Adatbányászat

Klaszterezés

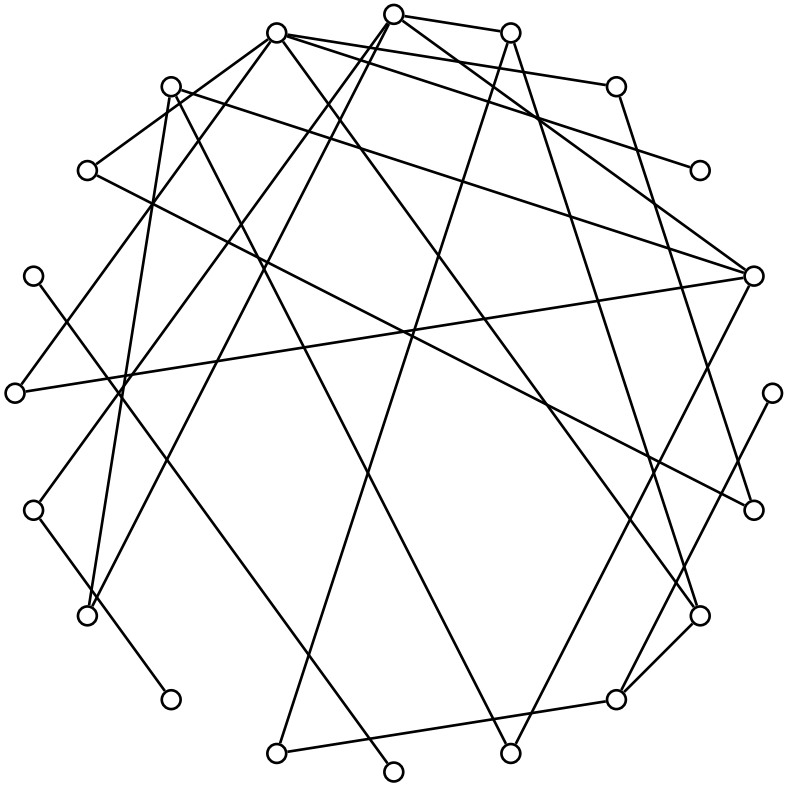
Szociális hálózatok

Szegedi Tudományegyetem

Szociális hálózatok

- Élei lehetnek címkézettek (pl. ellenség, barát), továbbá súlyozottak (pl. telefonbeszélgetés)
- Megjelenési formái
  - Ismeretségi, társszerzőségi gráf (Erdős-Bacon szám)
  - Levelezési gráf: $G = (V, E)$, hogy $(x, y) \in E \Leftrightarrow x, y \in V \wedge (\text{írt}(x, y) \wedge \text{írt}(y, x))$
    Mi lenne, ha az utolsó $\wedge$ helyén $\vee$ állna?
  - Valami egzotikusabb: páros (k-részű) gráfok ("social bookmarking")
- A hagyományos klaszterező eljárások nem (vagy legalábbis csak korlátozottan) működnek esetükben
- Okai
  - Skálázódás (hiánya)
  - Két csúcs között lévő kapcsolat mértékének számszerűsítése
  - Címkézett élek kezelése

Ilyenek a szociális hálózatok?

