
- Unutar nje imamo snažno povezanu jezgru te *in* i *out* komponente
- “Leptir kravata” struktura

Komplementarna kumulativna distribucija ulaznog stupnja (2012)



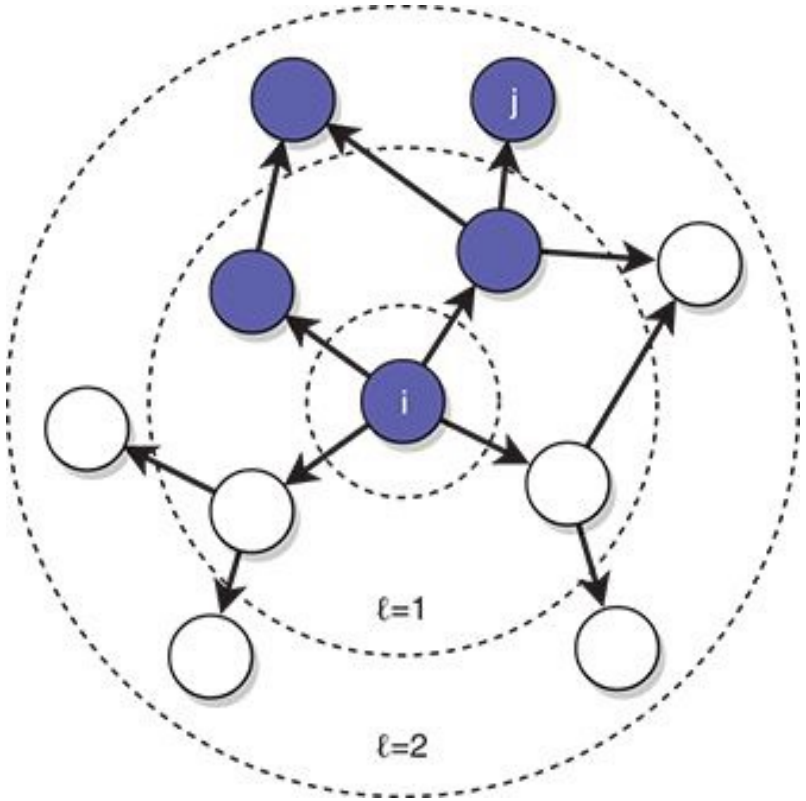
- Prosječan ulazni stupanj – 10-30
- Standardna devijacija red veličine veća -> velik faktor heterogenosti κ
- Oblik distribucije se ne mijenja od vremena kada je Web bio samo par godina star
- Distribucija izlaznog stupnja
  - Mnogo teža za analizu
  - Mnogo uža
  - Puno veza na druge čvorove (spam)

Tematska lokalnost



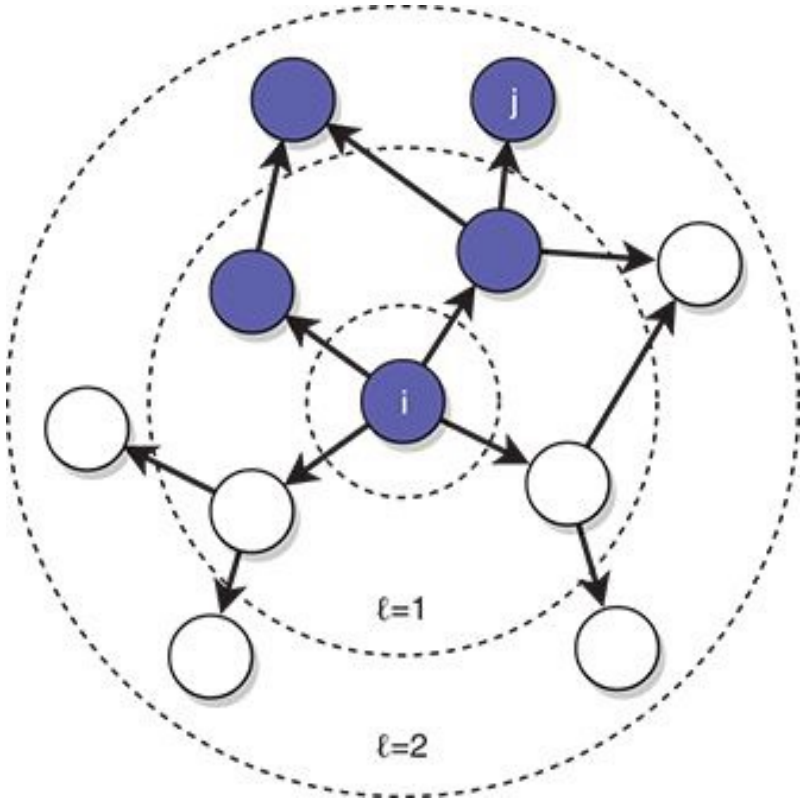
- Mjerenje
  - Izglednost da odredišna stranica u danoj udaljenosti od izvorišne ima istu temu
  - Usporedimo očekivanjem da stranica slučajno bude iste teme (ovisi o temi)
- Stranice unutar 1 – 2 veze od odredišne imaju red veličine veću izglednost u usporedbi sa slučajnom

Tematska lokalnost



- Homofilija – povezivanje sličnih čvorova
- Lokalnost po temi
  - Stranice s povezanim temama povezane ili imaju kratku udaljenost
  - Nove stranice -> povezivanjem s informacijama relevantnim za temu

Tematska lokalnost



- Mjerenje sličnosti teksta
- Više zajedničkih ključnih riječi -> jači dokaz o sličnosti
- Kosinusna sličnost
- Postupak:
  - BFS *crawl* od jedne ili više početnih
  - Mjerenje sličnosti obiđene i početne stranice, usrednjujući za sve početne i obiđene stranice
  - Prikaz promjene sličnosti s udaljenošću
- Tematski *drift* – manja izglednost da naletimo na sličnu stranicu s većom udaljenošću

