

TÁMOP-4.1.1.F-14/1/KONV-2015-0006

**„Munkaerő-piaci igényekre támaszkodó gyakorlatorientált képzések,
szolgáltatások a Szegedi Tudományegyetem fókuszában”**

Társadalmi és gazdasági hálózatok modellezése

London András

2015

SZÉCHENYI 



MAGYARORSZÁG
KORMÁNYA

Európai Unió
Európai Szociális
Alap



BEFEKTETÉS A JÖVŐBE

Társadalmi és gazdasági hálózatok modellezése

A kurzus anyagának rövid összefoglalója

London András

2015

Tartalomjegyzék

1. Bevezetés	1
2. A hálózatkutatás néhány alapfogalma	4
3. Véletlen gráfok, gráfmodellek	10
4. Közösségek hálózatokban	13
5. A sztochasztikus blokk modell	16

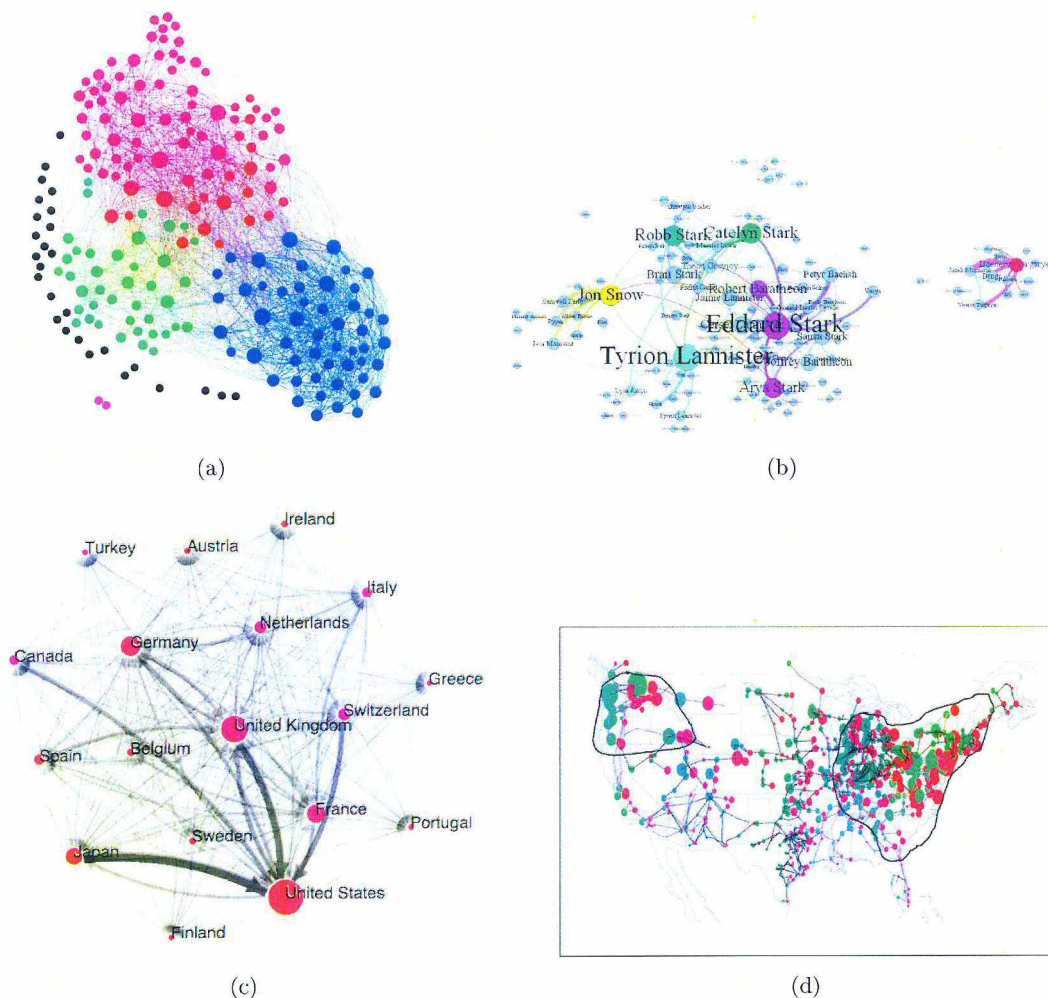
1. Bevezetés

A kurzus célja, hogy az alapképzésben részt vevő, elsősorban gazdaságinformatikus hallgatók megismerkedjenek a társadalmi, illetve gazdasági rendszerek hálózatokkal való modellezésének alapjaival. A gazdasági, ezen belül is elsősorban pénzügyi rendszerek modellezésének egyre nagyobb jelentősége miatt a modern hálózatelmélet oktatása egyre szükségesebbé válik. A kapcsolódó tananyag kiegészítés a képzésben kötelező közgazdaságtani és gazdaságinformatikai tárgyhoz, továbbá rendkívül hasznos lehet azok számára, akik pénzügyi, illetve adatelemzési munkakörben szeretnének elhelyezkedni. A tárgy elsősorban gráfelméleti és valószínűségelméleti alapismeretekre épít, és a hálózatelmélet alapfogalmainra alapozva indítva eljut olyan témakörökig, mint a gazdaság- és társadalomtudományokban is fontos szerepet játszó játékelmélet, döntésemélet vagy például az úgynevezett „fertőzési modellek” vizsgálata.

Motiváció, példák

Egy *hálózat* lényegében pontok (szereplők, entitások) halmaza, melyek élekkel (kapcsolatok, linkek) kapcsolódnak egymáshoz. A matematikában ezeket az objektumokat *gráfnak* nevezik és Leonhard Euler óta¹ rendkívül kutatott terület. Egészen a 20. század második feléig a legtöbb tapasztalati (valós) hálózat főleg kis méretű, szociológusok által konstruált társadalmi kapcsolati (barátok, munkahelyi kapcsolatok) háló volt. A modern számítógépek megjelenése lehetővé tette, hogy olyan óriási méretű hálózatokat lehessen vizsgálni, mint a WWW, online szociális hálók, biológiai rendszerek modell hálózatai, vagy olyan technológiai hálózatok mint egy elektromos ellátó rendszer, illetve különböző úthálózatok.

¹Lényegében az 1700-as évek első felében, a *Königsbergi hidak probléma* kapcsán alakult ki a gráfelmélet.



