Közösség detektálás gráfokban

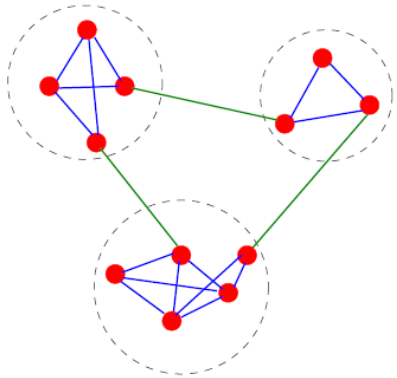
Önszervező rendszerek

Hegedűs István

- Célkitűzés:
  - valamilyen objektumok halmaza felett minták, csoportok detektálása csakis az egyedek közötti kapcsolatok struktúrájának a felhasználásával, valamint esetleges hierarchikus szerveződések feltárása

Közösség detekció

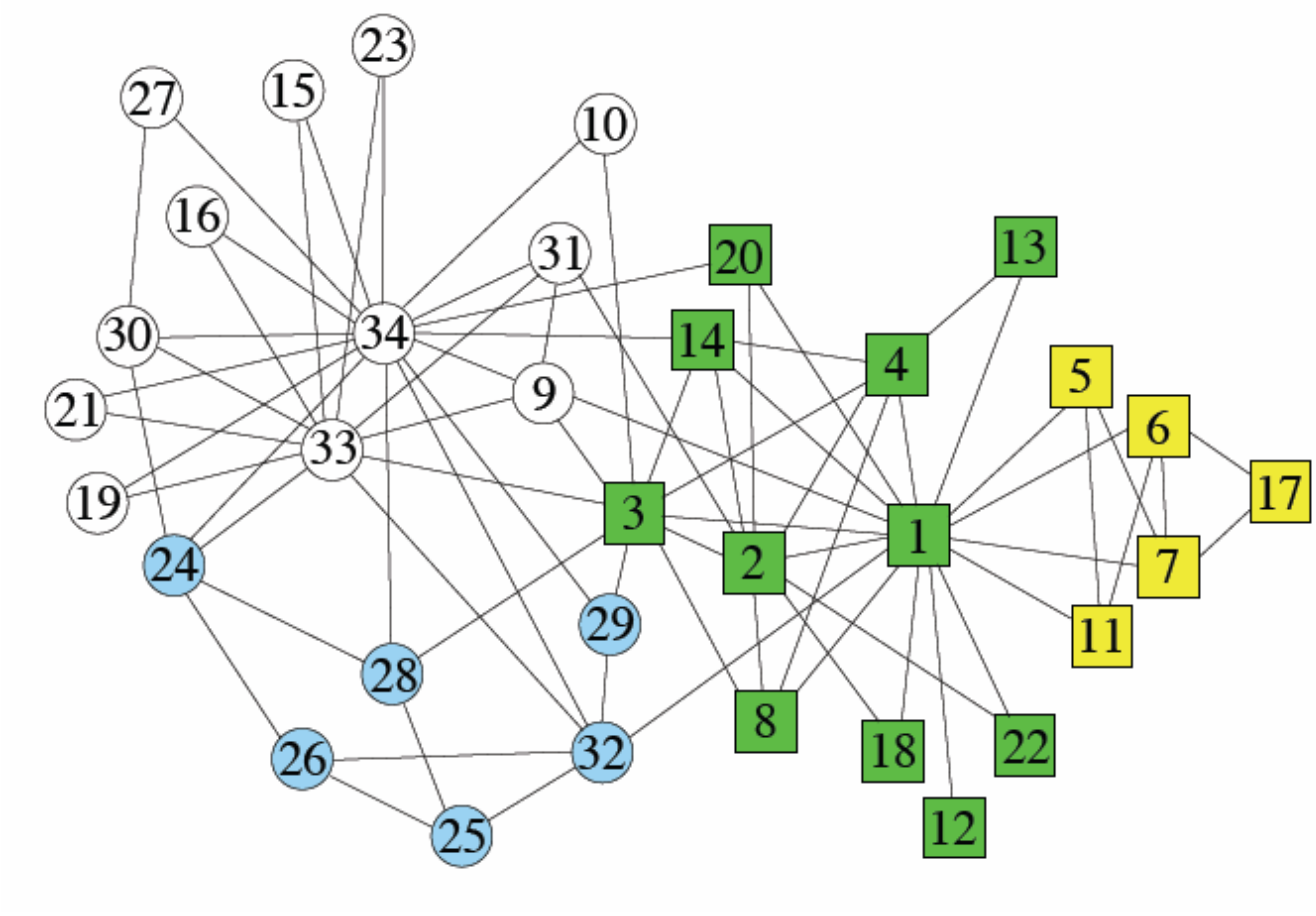
- Csoportosulások keresése gráfokban
- Jellemzők:
  - Csoportokon belül sok, kívül kevés él
  - Erdős-Rényi gráf modellben „esélytelen”
  - Közösségek:
    - Klaszterek
    - Modulok
    - Csoportok