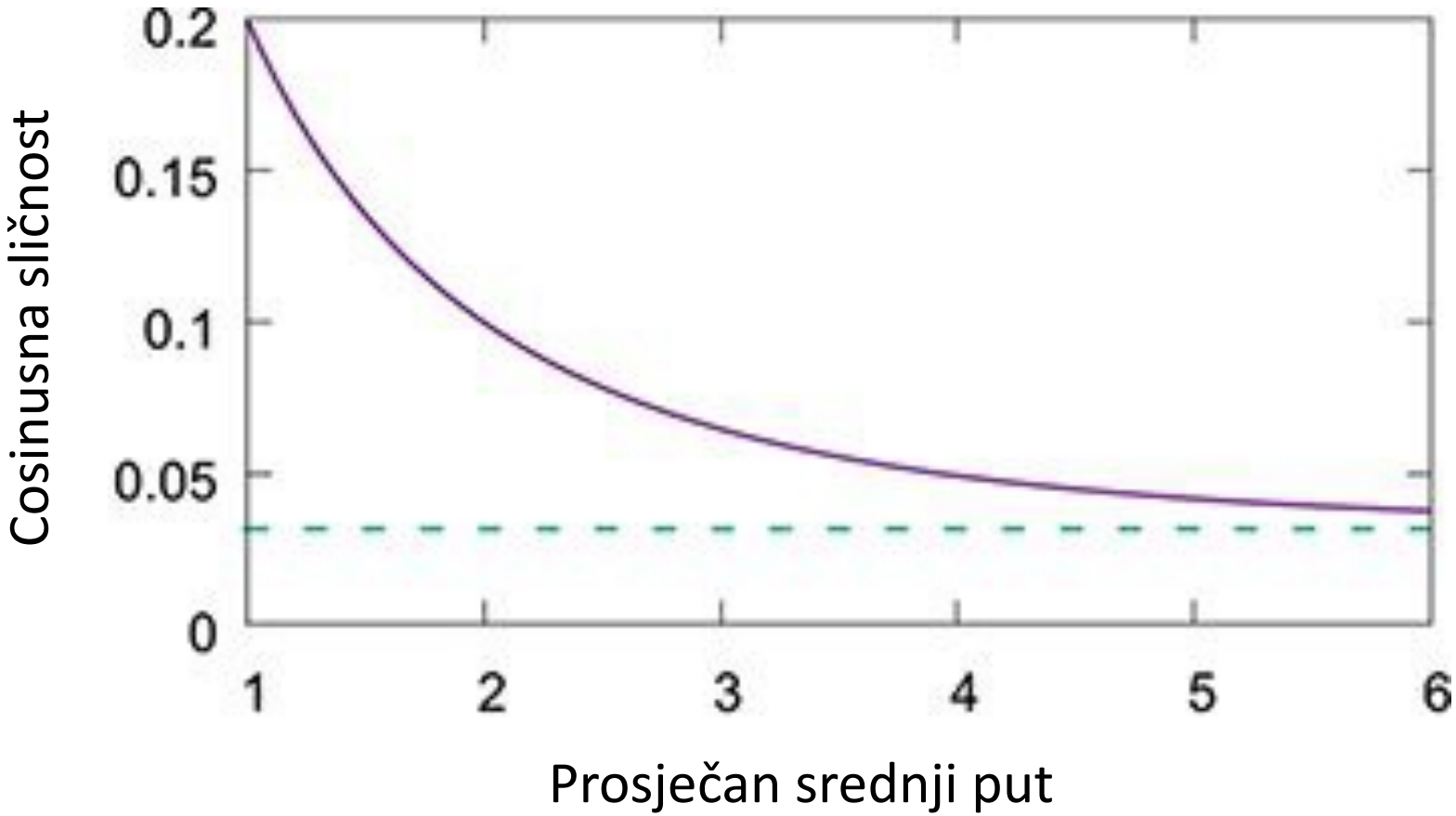
Kosinusna sličnost

- Mjerenje sličnosti između dva teksta (dokument, web stranica, …)
- Reprezentacija dokumenta vektorom $\vec{d} = \{w_{d,1}, \ldots, w_{d,n_t}\}$
  - $w_{d,t}$ - težina izraza *t* u *d*
  - n_t - ukupan broj izraza
  - Suvremeni NLP imaju slične vektore, no dimenzije su latentne reprezentacije umjesto riječi
- Težine
  - Obično proporcionalne frekvenciji pojavljivanja izraza
  - Uklanjanje stop riječi (veznici, članovi, …)
  - Umanjenje težine izrazima koji se često pojavljuju u raznim dokumentima

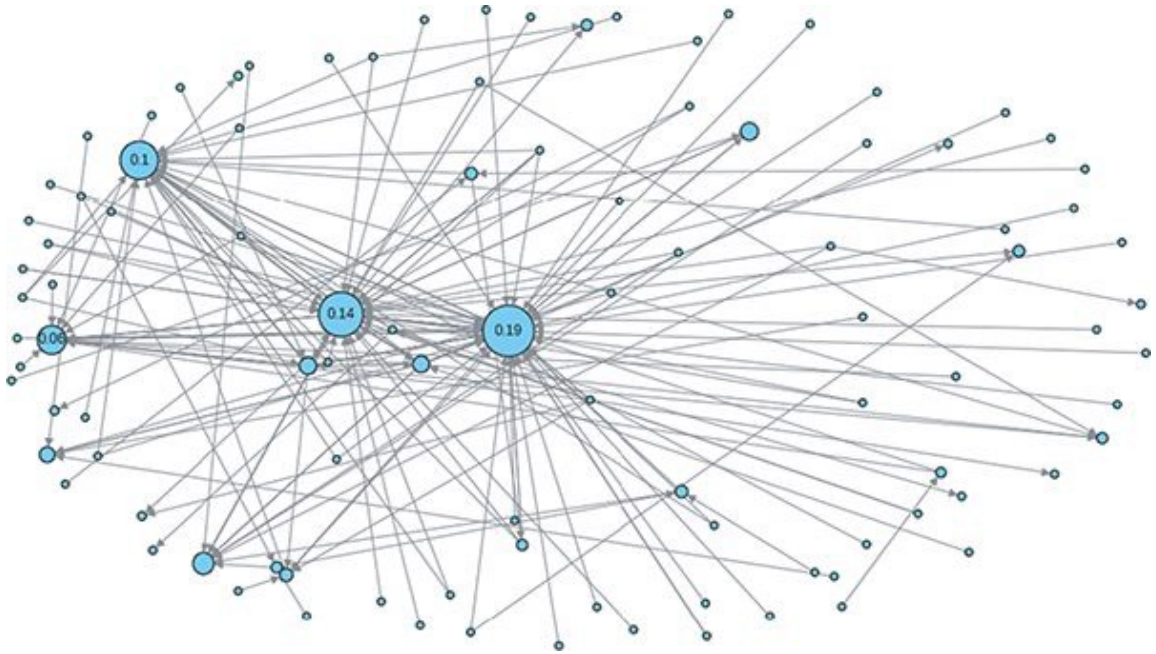
Kosinusna sličnost

- $\cos(\vec{d}_1, \vec{d}_2) = \frac{\vec{d}_1}{\|\vec{d}_1\|} \cdot \frac{\vec{d}_2}{\|\vec{d}_2\|} = \frac{\sum_t w_{d_1,t} w_{d_2,t}}{\sqrt{\sum_t w_{d_1,t}^2} \sqrt{\sum_t w_{d_2,t}^2}}$
- Ako su izrazi u $\vec{d}_1$ prisutni i u $\vec{d}_2$ cosinus će biti blizu 1
- Ako dva dokumenta ne dijele niti jedan izraz cosinus će biti 0
- $\|\vec{d}\| = \sqrt{\sum_t w_{d,t}^2}$ - normizacija normom vektora
- Normizacija - dugačkim dokumentima će norma smanjiti sličnost

