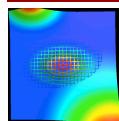


Nagy adathalmazok labor

2017-2018 őszi félév

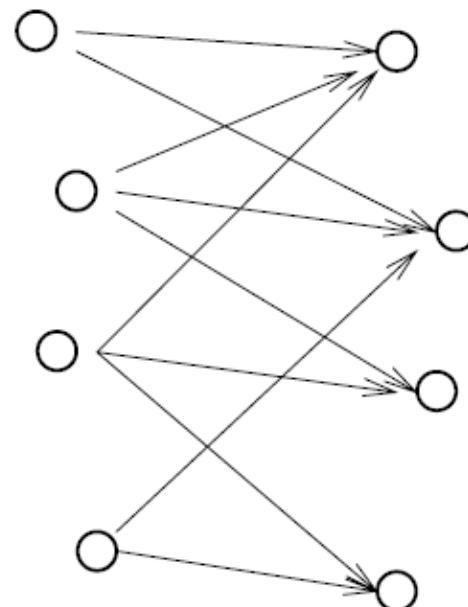
2017.11.15-21

1. Link analízis: HITS, PR
2. Hálózatok: ER, BA, WS, Broder
3. RF, AdaBoost,GBT

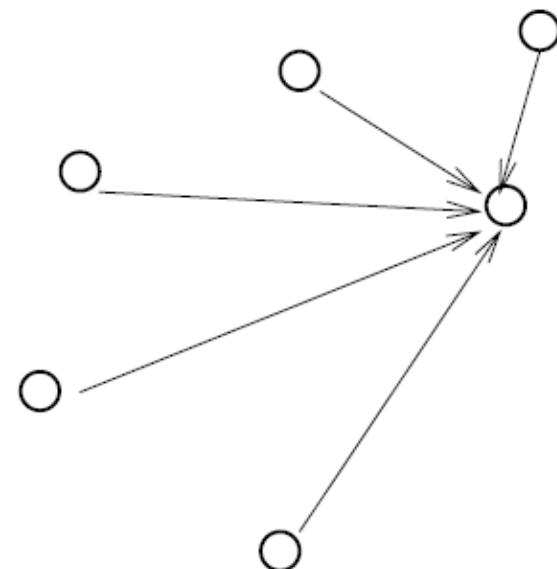


Web gráf: HITS bevezetés

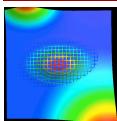
Hyperlink-Induced Topic Search (HITS)



hubs



authorities



Web gráf: HITS

Hyperlink-Induced Topic Search (HITS) Kleinberg '98

1. **Hubs:** gyűjtőoldalak, azon oldalak melyek jó authority oldalakra mutatnak
2. **Authorities:** maguk a releváns oldalak
Azok az oldalak akikre jó hub-ok mutatnak

A célunk meghatározni a két csoportot.

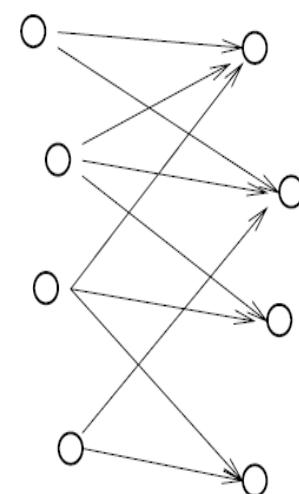
Minden oldalhoz hozzárendelünk egy nem negatív authority (x) és hub (y) értéket:

$$I(G, x, y): x^p \leftarrow \sum_{q:(q, p) \in E} y^q$$

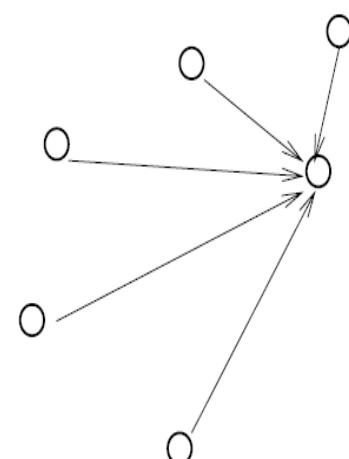
$$O(G, x, y): y^p \leftarrow \sum_{q:(p, q) \in E} x^q$$

$$\sum_p (x^p)^2 = 1$$

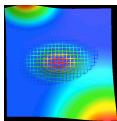
$$\sum_p (y^p)^2 = 1$$



hubs



authorities



Web gráf: HITS

HITS(G, k, q)

G : oldalak és élek halmaza, melyek a q keresés mag vagy bővítési halmazába tartoznak

k : konstans

z legyen egy n dimenziós valós vektor $(1; 1; 1; \dots; 1)$

Legyen $x_0 := z$:

Legyen $y_0 := z$:

for $i = 1, 2 \dots k$

$O(G, x_{i-1}; y_{i-1}) \rightarrow$ megkapjuk az új x' súlyokat

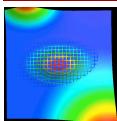
$I(G, x_{i-1}; y_{i-1}) \rightarrow$ megkapjuk az új y' súlyokat

Normalizáljuk x' -t, megkapjuk az új x -et

Normalizáljuk x' -t, megkapjuk az új x -et

Adjuk vissza $(x_k; y_k)$ -t.

Állítás: (x_1, x_2, x_3, \dots) és (y_1, y_2, y_3, \dots) sorozatok konvergálnak.



HITS

Bizonyítás: Legyen A az adjacencia mátrix a kiválasztott részgráfpon (“focus graph”, x:auth, y:hub)

$$\begin{aligned}x^{(k+1)} &= y^{(k)} A \\y^{(k+1)} &= x^{(k+1)} A^T\end{aligned}$$

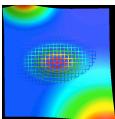
bontsuk ki:

$$x^{(k+1)} = x^{(1)} (A^T A)^k = x^{(1)} U W U^T$$

$$y^{(k+1)} = y^{(1)} (A A^T)^k = y^{(1)} V W V^T$$

ahol W diagonális.

Állítás: a normalizált $x^{(k)}$ és $y^{(k)}$ sorozat konvergál
(kezdővektor $x=y=(1, 1, \dots, 1)$)



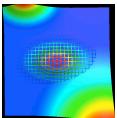
HITS

Állítás átfogalmazva :

$$(AA^T)^j y^{(1)} / \|(AA^T)^j y^{(1)}\| \text{ sorozat konvergál}$$

Segédtétel: ha egy $n \times n$ -es M mátrix pozitív szemidefinit szimmetrikus mátrix, melynek a sajátértékei $\lambda_1 > \lambda_2 \geq \lambda_3 \dots \geq \lambda_k \geq 0$ ($k < n$), akkor minden d dimenziós valós vektorra igaz, hogy kifejezhető $v = \sum_{i=1..k} a_i \omega^{(i)}$ ahol minden i -re $\|\omega^{(i)}\|=1$ és $\omega^{(i)\top} \omega^{(j)} = 0$, ha $i \neq j$.

Miután $\omega^{(i)}$ az i -dik sajátvektor: $M \omega^{(i)} = \lambda_i \omega^{(i)}$

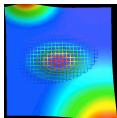


HITS

Miután $\mathbf{A}\mathbf{A}^T = \mathbf{M}$ pozitív szemidefinit szimmetrikus, így:

$$\frac{\mathbf{M}^j \mathbf{v}}{\|\mathbf{M}^j \mathbf{v}\|} = \frac{\sum_{i=1}^k \alpha_i \mathbf{M}^j \mathbf{w}^{(i)}}{\left\| \sum_{i=1}^k \alpha_i \mathbf{M}^j \mathbf{w}^{(i)} \right\|} = \frac{\sum_{i=1}^k \alpha_i \lambda_i^j \mathbf{w}^{(i)}}{\sqrt{\sum_{i=1}^k (\alpha_i \lambda_i^j)^2}}$$

$$\frac{\alpha_1 \lambda_1^j \mathbf{w}^{(1)} + \sum_{i=2}^k \alpha_i \lambda_i^j \mathbf{w}^{(i)}}{\sqrt{(\alpha_1 \lambda_1^j)^2 + \sum_{i=1}^k (\alpha_i \lambda_i^j)^2}} \cdot \frac{\frac{1}{\lambda_1^j}}{\frac{1}{\lambda_1^j}} = \frac{\alpha_1 \mathbf{w}^{(1)} + \sum_{i=2}^k \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^j \mathbf{w}^{(i)}}{\sqrt{\alpha_1^2 + \sum_{i=1}^k \left(\alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^j\right)^2}} \rightarrow \mathbf{w}^{(1)}$$



Random surfer “Stochasztikus szörfölő”

Feltételezés: véletlen séta az éleken.

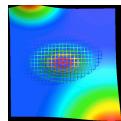
Minden pillanatban uniform módon egy hyperlinken tovább lépünk.

$$\Pr(i | j) = 1/d(j)$$

Vegyük az adjacencia mátrixát a gráfunknak.

