

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR

VÉLETLEN HÁLÓZATOK KLASZTEREZETTSÉGE

Babolcsay Barbara

Matematika BSc, Alkalmazott matematikus szakirány

Témavezető:
dr. Csiszár Villő



2023

Köszönetnyilvánítás

Szeretném megköszönni konzulensem, dr. Csiszár Villó
egész féléves támogatását, ötleteit és meglátásait.

Tartalomjegyzék

1. Bevezetés	3
2. A csoporterősség mértékét jellemző mutatók	4
2.1. Globális és átlagos lokális klaszterezettségi együttható	4
2.2. Diadikus klaszterezés háromszögmultiplicitással	6
2.3. Modularitás	8
3. Állandó csúcyszámú véletlen gráf generálása szabályozható klaszterezettségi együtthatóval	10
3.1. Newman általános algoritmus	10
3.2. A CONF-1 algoritmus	11
3.3. A CONF-2 algoritmus	12
3.4. A THROW-1 algoritmus	13
3.5. Empirikus eredmények	14
4. Valós hálózatok modellezése növekvő csúcyszámú véletlen gráffal	16
4.1. Történeti áttekintés, út a Barabási-Albert-modellig	16
4.2. A BA-modell és a ráépülő más modellek klaszterezettségi együtthatója	18
4.2.1. Az LCD-modell	18
4.2.2. Buckley-Osthus-modell/Móri modell	22
4.2.3. Egyéb modellek	26
5. A PA-osztály	27
5.1. A preferenciális kapcsolódás általánosítása	27
5.2. A T-osztály	29
5.3. A Polinomiális modell	30
5.4. Empirikus eredmények a Polinomiális modellről	34
6. Összegzés	37

1. Bevezetés

A 20. század második felében a természettudományokban és a szociológiában is előtérbe került a nagy kiterjedésű hálózatok, mint a társadalmi háló vagy a sejtek közti kapcsolatok vizsgálata. Az empirikus jellemzés mellett felmerült az igény ezen hálózatok minél valósághűbb matematikai modellezésére is, ez pedig az internet születése által újabb lendületre kapva az ezredfordulóra kinőtte magát a hálózatok kutatás területévé.

A kutatások egyik legfontosabb eredménye az volt, hogy a körülöttünk lévő világ rengeteg hálózata matematikai szempontból nagyon hasonló jellemzőkkel rendelkezik. Ez azt feltételezi, hogy kialakulásukat is hasonló belső szabályok alakították, így van esélyünk közös modellt találni, amely ezt az önszerveződést is demonstrálja. Ezen célok mentén kezdték el a véletlen gráfok különböző fajtáit alkalmazni a nagy hálózatok leképezésére.

A valóságban megfigyelt jellemzők közé tartozik a szoros kis csoportok kialakulására való viszonylag magas hajlam – gondoljunk csak a baráti társaságokra vagy az egy érdeklődési körbe tartozó, egymásra utaló weboldalakra. Természetesen a választott matematikai modelleknek ebből a szempontból is igazodnia kell a megfigyelésekhez. Ezen tulajdonságuk mérésére több mutatót is bevezettek, ezek közül a legszélesebb körben használtak a globális, illetve átlagos lokális klaszterezettségi együttható.

Ezen dolgozat célja röviden tárgyalni néhány mutatót és a közöttük lévő összefüggéseket, illetve bemutatni olyan véletlen modelleket, amelyek a globális és átlagos lokális klaszterezettségi együtthatójuk minél valóságosabbra hangolását célozzák. Áttekintünk állandó csúcsszámú véletlen gráfokat alkalmazó módszereket, ahol külső megszorítást alkalmazva pontosan beállítható lesz az elvárt klaszterezettségi együttható, majd megvizsgáljuk, hogy milyen lehetőségeink vannak növekvő csúcsszámú véletlen gráfok alkalmazásával, amikben az önszerveződés alakítja ki a végleges hálózatunkat. A bemutatott modellek elméleti analíziséről eredményeket és bizonyításokat láthatunk, alátámasztásukra pedig saját empirikus eredményeket is.

2. A csoporterősség mértékét jellemző mutatók

2.1. Globális és átlagos lokális klaszterezettségi együtt-ható

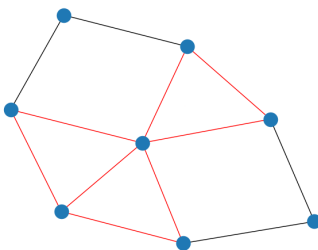
Az alapvető kérdésünk, amire számszerű választ kaphatunk ezen mutatóink segítségével az, hogy ha két csúcs rendelkezik közös szomszédal, akkor mekkora eséllyel fut köztük is él. Gyakorlati példán szemléltetve, ha két ember rendelkezik közös ismerőssel, mennyire valószínű, hogy ők is ismerik egymást, mennyire sűrű a társadalmi háló szövete ilyen tekintetben.