Primjer tematske lokalnosti weba

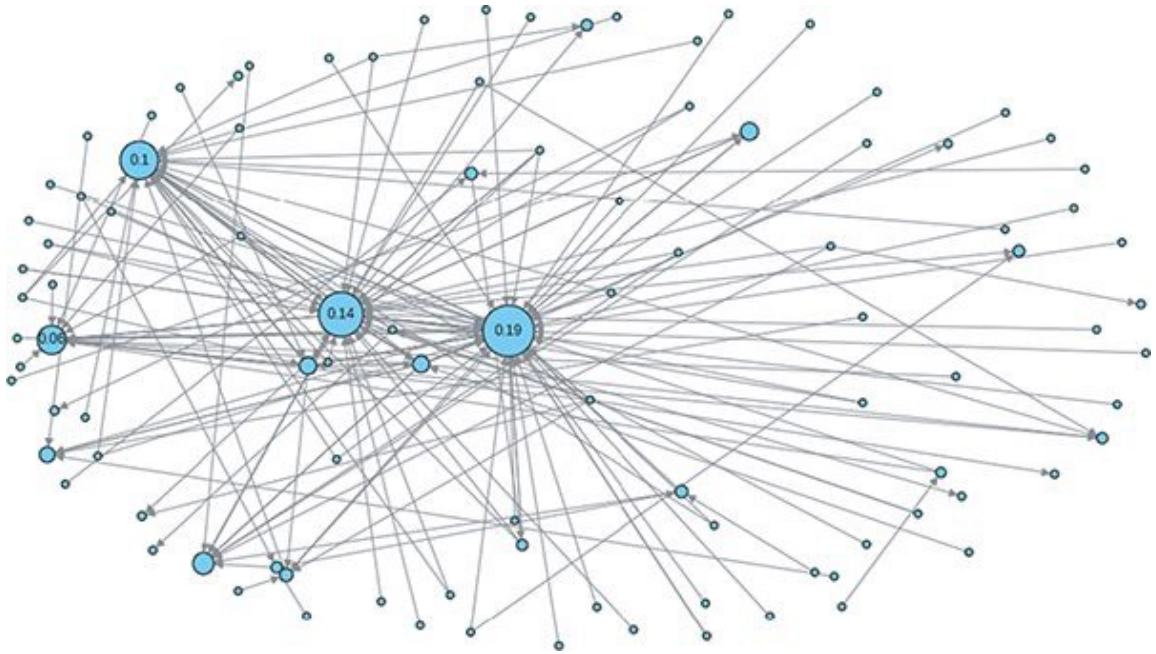


PageRank



- Algoritmi rangiranja
- Mrežna centralnost 1998 (PageRank - Google)
- Stranice s većim PageRankom su bolje rangirane

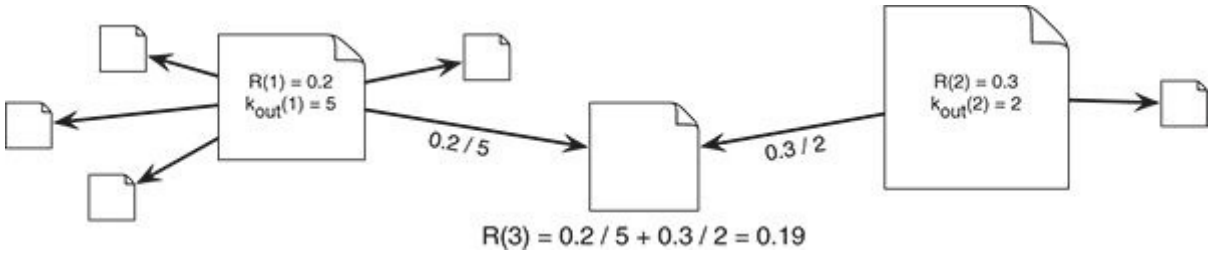
PageRank



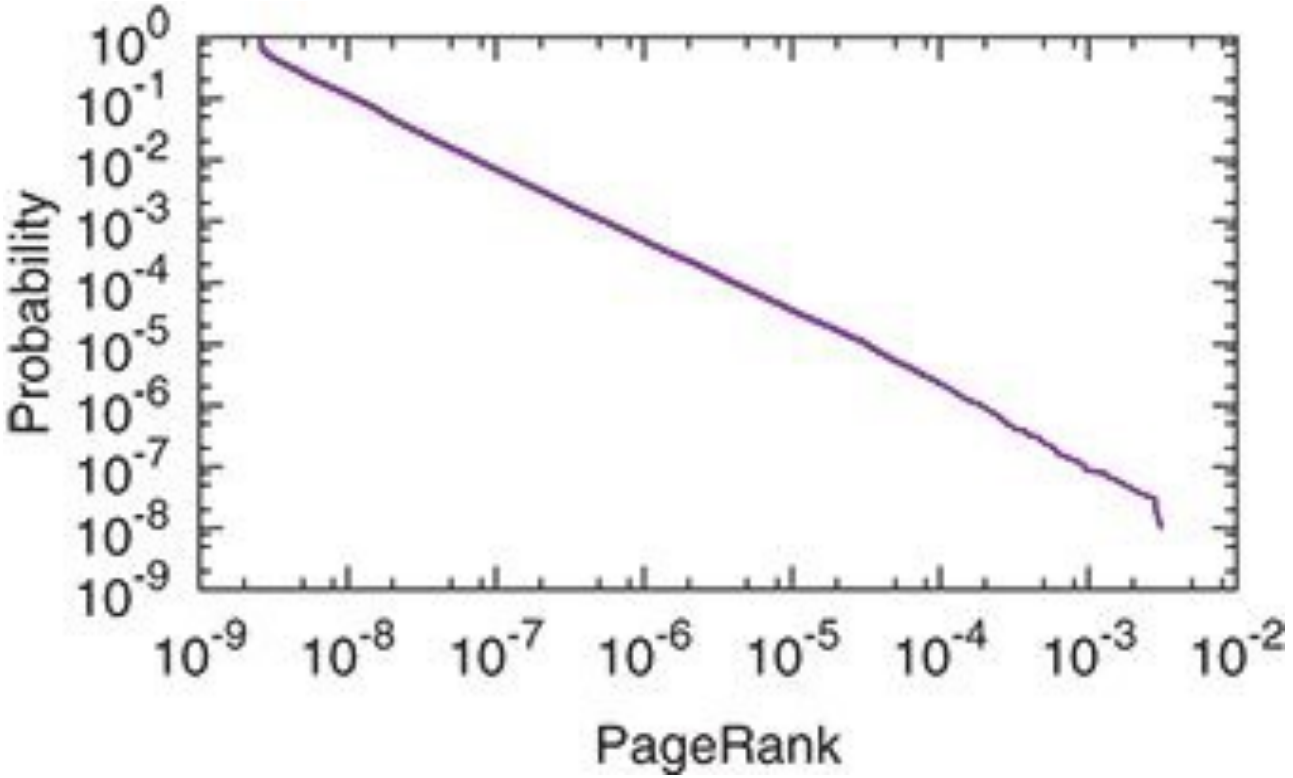
- Slučajna šetnja – svaki susjedni čvor ima jednaku vjerojatnost da bude posjećen
- Slučajni skokovi. Korisnik kreće u sasvim novu potragu (teleportacija)
- PageRank dio vremena koje provedemo stranici