- Felhasználási területek: biológia, bioinformatika, számítás tudomány, politika, …

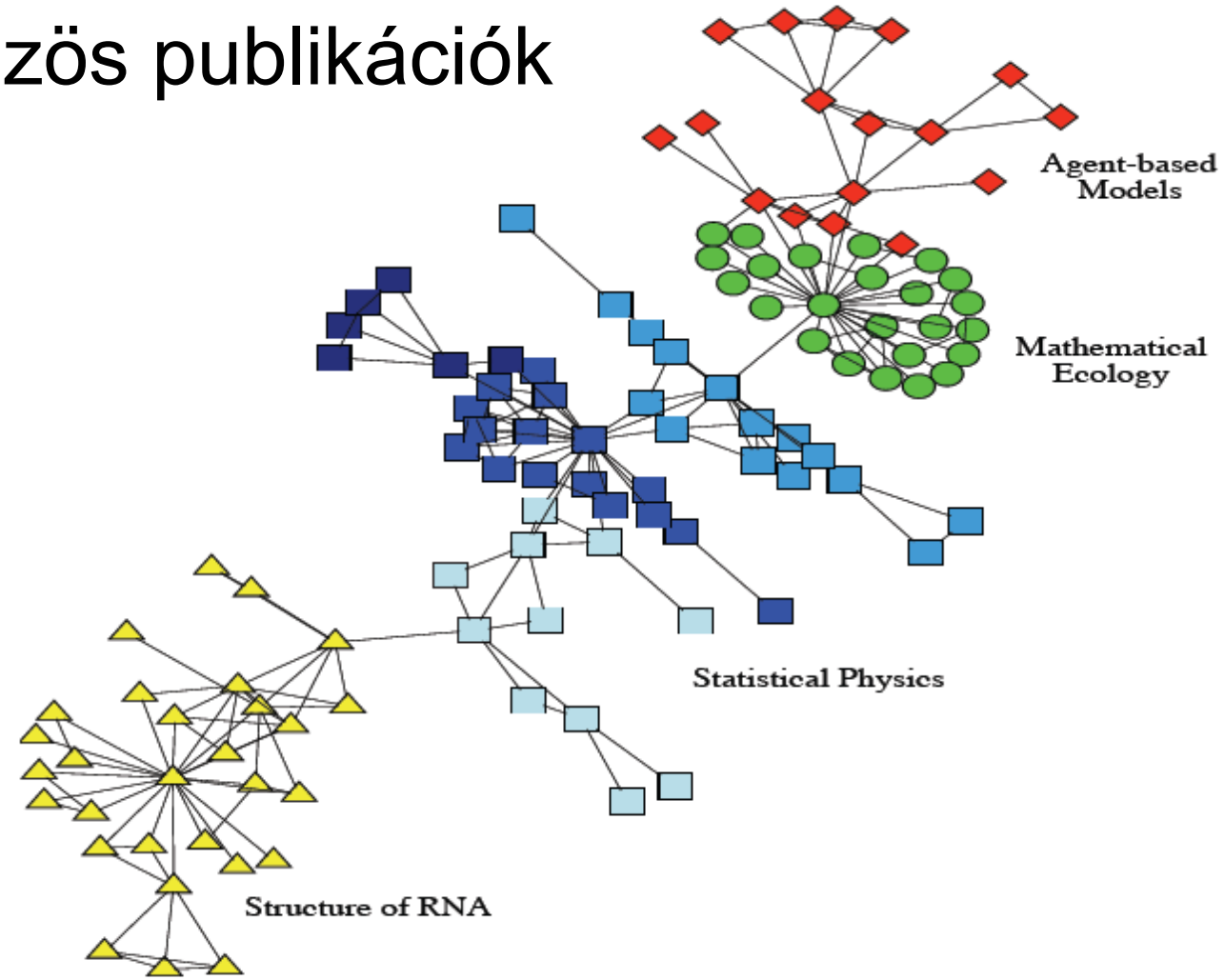
Zachary-gráf

- Szociális kapcsolatok