Cseréljük ki az értekeket az átmenet valószínűségekre: M

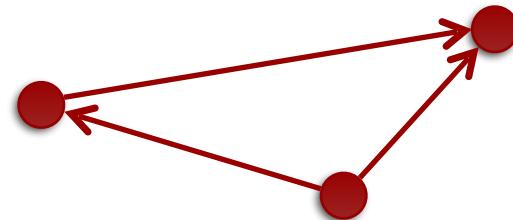
Sorösszeg?



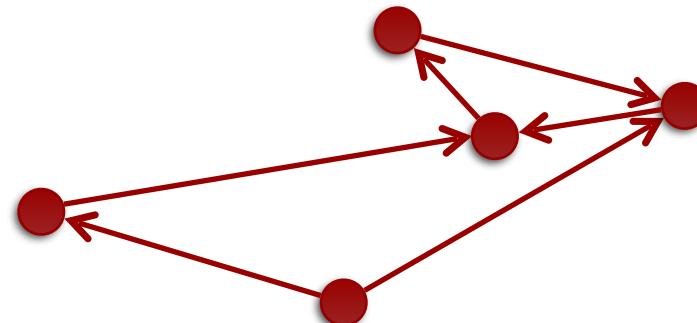
Random surfer “Stochasztikus szörfölő”

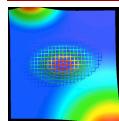
Milyen esetekben nem 1 a sorösszeg?

Zsákutca, forrás:



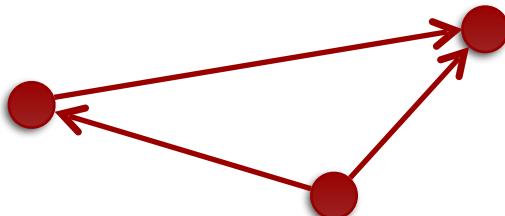
Pókháló:



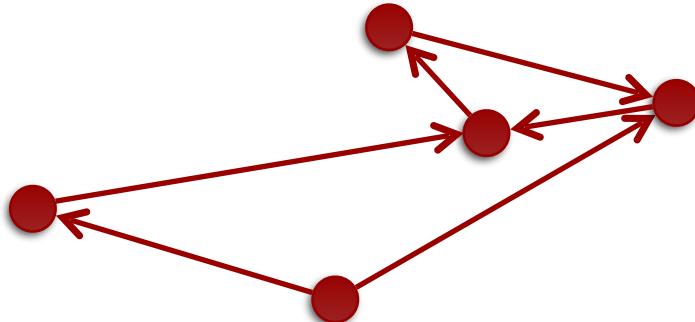


Random surfer “Stochasztikus szörfölő”

Zsákutca, forrás:



Pókháló:



Ergódikus:

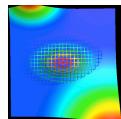
- erősen összefüggő
- aperódikus

Ha egy MC ergódikus
(Perron-Frobenius):

Létezik stacionárius eloszlás:

$$\pi^T M = \pi^T$$

Sőt: a legnagyobb sajátérték egyszeres!



Random surfer

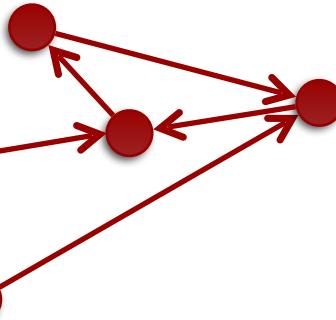
“Stochasztikus szörfölő”

Zsákutca, forrás:



Hogyan lehetjük könnyen ergódikussá a meglévő hyperlink gráfunkat?

Pókháló:



Ergódikus:

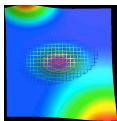
- erősen összefüggő
- aperódikus

Tényleg uniform a teleportáció?

Letezik stacionarius eloszlás:

$$\pi^T M = \pi^T$$

Sőt: a legnagyobb sajátérték egyszeres!



PageRank

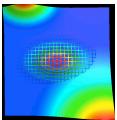
Larry Page , Sergey Brin , Rajeev Motwani és Terry Winograd 1998-as cikke.

Adott egy hyperlinkekkel összekötött dokumentumokból álló hálózat, mint pl. a WWW.

Általános szöveg alapú keresés során a releváns dokumentumok (a keresett szavakat tartalmazó dokumentumok) sorrendjét a legjobb illeszkedés alapján határozzák meg. (amelyik dokumentum leginkább illeszkedik a kérdésre az kerül előre)

Az algoritmus sok hibás vagy igazából nem releváns találatot hátra sorol, felhasználva a hivatkozások hálózatát.

Az elv egyszerű: amelyik oldalra többen és/vagy fontosabbak hivatkoznak, fontosabb mint amire kevesebben és/vagy kevésbe fontosak.

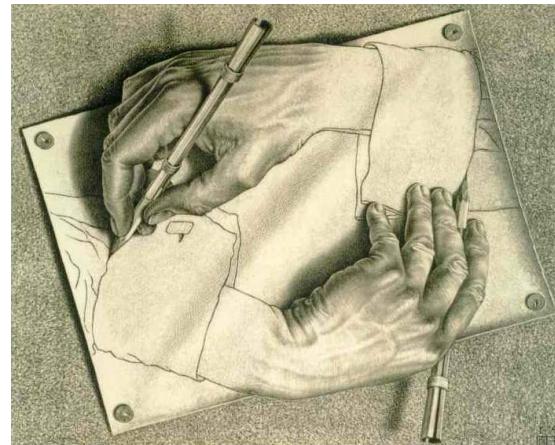


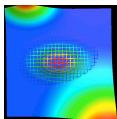
PageRank

Egy adott A dokumentum (oldal) PageRank értéke $PR(A)$:

$$PR(A) = \sum_{B \in I_A} \frac{PR(B)}{L(B)}$$

I jelöli az egy elemre hivatkozó oldalak halmazát, L(B) a B oldal kimeneti linkjeinek száma, PR(B) pedig B PageRank értéke.



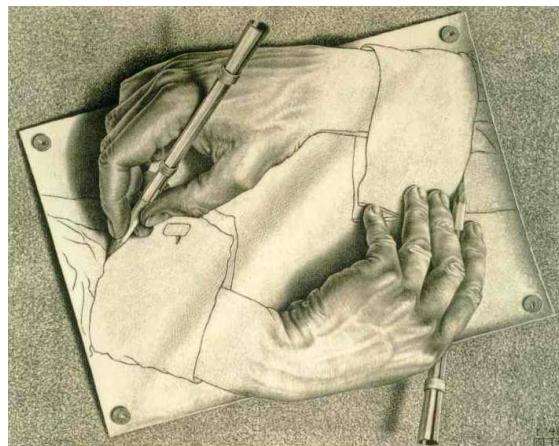


PageRank

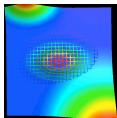
Egy adott A dokumentum (oldal) PageRank értéke $PR(A)$:

$$PR(A) = \sum_{B \in I_A} \frac{PR(B)}{L(B)}$$

I jelöli az egy elemre hivatkozó oldalak halmazát, L(B) a B oldal kimeneti linkjeinek száma, PR(B) pedig B PageRank értéke.



Mi a kapcsolat a sztochasztikus szörfölő modellel?



PageRank

“Random surfer” modell: folyamatosan csökkenő aktivitás → teleportation

$$PR(A) = 1 - d + d \sum_{B \in I_A} \frac{PR(B)}{L(B)}$$

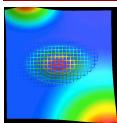
vagy

$$PR(A) = \frac{1-d}{N} + d \sum_{B \in I_A} \frac{PR(B)}{L(B)}$$

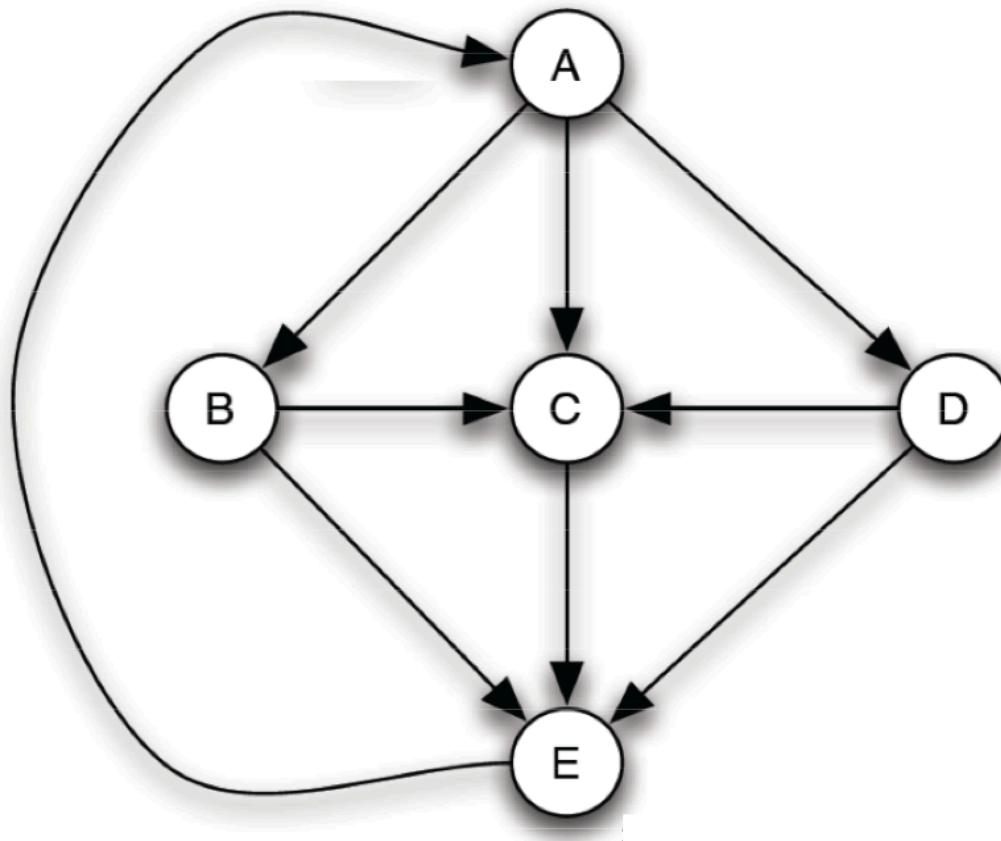
ahol N az összes oldal száma.