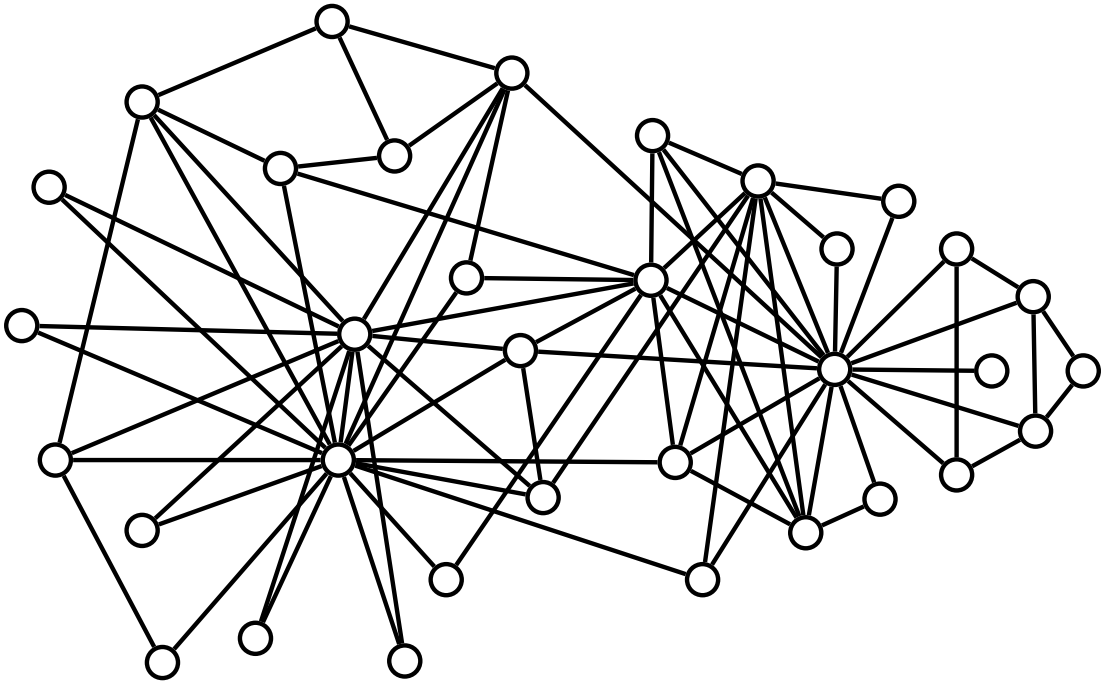


Nem mondhatnánk.

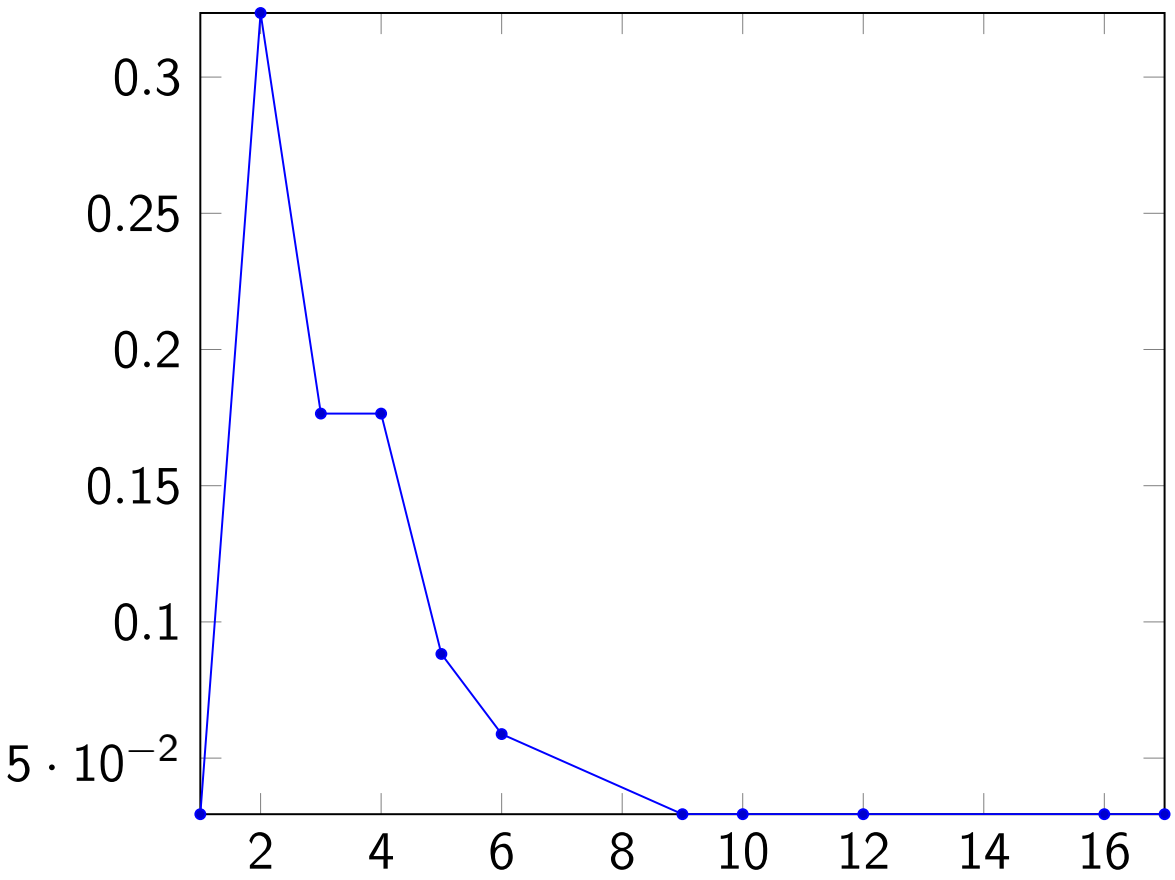
Ez sokkal inkább egy véletlen (Erdős-Rényi) gráf

