Szociális hálók

Szalai Péter

April 17, 2015

Miről lesz szó?

1. Megfigyelések
   - Kis világ
   - Power-law
   - Klaszterezhetőség
2. Modellek
   - Célok
   - Erdős-Rényi
   - Watts-Strogatz
   - Barabási modell
   - Copy-modell
3. Spektrál gráf partícionálás
   - Matek
   - SVD
   - Feladat

Section 1

Megfigyelések

Subsection 1

Kis világ

Kis világ jelenség

A kisérlet:

- Milgram
- 1969
- Nebraska $\rightarrow$ Boston

Kis világ jelenség

A kisérlet:

- Milgram
- 1969
- Nebraska $\rightarrow$ Boston

Eredmény:

- 20%-nak sikerült
- Átlagos lépésszám: 6.5
- Ugyanez facebookon:
  - 2008: 5.28
  - 2011: 4.74

Kis világ jelenség

A kisérlet:

- Milgram
- 1969
- Nebraska → Boston

Eredmény:

- 20%-nak sikerült
- Átlagos lépésszám: 6.5
- Ugyanez facebookon:
  - 2008: 5.28
  - 2011: 4.74

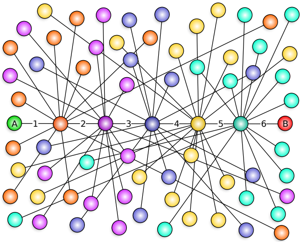


Figure: Six degrees of separation

Subsection 2

Power-law

Power-law eloszlás

Power-law eloszlás

Az X valváltozó eloszlása *power-law* eloszlás, ha:

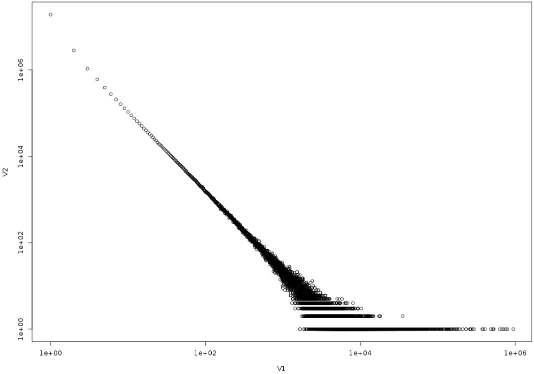
$$Pr(X = x) = Cx^{-\alpha}$$

Power-law eloszlás

Power-law eloszlás

Az X valváltozó eloszlása *power-law* eloszlás, ha:

$$Pr(X = x) = Cx^{-\alpha}$$



Power-law eloszlás

Power-law eloszlás

Az X valváltozó eloszlása *power-law* eloszlás, ha:

$$p(x) = Pr(X = x) = Cx^{-\alpha}$$

Vegyük a logaritmusát:

$log(p(x)) = log(C) - \alpha log(x) \implies log(y) = -\alpha log(x) + C'$
ami log-log skálán:

$$y' = -\alpha x' + C'$$

Power-law eloszlás

Power-law eloszlás

Az X valváltozó eloszlása *power-law* eloszlás, ha:

$$Pr(X = x) = Cx^{-\alpha}$$

- Empirikus adatban a leggyakoribb eloszlás.
- *"Örökifjú"*, $f(cx) = g(c)f(x)$ egyetlen megoldása.
- Először *Nature* biológiai vonatkozás
- Példák:
  - Vagyon eloszlás (Pareto 80-20)
  - Hashtagek népszerűség eloszlása
  - Szavak előfordulása
  - Ismeretségi gráfok fokszám eloszlás!!!

