Razvoj omrežij skozi čas — Verjetnostni induktivni razredi grafov

N. Kejžar[1] V. Batagelj[2]

[1]Fakulteta za družbene vede

[2]Fakulteta za matematiko in fiziko

Seminar na IBMI, 24. maj 2007

Kazalo

Uvod

Zgodovina:

- teorija grafov od 18. stoletja, Euler (osnova za analizo)
- analiza socialnih omrežij (Moreno, Cartwright, Bavelas)
- antropologija (Barnes, Mitchell)

Analiza lastnosti enot in odnosov/povezav med njimi.

Graf — ogrodje omrežij

Omrežja dodatno lahko vsebujejo:

- lastnosti enot (vozlišč v grafu): spremenljivke (spol, starost)
- lastnosti na povezavah: tip povezave (prijateljstvo, svetovanje), moč povezave (kako močno je prijateljstvo)

Definicija

Omrežje N je graf z dodatnimi lastnostmi:

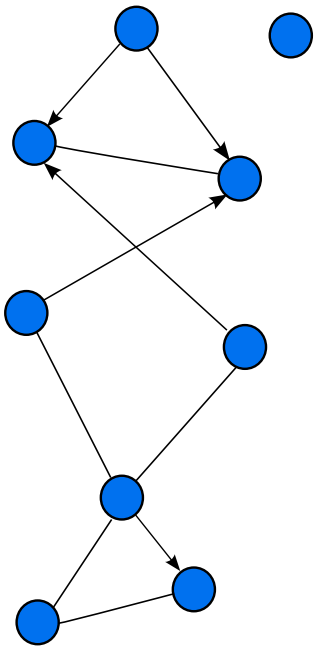
$$N = (V, L, \mathcal{P}_V, \mathcal{P}_L),$$

kjer je:

- V množica vozlišč
- L množica povezav (usmerjene ali neusmerjene)
- $\mathcal{P}_V$ množice lastnosti vozlišč
- $\mathcal{P}_L$ množice lastnosti povezav

Primeri

enota	**povezava**	**omrežje**
človek	prijateljstvo	socialno
raziskovalec	sodelovanje	
članek	citiranje	omrežje citiranj
podjetje	lastništvo	ekonomsko
protein	interakcija	biološko
strežnik	elektr. povezava	internetno
spletna stran	hiperpovezava	www omrežje



Glavne smeri raziskovanja

Glavne smeri raziskovanja analize omrežij:

1. analiza strukturnih lastnosti v omrežju
   - lastnosti enot in podomrežij (npr. centralnost enote, najbolj gosto povezano podomrežje)
   - lastnosti celega omrežja (npr. porazdelitev stopenj točk, koeficient grupiranja)
2. razvoj različnih modelov za generiranje omrežij
3. analiza procesov na omrežjih

Porazdelitev stopenj točk

stopnja točke $\equiv$ število povezav, ki imajo en konec v tej točki (vhodna stopnja, izhodna stopnja)

Definicija

$p_k \equiv$ verjetnost, da ima vozlišče v grafu stopnjo k

Prvi model omrežij: ER slučajni graf (Erdős, Rényi)
n točk, p verjetnost za obstoj povezave

$$p_k = \binom{n-1}{k} p^k \cdot (1-p)^{(n-1-k)},$$

Postavimo $p = c/n$, torej ko $n \to \infty$:

$$\lim_{n\to\infty} \binom{n}{x} \left(\frac{c}{n}\right)^x \left(1-\frac{c}{n}\right)^{n-x} = \frac{c^x e^{-c}}{x!}$$

porazdelitev konvergira proti Poisson(c).

Potenčna porazdelitev

$$p(x) = c \cdot x^{-\gamma},$$

$x > 0$, c — konstanta za normalizacijo

- logaritmiranje $\rightarrow$ linearna funkcija
- iskanje parametra γ iz podatkov
  1. logaritemsko kategoriziranje (velikost intervala za kategorijo se eksponentno povečuje) in prileganje premice na logaritmirane podatke
  2. prileganje premice na logaritmirano kumulativno porazdelitev

$$P(X > x) = \int_x^{\infty} c \cdot y^{-\gamma} dy = -\frac{c}{(\gamma - 1)} x^{-(\gamma - 1)}.$$

  3. MLE ocena

MLE ocena

Porazdelitvena funkcija, ko upoštevamo spodnjo mejo:

$$p(x) = \frac{\gamma - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\gamma}.$$

Logaritmirano največje verjetje:

$$L = \ln p(x|\gamma) = \sum_{i=1}^{n} \left(\ln(\gamma - 1) - \ln x_{min} - \gamma \ln \frac{x_i}{x_{min}} \right)$$

Ocena parametra:

$$\hat{\gamma} = 1 + n \left(\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right)^{-1}$$