Példa szociális hálózat

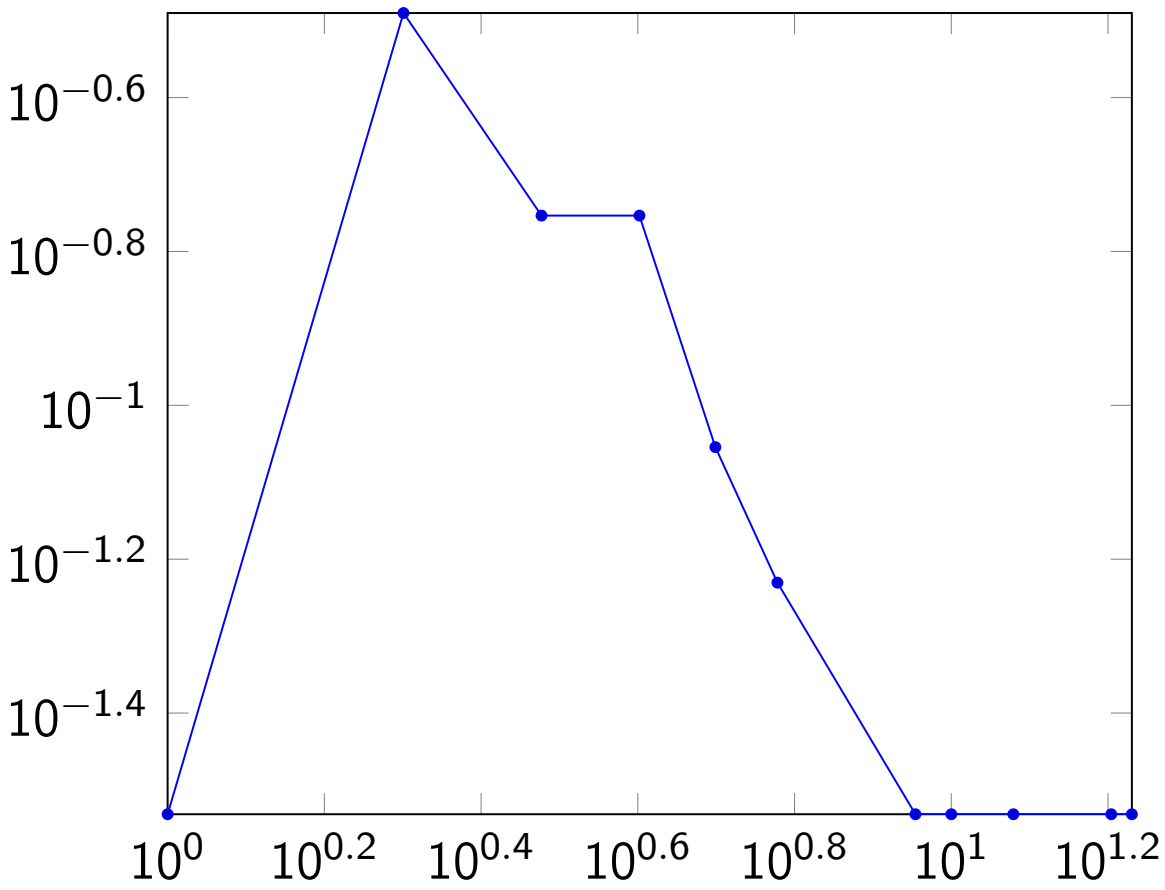
Zachary karate klub gráfja