Az algoritmus lehetőséget ad linkfarmok vagy mesterséges (fizetett) hivatkozások alapján manipulálásra.

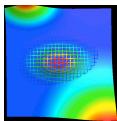
Épp ezért egy jó kereső emellett nem veszi figyelembe a már detektáltan ranking módosító honlapokat illetve linkekét -> web SPAM!



Page Rank



Órai feladat 1:
Mennyi lesz a PR értéke az
A,B,C,D,E pontoknak?

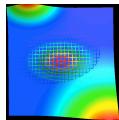


HITS vs. PageRank

	HITS	PageRank
Gráf	Keresésenként más	fix
Mértékek	Hub és auth. értékek	PR értékek

Ezeken kívül fontos különbség, hogy a PageRank a teljes gráf struktúráját próbálja feltérképezni, a HITS eredeti célja egy topik alapján kiválasztott részgráf pontjait rendezи két csoportba.

Mindkét módszer más más esetekben hatékony, de a HITS-et már maga a számításigénye miatt is ritkán használják a gyakorlatban.



Generatív hálózat modellek

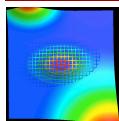
Jelenségek terjedésére nagy hálózatokban általában mint véletlen folyamatokra gondolunk

pl. járványok vagy szociális hálózatok

Közös bennük, hogy a hálózat struktúrája komplex

Alapvető cél, hogy generatív modellekkel szimuláljunk valós hálózatokat

Mit lehet tudni ezekről a hálózatokról? Mit lehet mérni?



Véletlen gráf avagy Erdős-Rényi

Forrás: Benczúr András



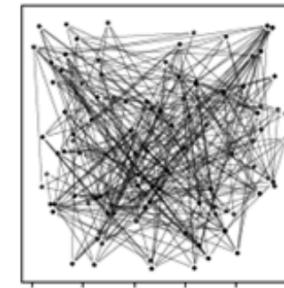
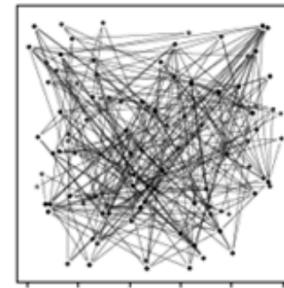
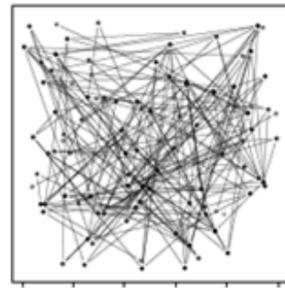
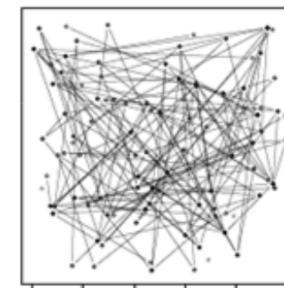
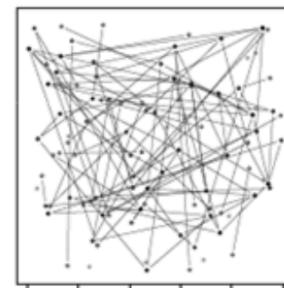
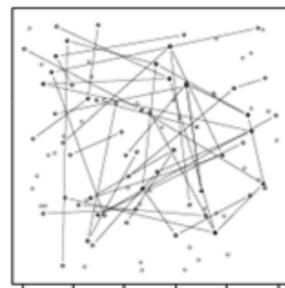
Erdős

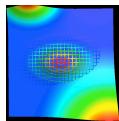


Rényi



Erdős–Rényi





Véletlen gráf avagy Erdős-Rényi

$G(n,p)$: n csúcs mellett p valószínűsséggel (függetlenül!) behúzunk egy élt

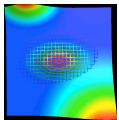
Sokat tudunk róla (ER 1959, Bollobás et al 2002):

Élek várható száma

$$|E| = \binom{n}{2} p = pn(n-1)/2$$

Átlagos fokszám

$$z = \frac{2|E|}{n} = \frac{2\binom{n}{2}p}{n} = (n-1)p$$



Véletlen gráf avagy Erdős-Rényi

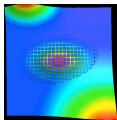
Fokszám eloszás Binomiális:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Ami ha n nagy Poisson (z legyen az átlagos fokszám)

$$P(k) = \frac{z^k e^{-z}}{k!}$$

Első kérdés: milyen az ismert hálózatainkban a fokszámeloszlás?



Véletlen gráf avagy Erdős-Rényi

Összefüggőség: ha ... , akkor majdnem biztosan

$np < 1$: legnagyobb összefüggő komponens $O(\log(n))$ nagyságrendű

$np = 1$: legnagyobb összefüggő komponens $O(n^{(2/3)})$ nagyságrendű

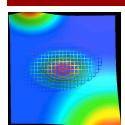
$np > 1$ konstans: $O(n)$ a legnagyobb összefüggő komponens és a második legnagyobb legfeljebb $O(\log(n))$ nagyságrendű

$np < (1-e) \log(n)$: van izolált csúcs

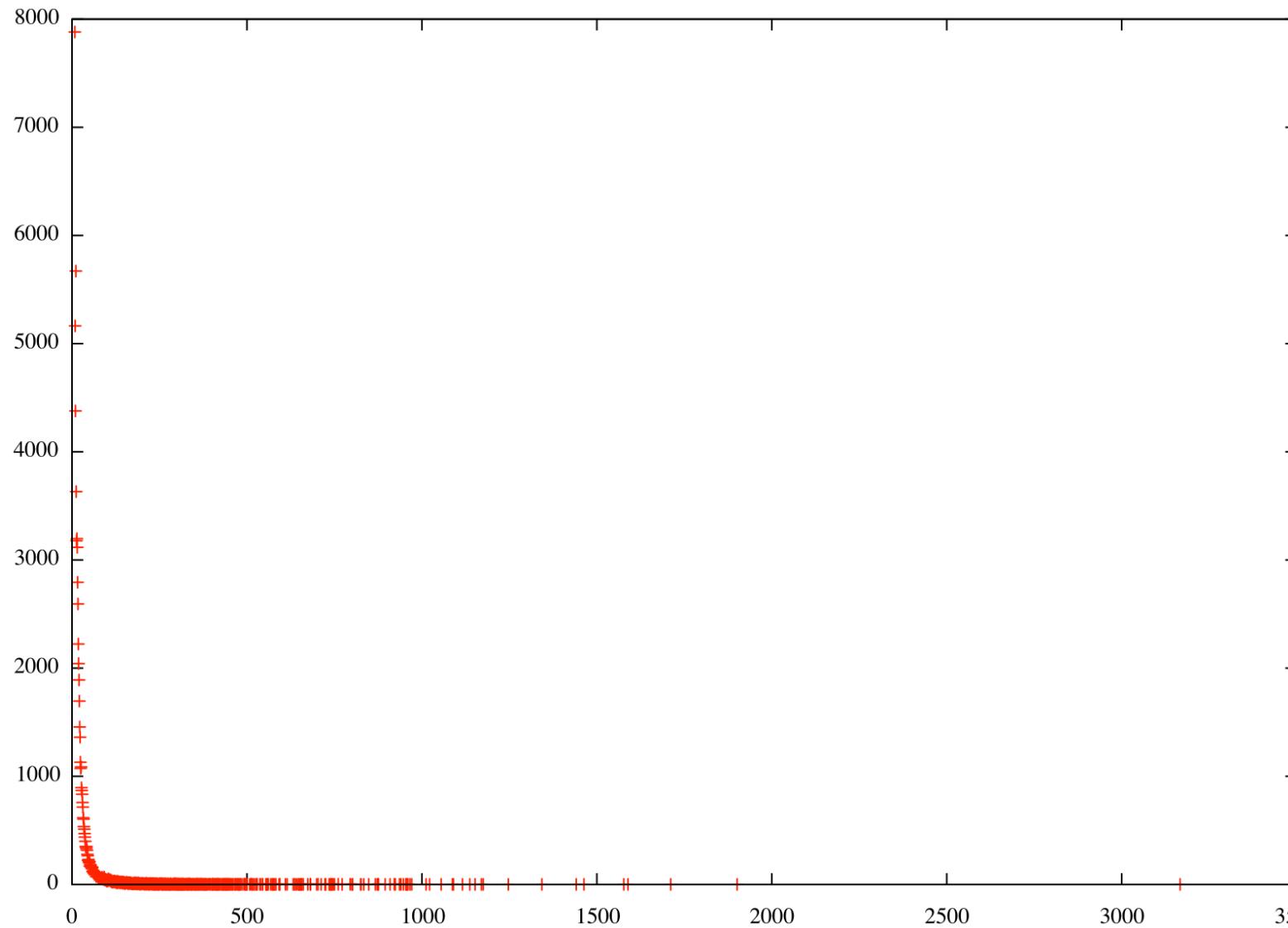
$np > (1+e) \log(n)$: összefüggő

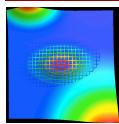
Átmérő:

$$\log(n)/\log(p(n-1))$$

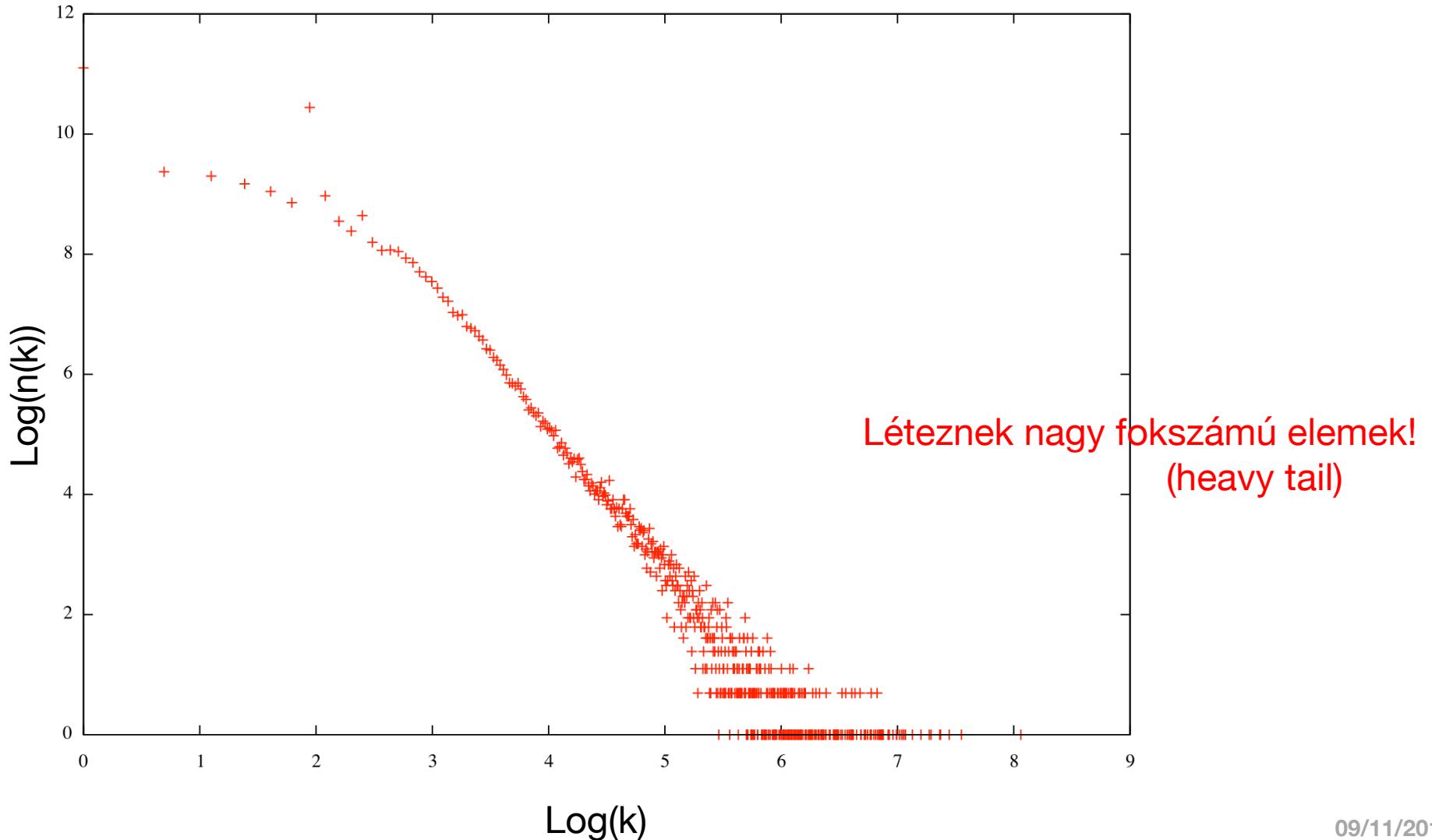


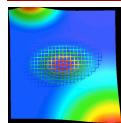
Wiki graph ($z=11.1667$)