1. ábra. Különböző valós rendszerek hálózatos reprezentációja

Az első ábrán csak néhány példát megemlítve, *komplex hálózatként* tekinthetünk online szociális hálózatokban megjelenő kapcsolati hálókra (a), sorozatok, könyvek szereplőinek kapcsolataira (b), a világ légitözeledésének hálózataira (c), országok kereskedelmi rendszerére, egy ország elektromos ellátási rendszerére (d), és a sort hosszan lehetne folytatni.

Miért modellezünk hálózatokkal?

Többek között

- Központi szerepük van az információ áramlásban
- Fontos szerepük van „fertőzések” terjedésének vizsgálatában
- Mit vásárolunk, milyen nyelven beszélünk, hogyan szavazunk, milyen oktatásban lesz részünk, sikereink leszünk-e szakmailag, ...

Kulcsfontosságú megérteni:

1. Milyen hálózati struktúrák jelennek meg a társadalomban és gazdaságban²
2. Hogyan hat a hálózati struktúra a szereplők viselkedésére
3. Hogyan befolyásolja a szerkezet a hálózatban végbemenő dinamikus folyamatok viselkedését

Mitől komplexek ezek a hálózatok?

- Sok egymással kapcsolatban álló és egymásra ható szereplő
- Adaptivitás: visszajelzés, szereplők közti kooperáció
- Növekednek és időben változnak (evolúció)
- Nincs linearitás: **Az egész több, mint a részek összessége!**

Gráfelméleti alapfogalmak

A $G := (V, E)$ pár egy **gráf**, ahol $V = \{1, 2, \dots, n\}$ a gráf pontjai (csúcsai) és $E \subseteq V \times V$ a gráf élei. Ha E elemei rendezetlen párok, akkor a gráf irányítatlan, ha rendezettek, akkor irányított gráfról beszélünk. Egy irányítatlan G gráf egyértelműen leírható az $A_G \in \mathbb{R}^{N \times N}$ **szomszédsági mátrixával**, melynek elemei

$$a_{ij} = \begin{cases} w_{ij}, & \text{ha } i \text{ és } j \text{ csúcs összekötött} \\ 0, & \text{különben} \end{cases}$$

Ha G súlyozatlan, akkor $w_{ij} = 1$. Irányított esetben természetesen általában $w_{ij} \neq w_{ji}$. Egy $i, e_1, j, e_2, k, \dots, e_\ell, \ell$ sorozat egy **séta** a gráfban, ha $i, j, \dots, \ell$ pontok, $e_1, e_2, \dots, e_\ell$ élek, továbbá az élek két végpontja a két mellette levő pont a sorozatban. Ha a pontok között nincs ismétlődés, akkor **útról** beszélünk. Ha az élek irányítottak és a sorozatban az élt megelőző pont az élt kezdőpontja, az őt követő pont pedig a végpontja, akkor egy $i - \ell$ irányított utat kapunk. Két pont között a **legrövidebb út** az összes lehetséges út közül a legrövidebb (természetesen több ilyen is létezhet, továbbá súlyozott gráf esetén ez a két pont közti út, amely élein lévő súlyok összege minimális)³ Egy **komponens** olyan részgráf⁴, melynek bármely két pontja közt van út, más szóval **összefüggő**. Irányított esetben az **erősen összefüggő komponens** fogalmát fontos definiálni. Azt mondjuk, hogy egy i és egy j pont elérhetők egymásból, ha $i - j$ és $j - i$ irányított út is létezik. Egy irányított gráf erősen összefüggő komponensei azok a részgráfok, amelyek minden pontjára igaz az elérhetőség⁵. Az $i \in V$ **pont foka** $k_i = \sum_{j=1}^n a_{ij}$, azaz a rá illeszkedő élek száma. Irányított esetben beszélhetünk be-, illetve kimenő fokról, melyek definíció szerint $k_i^{be} = \sum_{j=1}^n a_{ji}$, $k_i^{ki} = \sum_{j=1}^n a_{ij}$.

Az eddigi jelöléseink és definícióink alapján a következő néhány egyszerű észrevételt tehetjük. Az **élek száma**

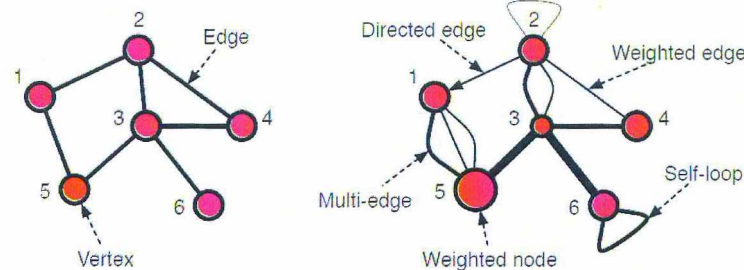
$$m = \frac{1}{2} \sum_{i=1}^N k_i = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_{ij},$$

²Természetesen a biológiai és technológiai rendszerek esetén is hasonló kérdések merülnek fel, de ezek vizsgálata nem tárgya a kurzusnak

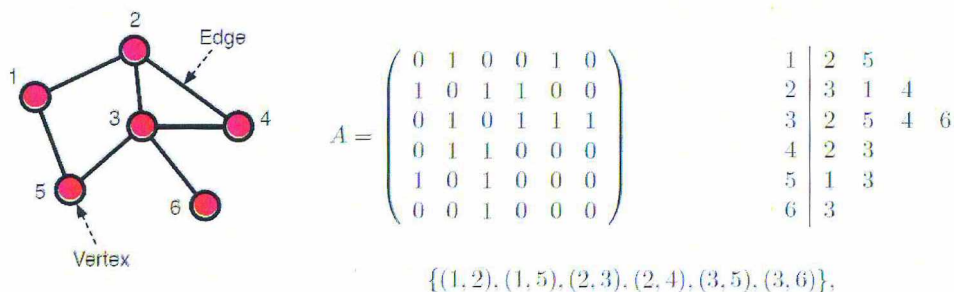
³Ismétlés: *Dijkstra, Ford-Bellman, Floyd-Warshall algoritmus*

⁴Olyan gráf, melynek ponthalmaza és élhalmaza az eredeti G gráf ponthalmazának, illetve élhalmazának részhalmaza

⁵Ismétlés: *mélyégi keresés, Tarján algoritmus*



2. ábra. Gráfok, alapfogalmak. Forrás: Aaron Clauset Network Analysis and modeling course



3. ábra. Gráfrepresentációk: szomszédsági mátrix, szomszédlista, éllista. Forrás: Aaron Clauset Network Analysis and modeling course

vagy másképp, a fokok összege egyenlő kétszer az élek száma. Az [átlagos fokszám](#)