Najkrajše poti

- l_{uv} — najkrajša pot med vozliščema u in v
- $l = \frac{1}{n(n-1)} \sum_{u<v} l_{uv}$ — povprečna najkrajša pot
- povezanost omrežja?
- $l^{-1} = \frac{1}{n(n-1)} \sum_{u<v} l_{uv}^{-1}$ — harmonična sredina (Latora et al)
- klika: $l = 1$
- ER slučajni graf:
  povprečno št. sosedov k, na razdalji l jih dosežemo pribl. k^l
  povprečje: $k^l \sim n$, $n = |V|$

$$l \sim \frac{\ln n}{\ln k}$$

small-world effect

Koeficient grupiranja — clustering coefficient

- gostota trikotnikov v grafu (močno povezane podskupine vozlišč v grafu)
- več mer:
  - za 1 vozlišče: $c_u = \frac{2x_u}{k(k-1)}$; x_u število povezav med sosedi u
  - popravljena verzija $c'_u = c_u \cdot \frac{k}{\Delta}$
  - za cel graf

$$c = \frac{6 \cdot \#\text{ trikotnikov}}{\#\text{ poti dolžine 2}}$$

- socialna omrežja imajo strukturo "core-periphery":
  - grupiranje glede na interes, članstvo v društvih ipd.
  - predstavitev kot 2-vrstno omrežje
  - razvrščanje v skupine, blockmodelling

Korelacija stopenj — degree correlation

- "measure of assortativity" (Newman) temelji na Pearsonovem korel. koeficientu
- verjetnost preostale stopnje vozlišča:

$$q_k = \frac{(k+1)p_{k+1}}{\sum_l l \cdot p_l}$$

- q_{jk} verjetnost za preostalo stopnjo slučajne povezave
- $q_{jk} = q_j q_k$, kjer korelacije stopenj ni
- mera:

$$r = \frac{1}{\sigma_q^2} \sum_{jk} (q_{jk} - q_j q_k)$$

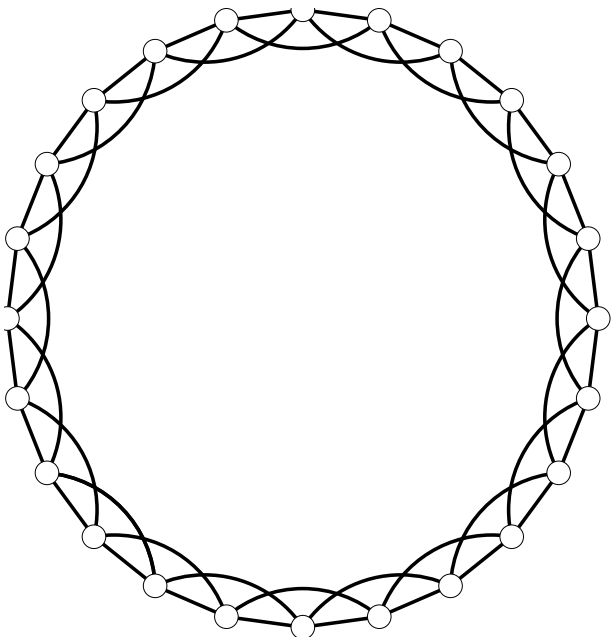
varianca preostale stopnje: $\sigma_q^2 = \sum_k k^2 q_k - (\sum_k k q_k)^2$

Korelacija stopenj — realna omrežja

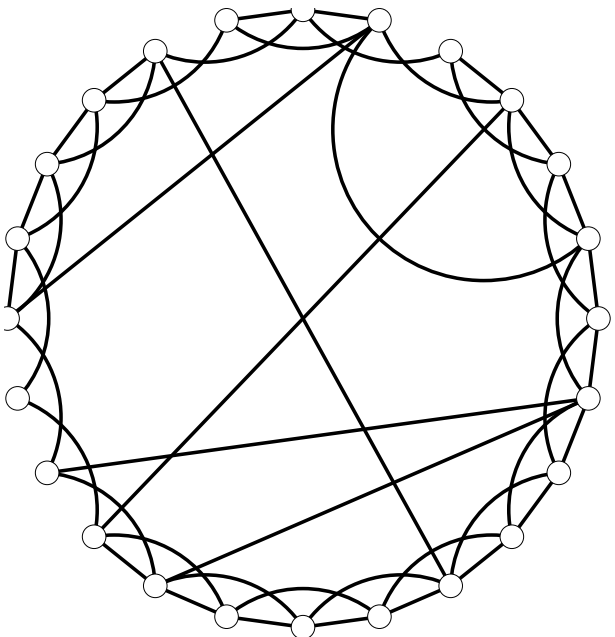
- socialna omrežja — pozitivna
  (pretvorba 2-vrstnega v enovrstno omrežje?)
- tehnološka (Internet, www)
  biološka omrežja (proteini, prehranjevalna veriga, nevronska omrežja) — negativna
- aplikacija v epidemiologiji — kako zaustaviti razširjanje bolezni