A globális klaszterezettségi együtt-ható a kérdést felülről, a teljes gráfot nézve vizsgálja:

2.1.1. Definíció (Globális klaszterezettségi együtt-ható). Legyen $G(V, E)$ egyszerű gráf és definiáljuk benne a globális klaszterezettségi együtt-hatót a következőképpen:

$$C_1(G) = \frac{3 \cdot \text{a háromszögek száma a gráfban}}{\text{a tripletek száma a gráfban}}.$$

A definícióban szereplő triplet egy olyan élpár, amelyek egy csúcsban találkoznak (magyarul a találó cseresznye elnevezés is használt). Ezt zártnak nevezzük, ha nem kapcsolódó végük között is fut él, nyitottnak, ha nem, tehát C_1 úgy is definiálható, mint a zárt tripletek aránya a gráfban az összeshez mérten.



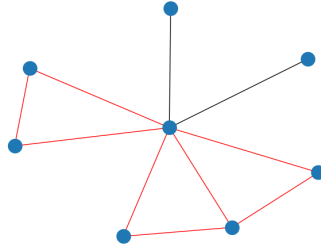
1. ábra. Zárt tripletek egy egyszerű gráfban

Egy másik megközelítési lehetőségünk, ha egy csúcs szomszédai között futó élek számát vizsgáljuk, ezt nevezzük lokális klaszterezettségi együttthatónak. A dolgozat további részében $\forall G(V, E)$ gráfra $V := \{1, \dots, n\}$, ha mást nem teszünk fel.

2.1.2. Definíció (Lokális klaszterezettségi együtttható). Legyen $G(V, E)$ egyszerű gráf, ekkor $\forall i \in V$ -re

$$c(i) = \begin{cases} \frac{T^i}{P_2^i} & \text{ha } d^i \geq 2 \\ 0 & \text{ha } d^i = 0, 1 \end{cases}$$

az i csúcs lokális klaszterezettségi együttthatója, ahol T^i az élek száma az i csúcs szomszédai között (azaz az i -t tartalmazó háromszögek száma), P_2^i pedig az i szomszédai közül kiválasztható csúcspárok száma, tehát ha i fokszáma d_i , akkor $P_2^i = \binom{d_i}{2} = \frac{d_i(d_i-1)}{2}$.



2. ábra. Zárt tripletek egy csúcs szomszédai közt

A lokális klaszterezettségi együttthatók kiátlagolásával kapjuk az átlagos lokális klaszterezettségi együttthatót.

2.1.3. Definíció (Átlagos lokális klaszterezettségi együtttható). Legyen $G(V, E)$ egyszerű gráf, ekkor

$$C_2(G) = \frac{1}{n} \sum_{i \in V} c(i).$$

Megfigyelhetjük, hogy a lokális és globális klaszterezettségi együttthatók között a következő összefüggés áll fenn:

$$C_1(G) = \frac{\sum_{i \in V} \binom{d_i}{2} c(i)}{\sum_{i \in V} \binom{d_i}{2}},$$

hiszen a definíció alapján:

$$c(i) = \frac{T^i}{\binom{d_i}{2}},$$

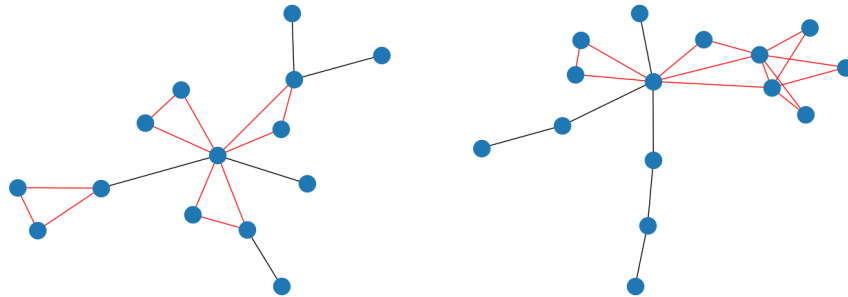
így az állítás

$$C_1(G) = \frac{\sum_{i \in V} T^i}{\sum_{i \in V} \binom{d_i}{2}},$$

ami egybeesik $C_1(G)$ definíciójával.

2.2. Diadikus klaszterezés háromszögmultiplicitással

Az előző részben bevezetett lokális klaszterezettségi együttható jellemzője, hogy egy csúcs előfordulását vizsgálja meg a gráf háromszögeiben. Ennek mintájára felvetődhet az a kérdés is, hogy egy csúcspár, azaz egy él hány háromszögben szerepel. Ezt a tulajdonságot megvizsgálva arról kaphatunk képet, hogy a gráf jellemzően különálló háromszögeket tartalmaz, vagy inkább olyanokat, amelyek osztoznak egy-egy élükön. A gráfunk ezen tulajdonságát diadikus klaszterezettségnek nevezzük. M. Á. Serrano and M. Boguñá [10] cikkükben foglalkoznak ezzel a kérdéssel, a diadikus klaszterezettség mérésére bevezetik az *edge multiplicity* fogalmát, melyet az egyértelműség kedvéért fordítsunk most az élek háromszögmultiplicitásának.