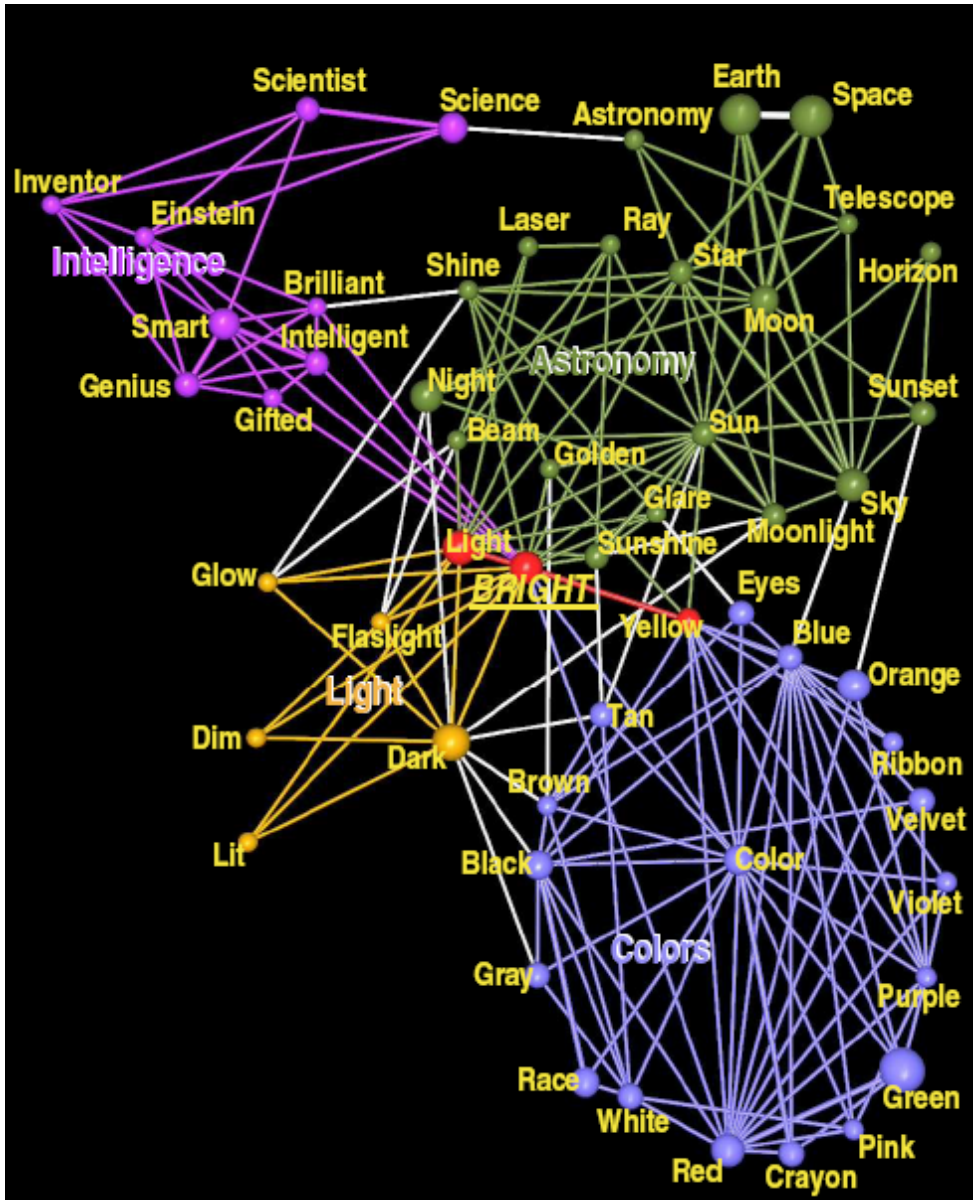
Santa Fe Institute

- Közös publikációk



Asszociációs kísérlet

- Példa a csoportok átfedésére
- Egy pont nem feltétlenül egy csoporthoz tartozik