Subsection 3

Klaszterezhetőség

Klaszterezhetőség

- Csoportulások
- Csoportok közotti kapcsolatok = Egyének közötti kapcsolatok
- Skálafuggetlen tulajdonság
- Klaszterezés problémája
- Később

Klaszterezhetőség

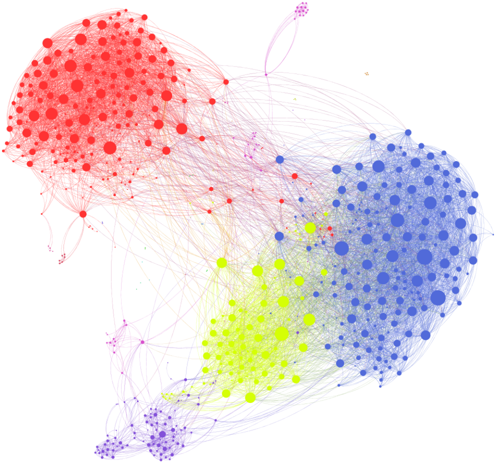


Figure: Facebook klaszterezve

Section 2

Modellek

Subsection 1

Célok

Hálózatmodellek

Miért?

- Jelenségek megértése *(Mi történik facebookon?)*
- Szimuláció
- Mérések

Validálás:

- Fokszám eloszlása power-law?
- Kis világ tulajdonság teljesül?
- Klaszterezhetőség?

Néhány hálózatmodell:

- Erdős-Rényi
- Watts-Strogatz
- Barabási-Albert
- Broder

Erdős-Rényi modell

$G(n, p) = (V, E)$, ahol $V = [1, \dots n]$ és $Pr((i, j) \in E)) = p$

Más megfogalmazás:

$G(N, M) = (V, E)$, ahol $|V| = N$ és $|E| = M$.

Kapcsolat a két modell között:

$$M = \binom{N}{2} p$$

Erdős-Rényi modell

$G(n,p) = (V, E)$, ahol $V = [1, \dots n]$ és $Pr((i,j) \in E)) = p$

Más megfogalmazás:

$G(N,M) = (V, E)$, ahol $|V| = N$ és $|E| = M$.

Kapcsolat a két modell között:

$$M = \binom{N}{2} p$$

Erdős vizualizáció

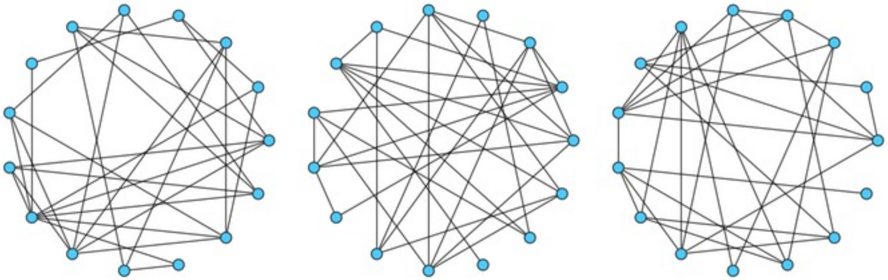


Figure: Erdő gráf azonos N, M esetén

ER gráf tulajdonságai

Néhány tulajdonság:

- Átlag fokszám $\frac{2|E|}{n} = \frac{2\binom{n}{2}p}{n} = (n-1)p$
- Fokszám-eloszlás: $Pr(d(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$
- Poisson $Pr(d(v) = k) = \frac{d(v)^k e^{-z}}{k!}$
- Összefüggőség:
  - $np < 1$ legnagyobb összefüggő komponens $O(log(n))$
  - $np = 1$ legnagyobb összefüggő komponens $O(n^{\frac{2}{3}})$
  - $np > 1$ legnagyobb összefüggő komponens $O(n)$ és a második $O(log(n))$
- Átmérő: $\frac{log(n)}{log(p(n-1))}$
- $Pr(G) = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|}$

Watts-Strogatz

Kifejezetten a kis világ tulajdonság modellezésére találták ki.

$G(n, k, p)$

1. Körön elhelyezünk n pontot.
2. A k távolságra lévőket összekötjük.
3. Minden élet p valószínűséggel átkötünk véletlen helyre.

Watts-Strogatz

Kifejezetten a kis világ tulajdonság modellezésére találták ki.

$G(n, k, p)$

1. Körön elhelyezünk n pontot.
2. A k távolságra lévőket összekötjük.
3. Minden élet p valószínűséggel átkötünk véletlen helyre.