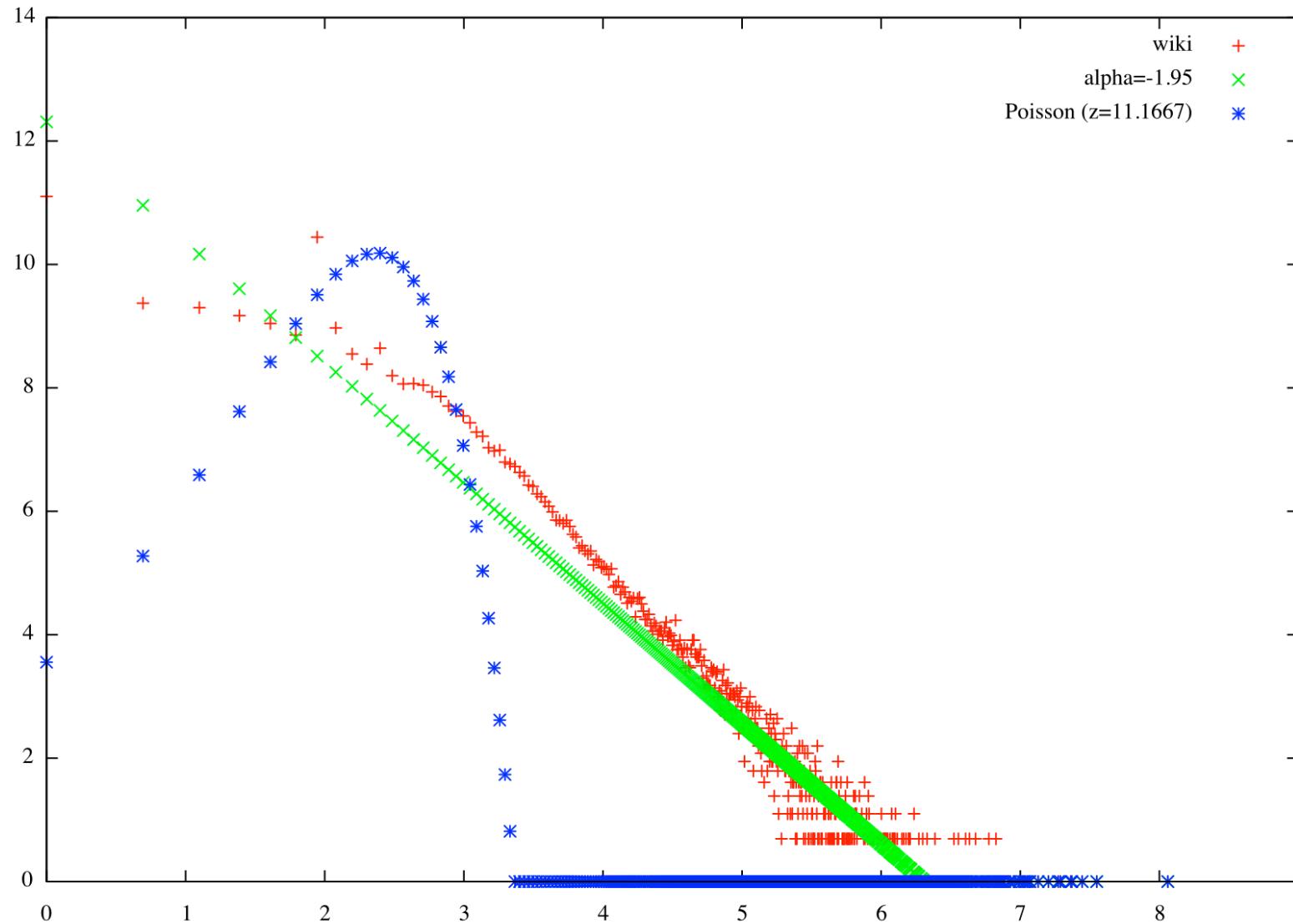


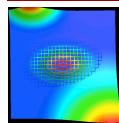
Sokat nem mondott... log-log skála



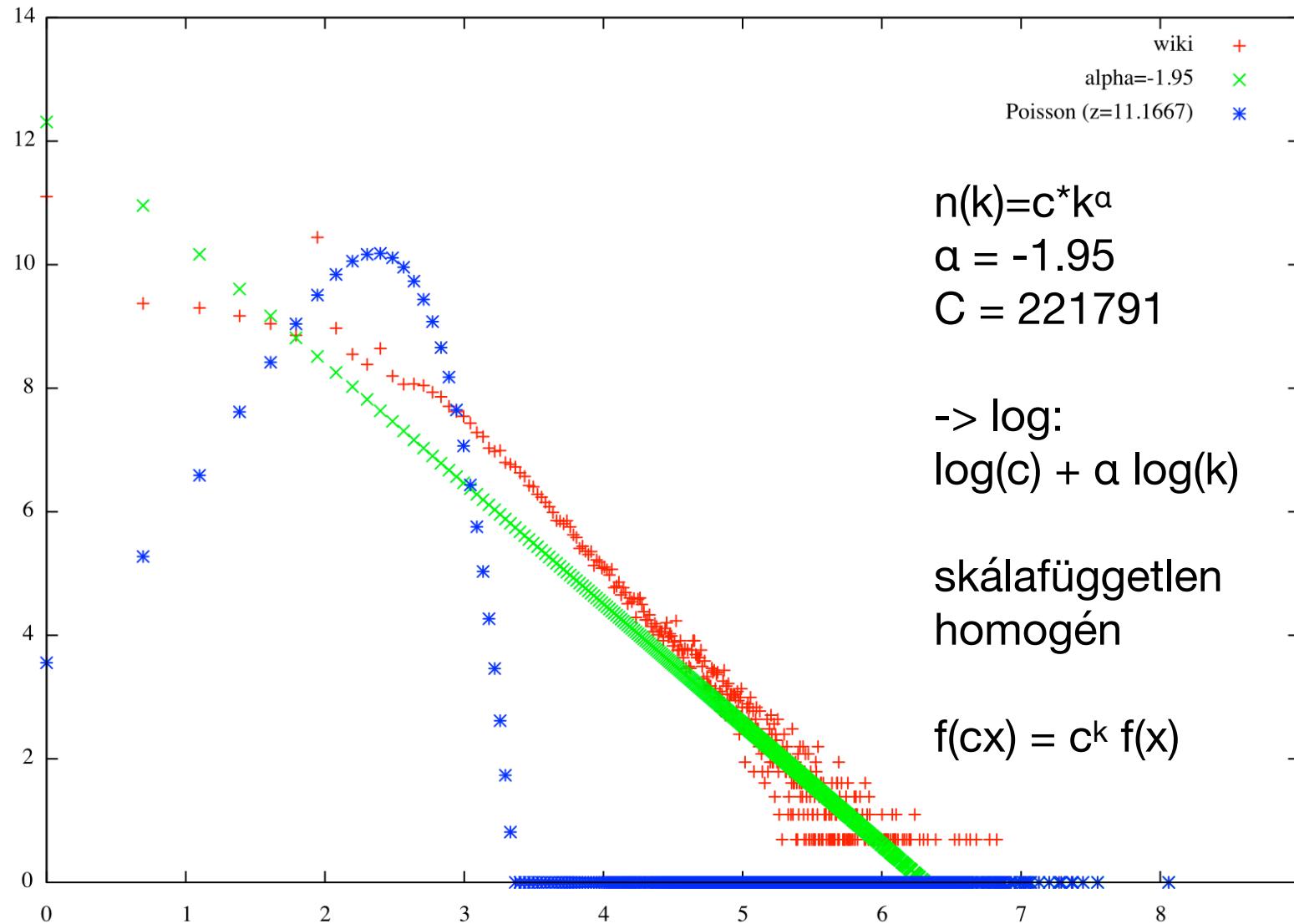


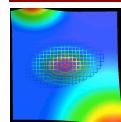
Wiki graph is hatványeloszlás (vagy?)!



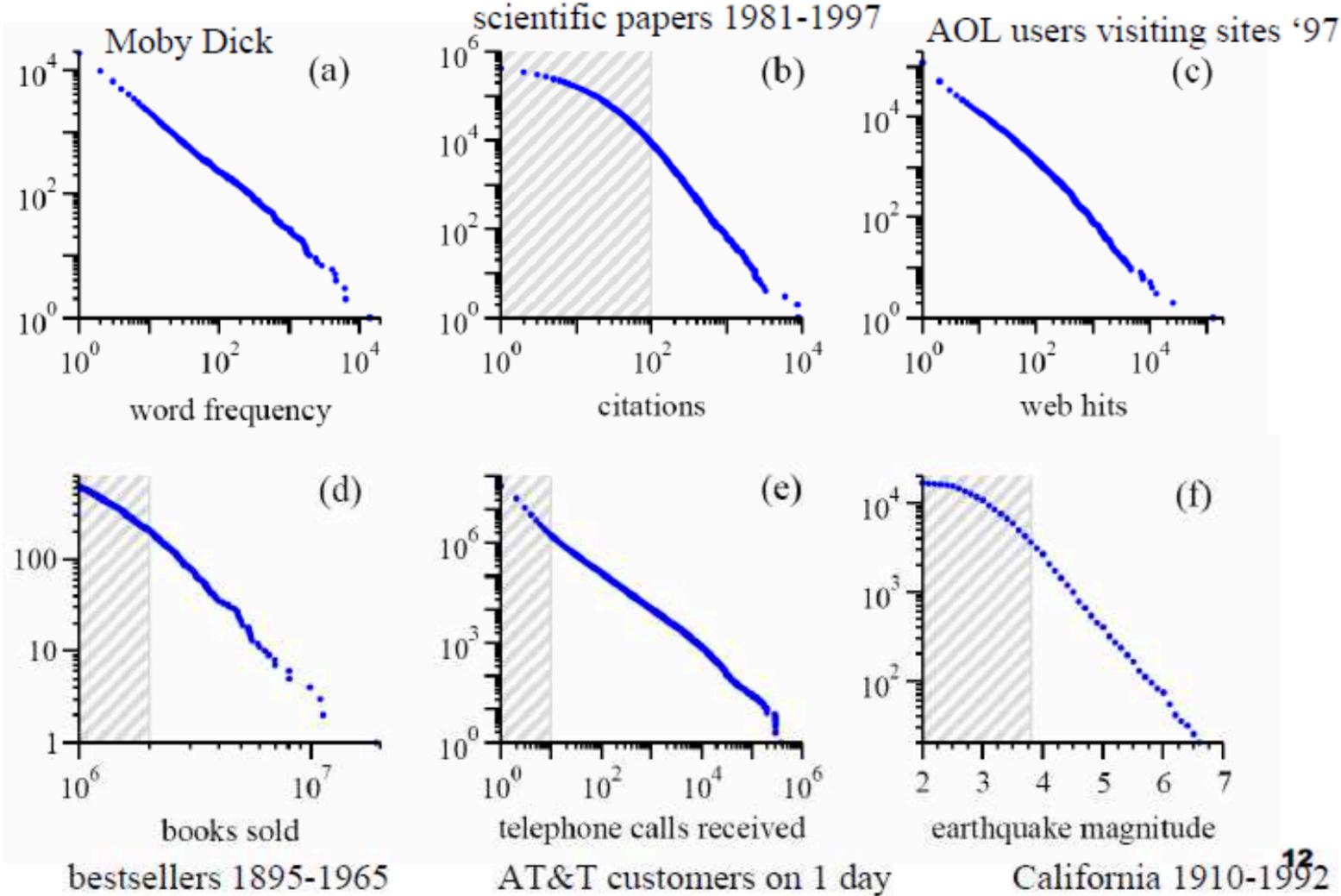


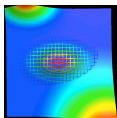
Wiki graph is hatványeloszlás (vagy)!





Példa hatványeloszlásokra (ábrák: Daniel Bilar)





Sokszorelfedezték

Pareto (1897): 80-20 szabály

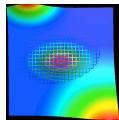
Yule (1925): evolúció

Zipf (1949): szóeloszlás

Simon (1955): Zipf alapján

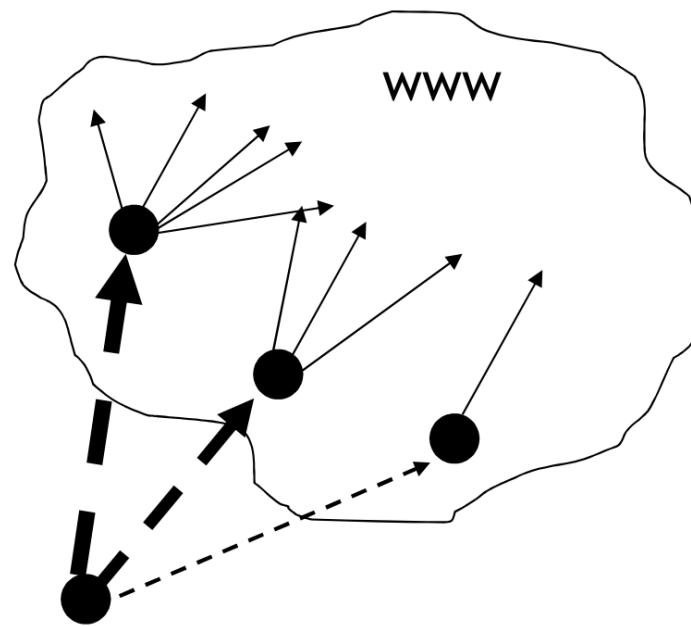
Price (1976): hivatkozási gráf!