A közösség detekció alapelemi

- „közösségen beül várhatóan több az él”
- Gráfok csúcshalmazának jellemzése
  - Klaszteren belüli eloszlás

$$\delta_{int}(\mathcal{C}) = \frac{\#\text{ internal edges of } \mathcal{C}}{n_c(n_c-1)/2}$$

  - Klaszterek közötti eloszlás

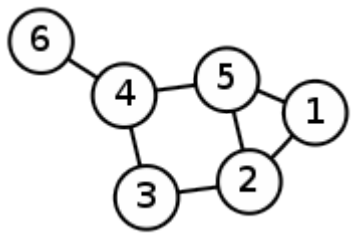
$$\delta_{ext}(\mathcal{C}) = \frac{\#\text{ inter-cluster edges of } \mathcal{C}}{n_c(n-n_c)}$$

Közösségek definíciói

- Lokális definíció (jellemző)
  - Mindenki mindenkinek a barátja
    - Klikk, n-klikk, n-klán, n-klub, k-plex, k-core
    - Élkapcsolat, lambda halmaz, relatív fokszám eloszlás
- Globális definíció
  - Random gráftól való eltérés (modularitás)
- Csúcs hasonlóság alapú definíció
  - Euklideszi távolság, koszinusz hasonlóság, átfedés, Pearson korreláció, élkapcsolat

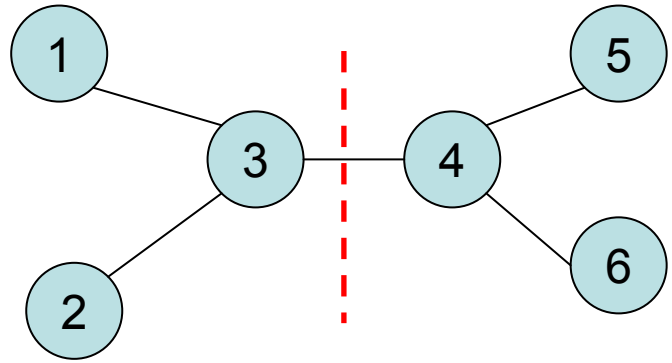
Módszerek: gráf partícionálás

- Csúcsok 2 halmazba osztása minimális klaszterek közötti élszám mellett + kiegyenlített
  - Kernighan-Lin algoritmus
    - Két részre osztás majd cserélgetés
  - Spektrális algoritmus
    - Laplace mátrix $L := (\ell_{i,j})_{n \times n}$. $\ell_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$
    - Legkisebb nem 0 sajátértékhez tartozó sajátvektor



$$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

– Spektrális algoritmus:



– L =

1	0	-1	0	0	0
0	1	-1	0	0	0
-1	-1	3	-1	0	0
0	0	-1	3	-1	-1
0	0	0	-1	1	0
0	0	0	-1	0	1

– V =

-0.1845	0.2887	-0.2780	0.6502	0.4647	-0.4082
-0.1845	0.2887	0.2780	-0.6502	0.4647	-0.4082
0.6572	-0.5774	-0.0000	0.0000	0.2610	-0.4082
-0.6572	-0.5774	-0.0000	-0.0000	-0.2610	-0.4082
0.1845	0.2887	0.6502	0.2780	-0.4647	-0.4082
0.1845	0.2887	-0.6502	-0.2780	-0.4647	-0.4082

– D =

4.5616	0	0	0	0	0
0	3.0000	0	0	0	0
0	0	1.0000	0	0	0
0	0	0	1.0000	0	0
0	0	0	0	0.4384	0
0	0	0	0	0	-0.0000

Módszerek: gráf partícionálás 2

– Maximális folyam – minimális vágás

– conductance

  • c → vágás költsége
    k → halmaz fokszáma

$$\Phi(\mathcal{C}) = \frac{c(\mathcal{C}, \mathcal{G} \setminus \mathcal{C})}{min(k_{\mathcal{C}}, k_{\mathcal{G}\setminus\mathcal{C}})}$$

– vágási arány

  • n → csúcsok száma

$$\Phi(\mathcal{C}) = \frac{c(\mathcal{C}, \mathcal{G} \setminus \mathcal{C})}{n_{\mathcal{C}} n_{\mathcal{G}\setminus\mathcal{C}}}$$

Módszerek: hierarchikus klaszterezés

- Nem tudjuk előre a klaszterek számát, nem is kell…
  - Agglomeratív algoritmusok
    - Single linkage
    - Complate linkage
    - Avarage linkage
  - Divizív algoritmusok