PageRank računanje

- Iterativna metoda
- Suma svih vrijednosti je 1
- Inicijalna vrijednost za svaki čvor je 1/N
- Slučajni skokovi s događaju s parametrom α, obično $\alpha \approx 0.15$
- U svakom koraku vjerojatnost skoka je α, a nastavka pretrage je $1 - \alpha$
- $R_t(i) = \frac{\alpha}{N} + (1-\alpha)\sum_{j \in pred(i)} \frac{R_{t-1}(j)}{k_{out}(j)}$
  - Prvi izraz je teleportacija do *i*
  - Drugi izraz je slučajan hod i uključuje vjerojatnosti dolaska u *i* iz svih njegovih prethodnika
  - Kada je $\alpha > 0$ konvergira u manje od 100 koraka



Komplementarna kumulativna distribucija PageRanka



- Distribucija slična distribuciji čvorova – zašto ne uzeti onda istu ?
- Nisu svi putevi isti – na važnost stranice utječu stranice koje povezuju na nju
- Između dvije stranice s istim ulaznim stupnjem, ona povezena stranicama s većim PageRankom pobjeđuje

Optimizacija za pretraživač

- Poboljšanje ranga
  - Prilagođenje opisa stranice i mogućnosti navigacije
  - Neetičke prilagodbe
    - Kreiranje velikog broja fiktivnih stranica koje povezuju jedna drugu i ciljnu stranicu
    - Kada suvremeni algoritmi naiđu na takve zloupotrebe – micanje iz indeksa pretraživanja

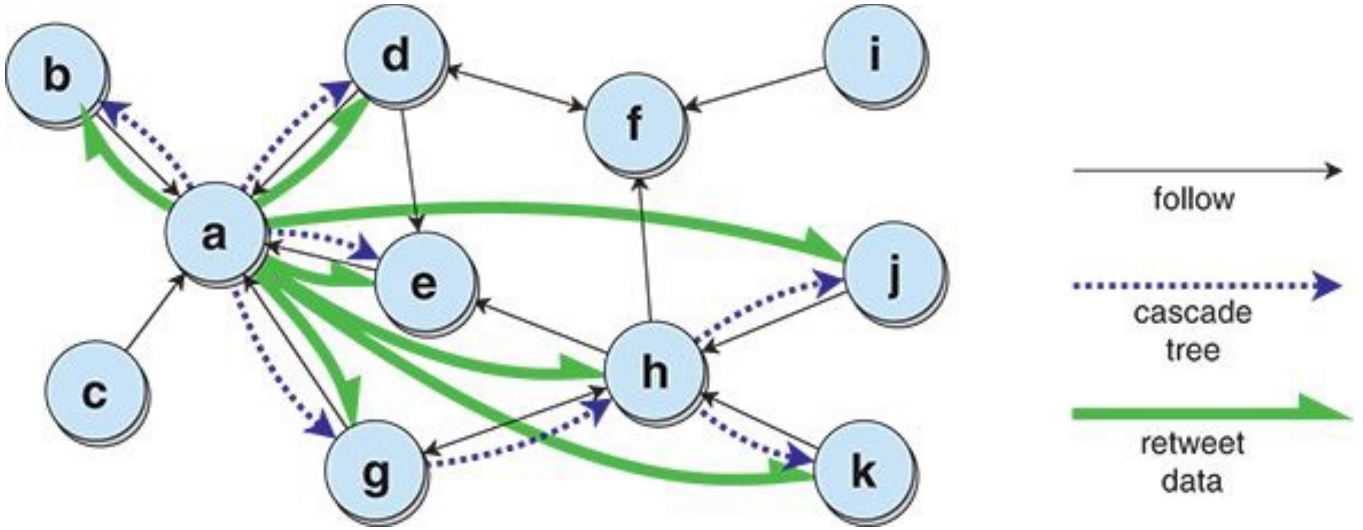
Težinske mreže

- Primjeri:
  - Dva korisnika mogu retweetati jedan drugoga proizvoljan broj puta
  - Mreža emailova
  - Mreže mozga – sinapse – različite razine signala
  - Internet – brzina prijenosa paketa između usmjerivača
- Mreže za koje mislimo da su netežinske
  - Facebook – nisu nam svi prijatelji isti
  - Mreža glumaca koji su glumili zajedno u filmovima
- Informacijske i transportne mreže
- Jačina veza, jačina ulaznih i izlaznih veza

Informacija i dezinformacija

- Mreže koje difuziraju informaciju
  - Čvorovi – ljudi
  - Veze – dijelovi informacije (ideje, koncepti, novosti, ponašanja) koji se prenose s osobe na osobu
  - Prijenosna jedinica informacije – *meme*
  - Korištenje # kod Twittera #FER

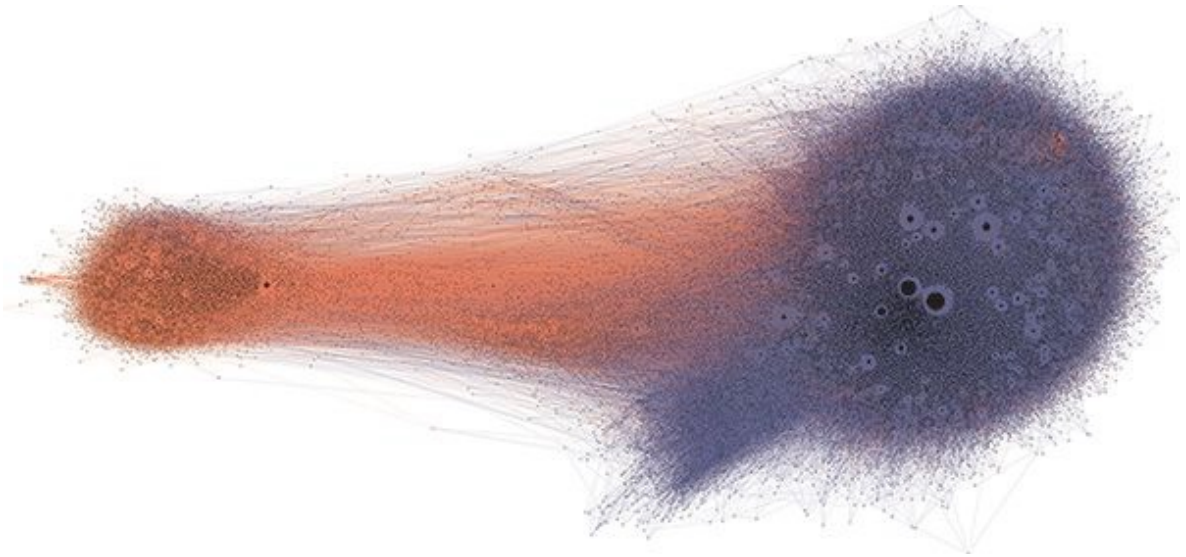
Twitter/X



- Difuzija u Twitter mreži:
  - Retweet, quoted retweet, spominjanje, odgovori
  - Retweet kaskada – usmjereno stablo tijeka širenja meme od kreatora do svih izloženih korisnika
  - Jedan meme može generirati više kaskadnih stabala
  - Agregacija stabala – šuma
  - Kaskadnu šumu možemo promatrati kao sloj u višeslojnoj mreži