és Barabási-Albert modell (1999): WWW gráf fokszámeloszlása

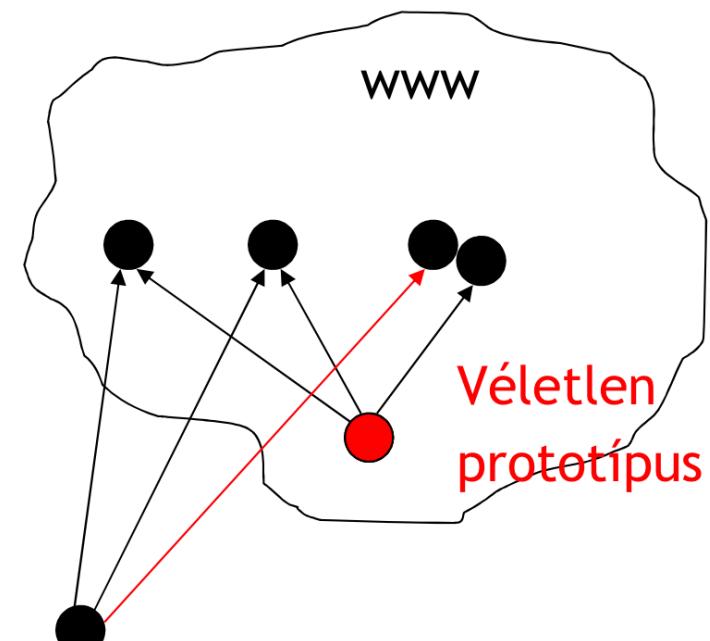


Örülünk Vincent, de kéne valamivel modelleznünk

Preferential attachment
(Albert, Barabási 1999)



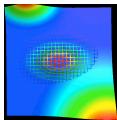
Evolving copy
(Broder et al., 2000)



Paraméter: m éellel kapcsolódik egy új pont

Paraméter: másolás minősége

09/11/2016



Barabási-Albert modell

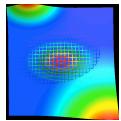
1. A modell minden lépésben egy új csúccsal bővíti a meglévő gráfot
2. Majd a meglévő fokszámok alapján kiválaszt m csúcsot és összeköti az új csúccsal

A kapott hatványeloszlás (i-dik időpillanatban keletkezett csúcs t időpontbeli fokszámára vontakoztatva):

$$P(k_i(t) = k) = m^2 t \left(\frac{1}{k^2} - \frac{1}{(k-1)^2} \right) \sim k^{-3}$$

1. Növekednie kell folyamatosan (ER nem!)
2. A növekedésnek BA szerint kell megtörténnie

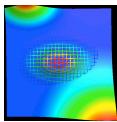
Ha bármelyik nem igaz, nem lesz PL!



Készen vagyunk?

Már van hatványeloszlásunk!

De egy valós hálózat más tulajdonságokkal is rendelkezik:



Készen vagyunk?

Már van hatványeloszlásunk!

De egy valós hálózat más tulajdonságokkal is rendelkezik:

Kis világ modell (Watts és Strogatz)

1. Alacsony átlagos távolság

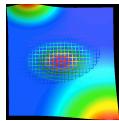
Régóta ismert jelenség, hogy pár emberen keresztül mindenkit ismerünk

Általában $\log(N)$ -el arányos

Nem szorosan:

Friendship paradoxon (Scott L. Feld 1991):

legtöbb embernek kevesebb barátja van, mint a barátainak átlagosan



Készen vagyunk?

Már van hatványeloszlásunk!

De egy valós hálózat más tulajdonságokkal is rendelkezik:

Kis világ modell (Watts és Strogatz)

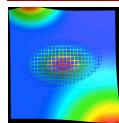
1. Alacsony átlagos távolság

2. Erős klaszterezettség:

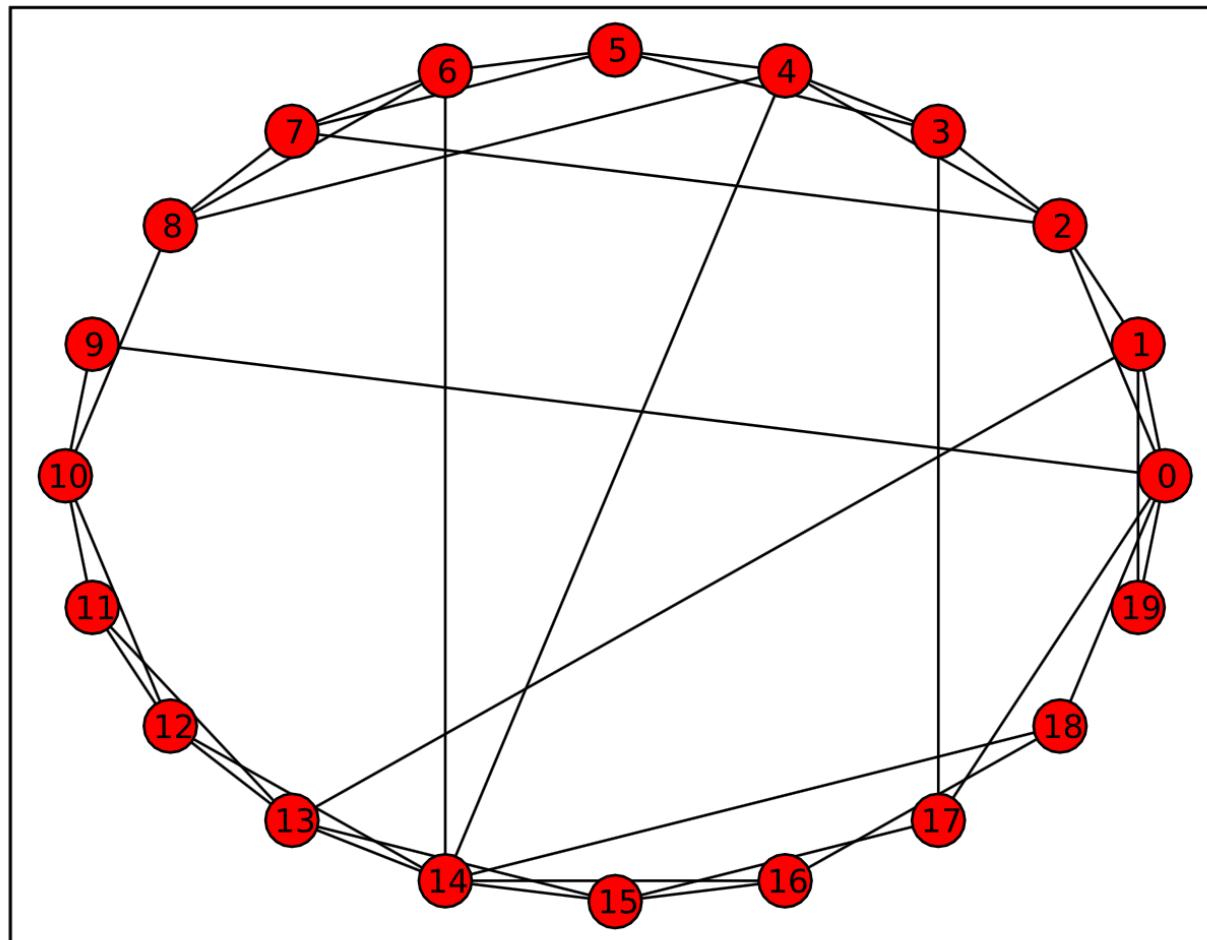
A csúcsok szomszédai átlagosan “erősen” összefüggnek

külön csúcsokra (k_i a kifok):

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$



Watts-Strogatz modell

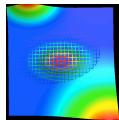


r reguláris gyűrű

Ebből véletlen veszünk el
és kötjük be ER szerűen
máshova

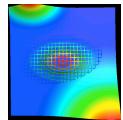
Eredmény: kis-világ

De nem hatványeloszlás!



Összefoglalva

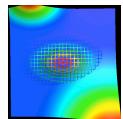
	Fokszámeloszlás	Klaszterezettség	Átlagos távolság
Valós hálózatok	Hatónyeloszlás	erősen	kicsi
Erdős-Rényi	Poisson	gyengén	kicsi
Barabási-Albert	Hatónyeloszlás	gyengén	kicsi
Watts-Strogatz	Poisson	erősen	kicsi
Broder et al.	Hatónyeloszlás	erősen	kicsi



Random Forest (Breiman, 2001)

Döntési fa vagy erdő?





Random Forest (Breiman, 2001)

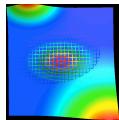
Döntési fá vagy erdő?

Bagging (Breiman , 1996):

- újramintavételezés -> modell aggregáció

Miért?

-> DT esetében?



Random Forest (Breiman, 2001)

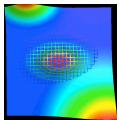
Zajos attribútumok?