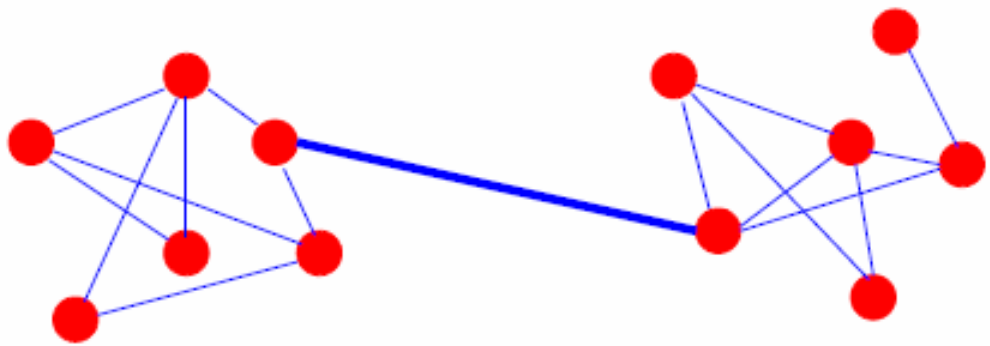
Módszerek: partíciós klaszterezés

- Előre definiált klaszterszám (k)
  - Minimum k-clustering (átmérő min)
  - k-clustering sum (átlagos átmérő min)
  - k-center (centertől való max távolság min)
  - k-median (centertől való átlagos táv. min)
  - k-means (centertől való négyzetes távolságösszegek min.)
  - Fuzzy k-means

$$\sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} ||\mathbf{x_j} - \mathbf{c_i}||^2$$

Divizív algoritmusok

- Hierarchikus: hasonlóság alapján vág
- Girvan Newman algoritmus
  - él központiság számítása
    - bármely két pont pár között a legrövidebb utak hányszor érintik ugyan azt az élet
  - a „legközpontibb” él mentén kell vágni a gráfot



Girvan-Newman algoritmus 2

- Random-walk alapú él-központiság
- Még ritka gráfok esetén is 10 000 csúcs a felső határa az algoritmusnak
- Könnyen adaptálható súlyozott élek esetén
- Tyler et al. gyorsítás
  - Monte Carlo módszerrel mintavételezett pontok közt mért csak központiságot
  - A közösséghatáron fekvő pontok bizonytalan kalszterezése → ebből adódik, hogy mely közösségek fednek át