Modeli za razvoj omrežij

- ER slučajni graf
- konfiguracijski model (določena *porazdelitev stopenj*)
- small-world model
  - majhna *povprečna najkrajša pot*
  - opazno *grupiranje*



1D mreža s 24 vozlišči



prevezovanje s $p_{sw} = 1/7$

"Preferential attachment model"

- model razvijajočega omrežja (stohastični proces)
- preferenca povezovanja na bolj priljubljena vozlišča
- zgradi graf s potenčno porazdelitvijo stopenj ($\gamma = 3$)
- Algoritem (1 časovni korak):
  - dodaj vozlišče
  - vozlišče poveži z m_{ba} ostalimi vozlišči, ki jih izbereš preferenčno na njihovo trenutno stopnjo:

$$p_u = \frac{k_u}{\sum_j k_j}$$

- bolj natančen opis:
  - začnemo z 2 povezanima vozliščema
  - v vsakem koraku dodamo vozlišče in povezavo
  - za $m_{ba} > 1$ združimo m_{ba} zaporedno dodanih vozlišč v eno

"Preferential attachment model" — variacije

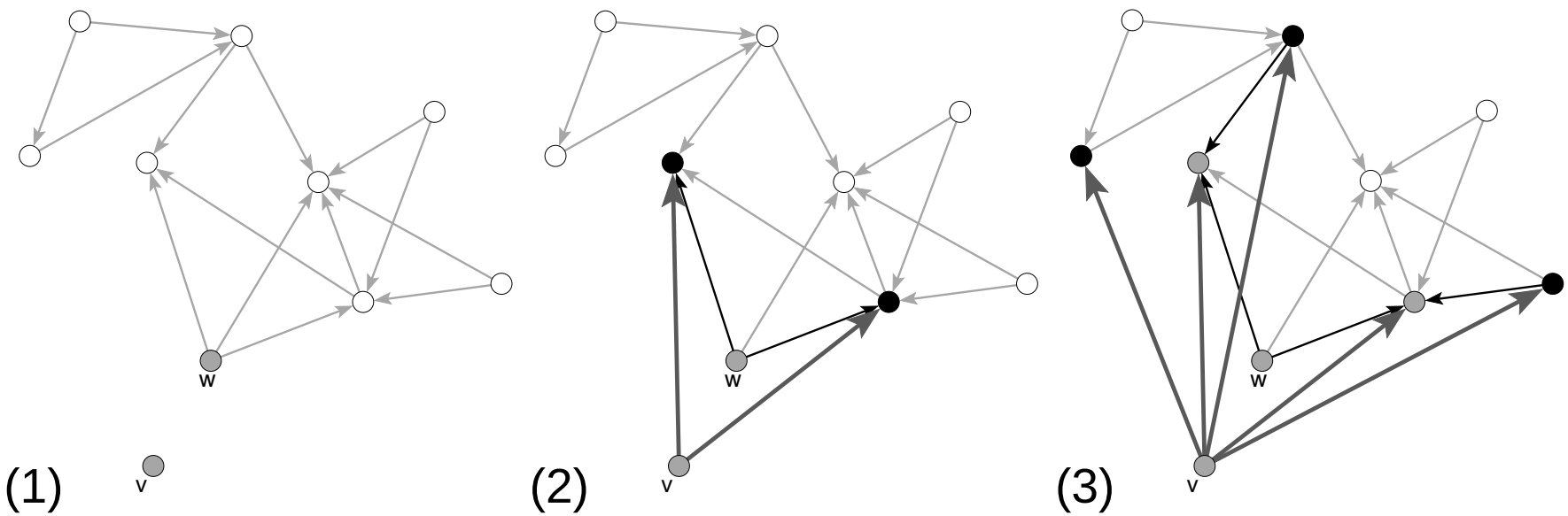
- nelinearna (k^α) preferenca povezovanja (Krapivsky, Redner)
  - $\alpha < 1$ — porazdelitev stopenj potenčna, pomnožena z raztegnjeno eksponentno
  - $\alpha > 1$ — 1 vozlišče dobi končen delež vseh povezav
  - $\alpha > 2$ — 1 verjetnost, da je vozlišče povezano z vsemi > 0
- prevezovanje, brisanje povezav, vozlišč med gradnjo omrežja
- z grupiranjem (Dorogovtsev) — novo vozlišče se poveže z vozliščema slučajno izbrane povezave