-> minden vágás előtt mintavételezés: m attribútum kiválasztása (M -ből)

Random Forest:

1. Bagging: 100-500
2. Fa építés
 1. minden levelet vágunk a mintavételezett attribútumhalmaz alapján
3. Aggregáció: pl. többbségi döntés, várható érték stb.

Mi a gyakorlati különbség a DT és az RF között?



AdaBoost

Freund and Schapire (1995):

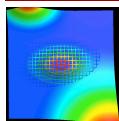
Adaptive boosting

"Weak" ("rosszul" teljesítő) modellek halmaza, lineáris kombinált

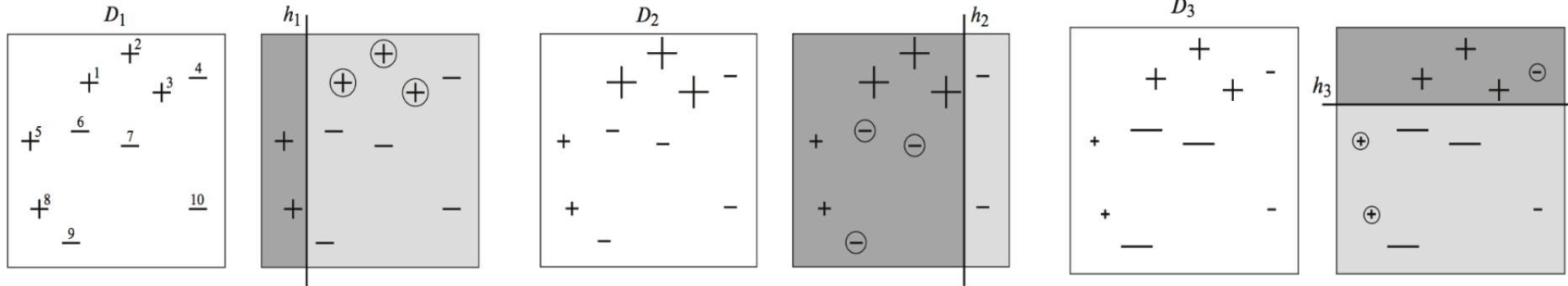
Meta osztályozó

Továbbfejlesztései: Gradient Boosting, Gradient Boosted Tree or LogitBoost.

Kaggle: xGBT



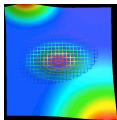
AdaBoost



$$H = \text{sign} \left(0.42 \begin{array}{|c|c|} \hline \text{+} & \text{-} \\ \hline \end{array} + 0.65 \begin{array}{|c|c|} \hline \text{-} & \text{+} \\ \hline \end{array} + 0.92 \begin{array}{|c|c|} \hline \text{-} & \text{-} \\ \hline \end{array} \right)$$

$$= \begin{array}{|c|c|c|} \hline \text{+} & \text{+} & \text{-} \\ \hline \text{+} & \text{-} & \text{-} \\ \hline \text{+} & \text{-} & \text{-} \\ \hline \end{array}$$

Fig.: Freund & Schapire



AdaBoost

Given: $(x_1, y_1), \dots, (x_m, y_m)$ where $x_i \in \mathcal{X}, y_i \in \{-1, +1\}$.

Initialize: $D_1(i) = 1/m$ for $i = 1, \dots, m$.

For $t = 1, \dots, T$:

- Train weak learner using distribution D_t .
- Get weak hypothesis $h_t : \mathcal{X} \rightarrow \{-1, +1\}$.
- Aim: select h_t with low weighted error:

$$\varepsilon_t = \Pr_{i \sim D_t} [h_t(x_i) \neq y_i].$$

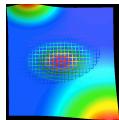
- Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right)$.
- Update, for $i = 1, \dots, m$:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right).$$



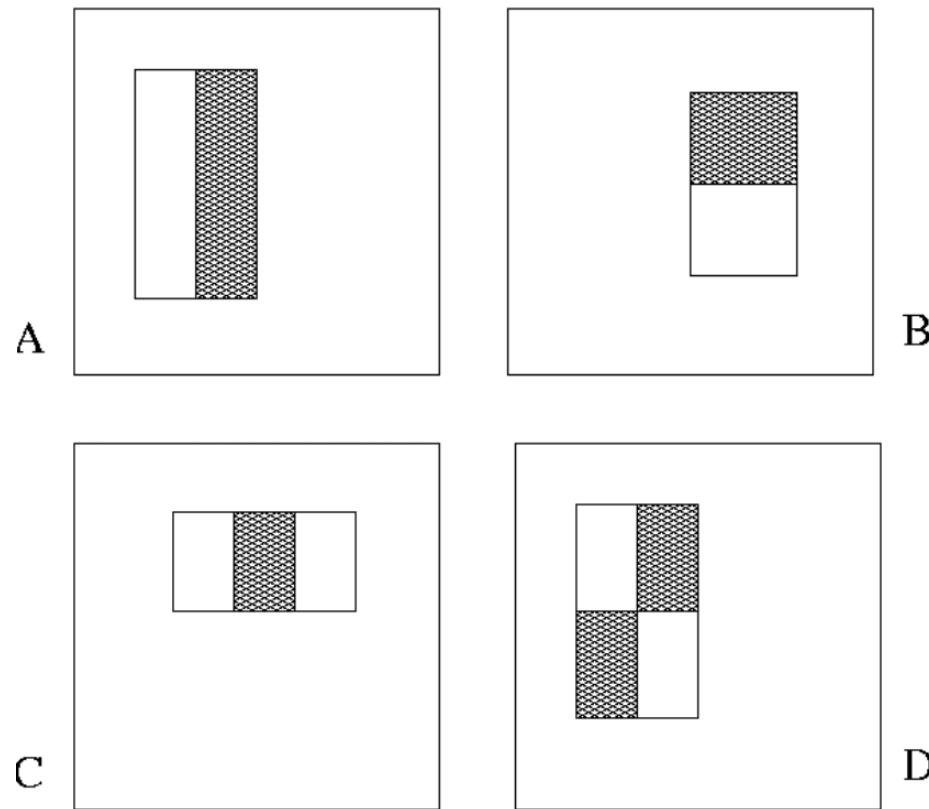
AdaBoost: Viola Jones

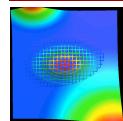
Haar like attribútumok
162k potenciális terület

AdaBoost Decision stump ->
<100 features

Cascade osztályozó

Hogyan lehet gyorsan
kiszámolni?





Viola Jones detector

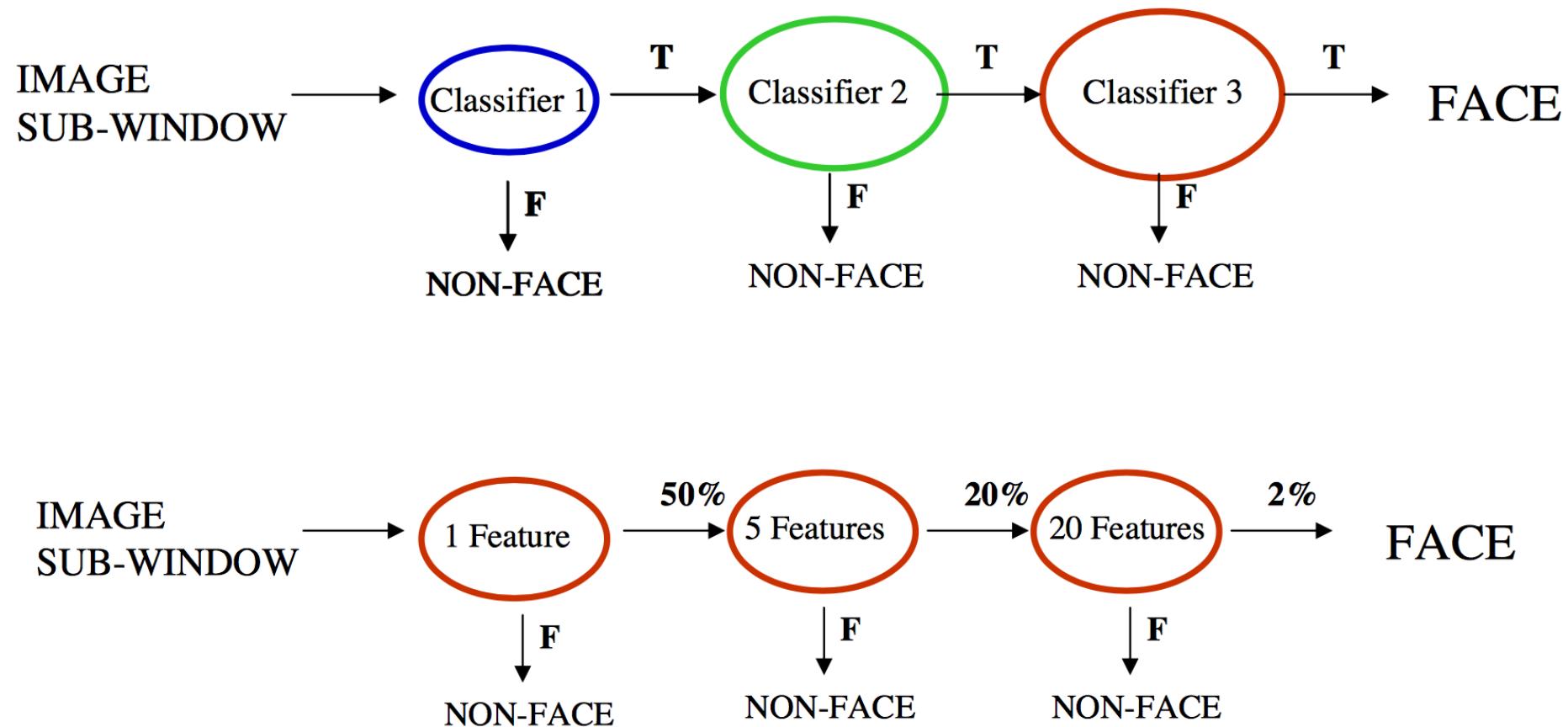
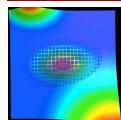