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^N k_i = \frac{2m}{n},$$

továbbá az [élsűrűség](#) (az összes lehetséges él mekkora része van a gráfban)

$$\rho = \frac{m}{\binom{n}{2}} = \frac{\langle k \rangle}{n-1}.$$

További olvasnivaló:

- Ismétlés: gráfelméleti alapok, valószínűség számítás alapok, algoritmusok
- Jackson, Matthew O. Social and economic networks. Vol. 3. Princeton: Princeton University Press, 2008 – I. fejezet
- Newman, Mark EJ. "The structure and function of complex networks." *SIAM review* 45.2 (2003): 167-256. – I. és II. fejezet

2. A hálózatkutatás néhány alapfogalma

Jelölje ℓ_{ij} a [legrövidebb utat](#) a hálózatban i és j pontok között. Ekkor $\Delta = \max_{i,j} \ell_{ij}$ a hálózat [átmérője](#), azaz az összes legrövidebb utak hosszának a maximuma. Hálózatok fontos jellemzője az [átlagos úthossz](#), ami a pontpárok közti legrövidebb úthosszak átlaga:

$$\langle \ell \rangle = \frac{1}{\binom{n}{2}} \sum_{i,j} \ell_{ij}.$$

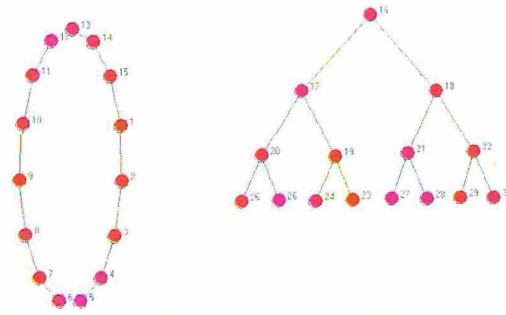
Érdemes gondolkozni rajta, hogy valós hálózatokban ez miért érdekes, milyen információt ad. A későbbiek folyamán erre a kérdésre még visszatérünk.

Fokszámeloszlás

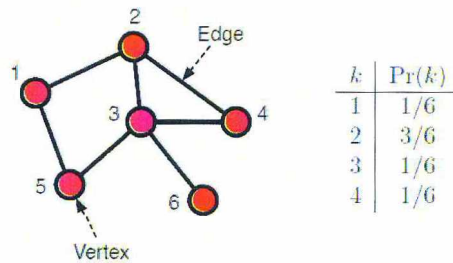
Komplex hálózatok egy újabb fontos globális jellemzője az úgynevezett fokszámeloszlás, ami azt mutatja meg, hogy mennyi az adott fokszámú pontok aránya a hálózatban az összes pont-hoz viszonyítva (ld. 3. ábra). Másképpen fogalmazva, mennyi annak a valószínűsége, hogyha véletlenül kiválasztjuk a hálózat egy pontját, annak a foka éppen k ,

$$P(k) = \mathbb{P}(\text{egy véletlenül választott pont foka } k).$$

Miért érdekes egy hálózat fokszámeloszlása? Milyen fokszámeloszlást követnek a valós hálózatok? Van-e ebben hasonlóság? **Kulcsfontosságú fogalom**, a későbbiekben részletesen tárgyalni fogjuk.



4. ábra. Mennyi egy n pontú kör és egy n pontú bináris fa átmérője?



5. ábra. Fokszámeloszlás. Forrás: Aaron Clauset Network Analysis and modeling course

Centralitási mértékek

Hálózatok vizsgálatánál alapvető kérdés, hogy melyek a hálózat „fontos” pontjai? Szubjektív lehet, hogy mit értünk fontosság alatt, ugyanakkor strukturális szempontból fontosnak tekintenénk egy pontot, ha például

- magas fokszerű
- a hálózat „központjában” van
- valamilyen dinamikus folyamat szempontjából fontos szerepet játszik (pl. fertőzés terjedés, véletlen bolyongás)

A **centralitás** fogalma a pontoknak ilyen tulajdonságait próbálja megfogni. Mivel nem egységes, hogy mit is értünk egy pont fontossága alatt, több különböző definíció (és kiszámító algoritmus) született. Összességében egy centralitási mérték definíciójánál az általános alapelv a következő:

„Egy pont minél centrálisabb annál fontosabb, minél kevésbé centrális annál kevésbé fontos.”

Fokszám centralitás

Az egyik legegyszerűbb centralitási mérték azon a feltevésen alapul, miszerint minél nagyobb fokszerű egy pont (gondolhatunk például, hogy egy ismertségi hálózat esetén annál fontosabb valaki, minél több ismerőse van), annál fontosabb, azaz $C_{\text{deg}}(i) = k_i = \sum_{j=1}^n a_{ij}$. Irányított hálózatok esetén pedig megkülönböztetjük a befok, illetve a kifok centralitást.

Betweenness (köztiség) centralitás

Két pont között általában több olyan út is létezik, aminek a hossza minimális. Egy adott pont fontosságát a fogalom úgy ragadja meg, hogy két pont közti legrövidebb utak hány százalékában van jelen ez a pont, összegezve az összes pontpárra. Másképpen, két tetszőleges pont közti legrövidebb utak közül véletlenszerűen választunk, mekkora eséllyel haladunk egy át egy adott harmadik ponton. Formálisan a (adott harmadik) pont **betweenness centralitása**

$$BC(k) = \sum_{i \neq k \neq j} \frac{\sigma_{ij}(k)}{\sigma_{ij}},$$

ahol σ_{ij} a i és j közötti legrövidebb utak száma, $\sigma_{ij}(k)$ pedig azon legrövidebb $i - j$ utak száma, melyek átmennek k -n.

Szorgalmi feladat: gondolkozzunk egy $O(nm)$ futási idejű BC számító algoritmuson (n a pontok, m pedig az élek száma)

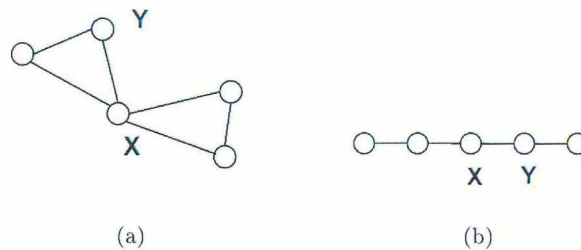
Closeness (közelség) centralitás

A **closeness centralitás** azt mutatja meg, hogy mennyire van a hálózat „középpontjában” egy pont, pontosabban, hogy átlagosan milyen hosszúak a pontból a hálózat többi pontjába induló legrövidebb utak. Definíció szerint

$$C(i) = \frac{n-1}{\sum_{j \neq i} \ell_{ij}},$$

ahol ℓ_{ij} az i és j közti legrövidebb út hossza⁶.