Model kopiranja

- stohastični proces kopiranja povezav
- potenčna porazdelitev stopenj
- grupiranje (motivacija www omrežje)
- Algoritem (1 časovni korak):
  - dodaj vozlišče (z neko verjetnostjo)
  - izberi vozlišče v;
    - z verjetnostjo p_c dodaj m_c povezav med v in naključno izbranimi vozlišči
    - z verjetnostjo $1 - p_c$ izberi naključno vozlišče u in kopiraj m_c njegovih povezav (poveži v s m_c sosedi vozlišča u)
- osnovna verzija:
  - dodamo vozlišče in ga izberemo (torej postane v)
  - $m_c = 1$ (v dobi 1 povezavo)

"Forest-fire model"

- naraščanje povprečne stopnje v omrežju
- krčenje povprečne najkrajše poti

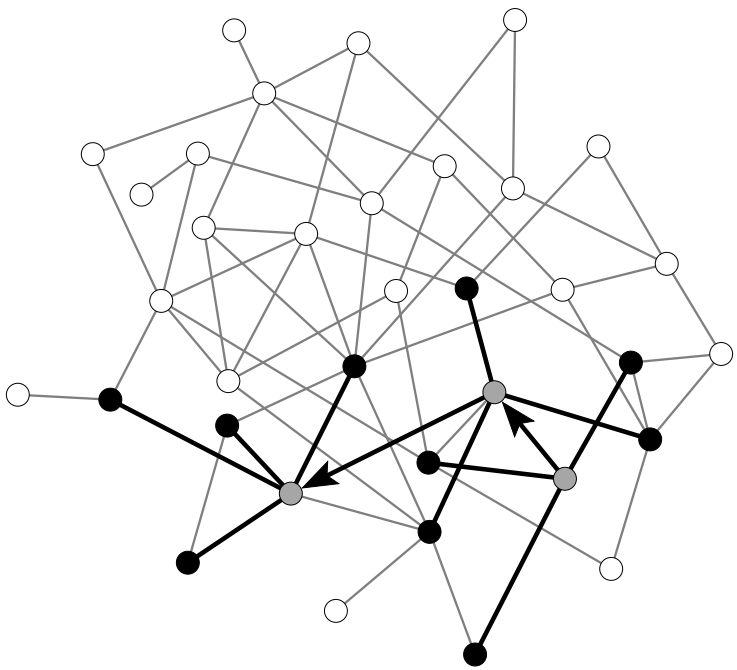


Analiza procesov na omrežjih

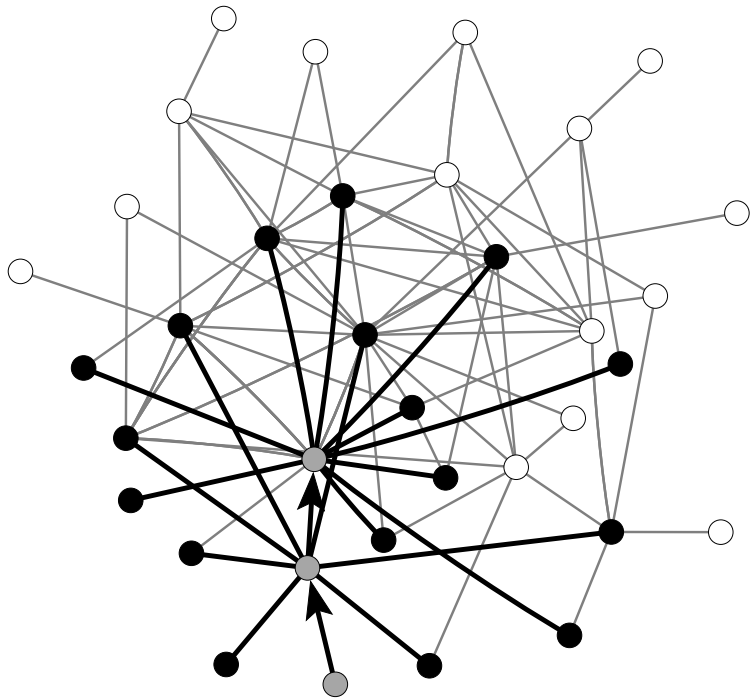
- odpornost omrežja na napade — Internet, www
  (povezanost omrežja)
  analiza zmanjševanja največjega povezanega podgrafa
  (krepke komponente)
- razširanje virusov (znanja) po omrežju
  le ničelna verjetnost prenosa bolezni ne bo povzročila izbruha epidemije v omrežju s potenčno porazdelitvijo stopenj ($\gamma < 3$)
- iskanje po www omrežju (Adamic et al) računsko hitrejše iskanje po potenčnih omrežjih

Analiza procesov na omrežjih — 2

v 3 korakih dosežemo:
13 vozlišč v ER modelu
20 vozlišč v potenčnem modelu



Erdős-Rényi model



potenčno omrežje

Matematična indukcija

1. za dokaze, da neka trditev drži
2. za definiranje razreda objektov
   - množico osnovnih objektov
   - množico pravil za generiranje

Pravila transformirajo objekt iz razreda v nov objekt razreda.
Eberhard je prvi induktivno opisal razrede grafov.
Transformacije razumemo kot implicitne časovne korake.