Fig.: Viola&Jones

14/05/2017



Viola Jones detector

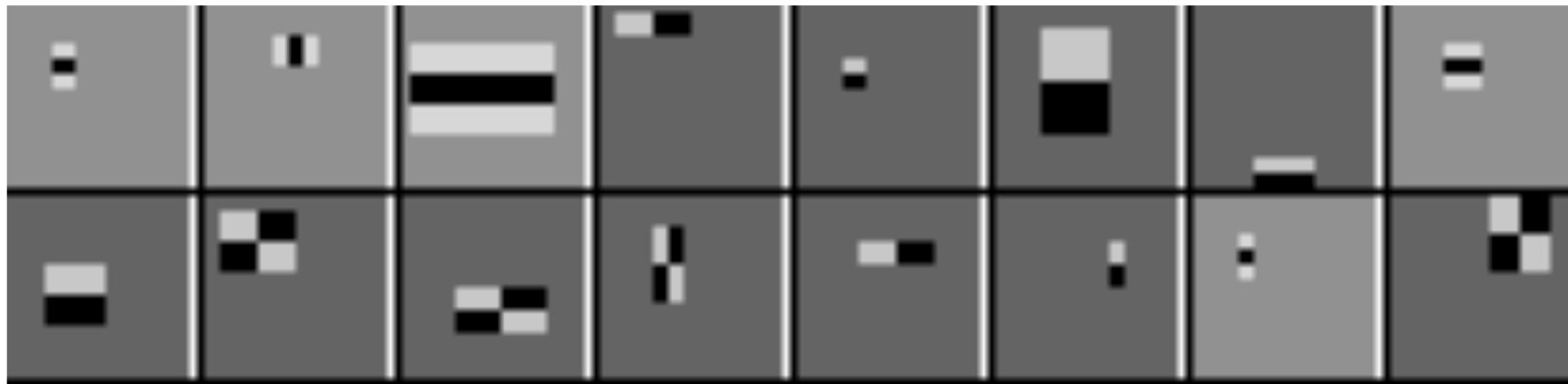
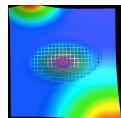


Fig.: Viola&Jones

14/05/2017



Viola Jones detector

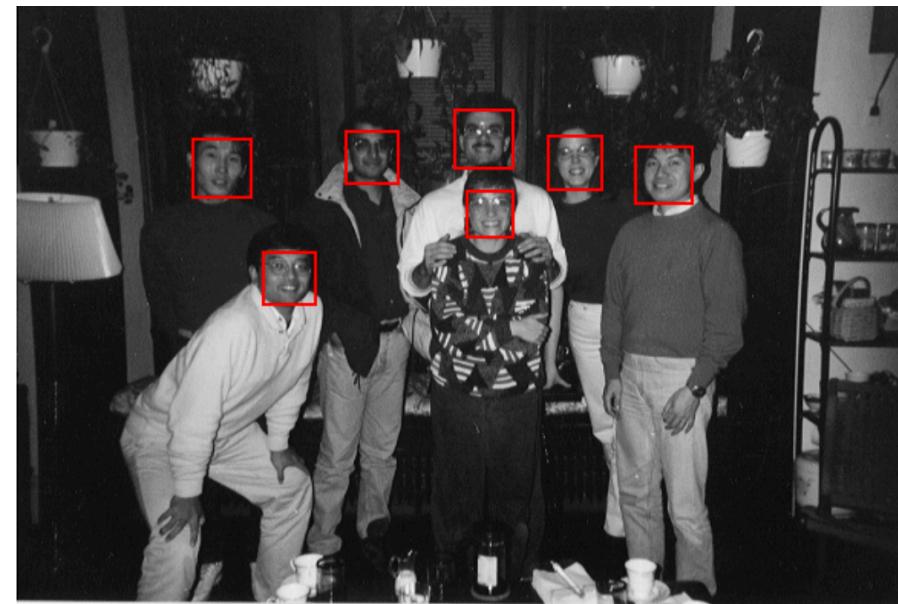
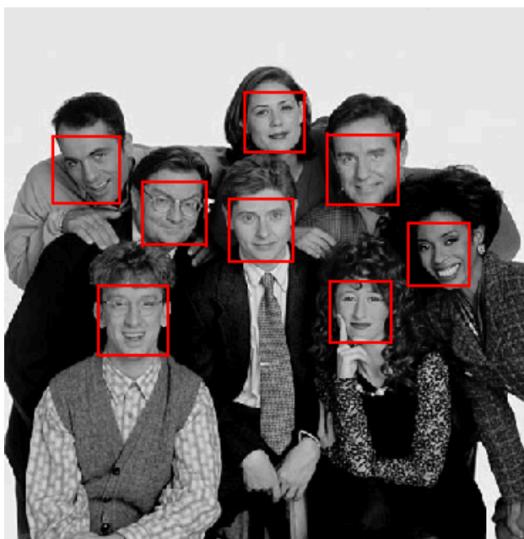
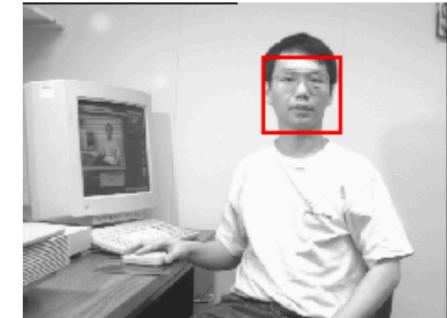
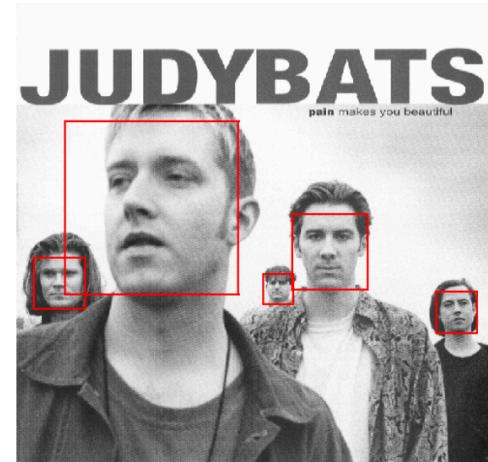
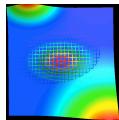


Fig.: Viola&Jones

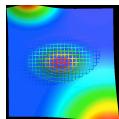
14/05/2017



Gradient Boosted Trees (Friedman, Hastie, Tibshirani)

AdaBoost: decision stump mint gyenge ("weak") osztályozó

Miért DS?



Gradient Boosted Trees

Regressziós fák! (DT + valós kimenet)

Predikció:

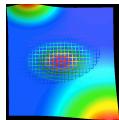
$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F}$$

Vagy (n a tanulóhalmaz számossága):

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

"Error" (bal) és komplexitás (jobb)

VC tételek -> alacsony komplexitás



Gradient Boosted Trees

Hogyan lehet optimalizálni?

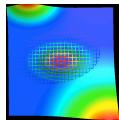
Iteratívan:

$$\begin{aligned}
 \hat{y}_i^{(0)} &= 0 \\
 \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
 \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
 &\dots \\
 \hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
 \end{aligned}$$

Tegyük vissza az objektív függvénybe 😊

src.: Chen

14/05/2017



Gradient Boosted Trees

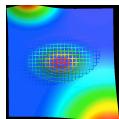
Eredmény:

$$Obj^{(t)} = \sum_{i=1}^n l \left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i) \right) + \Omega(f_t) + constant$$

Már csak egy loss függvényre van szükségünk.

RMSE?

Hogy lehet optimalizálni RMSE-re?



Gradient Boosted Trees

Megoldás: Taylor sorba fejtjük az objektív függvényt!

Eredmény:

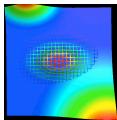
$$Obj^{(t)} \simeq \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + constant$$

ahol

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

és

$$h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)})$$



Gradient Boosted Trees

Komplexitás?

Attól függ.... De:

1. a pontok súlya
2. a fák mérete?

Próbáljuk ki a 'person.txt'-n:

scikit ensemble (AdaBoost és RandomForest)
xGBoost ("Kaggle's favourite flavor")