3. ábra. Gráfok kisebb ill. nagyobb háromszögmultiplicitással

2.2.1. Definíció (Élek háromszögmultiplicitása). [10] Legyen $G(V, E)$ egyszerű gráf, $\forall (ij) \in E$ -re az (ij) él háromszögmultiplicitásának nevezzük és m_{ij} -vel jelöljük azon háromszögek számát, amelyek az i és j csúcsokat is tartalmazzák.

Az újonnan bevezett multiplicitás és az előzőekben látott T^i mennyiségek között fennáll a következő összefüggés:

$$\sum_j m_{ij} a_{ij} = 2T^i,$$

ahol a_{ij} a gráf adjacenciamátrixának megfelelő eleme.

A gráfunk háromszögeinek összeszámlálása mellett az eloszlásukat is jellemezhetjük az egyes fokszámosztályok közötti megoszlásuk alapján. Erre alkalmas a lokális klaszterezettségi együttható és a háromszögmultiplicitás is.

A k fokszámú csúcsok osztályán (Γ_k) kiátlagolva a lokális klaszterezettségi együtthatót megkapjuk, hogy az ide tartozó csúcsok milyen arányban szerepelnek háromszögben:

$$\bar{c}(k) = \frac{1}{|\Gamma_k|} \sum_{i \in \Gamma_k} c(i).$$

Az élek háromszögmultiplicitását használva pedig már azt is vizsgálhatjuk, hogy két fokszám-osztályon hány háromszög osztozik, ha meghatározzuk a fokszámosztályok ($\Gamma_k, \Gamma_{k'}$) között futó élek átlagos multiplicitását:

$$\bar{m}_{kk'} = \frac{\sum_{i \in \Gamma_k} \sum_{j \in \Gamma_{k'}} m_{ij} a_{ij}}{|E_{kk'}|},$$

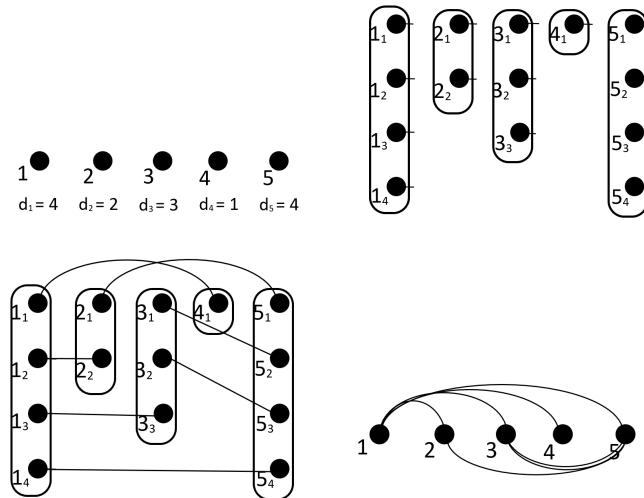
ahol $E_{kk'}$ a két fokszámosztály, Γ_k és $\Gamma_{k'}$ között futó összes él. Látható, hogy $k = k'$ választása esetén ez a formula is alkalmas az egy osztályon belül futó élek átlagos multiplicitásának meghatározására is, ezesetben $|E_{kk'}| = 2|E_{\Gamma_k}|$ -ra módosul, vagyis a k fokszámosztályban futó élek számának kétszeresére.

A diadikus klaszterezettséget tovább általánosítva akár azt is vizsgálhatjuk, hogy a háromszögek eloszlása hogyan alakul három különböző fokszámosztály felett. Ehhez az ún. triadikus klaszterezéshez már három csúcsunkat vizsgálnánk.

2.3. Modularitás

A gráf modularitásának vizsgálatakor arról kaphatunk képet, hogy mennyire sikeresen tudunk meghatározni részgráfokat úgy, hogy bennük sok, két különböző között viszont kevés él fusson. A cél, hogy megfelelően partícionáljuk a gráfunkat, majd megvizsgáljuk, hogy a kialakuló részhalmazokon (klasztereken) belül mennyivel több él fut, mint egy véletlen gráfnál várható lenne. Véletlen gráfok modularitásával Jung Attila foglalkozott szakdolgozatában [6], az alábbiakban ezen dolgozat alapján tekintünk be ebbe a témakörbe. A pontosabb definícióhoz vezessük be a konfigurációs modell fogalmát:

2.3.1. Definíció (Konfigurációs modell). Legyen V csúcsok egy halmaza és d_i – az i csúcs fokszáma – adott $\forall i \in V$ -re (az egyszerűség kedvéért tegyük fel, hogy ezek összege páros). Ennek megfelelően hozzuk létre az $L_i = \{i_1, i_2, \dots, i_{d_i}\}$ halmazokat a csúcsok lemásolásával és illesszünk minden csúcsra egy élcsontot (olyan élt, amelynek másik vége szabad). Ezután hozzuk létre a $G(V, E)$ véletlen gráfot úgy, hogy véletlenszerűen választunk az élcsontok közül párokat, amiket összekötünk, egészen addig, amíg azokból van szabad.