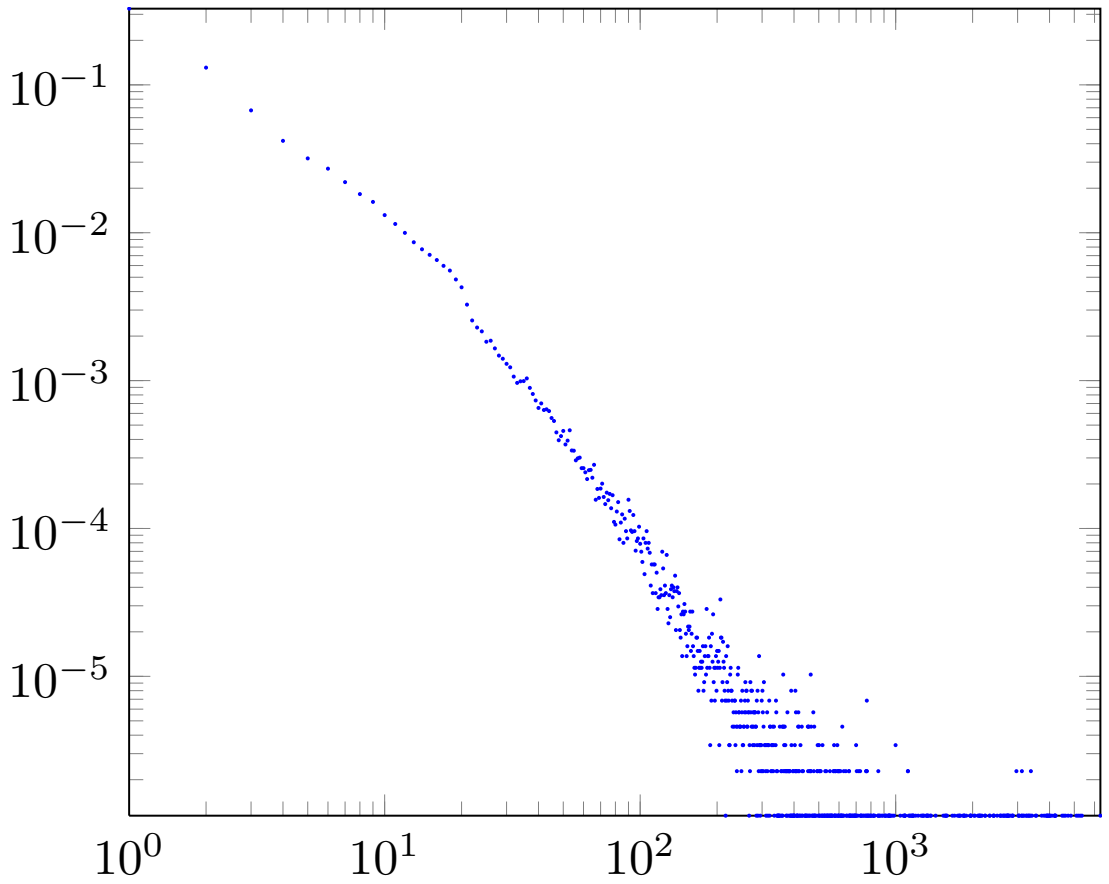
Kisvilággráfok sajátosságai – fokszámeloszlás lineáris skálán