Modularitás

- Random gráftól való különbözőség

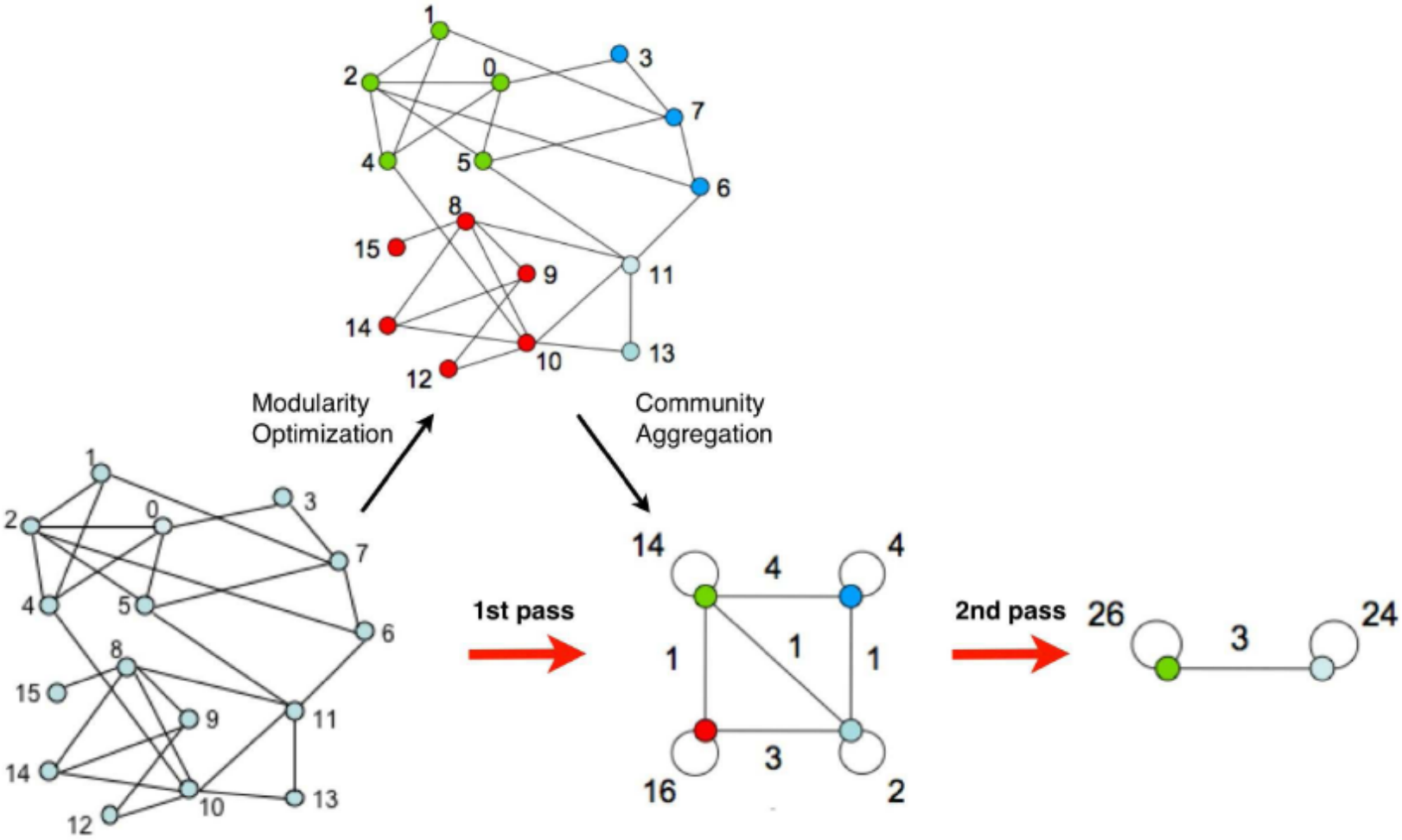
$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij})\, \delta(C_i, C_j)$$

  - A: szomszédsági mátrix
  - P: szomszédsági valószínűség
  - m: élek száma
  - Közelítése: $$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$
  - k: fokszám
  - súlyozott élekkel $$Q_w = \frac{1}{2W} \sum_{ij} \left(W_{ij} - \frac{s_i s_j}{2W} \right) \delta(C_i, C_j)$$

Modularitás alapú algoritmusok

- Mohó módszer (Newman):
  - Kezdetben minden csúcs külön klaszter
  - Nincs egy él sem behúzva
  - Adjuk hozzá azt az élet amely növeli a modularitást
  - Modularitás mindig az eredeti gráf alapján vannak számolva
    - Ha egy klaszteren belül húzunk be egy élet, az nem változtat a modularitáson
    - → csúcshalmazok egyesítése
  - A módszernek sok javítása született, mind sebesség, mind az optimum közelítése szempontokból

- Blondel et al. 2008
  - Minden csúcs külön közösség
  - Majd sorban a csúcsokhoz hozzáveszi a szomszédait, amíg a modularitás nem csökken (iteratívan) → hierarchia



Modularitás alapú algoritmusok

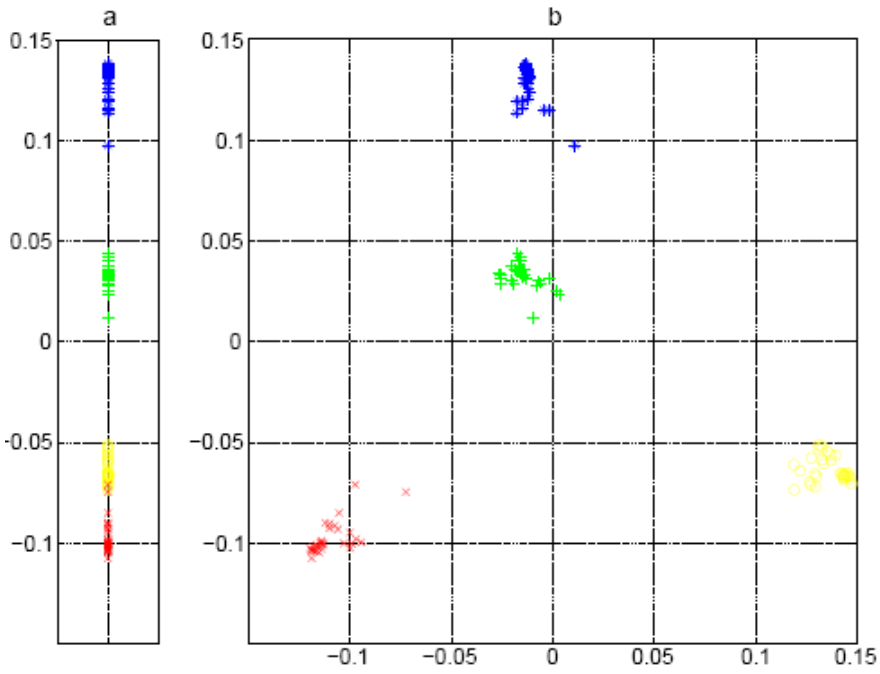
- Szimulált hűtés
  - Lokális mozgatás: véletlen módon egy csúcs átkerül egy másik csoportba
  - Globális mozgatás: csoportok vágása és egyesítése
  - Jó közelítést ad a globális optimumra, de lassú
- Genetikus algoritmusok
  - Modularitást használva fitness értékként

