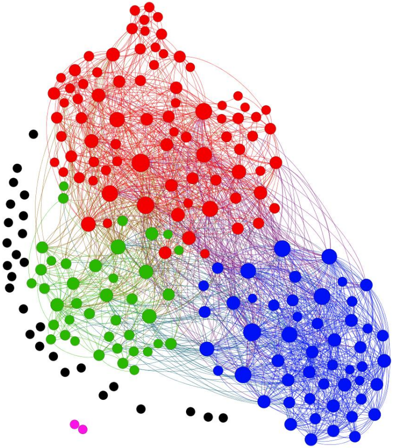
Hálózattudomány

SZTE Informatikai Intézet
Számítógépes Optimalizálás Tanszék
Előadó: London András

5. Előadás

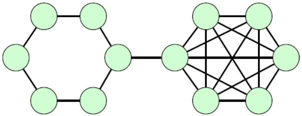
Közösségek hálózatban



ábra. Facebook kapcsolati háló

Közösségek hálózatban

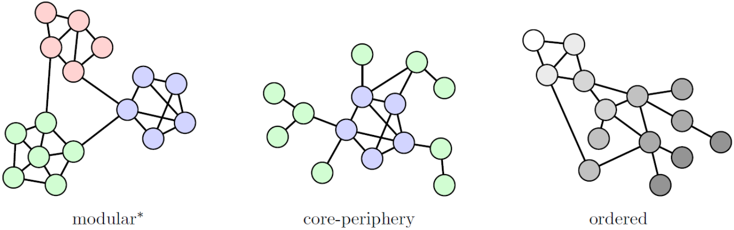
- Egyszerű mértékek : átlagos fokszám, fokszámeloszlás, klaszterezettség, átlagos úthossz $\longrightarrow$ sok információ a hálózatról DE
- elrejthetik az eloszlás heterogenitását



ábra. Háromszögek száma nagy, de azok csak a hálózat egy részén találhatók

- Mi a magas szintű szerveződés mintázata ?
- Kis méretű hálózat : szemmel látható (?) ; nagy méretű hálózat : kvantitatív eszközök szükségesek

Magas szintű szerveződési mintázatok



ábra. Közösségszerkezet, mag-periféria szerkezet, rendezett (lineáris hierarchia) szerkezet (forrás: Aaron Clauset, Network analysis and modelling course)

Asszortatív kapcsolódás

- Bizonyos attribútumok (tulajdonságok, jellemzők) kapcsolatára az meglévő élek világítanak rá $\longrightarrow$ pl. ismerőseink egy részének közös jellemzője a középiskola, ahová jártak
- Társadalmi hálózatokban jellemző, hogy olyan ismerőseink vannak akik hasonlítanak ránk. Pl. életkor, nyelv, születési hely, végzettségi szint, anyagi helyzet

Fontos kérdés: (1) az él a hasonlóság miatt létezik, vagy (2) az él létezése (ismeretség) miatt hasonlóvá válnak az attribútumok? (ld. pl. politikai beállítottság)

(*Egy modellezési lehetőség: Fitness modell (a későbbiekben tárgyaljuk, esetleg projektfeladat)*)

A modularitás függvény

- Az adott gráf mennyire tér el egy ugyanolyan fokszámeloszlású véletlen gráftól?

  Modulartás $=$ #{közösségen belüli élek} – $\mathbb{E}$[#{közösségen belüli élek}) egy azonos fokszámeloszlású véletlen gráfban]

- Newman-modularitás [1]

$$Q = \frac{1}{2m} \sum_{i,j} (a_{ij} - e_{ij}) \delta(C_i, C_j),$$

  δ a Dirac-delta függvény ($\delta(C_i, C_j) = 1$, ha $i = j$ és 0 különben), e_{ij} az i és j közti élek számának várható értéke (általában 0 és 1 között) egy véletlen (null-modell) gráfban

[1] Newman, *Physical Review E*, 2004

De mi legyen a null-modell?

- A síma véletlen gráf $G(n,p)$ általában nagyon „messze van” a valós hálózatoktól.
- Használjuk a konfiguráció modellt

Azaz ha az eredeti gráf fokszámsorozata $(k_1, k_2, \ldots, k_n)$ akkor

$$e_{ij} = \frac{k_i k_j}{2m}$$

így

$$Q = \frac{1}{2m} \sum_{i,j} (a_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j),$$

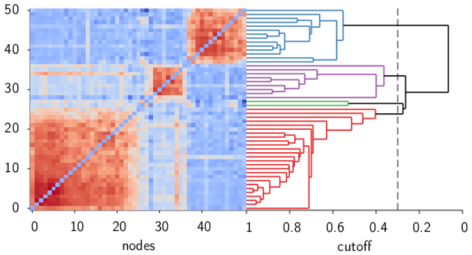
A cél pedig a pontok osztályozása (szétosztása) $C_1, \ldots, C_k$ $(k = ?)$ osztályokba (klaszterekbe), hogy Q minél nagyobb (maximális) legyen.

Modularitás maximalizálás

- Ha S a G gráf pontjainak összes lehetséges partícionálása (klaszterezése) akkor egy $f : S \to \mathbb{R}$ függvény méri egy adott $P \in S$ felosztás „jóságát" (Mitől jó?)
- Q egy lehetséges ilyen függvény $\longrightarrow$ „minél nagyobb, annál jobban klaszereztünk"
- S mérete exponenciálisan nagy (miért?), továbbá Q-t maximalizálni NP-nehéz
- Léteznek jól működő heurisztikák

Egy mohó algoritmus

1. Kezdetben minden pont egy önálló közösség
2. Majd mohó módon olvasztunk össze közösségeket, aszerint, hogy a lépés minél jobban növeli Q értékét



ábra. Közösségek és hierarchikus szerkezet

Az eljárás implementálásra számos különböző technika létezik (ld. pl. *single linkage, average linkage, etc.*)

Algoritmusok sokasága

- Modularitás optimizálás számos változata
- Más kiértékelő függvények használata (Mit tartunk fontosnak közösségkeresés esetén?)
- Sztochasztikus blokk modell (tárgyaljuk részletesebben)
- Átfedő közösségek keresése
- Spektrális módszerek, dinamikus közösségkeresés
- ...

Jegyzet, további olvasnivaló:

- Véletlen gráfok: Jackson könyv IV. fejezet, Newman cikk IV. szakasz
- Közösségek: Newman III. szakasz
- Közösségkeresés összefoglaló cikk: Santo Fortunato (2010): Community detection in graphs, *Physics Reports*

Feladat:

- Próbáljunk ki különböző közösségkereső algoritmusokat egy valós gráfon, nézzük meg a különbségeket