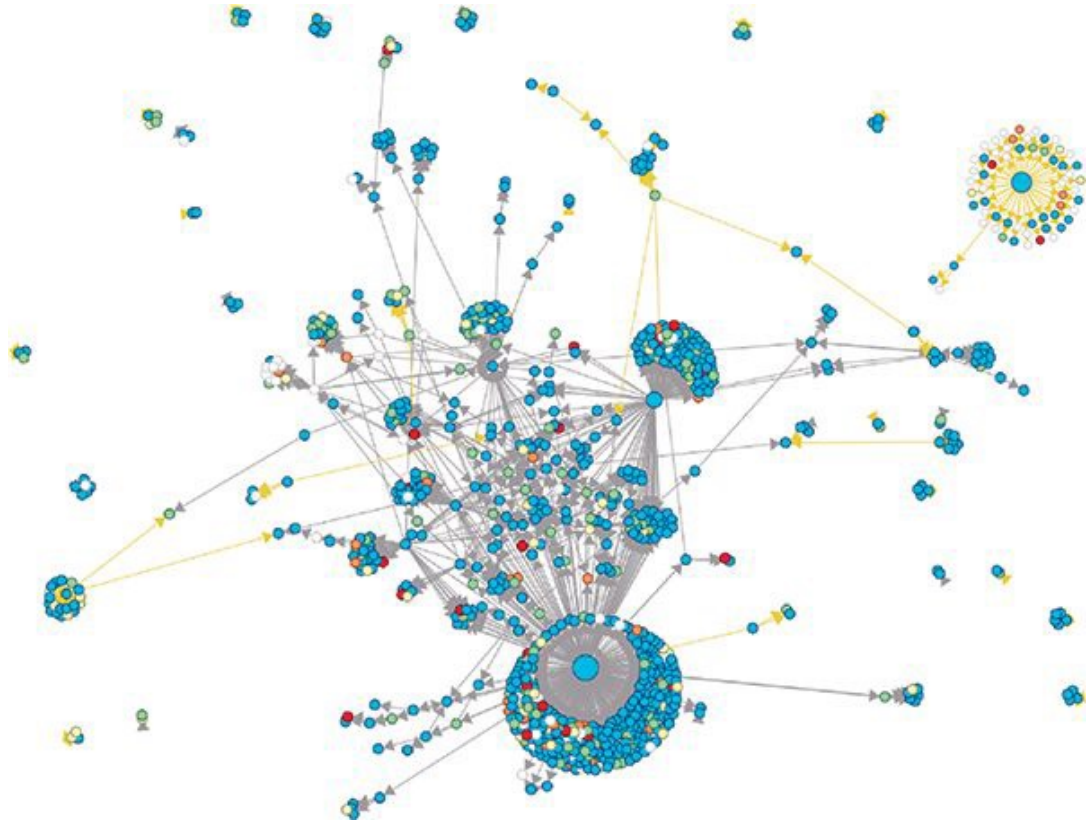
Primjer difuzne mreže – 2016 izbori u US

Primjer difuzne mreže – 2016 izbori u US



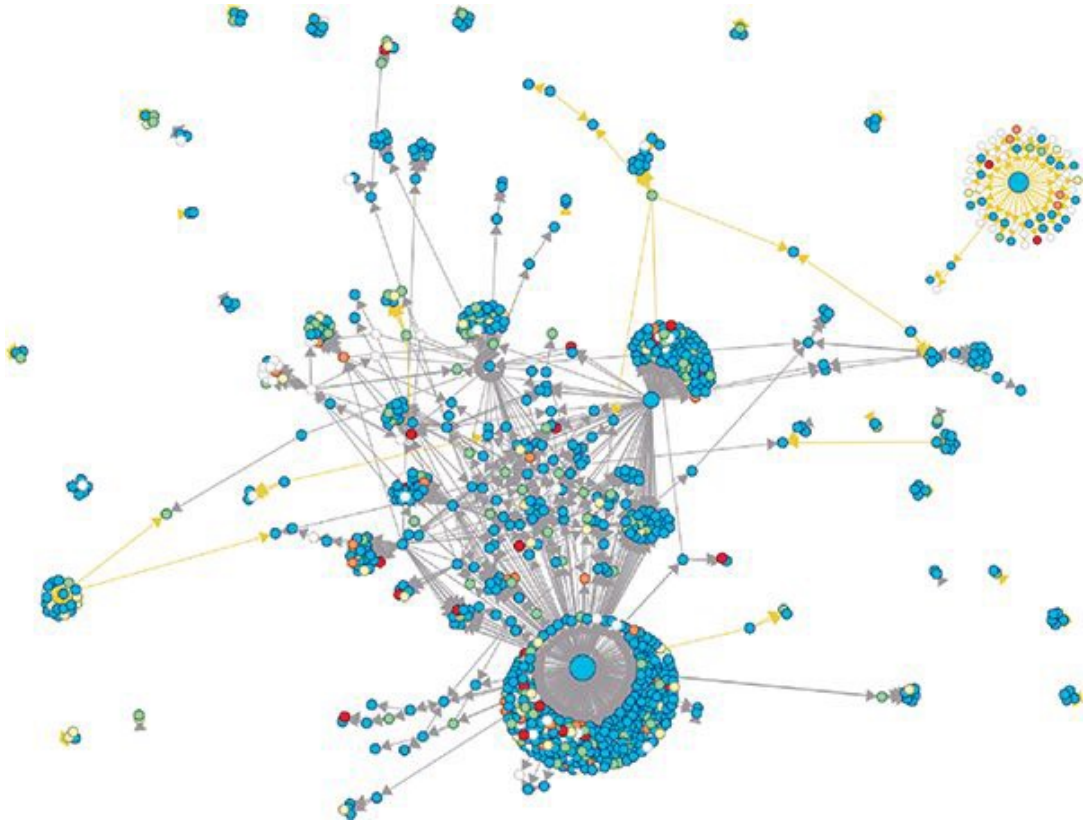
- Podmreža retweetova članaka vezanih uz izbore 2016
- Čvorovi – Twitter računi,
- Veze - retweet linka na članak (crveni – provjera informacije, ljubičasta niska vjerodostojnost)
- k = 5 jezgra pune mreže
- Korisnici koji šire dezinformacije dijele jako mala članaka iz izvora koji provjeravaju istinitost