Induktivni razredi grafov

Induktivni razred grafov

ICG (inductive class of graphs) je definiran kot $\mathcal{I} = (\mathcal{B}; \mathcal{R})$ (Curry, 1963):

1. množica $\mathcal{B}$ osnovnih objektov, baza ICG,
2. množica $\mathcal{R}$ pravil za generiranje, ki so določena z **levim elementom** (delom grafa), na katerem uporabimo pravilo (ga spremenimo v **desni element**)

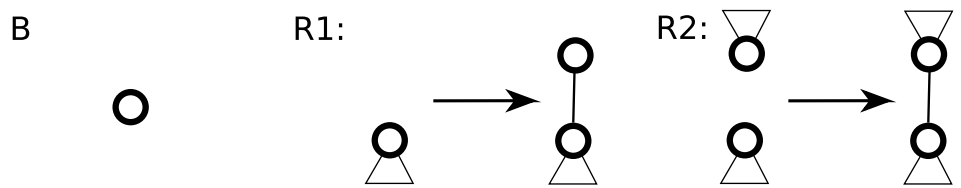
Induktivni razredi grafov — 2

Induktivni razred je sestavljen iz grafov, ki jih dobimo iz baze in končnega števila korakov — uporabe pravil za generiranje.
Lepe lastnosti pravil za generiranje:

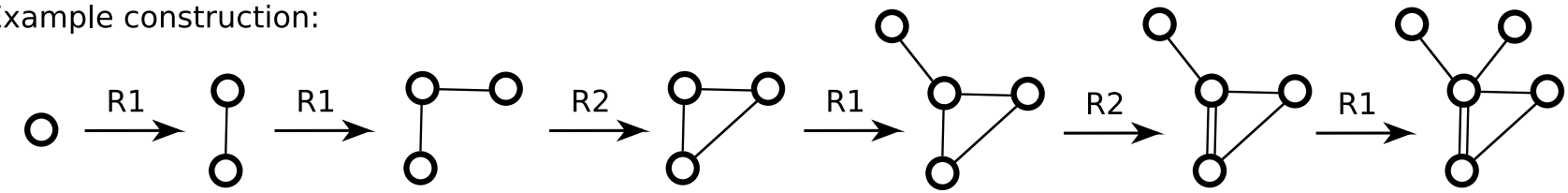
- lokalnost (levi element pravila je povezan)
- razširljivost (pravilo poveča neko lastnost grafa; npr. število vozlišč v grafu)

Primer

$\mathcal{I} = (\mathcal{B}; \mathcal{R})$, $\mathcal{B} = \{B\}$, $\mathcal{R} = \{R1, R2\}$



Example construction:



Verjetnostni ICG

Verjetnostni induktivni razred grafov

PICG (probabilistic inductive class of graphs) je definiran kot $\mathcal{P} = (\mathcal{B}; \mathcal{R}; f_B; f_R; \mathcal{F}_L)$:

1. množica $\mathcal{B}$ osnovnih grafov, baza,
2. množica $\mathcal{R}$ pravil za generiranje
3. verjetnostna porazdelitev f_B, ki določa, kako so osnovni grafi izbrani iz množice $\mathcal{B}$,
4. verjetnostna porazdelitev f_R, ki določa, kako so uporabljena pravila iz množice $\mathcal{R}$,
5. množica $|\mathcal{R}|$ verjetnostnih porazdelitev — $\mathcal{F}_L$, ki določa, kako so izbrani levi elementi vsakega pravila v množici $\mathcal{R}$.

Verjetnostni ICG

Verjetnostni induktivni razred grafov

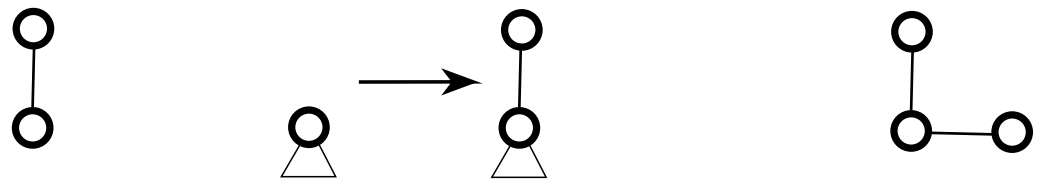
PICG (probabilistic inductive class of graphs) je definiran kot $\mathcal{P} = (\mathcal{B}; \mathcal{R}; f_B; f_R; \mathcal{F}_L)$:

1. množica $\mathcal{B}$ osnovnih grafov, baza,
2. množica $\mathcal{R}$ pravil za generiranje
3. verjetnostna porazdelitev f_B, ki določa, kako so osnovni grafi izbrani iz množice $\mathcal{B}$,
4. verjetnostna porazdelitev f_R, ki določa, kako so uporabljena pravila iz množice $\mathcal{R}$,
5. množica $|\mathcal{R}|$ verjetnostnih porazdelitev — $\mathcal{F}_L$, ki določa, kako so izbrani levi elementi vsakega pravila v množici $\mathcal{R}$.