⁶Számolás: a Floyd-Warshall algoritmus segítségével



6. ábra. Mennyi X és Y betweenness, closeness és harmonikus centralitás értéke az egyes hálózatok esetén?

Harmonikus centralitás

Két alapvető probléma merül fel closeness centralitással kapcsolatban. Az első, hogy a valós hálózatok átmérője általában kicsi, vagyis a closeness értékek csak egy szűk tartományban változnak, vagyis nem adnak túl sok információt a pontok hálózatban betöltött szerepeinek különbségeiről. Másik gond, hogy nem összefüggő hálózat esetén nem számolható, ugyanis bizonyos ℓ_{ij} értékek nullák. A **harmonikus centralitás**, mely kiküszöböli ezeket a problémákat, definíció szerint

$$C_h(i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{\ell_{ij}},$$

ahol $\ell_{ij} = \infty$, ha nincs a gráfban $i - j$ út.

Sajátérték centralitás

A **sajátérték centralitás** bevezetésének alapgondolata, hogy **nem minden szomszéd egyforma súllyal számít** a centralitás kiszámításánál. Az alapfeltevés, hogy minél fontosabb egy szomszéd, annál nagyobb mértékben járul hozzá az adott pont fontosságához. A centralitást kiszámító rekurzív formula

$$x_i^{(t+1)} = \sum_{j=1}^n w_{ij} x_j^{(t)},$$

illetve átírva mátrixos formában:

$$A\mathbf{x} = \lambda_1 \mathbf{x},$$

ahol λ_1 az A mátrixhoz tartozó legnagyobb sajátérték.⁷

PageRank

A sajátérték centralitás egyik problémája, hogy irányított gráfok esetén nem túl hatékony. Általánosan, ha egy pont nincs benne egy erősen összefüggő komponensben (egymaga alkot egy komponenset), akkor a sajátérték centralitása nulla lesz. Sőt, ha a gráfban nincs irányított kör, akkor csak lesz nem nulla egy pont centralitása, ha a kifoka nulla. A **PageRank centralitás** kiküszöböli ezt a problémát, úgy, hogy minden pont eleve kap egy kis centralitás értéket. (A modellt a véletlen bolyongás kapcsán részletesebben fogjuk tárgyalni; A PageRank a Google keresőmotorjának is egy alapeleme, lényegében egy interneten „véletlen szörföző” viselkedését

⁷ld. részletesebben: *Perron-Frobenius tétel*

szimulálja ⁸.) A PageRank rekurzió:

$$PR(i) = \frac{1-\lambda}{n} + \lambda \sum_{j \in N^+(i)} \frac{PR(j)}{k^{ki}(j)},$$

ahol $\lambda \in [0, 1]$ egy paraméter (ugró faktor), $N^+(i)$ az i pont azon szomszédjai, amely pontokból megy él i -be. Mátrixos formában, A PageRank mátrix felírható

$$R = \lambda P + (1 - \lambda)U$$

alakban, ahol U egy $n \times n$ mátrix, melynek minden eleme $1/n$. Mivel a mátrixhoz tartozó legnagyobb sajátérték 1 (ld. *Perron-Frobenius tétel sztochasztikus mátrixokra*), így a PageRank vektor (mely az egyes pontok PR értékeit tartalmazza) a következő sajátérték egyenlet megoldása:

$$\mathbf{PR} = \mathbf{PR}R = \mathbf{PR}(\lambda P + (1 - \lambda)U).$$

Ezt átalakítva

$$\begin{aligned} \mathbf{PR} &= \mathbf{PR}R = \mathbf{PR}(\lambda P + (1 - \lambda)U) = \\ &= \lambda \mathbf{PR}P + (1 - \lambda)\mathbf{PR}U = \\ &= \lambda \mathbf{PR}P + (1 - \lambda)\mathbf{PR}\mathbb{1}\mathbb{1}^T \frac{1}{N} \\ &= \lambda \mathbf{PR}P + (1 - \lambda)\mathbb{1}^T \frac{1}{N} \end{aligned}$$

felhasználva, hogy $U = \mathbb{1}\mathbb{1}^T \frac{1}{N}$ és $\mathbf{PR}\mathbb{1} = 1$. Innen kapjuk, hogy

$$\mathbf{PR} = \frac{1-\lambda}{N} \mathbb{1}(I - \lambda P)^{-1},$$

majd ezt átírhatjuk a

$$\mathbf{PR} = \frac{1-\lambda}{N} \mathbb{1} \sum_{n=0}^{\infty} (\lambda P)^n$$

alakra. Ez egy módszert is ad a PageRank számolására, nevezetesen a [hatványmódszert](#), miszerint P mátrix hatványait addig számoljuk, amíg a PageRank értékek nem konvergálnak, azaz

$$|\mathbf{pr}^{(k)} - \mathbf{pr}^{(k-1)}| < \varepsilon.$$

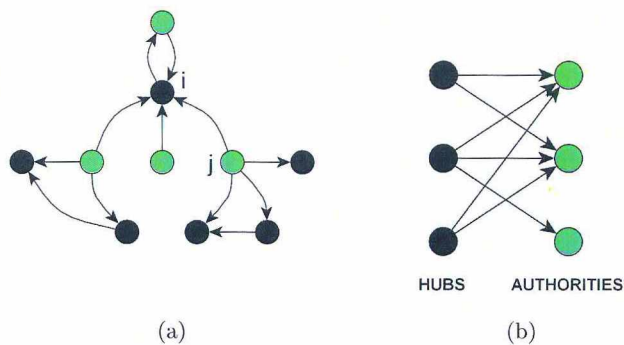
HITS (Hyperlink Induced Topic Search)

Join Kleinberg nevéhez fűződik az algoritmus ami ⁹ a [PageRank egy "finomított" változatának tekinthető](#). A gráf pontjainak rangsorolásánál megkülönböztet ún. Hub, illetve Authority típusú pontokat, miszerint

- Jó Authority pont, amibe sok link mutat
- Jó Hub az, amiből sok link megy jó Authority pont felé

⁸Brin & Page, *Computer networks and ISDN systems*, 1998

⁹Kleinberg, *Journal of the ACM*, 1999



7. ábra. (a) Az i pont PR értéke a szomszédjai (zöld pontok) PR értékeiből számítható. (b) Hubok és authority-k egy gráfban.