4. ábra. A konfigurációs modell generálása

A konfigurációs modell segítségével az elvárt fokszámeloszlásnak megfelelő véletlen gráfot kapunk, amely esetleg tartalmazhat párhuzamos éleket és hurokéleket, de ezek redukálásával a gráf egyszerűvé tehető.

2.3.2. Definíció (Gráfpartíció modularitása). [6]

Legyen $G = (V, E)$ a vizsgált gráf és $G^{konf} = (V, E^{konf})$ a G fokszámeloszlásához létrehozott konfigurációs modell, illetve A és A^{konf} G -nek és G^{konf} -nak egy-egy partíciója, amelyek V csúcsait azonosan partícionálják. Ekkor

$$q_A(G) = \frac{\sum_{A_i \in A} |E_{A_i}|}{|E|} \mathbb{E} \left(\frac{\sum_{A_i^{konf} \in A^{konf}} |E_{A_i^{konf}}|}{|E|} \right)$$

az A partíció modularitása G -ben. Azaz a G gráfunk A partíciójának részalmazzaiban összesen futó élek és az összes él arányának és a konfigurációs modell által létrehozott gráfokban ugyanennek az arányának a várható értékének a különbsége.

Megmutatható, hogy

$$\mathbb{E} \left(\frac{\sum_{A_i^{konf} \in A^{konf}} |E_{A_i^{konf}}|}{|E|} \right) = \frac{\sum_{A_i \in A} vol^2(A_i)}{vol^2(G)},$$

ahol vol az adott részalmaz csúcsainak fokszámmösszege. Vagyis a létrehozott konfigurációs modellben a belül futó élek hányadának várható értékét a klaszterek fokszámmösszege határozza meg. Ez alapján a partíció modularitása így is definiálható:

$$q_A(G) = \frac{\sum_{A_i \in A} |E_{A_i}|}{|E|} \frac{\sum_{A_i \in A} vol^2(A_i)}{vol^2(G)}.$$

2.3.3. Definíció (Gráf modularitása). Egy gráf modularitása az összes partíciójának modularitásának a maximuma:

$$q(G) = \max_A q_A(G).$$

3. Állandó csúcsszámú véletlen gráf generálása szabályozható klaszterezettségi együtthatóval

3.1. Newman általános algoritmusa

Az előző fejezetben bevezetett konfigurációs modell segítségével már képesek vagyunk egy adott fokszámeloszlást követő véletlen gráfot létrehozni. Ha az egyes csúcsokra nézve nem csak azt határozzuk meg, hogy hány él kapcsolódjon hozzájuk, hanem azt is, hogy a csúcsok hány háromszögben szerepeljenek, akkor beállíthatjuk a gráf klaszterezettségi együtthatóját is.

Az így kapott modell előnye, hogy tetszőleges fokszámeloszlás és klaszterezettség jó közelítéssel előállítható. Ezek a tulajdonságok azonban nem önszerveződés, hanem külső megkötések hatására alakulnak ki, valamint az állandó csúcsszám miatt a modell nem tükrözi a valós hálózatok dinamikus kialakulását.

A konfigurációs modellen alapuló felépítést Newman veti fel [7]-ben, N. Parikh [5]-ban pedig az algoritmus négy verzióját mutatja be. Ezek alap gondolata a következő:

Legyen d_i az i csúcs fokszáma, t_i a csúcsot tartalmazó háromszögek száma és $s_i = d_i - 2t_i$, a csúcshoz kapcsolódó, különálló élek száma. Az egyszerűség kedvéért tegyük fel, hogy $\sum_i s_i$ kettővel, $\sum_i t_i$ hárommal osztható. A konfigurációs modellben látott módon minden csúcshoz kapcsoljunk t_i db cseresznyecsonkot és s_i db élcsonkot. Ezek után három csúcs véletlen választásával megkonstruáljuk a háromszögeinket – összekötjük a cseresznyecsonkjait – két csúcs választásával pedig a fennmaradó, különálló éleinket – az élcsonkok összekötésével. Az algoritmus verziói ennek megvalósításában térnek el egymástól.