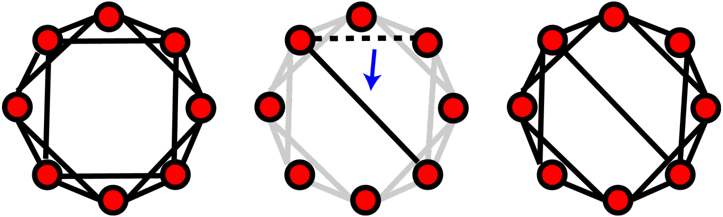


Figure: n=8,k=2

Watts-Strogatz

Kifejezetten a kis világ tulajdonság modellezésére találták ki.

$G(n, k, p)$

1. Körön elhelyezünk n pontot.
2. A k távolságra lévőket összekötjük.
3. Minden élet p valószínűséggel átkötünk véletlen helyre.

Tulajdonságok:

- $G_{WS}(n, k, p)$ közel azonos a $C_n + 2C_{\frac{n}{2}} + \cdots + kC_{\frac{n}{k}} + G_{ER}(n, p)$ gráffal.
- Poisson $+$ konstans fokszámeloszlás.e
- Részgráfok ritkák a függetlenség miatt.
- Grid modell $\left(\frac{[d(u,v)^{-r}]}{\sum_{v;v\neq u}[d(u,v)^{-r}]}\right)$.Minden decentralizált algoritmus $\Omega(n^{\frac{r-2}{r-1}})$ lépést használ.

Barabási-Albert modell

1. Kiindulunk egy G_0 gráfból m_0 éllel.
2. Felveszünk egy új pontot.
3. Bekötjük m helyre.
4. Bekötés valószínűsége: $p(v) = \frac{d(v,t)}{\sum_{w \in V} d(w,t)}$.

Tulajdonságok:

- Fokszámeloszlás: hatvány.
- Kis világ tulajdonság: teljesül
- Klaszterezhetőség: nem teljesül.

Broder

Másolásos modell

1. Kiindulunk egy G_0 gráfból m_0 éllel.
2. Felveszünk egy új pontot.
3. Válsztunk egy már meglévő pontot.
4. Ha d élet kell behúzni akkor α valószínűséggel másolunk egy élet és $(1 - \alpha)$-val egyenletesen választunk egy csúcsot.

Tulajdonságok:

- Fokszámeloszlás: hatvány.
- Kis világ tulajdonság: teljesül
- Klaszterezhetőség: teljesül.

Section 3

Spektrál gráf partícionálás

Négyzetes hiba *(MSE)* Mátrixra

$\|A - B\|_?$

Négyzetes hiba *(MSE)* Mátrixra

$\|A - B\|_?$
$(A - B)^2$ nem egy szám $\implies$ nem jó.

Négyzetes hiba *(MSE)* Mátrixra

$\|A - B\|_?$
$(A - B)^2$ nem egy szám $\implies$ nem jó.
Megoldás:

Frobénius norma:

$$\|A\| = \sum_{i,j} a_{ij}^2$$

MSE: $\|A - B\|_F^2 = \sum_{i,j}(a_{ij} - b_{ij})^2$

SVD - *Singular value decomposition*

Tetszőleges A $m \times n$-es mátrixhoz létezik $Q \in \mathbb{R}^{m \times n}$, $D \in \mathbb{R}^{n \times n}$ és $P \in \mathbb{R}^{n \times n}$ mátrixok, hogy

$$A = Q \times D \times P,$$

ahol:

- Q és P ortonormált.
- D pedig diagonális.

Ekkor:

$$a_{ij} = \sum_k q_{ik} d_k p_{jk}$$

Továbbá: $\|A\|_F = \sum_{i,j} a_{ij}^2 = \sum_i d_i^2$

Legjobb k-rangú közelítése egy mátrixnak:

- Vegyük a k legnagyobb szinguláris értéket.
- A többit töröljük (0-ra cseréljük) és a hozzá tartozó vektorokat is.

Miért szinguláris?

$$AA^T p_i = P^T D^T Q^T Q D P p_i = P^T D^2 P(0, \ldots, 1, \ldots, 0)^T = d_i^2 v_i$$

- d_i-k a sajátértékek gyökei.
- P oszlopai az AA^T sajátvektorai.

Mire jó az SVD?