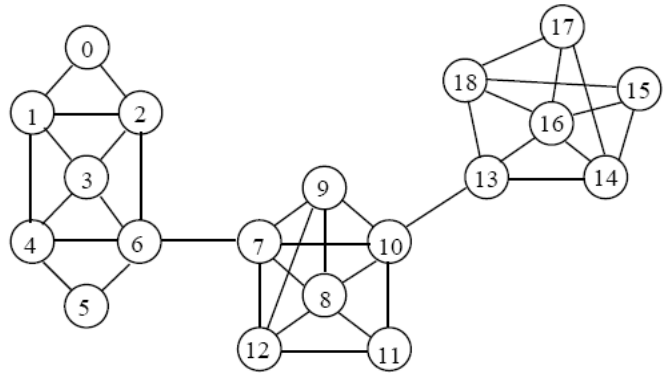
Modularitás alapú algoritmusok

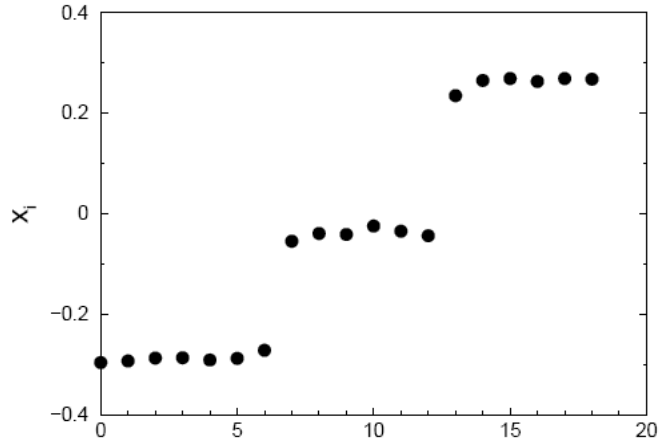
- Spektrális módszer:
  - modularitás: $$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$
  - legyen: $$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$
  - Vegyük a Laplace mátrix helyett a B mátrixot
    - Vágjuk ketté a gráfot rekurzívan, amíg nő a modularitás (legkisebb nem 0 sajátértékhez tartozó sajátvektor alapján)
    - Vegyük a k legkisebb (nem 0) sajátértékhez tarozó sajátvektorból álló nXk méretű mátrixot. Ennek a sorai reprezentálnak egy k dimenziós pontot. Klaszterezzük ezeket k klaszterbe.

Spektrális módszerek

- Laplace mátrix sajátvektorok + klaszterezés (bal)
- Jobb-sztochasztikus szomszédsági mátrix (jobb) + sajátvektorok lépcsősek







Címke terjesztés

- Legyen minden csúcsnak egyedi címkéje
- Iteratívan (minden csúcsra):
  - legyen a csúcs címkéje a leggyakoribb a szomszédos címkék közül
  - ha több leggyakoribb van, akkor véletlenszerűen közülük
  - folytassuk, amíg már csak kevés csúcs címkéje változik az iterációk során

Klikk perkoláció

- Vegyük a k méretű klikkeket a gráfban
  - Két k méretű klikk szomszédos ha k-1 csúcsuk megegyezik
  - Egy k-klikk lánc a szomszédos klikkek sorozata
  - A közösségek pedig a leghosszabb láncok uniója