Az első két verzióban (CONF-1 és CONF-2) a háromszögek konstruálásakor három, a különálló élek konstruálásakor két különböző csúcsot választunk. Ezt CONF-1-ben azonos valószínűséggel tesszük, míg CONF-2 esetében előnyben részesülnek azok a csúcsok, amiknek még több hiányzik t_i vagy s_i , azaz az elvárt értékek eléréséhez. Amennyiben még egyik kiválasztott csúcs sem érte el a számára meghatározott t_i vagy s_i értéket, behúzzuk a megfelelő éleket, ellenkező esetben megismételjük a választást.

Ezt folytatjuk addig, amíg létre nem jött az összes elvárt háromszög ($T := \frac{\sum_i t_i}{3}$), majd az összes különálló él ($S := \frac{\sum_i s_i}{2}$).

Mivel mindkét esetben különböző csúcsokra van szükség egy-egy lépés végrehajtásához, problémát okozhat, ha már nem marad elegendő, szabad cseresznye- vagy élcsomkkal rendelkező csúcsunk. A CONF algoritmusok ezekben félbeszakadnak, T db háromszög vagy S db különálló él behúzása előtt. Ezt orvosolandó, a THROW-1 és THROW-2 nevet viselő algoritmusokban a csúcsok választásának valószínűsége mindig a t_i és s_i értékekkel arányos, így akkor is kiválasztásra kerülhetnek, ha már elérték ezeket az értékeket. Ezek mellett THROW-2-ben megengedjük azt is, hogy egy lépésben ne csak különböző csúcsokat választhassunk, ez azonban nem okoz jelentős eltérést, így THROW-2 tárgyalásától eltekintünk.

3.2. A CONF-1 algoritmus

Az alábbiakban olvasható a CONF-1 algoritmus pszeudokódja. Az algoritmus inputként kéri a csúcsok számát: n , a t_i -ket tartalmazó t listát és az s_i -ket tartalmazó s listát, majd visszaadja a csúcsok és élek halmazát (E, V) .

Az algoritmus első felében három különböző csúcsot választ $\frac{1}{n}$ valószínűséggel és hozzáadja a köztük futó éleket E -hez, ha még nem szerepelnek benne. Ezt addig folytatja, amíg el nem érjük T -t vagy amíg már nem találunk elég különböző csúcsot, ahova még húzható él. A második felében ugyanez történik, csak két csúcs választásával a különálló élekhez, amíg el nem érjük S -et vagy amíg már nem találunk két különböző csúcsot, ahova még húzható él.

Annak mérésére, hogy az i csúcsnak még mennyi hiányzik t_i illetve s_i eléréséhez, bevezetjük a rt_i és rs_i változókat, amiket folyamatosan frissítünk.

CONF-1(n, s, t)

-
- 1: $V := \{1, 2, \dots, n\}$
 - 2: $E := \{\}$
 - 3: $rt := t$
 - 4: $rs := s$
 - 5: $T := \frac{\sum_i t_i}{3}$
 - 6: $S := \frac{\sum_i s_i}{2}$

7: **Ciklus amíg** $T > 0$ és $\sum_{x \in rt} (x > 0) > 2$
 8: $(u, v, w) :=$ "három különböző csúcs azonos valószínűséggel választva"
 9: **Ha** $rt[u] > 0$ és $rt[v] > 0$ és $rt[w] > 0$ **akkor**
 10: $E := E \cup \{(u, v), (v, w), (u, w)\}$
 11: $rt[u] := rt[u] - 1$
 12: $rt[v] := rt[v] - 1$
 13: $rt[w] := rt[w] - 1$
 14: $T := T - 1$
 15: **Elágazás vége**
 16: **Ciklus vége**
 17: **Ciklus amíg** $S > 0$ és $\sum_{x \in rs} (x > 0) > 1$
 18: $(u, v) :=$ "két különböző csúcs azonos valószínűséggel választva"
 19: **Ha** $rs[u] > 0$ és $rs[v] > 0$ **akkor**
 20: $E := E \cup \{(u, v), (v, w)\}$
 21: $rs[u] := rs[u] - 1$
 22: $rs[v] := rs[v] - 1$
 23: $S := S - 1$
 24: **Elágazás vége**
 25: **Ciklus vége**

3.3. A CONF-2 algoritmus

A második algoritmust CONF-1 módosításával kapjuk, az egyetlen eltérés a csúcsok súlyozása azok véletlen választásakor. Az előző azonos valószínűségek helyett itt a háromszögeképésnél

$$\forall i \in V\text{-re } p_t(i) := P(i\text{-t választjuk}) = \frac{rt_i}{\sum_j rt_j},$$

és az élképésnél

$$\forall i \in V\text{-re } p_s(i) := P(i\text{-t választjuk}) = \frac{rs_i}{\sum_j rs_j},$$

valószínűségeket használjuk. Ezek minden lépésben változnak, ahogy a csúcsok r_s és r_t értékei is.