A **HITS algoritmus** a következő:

Input G irányított gráf

Output a pontok Hub és Authority értékei

- 1: Kezdetben minden pont értéke 1
- 2: **repeat**
- 3: **for all** hub $i \in H$ **do**
- 4: $h_i = \sum_{j \in F(i)} a_j$ % $F(i)$: azon pontok, melyekből megy el i -be
- 5: **end for**
- 6: **for all** authority $i \in A$ **do**
- 7: $a_i = \sum_{j \in B(i)} h_j$ % $B(i)$: azon pontok, melyekbe megy el i -ből
- 8: **end for**
- 9: **until** konvergál
- 10: Normálás

Szorgalmi feladat: Írjuk át a HITS-et mátrixegyenletek formájába és gondoljuk át a kapott sajátérték egyenleteket.

Szoftverek hálózatelemzéshez

Néhány ingyenes program hálózat vizualizációhoz és elemzéshez

- Cytoscape (GUI)
- Gephi (GUI)
- iGraph (R, C++, Python)

Feladatok:

1. Egy hálózat (pl. a Zachary-féle karate klub) pontjainak különböző centralitásai és vizualizáció.
2. Ismétlés: mátrixok, egyenletrendszerek, sajátérték-sajátvektor

További olvasnivaló

- Jackson könyv 2. fejezet

3. Véletlen gráfok, gráfmodellek

Véletlen gráfok

- Milyen általános **közös tulajdonságai** vannak „tipikus” gráfoknak?
- Tudjuk-e valamilyen **modellel közelíteni** a valóságban megjelenő hálózatokat?
- A különböző területeken (társadalom, gazdaság, biológia, technológia) megjelenő hálózatok modelljei között mik a **legfontosabb különbségek/hasonlóságok**?

Fontos kérdés a hálózatkutatásban, hogy **mely gráfok az érdekesek** és mi alapján különböztetjük meg az érdekeset a nem érdekestől? Egy referencia pont hálózatelemzésnél a **véletlen gráf**. A gráfmodelleket két fő típusba sorolhatjuk:

- **Konstruktív modellek**: adott szabályok mentén bizonyos típusú hálózatot hoz létre; pl. „preferential attachment” algoritmus
- **Generatív modellek**: szabad paraméterek felhasználásával generál hálózatot; pl. az élvalószínűség adott

Az Erdős-Rényi modell ¹⁰

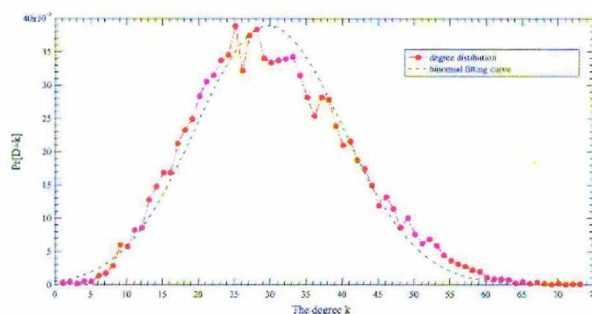
A modell egy olyan n pontú gráfot definiál, melyet $G(n, p)$ -vel jelölünk, amelyben **minden élt** $p \in [0, 1]$ **valószínűséggel húzunk be** (és $1 - p$ -vel nem húzunk be), minden pontpár esetén egymástól függetlenül. A következő néhány egyszerű észrevételt tehetjük a gráffal kapcsolatban:

- Az élek száma várható értékben: $\binom{n}{2}p$
- Átlagos fokszám: $\bar{k} = (n - 1)p$
- **fokszámeloszlás**:

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

azaz **binomiális**.

Fontos és rendkívül intenzíven vizsgált matematikai modell. (További olvasnivaló a véletlen gráfról: Bollobás Béla. Random graphs. Springer New York, 1998.; $G(n, m)$ modell, melyet Gilbert definiált még Erdős és Rényi előtt)



8. ábra. Egy generált ER gráf fokszámeloszlása és az illesztett binomiális eloszlásgörbe

¹⁰Erdős & Rényi, 1959

A Konfiguráció modell

Egy másik fontos gráfmodell a **konfiguráció modell**, melyet $G(n, \mathbf{k})$ -val jelölünk, ahol n a gráf pontjainak a száma, míg $\mathbf{k} = (k_1, k_1, \dots, k_n)$ a gráfhoz tartozó **fokszámsorozat** ($\sum_i k_i$ páros!). Speciális esetben, ha minden k_i egyenlő, akkor egy reguláris gráfot kapunk; ha k_i *Poisson-eloszlású* véletlen változó c/n várható értékkel, akkor a $G(n, p)$ -hez „kerül közel” a konfiguráció. A kérdés az, hogyan generálnánk le a gráfot, ha adott n és $\mathbf{k}$? A modellben a következő valószínűségek definiálásával tesszük ezt meg:

$$\mathbb{P}(i \text{ és } j \text{ összekötött}) = \frac{k_i k_j}{2m}. \quad (1)$$

Ha adott egy gráf, és ezzel együtt annak fokszámsorozata, akkor ezt a fokszámsorozatot használva egy olyan véletlen gráfot tudunk generálni, mely az eredeti gráf fokszámait megőrzi. Ezáltal a modell lehetőséget biztosít valós hálózatok vizsgálatára azáltal, hogy kvantitatívan meghatározza, hogy *egy megfigyelt hálózati struktúra mennyire pusztán a fokszámok következménye*. A megfigyelt mintázatok a hálózatokban mennyire magyarázhatók pusztán a fokszámok ismeretében? A konfiguráció modell fontos eszköz lesz közösségek vizsgálatánál, nevezetesen a *Newman-modularitás* bevezetésénél (4. fejezet).