Zaraznost *memea*



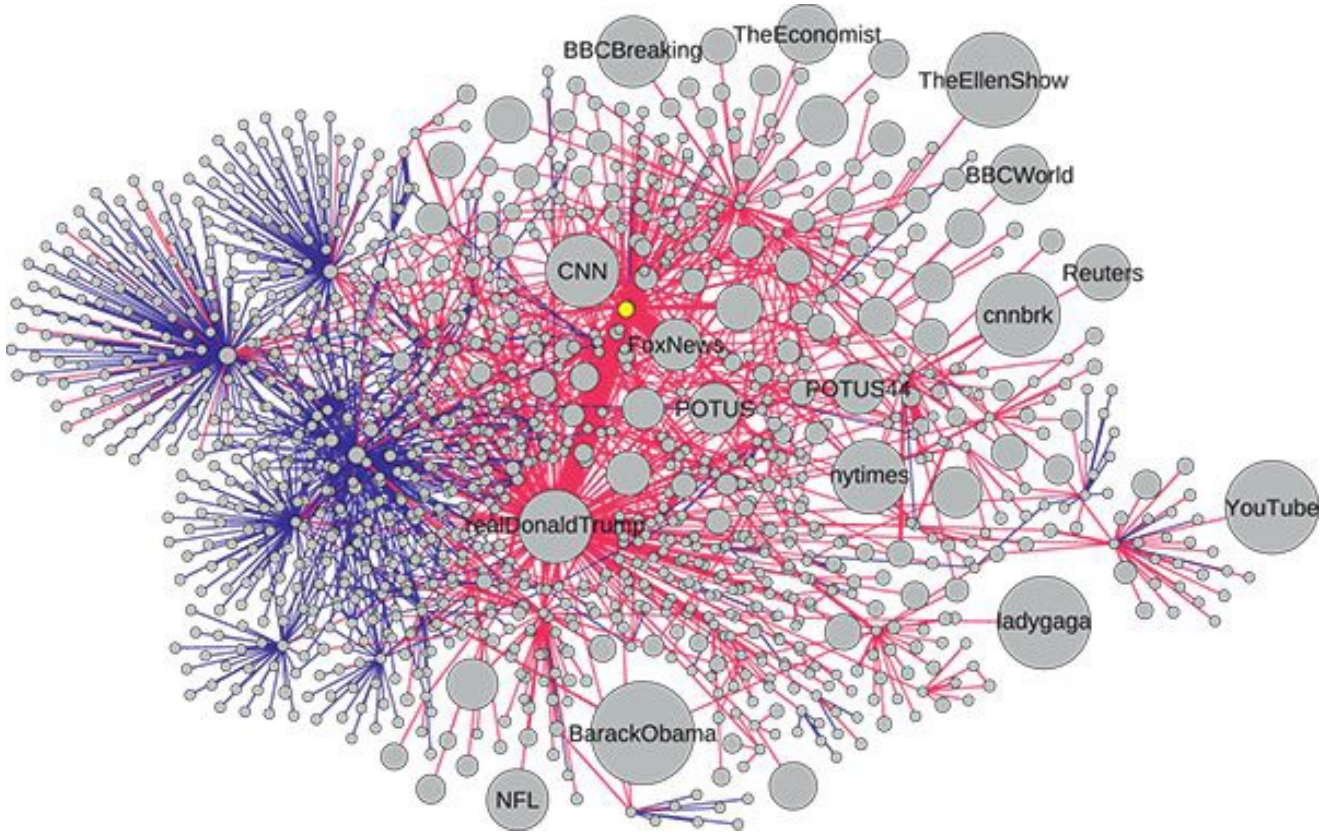
- Mjerenje zaraznosti:
  - Broj izloženih korisnika
  - Struktura mreže. *Meme* koji šire popularne osobe može biti više vezan uz to osobu nego *meme*
  - Duboka mreža *retweetova* može govoriti o široj privlačnosti poruke

Retweet mreža dva članka Bijelih Kaciga u Siriji



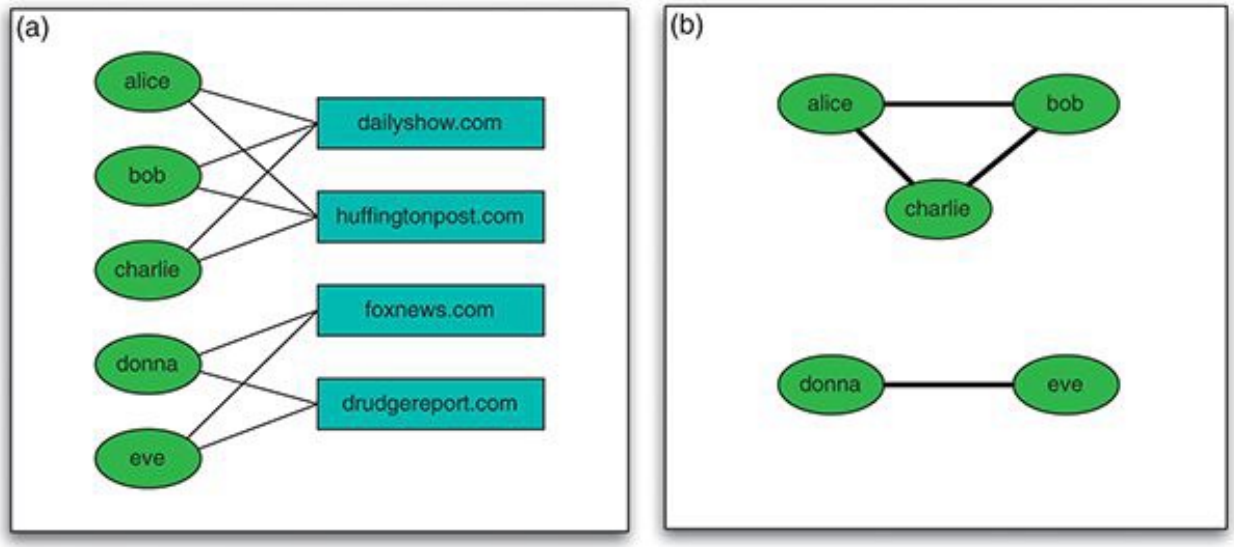
- Kampanja dezinformacijama
  - Lažna povezanost s teroristima
  - Druge teorije zavjere
  - Sive veze – dezinformacija
  - Žute veze – provjerena tvrdnja
  - Plavi čvorovi – ljudi
  - Crveni čvorovi – botovi
- DeepFake