Kisvilággráfok sajátosságai – fokszámeloszlás log-log skálán



Kisvilággráfok sajátosságai – fokszámeloszlás egy valós nagyságú példán



Kisvilággráfok sajátosságai – Hatványeloszlás

- Fokszámok eloszlása hatványeloszlást követ (valami barátságos eloszlás helyett)
  - $p(x) \propto \exp -(x-\mu)^2$ helyett $p(x) \propto \frac{1}{x^\beta}$
  - Milyen hatása lesz ennek $\log p(x)$ alakulására?

Kisvilággráfok sajátosságai – Kis átmérő

- $d(G) = \frac{2}{|V|(|V|-1)} \sum_{u,v \in V} |u \rightsquigarrow v| \approx log(|V|)$
- Kis átmérő: alacsony a pontpárok közötti átlagos távolság (Milgram, Karinthy)

Kisvilággráfok sajátosságai – Klaszterezettség, összefüggőség

- ha A csúcs szomszédja B és C is, akkor B és C csúcsok "véletlennél nagyobb" valószínűséggel szomszédok
- $X \in G = (V, E)$ csúcsának klaszterezettsége: $c(X) = 2\frac{\text{X szomszédjai között menő élek száma}}{|\text{X szomszédjai}|(|\text{X szomszédjai}|-1)}$
- G klaszterezettsége: a csúcsok klaszterezettségének átlaga
- Hogy jelenik meg mindez a szomszédsági gráfban?
- Néhány további nevezetes gráf típus/modell
  - Watts-Strogatz
  - Páros (k-részes) gráfok
  - Rácsok

Mi is a célunk tulajdonképp? – Gráfok vágásai

- Egy $G = (V, E)$ hasonlósági gráf csúcsainak olyan particionálására ($S \cap T = \emptyset \wedge S \cup T = V$) vágyunk, ami **minimális** él eltávolításával ($Cut(S, T)$) jár $\Rightarrow$ minimális vágás
  - továbbá közel azonos méretű részhalmazokat eredményez $\Rightarrow$ **Normalizált** vágás $= \frac{Cut(S,T)}{Vol(S)} + \frac{Cut(S,T)}{Vol(T)}$, ahol $Vol(S)$, illetve $Vol(T)$ azon élek számát adja meg, amely legalább egy csúcsa S-ben, illetve T-ben található
- Laplace mátrix: $L = D - A$
  - $a_{ij} = \begin{cases} 1 & \text{, ha } (i,j) \in E \\ 0 & \text{, különben} \end{cases}$ $\quad d_{ij} = \begin{cases} \sum_{k=1}^{|V|} a_{ik} & \text{, ha } i = j \\ 0 & \text{, különben} \end{cases}$
- Állítás: L szimmetrikus és pozitív szemidefinit. Biz.: táblán
  - $x^\intercal L x = \frac{1}{2} \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} a_{ij}(x_i - x_j)^2 \Rightarrow L$ sajátértékei valósak és ≥ 0