Kisvilág-gráfok

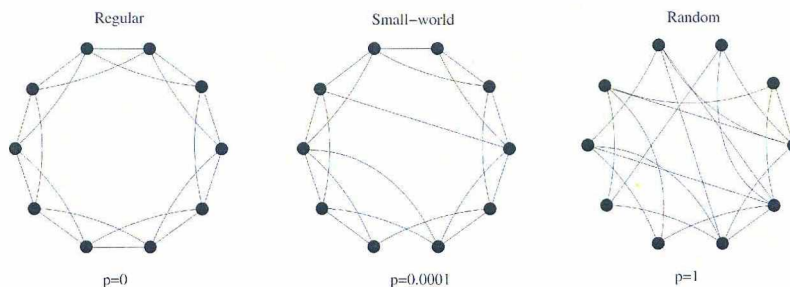
Stanley Milgram (1933-84) kísérlete 1967-ben az volt, hogy véletlenül kiválasztott emberek próbáljanak egy levelet eljuttatni egy általuk vélhetően ismeretlen bostoni orvosnak. A levelet egy ismerősüknek továbbíthatják néhány alapvető információval a bostoni orvostól. Ha az ismerősük személyesen ismeri az orvost, akkor egyenesen postázhatja neki a levelet, ha nem, akkor egy ismerősnek kell továbbítani, tovább folytatva a levelezési láncot. A kísérlet eredménye az volt, hogy a 64 célba érkező levél az USA 64 különböző pontjáról átlagosan 5.5 levélváltás után célba ért.

A kisvilág gráfok legfontosabb jellemzői:

- **Kicsi átmérő:** A legtávolabbi pontok sincsenek túl messzire egymástól (ld. Milgram-kísérlet)
- **A háromszögek száma nagy** („A barátom barátainak nagy részét én is ismerem”). Ezzel kapcsolatos fogalom az ún. klaszterezettség (clustering coefficient), mely definíció szerint

$$C = \frac{3 \times \text{háromszögek száma}}{\text{összefüggő ponthármak száma}}$$

Szorgalmi feladat: számoljuk ki a háromszögek várható számát a véletlen gráfban, illetve nézzünk meg néhány valós társadalmi hálózat klaszterezettségét.



9. ábra. Reguláris gráf ($p = 0$), kisvilág gráf ($p > 0$), véletlen gráf ($p = 1$). Forrás: Watts & Strogatz, *Nature*, 1998

A Watts-Strogatz modell ¹¹

A modell lényegében egy algoritmus, mely kisvilág tulajdonságú gráfot generál a következők szerint:

1. Kiindul egy 4-reguláris gráfból (minden pont foka 4)
2. Minden élt p valószínűséggel átdrótoz (azaz (i, j) él esetén választunk véletlenül egy k pontot, p valószínűséggel töröljük (i, j) -t és behúzzuk (i, k) -t)

Ez eljárás egy $\sim \log(n)$ átmérőjű gráfot hoz létre, továbbá nagy klaszterezettség érték jellemző.

A Barabási-Albert modell ¹²

A konstrukció, lényegében azt modellezi, hogy időben **hogyan fejlődhet ki egy hálózat**. A modell az ún. **preferenciális kapcsolódáson** (*preferential attachment*) alapul és a következő egyszerű algoritmussal adható meg:

1. kezdetben egy összefüggő G_0 gráf n_0 ponton
2. t időpontban hozzáadunk G_t -hez egy új v pontot úgy, hogy

$$\mathbb{P}(v\text{-t összekötjük egy meglévő } i\text{-vel}) = \frac{k_i}{\sum_j k_j}.$$

A 6. fejezetben részletesen fogjuk tárgyalni a modellt.

¹¹Watts & Strogatz, *Nature*, 1998

¹²Barabási & Albert, *Science*, 1999

4. Közösségek hálózatokban

A hálózatra vonatkozó egyszerű mértékek, mint az átlagos fokszám, a fokszámeloszlás, a klaszterezettség vagy az átlagos úthossz sok információt adnak a vizsgált rendszerről **de elrejtetik az eloszlások heterogenitását**. Például azt, hogy a háromszögek vagy a fokszámok nem egyenletesen helyezkednek el a hálózatban, hanem bizonyos részeken magas a háromszögek száma vagy a fokszám, a hálózat más részein pedig kis fokú pontokat és fa szerű részgráfot látunk (10. ábra).

Természetesen merülnek a kérdés, hogy mi a hálózat „magasabb szintű szerveződésének” mintázata? Gondolhatunk erre úgy, hogy egy kisméretű hálózat esetén ez szemmel látható (feltételezve, hogy aki lerajzolta a gráfot ismerte ezeket a mintázatokat és szemléletes ábrát adott (10-11. ábra); ugyanakkor nagyméretű (akár több ezer, vagy milliós nagyságrendű pont és él) hálózatok esetén ezen mintázatok meghatározásához kvantitatív eszközök szükségesek.

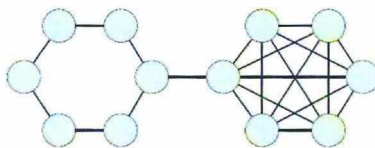
Asszortativitás

A hálózat pontjai között bizonyos **attribútumok** (tulajdonságok, jellemzők) azonosságára megélő élék világítanak rá, például ismerőseink egy részének közös jellemzője a középiskola, ahová jártunk, más részének a munkahely vagy egyéb közösség, ahová tartozunk. Társadalmi hálózatokban az is jellemző, hogy olyan ismerőseink/barátaink vannak akik valamiben **hasonlítanak** ránk, például életkor, beszélt nyelv, születési hely, végzettség, anyagi helyzet, érdeklődési kör, stb. A hálózatkutatás szempontjából érdekes és fontos kérdés, hogy egy ilyen hálózatban

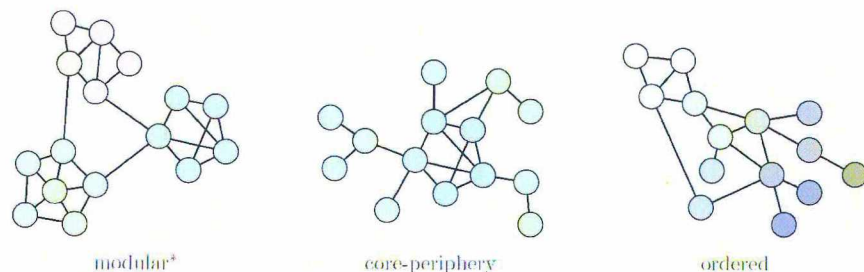
- az él a hasonlóság miatt létezik (asszortatív kapcsolódás)
- az él létezése (ismertség/barátság) miatt válnak hasonlóvá bizonyos attribútumok, pl. érdeklődési kör vagy politikai beállítottság

A hálózat heterogenitása globális szinten adódhat ilyen típusú asszortatív kapcsolódás miatt, illetve abból, hogy a vizsgált pontok olyan diszjunkt csoportokra oszlanak melyek egyenként meglehetősen homogének. Hálózatokban az ilyen mintázatot moduláris-, vagy **közösségszerkezetnek** hívják, ahol a közösségek homogén építő kövei egy amúgy heterogén struktúrának. Gráfos megközelítésben azt mondhatjuk, hogy a közösségek olyan részgráfok, amelyekben a kapcsolatsűrűség nagy, míg az egyes közösségek közötti menő élék száma relatíve kicsi.