Omejimo se na preproste definicije PICG, kjer je levi element pravil vedno možno najti.

"Preferential attachment model" kot PICG

B R after 1 step



$\mathcal{P} = (\mathcal{B}; \mathcal{R}; f_B; f_R; \mathcal{F}_L):, \mathcal{B} = \{B\}, \mathcal{R} = \{R\}$

- $f_B = 1$
- $f_R = 1$
- $f_u = \frac{deg(u)}{\sum_v deg(v)}, \qquad \forall u \in V, f_u \in \mathcal{F}_L$

Pričakovano število vozlišč v PICG

Pričakovana sprememba v številu vozlišč mora biti pozitivna (sicer graf izumre).
Verjetnost za število vozlišč (rekurzivni izraz):

$$p_t\left(N = n\right) = \sum_{R_i \in \mathcal{R}} r_i p_{t-1}\left(N = n - \Delta n_i\right)$$

- N — slučajna spremenljivka za število vozlišč
- t — časovni korak
- $r_i \equiv f_R(R_i)$ — verjetnost za izbiro pravila R_i
- Δn_i število vozlišč, ki jih pravilo R_i doda grafu

Začetna vrednost za grafe iz baze:

$$p_0\left(N = n\right) = \sum_{B_i ; |B_i| = k} f_B(B_i)$$

Pričakovano število vozlišč

Pričakovano število vozlišč v času t:

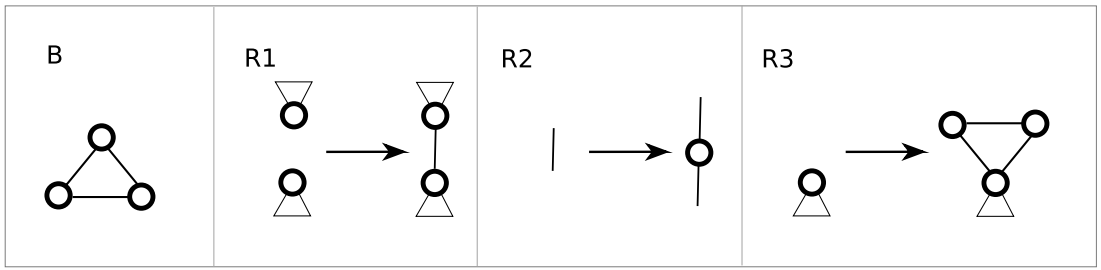
$$E_t[N] = \sum_{i=0}^{\infty} i p_t(i)$$

Da se pokazati, da v limiti $t \to \infty$:

$$\lim_{t\to\infty} \frac{E_t[N]}{t} = \sum_{R_i \in \mathcal{R}} r_i \Delta n_i$$

Podobno velja za število povezav.

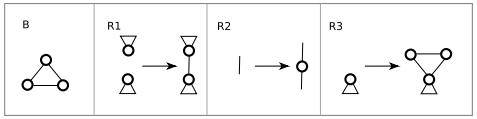
Uporaba — 2-povezavno-povezan graf



$$\mathcal{P}_{2E} = (B; \{R1, R2, R3\}; 1; f_R, \mathcal{F}_L)$$

- $f_R = \begin{pmatrix} R1 & R2 & R3 \\ q & r & s \end{pmatrix}$
- $\mathcal{F}_L = \{f_{R1}, f_{R2}, f_{R3}\}$
  - $f_{R1}(u, v) = \frac{1}{n(n-1)}$
  - $f_{R2}(e) = \frac{1}{m}$
  - $f_{R3}(u) = \frac{1}{n}$

... 2-povezavno-povezan graf



Iz rekurzivne relacije:

$$p_t(n) = qp_{t-1}(n) + rp_{t-1}(n-1) + sp_{t-1}(n-2)$$

dobimo

$$p_t(n) = \sum_{j=\lfloor \frac{n}{2} \rfloor}^{n-2} \binom{j-1}{n-j-2} \binom{t}{t-j+1} q^{t-j+1} r^{2j-n+1} s^{n-j-2}$$