Az algoritmus pszeudokódja a következő:

CONF-2(n, s, t)

- 1: $V := \{1, 2, \dots, n\}$
- 2: $E := \{\}$
- 3: $rt := t$
- 4: $rs := s$
- 5: $T := \frac{\sum_i t_i}{3}$
- 6: $S := \frac{\sum_i s_i}{2}$
- 7: **Ciklus amíg** $T > 0$ és $\sum_{x \in rt}(x > 0) > 2$
- 8: $(u, v, w) :=$ "három különböző csúcs p_t valószínűséggel választva"
- 9: **Ha** $rt[u] > 0$ és $rt[v] > 0$ és $rt[w] > 0$ **akkor**
- 10: $E := E \cup \{(u, v), (v, w), (u, w)\}$
- 11: $rt[u] := rt[u] - 1$
- 12: $rt[v] := rt[v] - 1$
- 13: $rt[w] := rt[w] - 1$
- 14: $T := T - 1$
- 15: **Elágazás vége**
- 16: **Ciklus vége**
- 17: **Ciklus amíg** $S > 0$ és $\sum_{x \in rs}(x > 0) > 1$
- 18: $(u, v) :=$ "két különböző csúcs p_s valószínűséggel választva"
- 19: **Ha** $rs[u] > 0$ és $rs[v] > 0$ **akkor**
- 20: $E := E \cup \{(u, v), (v, w)\}$
- 21: $rs[u] := rs[u] - 1$
- 22: $rs[v] := rs[v] - 1$
- 23: $S := S - 1$
- 24: **Elágazás vége**
- 25: **Ciklus vége**

3.4. A THROW-1 algoritmus

Ebben a verzióban CONF-2-höz hasonlóan szintén súlyozzuk a csúcsainkat, ez most a t_i és s_i értékek szerint történik:

$$\forall i \in V\text{-re } p_t(i) := P(i\text{-t választjuk}) = \frac{t_i}{\sum_j t_j},$$

és

$$\forall i \in V\text{-re } p_s(i) := P(i\text{-t választjuk}) = \frac{s_i}{\sum_j s_j}.$$

Vegyük észre, hogy CONF-2-vel ellentétben ezek az értékek a futás alatt állandóak maradnak. Ennek következtében ha egy csúcs már elérte a tőle elvárt például t értéket, de még nem hoztunk létre T háromszöget, akkor semmi nem tiltja, hogy a csúcs további háromszögekben szerepeljen.

THROW-1(n, s, t)

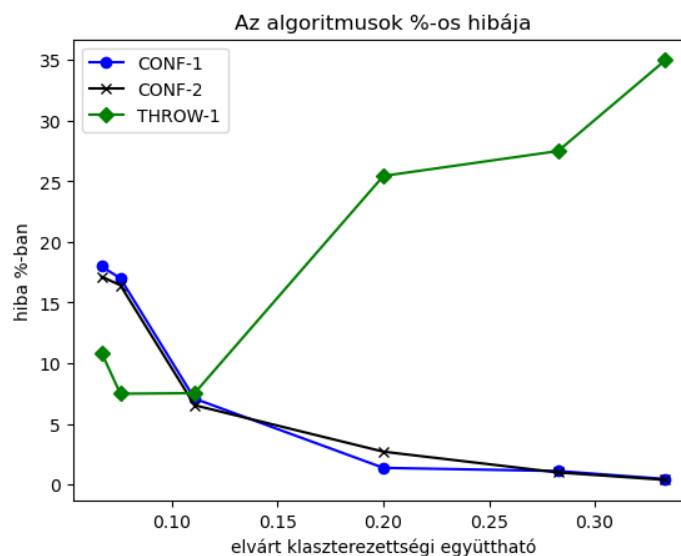
-
- 1: $V := \{1, 2, \dots, n\}$
 - 2: $E := \{\}$
 - 3: $P := \left(\frac{t_i}{\sum_j t_j} \forall i\right)$
 - 4: $Q := \left(\frac{s_i}{\sum_j s_j} \forall i\right)$
 - 5: $T := \frac{\sum_i t_i}{3}$
 - 6: $S := \frac{\sum_i s_i}{2}$
 - 7: **Ciklus amíg** $T > 0$
 - 8: $(u, v, w) :=$ "három különböző csúcs P választva"
 - 9: $E := E \cup \{(u, v), (v, w), (u, w)\}$
 - 10: $T := T - 1$
 - 11: **Ciklus vége**
 - 12: **Ciklus amíg** $S > 0$
 - 13: $(u, v) :=$ "két különböző csúcs Q valószínűséggel választva"
 - 14: $E := E \cup \{(u, v), (v, w)\}$
 - 15: $S := S - 1$
 - 16: **Ciklus vége**

3.5. Empirikus eredmények

A fent ismertetett három algoritmust empirikus vizsgálat céljából meg is valósítottam, $n = 1000$ csúcsú gráfra, 100 tesztesetet figyelembe véve. Az evárt C_1 értékek egy részéhez különböző paraméterű, hatványeloszlást követő véletlen gráfokat generáltam, $C_1 = 1, \frac{1}{3}$ és $\frac{2}{30}$ esetekben uniform gráfokat tekintettem. A véletlen tesztekre kapott átlagos, megfigyelt globális klaszterezettségi együtthatók az alábbi táblázatban láthatók.