Megjegyezzük, hogy attribútumok hasonlósága alapján generált hálózat modellezésére egy lehetőség az úgynevezett *Fitness modell* (a válogatott fejezetekben még visszatérünk rá.)



10. ábra. Háromszögek száma nagy, de azok csak a hálózat egy részén találhatók. Forrás: Aaron Clauset, Network analysis and modelling course



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

A modularitás függvény

Hálózatok közösségszerkezetének vizsgálatához egy lehetséges eszköz az úgynevezett **modularitás függvény** használata. A függvény tulajdonképpen azt méri, hogy **az adott gráf mennyire tér el egy ugyanolyan fokszámeloszlású véletlen gráftól**, azaz

Modulartás = $\#\{\text{közösségen belüli élek}\} - \mathbb{E}[\#\{\text{közösségen belüli élek}\}]$ egy, a hálózathoz valamiben hasonló véletlen gráfban],

ahol $\#$ a halmaz elemszámát, $\mathbb{E}$ -vel pedig a várható értéket jelöli. A **Newman-modularitás**¹³, mely az első ilyen függvény volt közösségszerkezet vizsgálatára a

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

függvény, ahol δ a *Dirac-delta függvény* ($\delta(C_i, C_j) = 1$, ha $i = j$ és 0 különben), p_{ij} pedig annak a valószínűsége, hogy i és j össze van kötve egy véletlen (null-modell) gráfban. A kérdés persze az, hogy mi legyen ez a véletlen null-modell gráf? Az Erdős-Rényi $G(n, p)$ véletlen gráf általában nagyon „messze van” a valós hálózatoktól, vagyis ezzel összehasonlítani egy valós gráfot nem biztos, hogy célszerű. Ha a konfiguráció modellt használjuk és ha az eredeti gráf fokszámsorozata $(k_1, k_2, \dots, k_n)$ akkor $p_{ij} = k_i k_j / 2m$, így

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

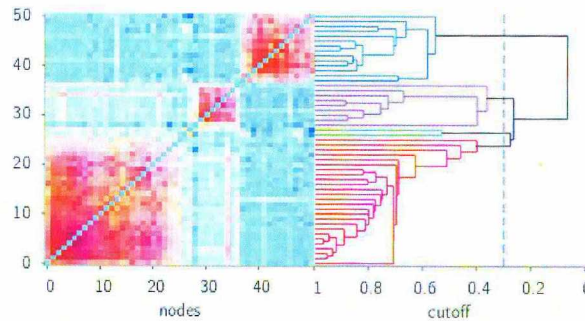
A cél pedig a **pontok osztályozása** (szétosztása) $C_1, \dots, C_k$ ($k = ?$) osztályokba (klaszterekbe), hogy **Q minél nagyobb (maximális) legyen**, ezáltal meghatározva a hálózat közösségszerkezetét.

Néhány algoritmus modularitás maximalizálásra

Ha S a G gráf pontjainak **összes lehetséges partícionálása (klaszterezése)** akkor egy $f : S \rightarrow \mathbb{R}$ függvény méri egy adott $P \in S$ felosztás „jóságát” (Persze kérdés mitől jó a felosztás, vagy még inkább az, hogy mit értünk jó felosztás alatt¹⁴). A Newman-féle Q modularitás függvény

¹³Newman, *Physical Review E*, 2004

¹⁴Itt említjük meg Jon Kleinberg egy eredményét, miszerint néhány egyszerű elvárást (axiómát) felállítva megmutatható, hogy nincs olyan klaszterező függvény amely minden elvárásnak eleget tenne. (Kleinberg, Jon. "An impossibility theorem for clustering." *Advances in neural information processing systems* (2003): 463-470.)



12. ábra. Közösségek és hierarchikus szerkezet

egy lehetséges ilyen függvény, az alapgondolat pedig az, hogy Q „minél nagyobb, annál jobban klasztereztünk”. Vegyük észre, hogy S mérete exponenciálisan nagy (miért?), továbbá Q -t maximalizálni NP-nehéz probléma, de mégis léteznek jól működő heurisztikák, továbbá a témában továbbra is nagy számban jelennek meg új tudományos eredmények. Itt egy mohó algoritmust említünk (illusztrálva a 12. ábrán):

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

Az eljárás implementálásra számos különböző technika létezik (ld. pl. *single linkage*, *average linkage*, *k-means* etc.). Megemlítünk még néhány további lehetőséget közösségek keresésére:

- Modularitás optimalizálás számos változata
- Más kiértékelő függvények használata (Mit tartunk fontosnak közösségkeresés esetén?)
- Sztochasztikus blokk modell (a következő fejezetben 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

5. A sztochasztikus blokk modell

Maradva a hálózatok magas szintű strukturális mintázatainak feltárása témakörben egy **véletlen generatív modellt** fogunk megvizsgálni, ami a véletlen gráfmodell egy finomabb változatának tekinthető. Ennek érdekében egy $\mathbb{P}(G|\theta)$ véletlen eloszlást fogunk definiálni az összes n pontú gráf halmazán, ahol θ kódolja az adott mintázatot, $\mathbb{P}(G|\theta)$ pedig megmondja, hogy milyen gyakran látjuk ezt a mintázatot a hálózatokban. Egyrészt ha adott θ , tudunk generálni egy gráfot (a véletlen segítségével) ezzel a mintázattal, másrészt visszafelé, ha adott G valós vagy szintetikus hálózat, meg tudjuk határozni a legvalószínűbb θ -t, ami a hálózat létrejöttét kontrollálta. De mire is jó ez?

- Explicit generálhatunk adott hálózatokat (nem egy algoritmus)
- Strukturával kapcsolatos hipotéziseket tudunk ellenőrizni vele
- Modellek „jóságát” ellenőrizhetjük (mennyire közelíti a modell a valóságot?)
- Hiányzó minták vagy jövőbeli struktúrák feltárására is egy lehetőség

Modell definíció

A modell először a 80-as években jelent meg szociológia témájú folyóiratban¹⁵. Napjainkban számos alkalmazási területe van, gyakran használják gépi tanulásban, komplex rendszerek vizsgálatához és a statisztikus fizikában. Léteznek általánosításai irányított és súlyozott gráfokra is, de itt csak a legegyszerűbb esetet tárgyaljuk.