Torej je pričakovana sprememba števila vozlišč:

$$\frac{E_t[N]}{t} \approx r + 2s$$

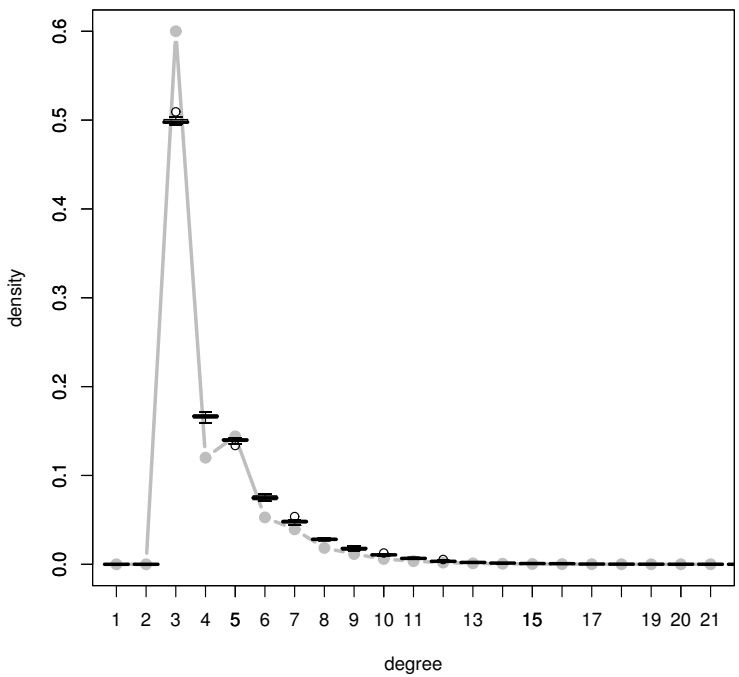
Podobno — pričakovana sprememba števila povezav:

$$\lim_{t\to\infty} \frac{E[M]}{t} = (q + r) + 3s$$

... 2-povezavno-povezan graf

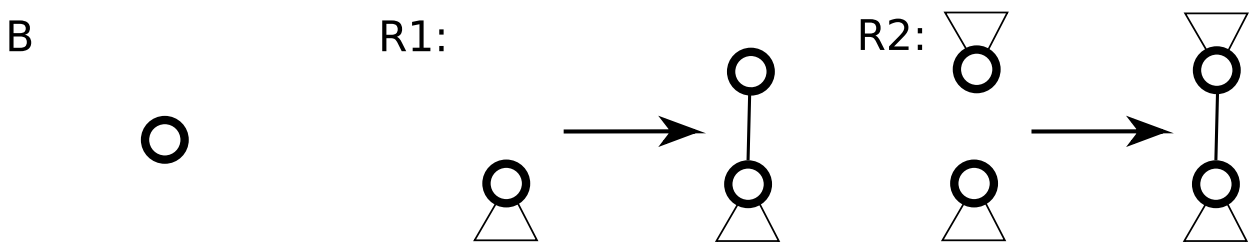
Porazdelitev stopenj (p_k — verjetnost, da ima vozlišče stopnjo k):

$$(n+r+2s)p'_k = np_k + q\,(p_{k-1} - p_k) + r\,(\delta_{k2}) + s\,(p_{k-2} + 2\delta_{k2} - p_k)$$



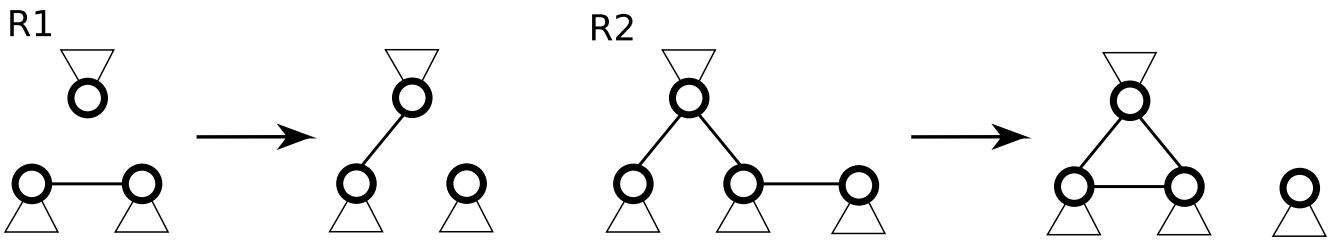
Model razširanja govoric

- baza — oseba z "informacijo"
- $R1$ — oseba, ki ve, posreduje govorico nekomu, ki tega še ne ve
- $R2$ — osebi, ki vesta za govorico, se o njej pogovorita