Elvárt $C_1(G)$	CONF-1	CONF-2	THROW-1
1	1	1	0.33787
0.33333	0.33486	0.33462	0.21674
0.28260	0.28575	0.28542	0.20491
0.20	0.20275	0.20542	0.14912
0.11070	0.11854	0.11794	0.10236
0.07560	0.08845	0.08802	0.08126
0.06667	0.07866	0.07810	0.07392

Az alábbi diagramon a fenti eredményeket ábrázoltam, a célszerűség kedvéért a $C_1(G) = 1$ eset nélkül. Szembetűnő, hogy amíg nagyobb elvárt klaszterezettségi együttható esetén a CONF-algoritmusok adnak jelentősen jobb közelítést – kicsivel az elvárt érték felett maradván –, a kevés háromszöggel rendelkező esetekben a THROW kerül hozzá közelebb.



4. Valós hálózatok modellezése növekvő csúcsszámú véletlen gráffal

4.1. Történeti áttekintés, út a Barabási-Albert-modellig

Az előző fejezetben állandó csúcsszámú gráfokat alkalmazó modelleket láthattunk. Ezen tulajdonságuk összhangban van az Erdős-Rényi-féle véletlen gráf-moddal, aminek módosított változatait egészen a 20. század végéig próbálták a valós hálózatok leírására is alkalmazni, habár megalkotóit eredetileg tisztán matematikai vizsgálatuk foglalkoztatta. A modell definíciójának két változata létezik:

4.1.1. Definíció (Erdős-Rényi-modell, A változat). Legyen $G(n, M)$ az a véletlen gráf, melyet egyenletes eloszlás szerint választunk az n csúcsú, M élű gráfok közül.

4.1.2. Definíció (Erdős-Rényi-modell, B változat). Legyen $G(n, p)$ az a véletlen gráf, melyben minden éllet p valószínűséggel húzunk be az n csúcs között.

Az informatika fejlődésével azonban egyre több adat begyűjtése vált lehetségessé, megmutatva a nagy hálózatok olyan tulajdonságait, amelyek nem voltak megmagyarázhatóak ezen modelltől kiindulva. Ilyenek például a hatványfüggvényt követő fokszámeloszlás ($\forall i \in V$ -re $P(d_i = k) \sim k^{-\gamma}$) vagy a megfigyelt klaszterezettség. Láttuk, hogy az előző fejezetben tárgyalt algoritmusok ezek megoldására megszorító követelményeket támasztanak, jó közelítéssel kikényszerítve az általunk elvárt paramétereket.

1999-ben Barabási Albert-László és Albert Réka azonban egy merőben új megközelítést ajánlott, amely a hálózaton belüli önszerveződés szabályaira építkezik. Két törvényszerűséget fogalmaztak meg, melyek általánosan igaznak bizonyultak a valós hálózatokra.

Az első változás annak a figyelembevételére, hogy a hálózat kialakulása során a csúcsok folyamatosan, egymást követően csatlakoznak a gráfba, így kapunk egy növekvő véletlen hálózatot. Precízen: kiindulunk m_0 pontból, majd minden lépésben hozzáadunk egy csúcsot, illetve a hozzá tartozó $m (< m_0)$ él is behúzzuk az új és egy már szereplő csúcs közé. Ezt ismételve t lépés után a gráfunk már $m_0 + t$ csúcsból és tm élből fog állni.

A növekvő csúcsszám hatására a korábban bekerült csúcsoknak már előnyük lesz az élszerzésben a később jövőkkel szemben, de emellett egy másik törvényszerűséget, a népszerűségi kapcsolódást is figyelembe vesszünk. Eszerint a valós hálózatokban a több kapcsolattal rendelkező csúcsok nagyobb eséllyel szereznek újabbakat, mint a kevésbé rendelkezők. Ezt lemodellezve egy behúzandó él végpontjának véletlen választásakor nagyobb valószínűséget rendelünk a nagyobb fokszámú csúcsokhoz.