Az SBM egyszerűen egy $\theta = (k, z, M)$ hármas, ahol

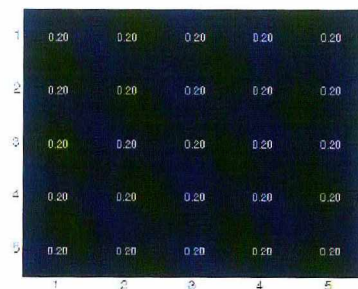
1. k a **csoportok (közösségek/ pont osztályok) száma** a hálózatban
2. z egy n hosszú vektor, ahol z_i megadja, hogy az i **pont melyik csoportba tartozik**
3. M egy $k \times k$ blokk mátrix, ahol M_{uv} megadja annak **a valószínűségét, hogy egy u csoportbeli és egy v csoportbeli pont kapcsolódik egymáshoz**

A modellben először k -t kell fixálni, továbbá megjegyezzük, hogy az azonos csoportban lévő pontok sztochasztikusan ekvivalensek. Hálózatot SBM-mel a következő lépésekkel tudunk generálni:

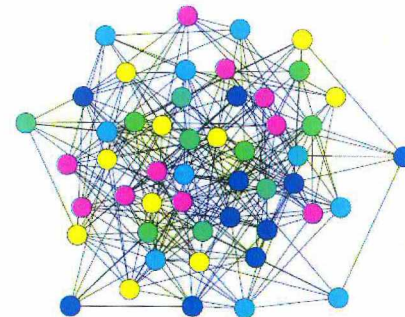
1. Megadjuk a (k, z, M) hármas
2. Minden (i, j) pontpárra feldobunk egy érmét: M_{z_i, z_j} valószínűséggel behúzzuk (i, j) élt, $1 - M_{z_i, z_j}$ -vel nem

Szemben a $G(n, p)$ -vel aminek két paramétere van, és $G(n, k)$ -val aminek $1 + n$, az SBM-nek $1 + n + \binom{k}{2}$ paramétere van, ez a nagy szabadsági fok pedig lehetőséget ad rengeteg különféle hálózat generálására. A $G(n, p)$ Erdős-Rényi gráf lényegében az SBM egy speciális esete, ahol $k = 1$, azaz egy csoport van, $M \equiv p$, vagyis $M_{z_i, z_j} = p$ mivel $z_i = z_j$ minden pontpárra.

¹⁵ Holland, Laskey, and Leinhardt, „Stochastic blockmodels: First steps.” *Social Networks*, 5(2), 109-137 (1983)

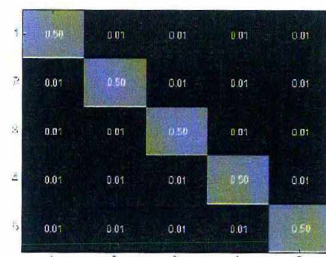


random graph block matrix

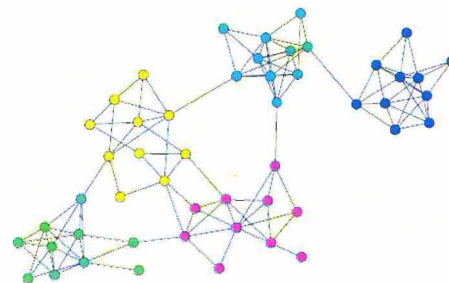


random graph

13. ábra. A blokkmátrix és a realizált gráf (Forrás: Aaron Clauset, Network Analysis and Modelling course)

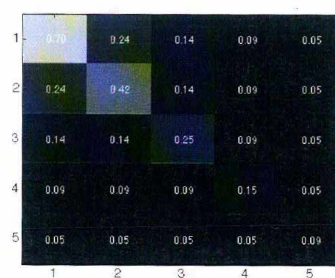


assortative block matrix

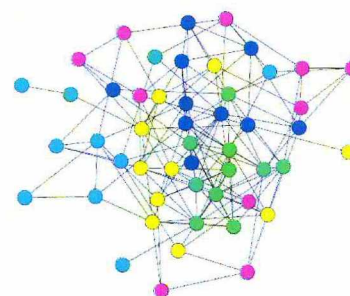


assortative communities

14. ábra. Gráf közösségszerkezettel (Forrás: Aaron Clauset, Network Analysis and Modelling course)



core-periphery block matrix



core-periphery network

15. ábra. Gráf mag-periféria szerkezettel (Forrás: Aaron Clauset, Network Analysis and Modelling course)

Maximum likelihood becslés

Adott egy (valós) hálózat, a kérdés az mi az a blokkmodell, ami legjobban közelíti ezt a struktúrát? Feltéve M -et és z -t (azaz θ -t) a **likelihood függvény**

$$\mathcal{L}(G|M, z) = \prod_{(i,j) \in E} M_{z_i, z_j} \prod_{(i,j) \notin E} 1 - M_{z_i, z_j}$$

A cél, hogy válasszuk M -et és z -t úgy, hogy $\mathcal{L}$ maximális legyen. Példaként tekintsük a következőt. Legyen N_u a pontok száma az u csoportban (blokkban), ekkor $N_{uv} = N_u N_v$ a lehetséges élek száma u és v között, továbbá legyen E_{uv} a ténylegesen létező élek száma (megfigyelés) u és v blokkok között. Ekkor az egyszerű becslésünk az élvalószínűsége a blokkok között: $\hat{M}_{uv} = E_{uv}/N_{uv}$. A likelihood függvényünk így

$$\mathcal{L}(G|M, z) = \prod_{u,v} M_{uv}^{E_{uv}} (1 - M_{uv})^{N_{uv} - E_{uv}} = \prod_{u,v} \left(\frac{E_{uv}}{N_{uv}} \right)^{E_{uv}} \left(1 - \frac{E_{uv}}{N_{uv}} \right)^{N_{uv} - E_{uv}}$$

Vegyük mindkét oldal logaritmusát, ekkor egyszerű számolással

$$\log \mathcal{L} = \sum_{u,v} E_{uv} \log E_{uv} + (N_{uv} - E_{uv}) \log(N_{uv} - E_{uv}) - N_{uv} \log N_{uv}$$

adódik, ami csak a z által indukált N_{uv} és E_{uv} mennyiségektől függ csak.