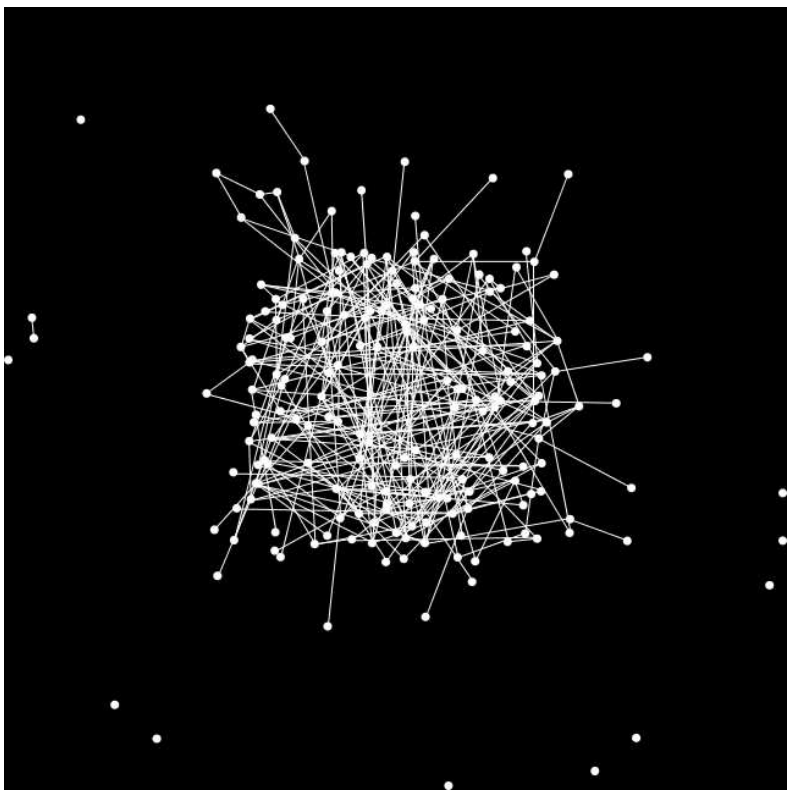
Model za spreminjanje strukture omrežja znanstev

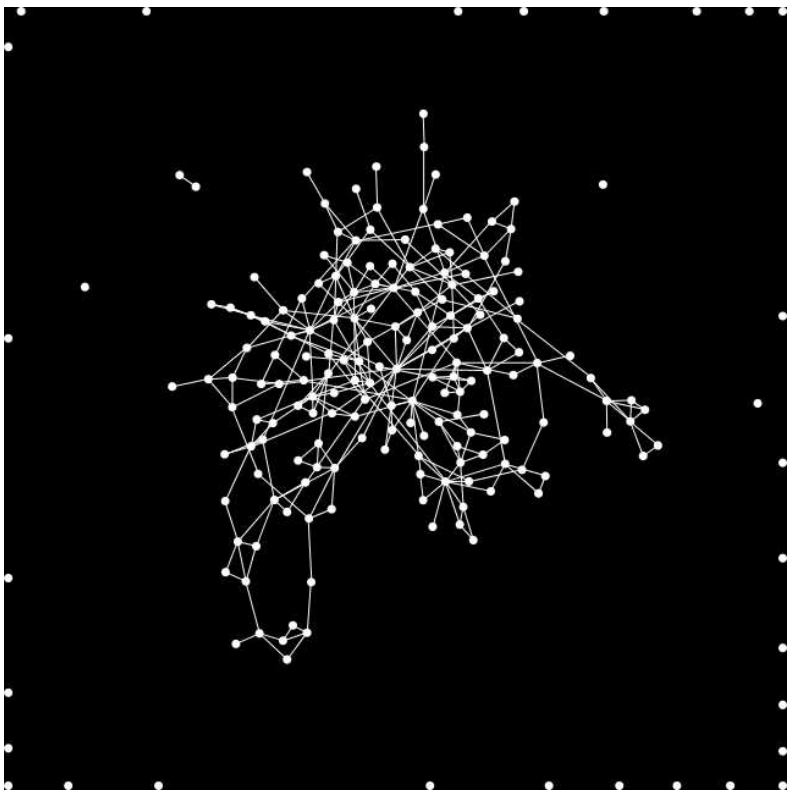
- baza — graf z n vozlišči in m povezavami (npr. ER slučajni graf)
- $R1$ — prekinitev znanstva za bolj "koristno" znanstvo (npr. isti pogledi, zanimanja)
- $R2$ — kreacija "močnejšega" odnosa s tem, da se zgradi trikotnik (npr. prijateljstvo)



Model za spreminjanje strukture omrežja znanstev

Levo: baza; ER model, 200 vozlišč in 200 povezav
Desno: graf po 1000 korakih ($p_1 = 0.28$ in $p_2 = 0.72$)





Zaključek

- PICG je orodje, ki nam omogoča formalen opis precej modelov omrežij
- vanj lahko vključimo dodajanje, brisanje, lastnosti povezav in vozlišč
- več v članku na `arXiv:math/0612778v1 [math.DS]`