Mreža lažnih vijesti o izbornoj prevari izbora 2016 od strane ilegalnih stranaca



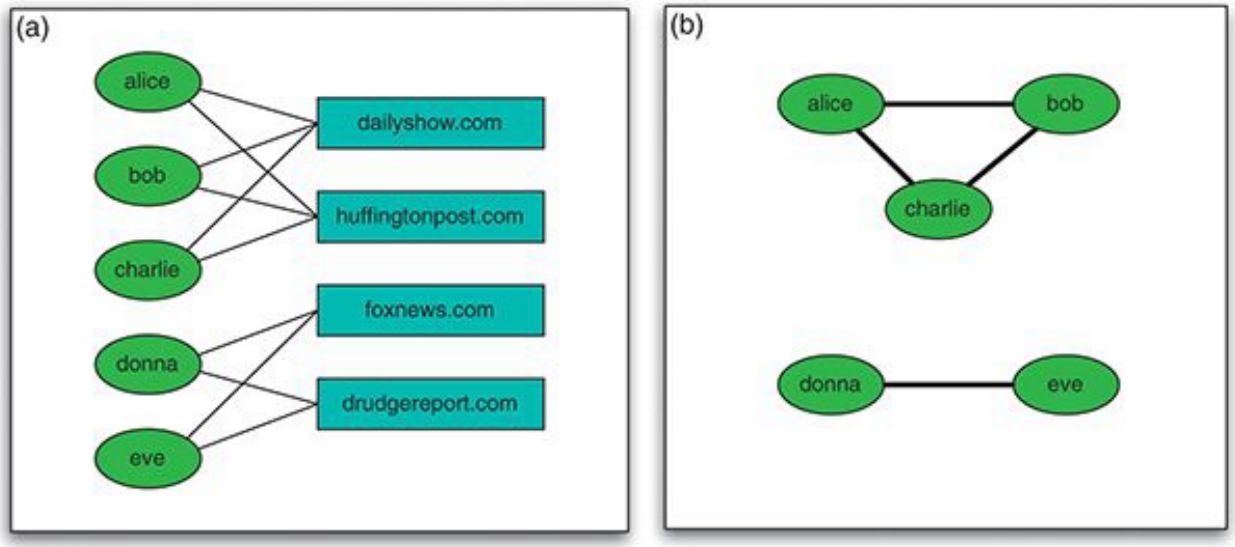
- Plavo - Citirani *tweetovi* i *retweetovi*
- Crveno – spominjanje i odgovori
- Širina linka – težina
- Mali žuti čvor – *bot* koji je sistematski širio dezinformacije u odgovorima na informacije koje su spominjale US predsjednika

Mreže zajedničkog pojavljivanja



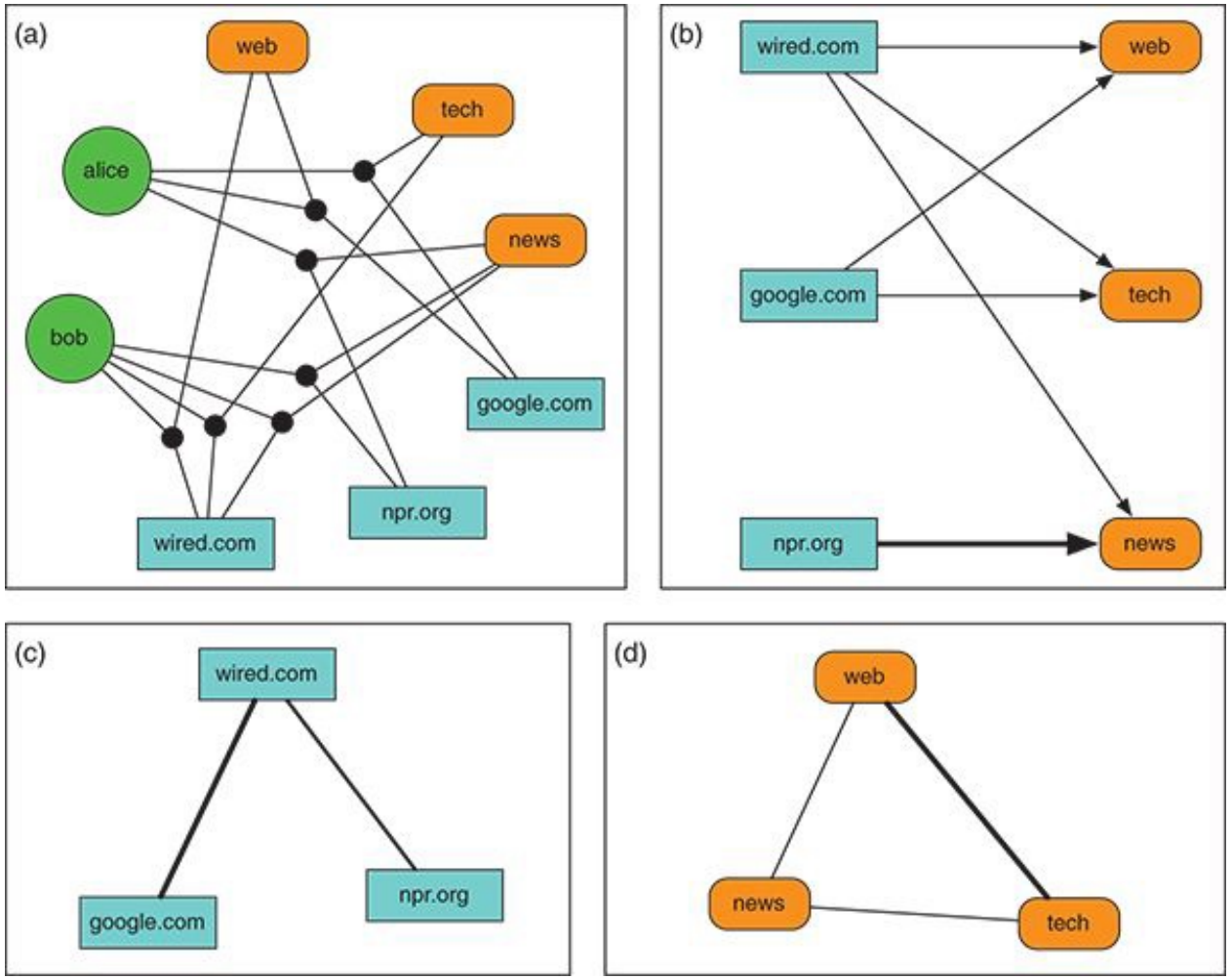
- Veze između različitih vrsta entiteta
- Usmjerena mreža
  - Svi izvorišni čvorovi na jednu stranu
  - Svi odredišni na drugu
  - Isti čvor može biti na obje strane (duplikat)
  - Mreža citata (rad koji citira i citirani rad) -> nova mreža u svakoj od grupa
  - Radovi mogu biti ko-citirani i biti ko-referencirani (nekoliko radova citira jedan ili više radova)
- Bipartitne mreže – način reprezentacije ovakvih mreža

Bipartitna mreža -> Mreža zajedničkog pojavljivanja



- Veze između dva čvora različitog tipa
- Mreža ko-zvijezda u filmovima
  - Kreiramo težinsku mrežu iz bipartitne (projekcija) – mreža zajedničkog pojavljivanja
- Bipartitna mreža „likeanja” (a)
- Mreža zajedničkog pojavljivanja (b)
- Koriste se za preporuke i ciljano oglašavanje