Spektrálklaszterezés alapja és a Rayleigh-Ritz tétel

- Lehet-e $\lambda_1 = 0$? Ha igen, mikor?
- Lehet-e $\lambda_i = 0$, $i > 1$? Ha igen, mikor?
- Végső soron $x^\intercal L x = \frac{1}{2} \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} a_{ij}(x_i - x_j)^2$ minimalizálása a cél, úgy, hogy $\|x\| = 1$, és $x \perp \vec{1}$
- Rayleigh-Ritz tétel: $\forall M = M^\intercal$-re $\min \frac{x^\intercal M x}{x^\intercal x} = \lambda_1$, valamint $\arg\min \frac{x^\intercal M x}{x^\intercal x} = x_1$, ahol λ_1 M legkisebb sajátértéke, x_1-pedig a hozzátartozó sajátvektor
  - Továbbá minden sajátvektorra $x_i = \arg\min_{x \perp x_1, \ldots, x_{i-1}} \frac{x^\intercal M x}{x^\intercal x}$
- Mit von mindez az általános eredmény maga után a Laplace mátrixra nézve?

Spektrálklaszterezés

- Minimalizálni akarjuk $\frac{x^\intercal L x}{x^\intercal x}$-t, hogy $\|x\| = 1$ és $x \perp \vec{1} \Rightarrow$ ez épp x_2, az ún. Fiedler-vektor lesz (Rayleigh-Ritz tétel értelmében)
- Egy lehetséges (korántsem egyedüli) forgatókönyv, hogy a pontokat kettéválasztjuk a pozitív és negatív értéket felvevő sajátvektorkomponenseknek megfelelően
  - Egyáltalán biztosak lehetünk benne, a Fiedler-vektor fog > 0 és < 0 elemeket tartalmazni egyszerre?
- A 0 mentén való szeparáció helyett hagyatkozhattunk volna a mediánra is, vagy további sajátvektorokra is akár
- A probléma visszavezethető még többek között: minimális vágásra, véletlen sétákra, ...

Girvan-Newman algoritmus

- (a, b) él köztessége = azon pontpárok száma, amelyek között menő legrövidebb út tartalmazza (a, b) élet
- A magas érték jó vagy rossz?
- A gráf minden X pontjából indulva határozzuk meg (X, b) csúcspárok köztességét (szélességi bejárás)
- A szélességi bejárás után lássuk el az egyes csúcsokat azokkal az értékekkel, hogy a gyökérből hányféle legrövidebb út vezet hozzájuk
  - A szélességi bejárás azonos szintjén "belüli" élek nem lehetnek részei a gyökérelemből induló legrövidebb utaknak
  - Ellenben a többi, ún. DAG-élek legalább egy legrövidebb úton fekszenek
  - A gyökérhez 1 értéket társítva haladjunk szintenként, és egy gyerekcsúcs értéke legyen a szülei értékeinek összege

Girvan-Newman algoritmus – folytatás

- Az aktuális X gyökérelem viszonyában lássuk el a gráf egyes éleit köztességértékekkel (alulról felfele haladva)
- Minden élhez határozzuk meg, hogy adott X gyökérelem viszonyában a legrövidebb utak mekkora hányada megy rajtuk keresztül
  - A levelekhez 1 értéket társítunk
  - Nemlevél csúcsok $:= 1 +$ kimenő DAG élek értékeinek összege
  - Egy csúcs köztességének "felterjesztése" során a szülők felé vezető éleken a szülők a hozzájuk X-ből vezető legrövidebb utak arányában osztozkodnak
    - Formálisabban: amennyiben egy nemlevél Z csúcs szülei $Y_1, \ldots, Y_k$, $\pi^*_{Y_i}$ pedig az Y_i-ig menő legrövidebb utak száma, úgy az (Y_i, Z) élhez társított köztességérték egyezzen meg Z csúcs köztességértékének $\frac{\pi^*_i}{\sum_{j=1}^{k} \pi^*_j}$-szeresével