- Mult óra vége.
- Ajánló rendszerek (Robi)
- Spektrál felbontás

Subsection 3

Feladat

Spektrál klaszterezés hasonlósági gráfokra

Legyen $G(V, E)$ egy gráf és X_G a gráfot reprezentáló mátrix. Preferenciák gráf klaszterezésnél:

- Hasonlók egy klaszterbe kerüljenek.
- A különböző klaszterekbe kerülők ne hasonlítsanak.

Tehát a célunk:

1. Maximalizálni az csoporton belüli össz súlyt.
2. Minimalizálni a csoportok közöttit.

Legyen a két klaszter G_A és G_B és A, B hozzájuk tartozó ponthalmaz.

$$cut(A, B) = \sum_{i \in A, j \in B} x_{ij}$$

Spektrál klaszterezés hasonlósági gráfokra

Legyen $G(V,E)$ egy gráf és X_G a gráfot reprezentáló mátrix.
Preferenciák gráf klaszterezésnél:

- Hasonlók egy klaszterbe kerüljenek.
- A különböző klaszterekbe kerülők ne hasonlítsanak.

Tehát a célunk:

1. Maximalizálni az csoporton belüli össz súlyt.
2. Minimalizálni a csoportok közöttit.

Legyen a két klaszter G_A és G_B és A, B hozzájuk tartozó ponthalmaz.

$$cut(A,B) = \sum_{i \in A, j \in B} x_{ij}$$

$$min\ cut(A,B)$$

Spektrál klaszterezés hasonlósági gráfokra

Legyen $G(V, E)$ egy gráf és X_G a gráfot reprezentáló mátrix.
Preferenciák gráf klaszterezésnél:

- Hasonlók egy klaszterbe kerüljenek.
- A különböző klaszterekbe kerülők ne hasonlítsanak.

Tehát a célunk:

1. Maximalizálni az csoporton belüli össz súlyt.
2. Minimalizálni a csoportok közöttit.

Legyen a két klaszter G_A és G_B és A, B hozzájuk tartozó ponthalmaz.

$$cut(A, B) = \sum_{i \in A, j \in B} x_{ij}$$

$$min\ cut(A, B)$$

Nem elég. Lehet, hogy elfajuló.

Spektrál klaszterezés hasonlósági gráfokra

Megoldás

$$min\left(\frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}\right),$$

ahol $vol(A)$ az A-ból induló össz élsúly.

Laplace mátrix

Legyen $D(i,i) = \sum_j x_{ij}$, fokszám mátrix. Ekkor

$$L = D - X,$$

a *Laplace* mátrix.

Számolás

Legyen f egy n dimenziós vektor *(csúcsok száma)*. Ekkor:

$$f^T Lf = f^T Df - f^T Xf = \sum_i d_i f_i^2 - \sum_{i,j} f_i f_j w_{ij} =$$

$$\frac{1}{2}\left(\sum_i \left(\sum_j x_{ij}\right) f_i^2 - 2\sum_{ij} f_i f_j x_{ij} + \sum_j \left(\sum_i x_{ij}\right) f_j^2\right) =$$

$\frac{1}{2}\sum_{ij} x_{ij}(f_i - f_j)^2$

Miért jó ez?

$\frac{1}{2}\sum_{ij} x_{ij}(f_i - f_j)^2$

Jelentése $f \in \{-1, 1\}^n$ esetén?

$$min_p \left(\sum_{i,j} x_{ij}(p_i - p_j)^2 \right) = min_p(p^T L p)$$

- $p = (1\ 1\ \ldots 1)^T$ a minimumot adja.
- Legyen plusz feltétel $\sum_i p_i = 0$, jelentése, hogy $|A| = |B|$.

Gyakorlatban

Normalizáljuk L-t:

$$L' = D^{0.5}(D - A)D^{0.5}$$

A felbontás második sajátvektora a jó klasszifikáló.

Más megoldás

Klikk perkoláció (második óra).

NP-nehéz probléma.

Miről volt szó?

1. Megfigyelések
   - Kis világ
   - Power-law
   - Klaszterezhetőség
2. Modellek
   - Célok
   - Erdős-Rényi
   - Watts-Strogatz
   - Barabási modell
   - Copy-modell
3. Spektrál gráf partícionálás
   - Matek
   - SVD
   - Feladat