Folksonomy

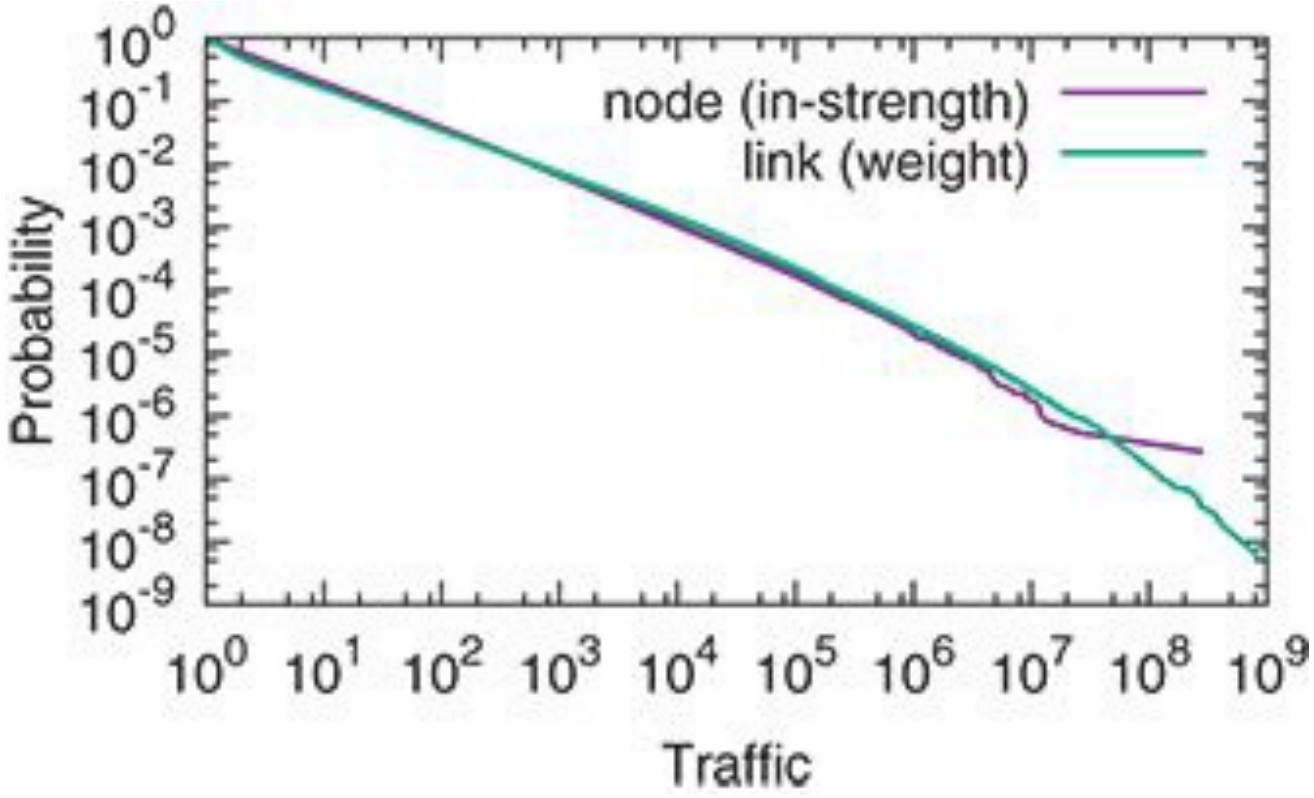


- Bipartitne mreže isto mogu imati težine.
- Sustavi ocjenjivanja – koliko nam se npr. sviđa knjiga ili film
- Označavanje – označavanje izvora (identificiranog URL-om) s jednom ili više oznaka (npr. YouTube)
- Osnovni element je trojka *(u,r,t)* gdje korisnik *u*, označava resurs *r* s oznakom (*hashtag* kod Twittera)
- Agregirano po puno korisnika dobijemo skup koji nazivamo *folksonomy*

Heterogenost težina

- Težina veze može nositi informaciju o procesu ili odnosima modeliranim mrežom
- Različite težine mogu predstavljati različito udruživanje
- Mreže prometa
  - Putnici ili letovi između aerodroma
  - Auti između križanja
  - Internet (paketi između usmjerivača)
  - Wikipedia (klikovi između članaka)

Mreža Web prometa



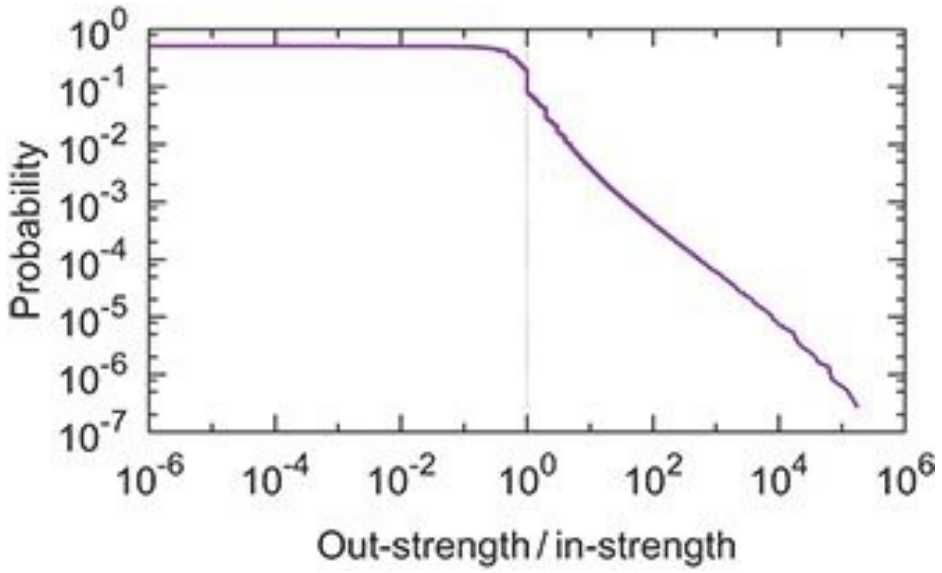
- Ulazna snaga čvora – ukupan broj klikova na stranicu
- Težina veze – ukupan broj klikova na hiperlink
- Obje distribucije heterogene (otežani rep)

Postoji li korelacija između PageRanka i ulazne snage ??

Usporedba PageRanka i ulazne snage

- Može li PageRank predvidjeti promet?
- NE, usprkos sličnoj distribuciji
- Korelacija je zapravo slaba
- Pretpostavka PageRank modela – model slučajnog surfera
- Inicijalne pretpostavke PageRank modela su narušene

Koje pretpostavke PageRanka su najmanje realistične?



- Omjer ulazne i izlazne snage čvorova
  - Osim za početne i krajnje čvorove jednog pretraživanja omjer treba biti 1
- Teleportiranje ne favorizira niti jedan čvor
  - Jednaka vjerojatnost za svaki čvor da će biti izvor ili odredište skoka (i s ovime omjer je blizu 1)
- Očekujemo usku distribuciju oko 1
- U stvarnosti krećemo od manjeg skupa stranica
- Većina čvorova nije zanimljiva i na njima završavamo pretragu i skačemo dalje
- Slučajna teleportacija je nerealistična