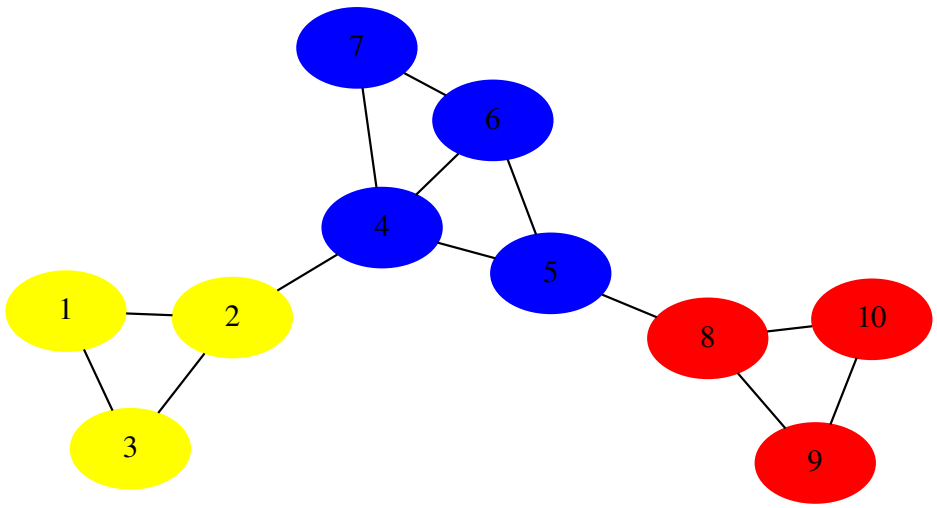
Köztességalapú particionálás

- A szélességi keresést ismételjük meg minden (minél több) csúcsból kiindulva, a csúcsokhoz társított köztességértékeket pedig aggregáljuk
- Az élekhez köztességértékeit tekinthetjük egyfajta távolságnak (még ha metrikának nem is)
- Iteratív módon összevonhatjuk a legkisebb köztességű élek végén lévő pontokat
- Eljárhatunk fordítva is: az eredeti gráfot a legnagyobb köztességértékű éleket elhagyva komponensekre bonthatjuk

A korábbi ismereteink is kapóra jönnek – Véletlen séták

- (extrémmód) Perszonalizált PageRank (PRP) segítségével
  - "Elfogult" teleportálás: $(1-\beta)$ valószínűséggel induljunk újra egy megadott pontból
- Rendezzük a csúcsokat PRP szerint és vegyük a normalizált vágások értékeit a k legmagasabb PRP értékű csúcsok mentén
- A normalizált vágások lokális minimumai a klaszterhatárok

csúcs	PRP(1)	Vágás
1	0,24	2/2+2/13≈1,15
2	0,22	3/4+3/12=1
3	0,17	**1/4+1/10=0,35**
4	0,12	3/7+3/9≈0,76
6	0,07	4/9+4/8≈0,94
5	0,06	3/10+3/6=1,5
7	0,05	**1/10+1/4=0,35**
8	0,03	2/12+2/3≈0,83
10	0,02	2/13+2/2≈1,15



Közösségek mint gyakori elemhalmazok

- Gyakori elemhalmazok keresése páros gráfokon
  - Kosarak: ?, Termékek: ?
  - K_{st}: páros gráf azon $s+t$ csúcsú részgráfja, melyben a csúcsok között megfigyelhető $s \cdot t$ él mindegyike jelen van $\Rightarrow$ s-gyakori t-elemhalmazokat keresünk
- Tfh. a páros gráfunk mindkét oldalán n csúcs található és d a csúcsok átlagos fokszáma $\Rightarrow \sum_{i=1}^{n} \binom{d_i}{t} = \Omega\left(n\binom{d}{t}\right)$ termék t-s található a "tranzakciós adatbázisban"
  - $\frac{n\binom{d}{t}}{\binom{n}{t}}$ egy termék t-s előfordulásának várható értéke
  - Ez $n \gg d \gg t$ esetén $\approx n\left(\frac{d}{n}\right)^t$, vagyis $n\left(\frac{d}{n}\right)^t \geq s \Rightarrow \exists K_{st}$