4.1.3. Definíció (Barabási-Albert-modell/Preferential Attachment model). [1] Legyen $G(V, E)$ egy egyszerű gráf, ahol $|V| = m_0, |E| = 0$, amit a t . lépésben növeljük a következő módon:

$$V_t := V_{t-1} \cup \{t\} \text{ (így } |V_t| := |V_{t-1}| + 1),$$

$$E_t := E_{t-1} \cup \{(t, i_1), \dots, (t, i_m)\} \text{ (így } |E_t| := |E_{t-1}| + m),$$

ahol $\forall k \in \{1, \dots, m\}$ -re és $\forall j \in \{1, \dots, m_0 + t\}$ -re

$P((t, i_k) = (t, j)) = \frac{d_j^*}{\sum_i d_i^*}$, ahol d_i^* az i csúcs fokszáma az aktuális lépést megelőzően.

Megmutatható, hogy az így kapott dinamikus modellben kellően sok lépés után kialakul a kívánt hatványfüggvény szerinti eloszlás, más külső megkötések bevezetése nélkül. Ez alapján alkalmasnak tűnt a valós hálózatok modellezésére, így az új évezredben már kiindulási pontként szolgált a hálózatkutatók számára.

A Barabási által bevezetett modell alapján több klasszikus véletlen gráf modellt is megalkottak. Ezek bizonyos szempontól hasonló viselkedést mutatnak, hiszen az őket felépítő logika megegyezik, de a különféle elemek módosításával, egyéb tényezők figyelembe vételével különböző megközelítésekhez jutunk. A következő fejezetekben két klasszikus modell bemutatása és vizsgálata következik, elsősorban a klaszterezettségi együtt-hatójuk szempontjából.

4.2. A BA-modell és a ráépülő más modellek klaszterezettségi együtthatója

4.2.1. Az LCD-modell

2001-ben Bollobás és Riordan fogalmazták meg a BA-modell egy matematikailag precíz változatát. Az eredeti megfogalmazáshoz illeszkedve él- és csúcsszám szerinti indukcióval írható le. Az így létrejövő gráfunk nem lesz egyszerű gráf. A szerzők megmutatták, hogy az így létrejövő véletlen gráf fokszámoszlása a hatványeloszláshoz tart $\gamma = 3$ paraméterrel, megfelelően a támasztott elvárásoknak.

4.2.1. Definíció (LCD-modell). [2] Azt mondjuk, hogy a $G_{n,m}(V, E)$ véletlen gráfunk (ahol $|E| = n$ és $m > 0$ egész) az LCD-modellbe tartozik, ha azt a következőképpen konstruáltuk: kezdetben $m := 1$ és induljunk ki a $G_{1,1}$ gráfból, azaz az egy csúcsot és egy hurokért tartalmazó gráfból.

Ezek után tegyük fel, hogy $G_{t-1,1}$ -et már ismerjük. Ekkor az indukciós lépésben hozzáadjuk a gráfunkhoz a t csúcsot, illetve egy közte és az i csúcs között futó élt. Az i csúcs kiválasztása véletlenszerűen történik ezen valószínűséggel:

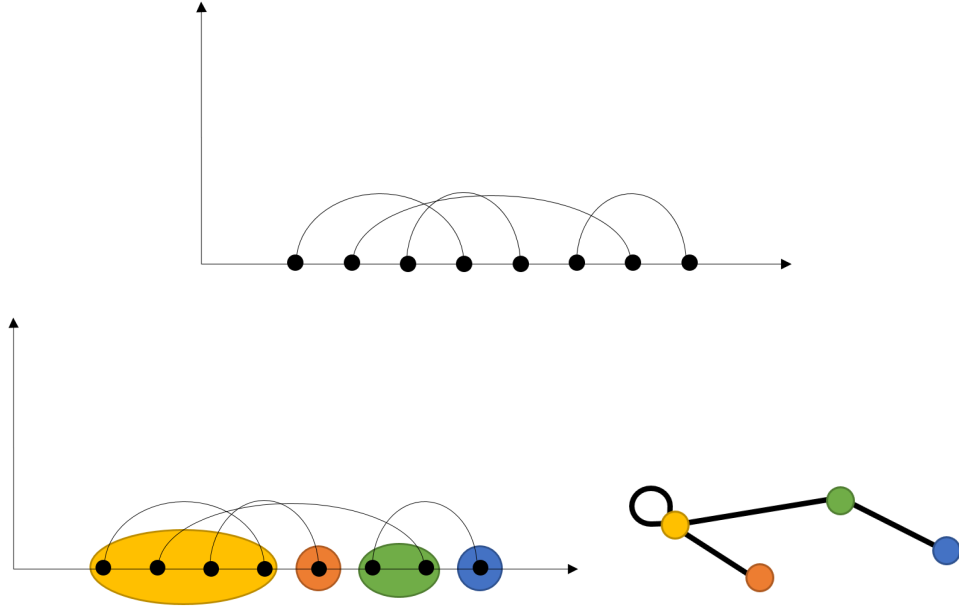
$$P(i = s) = \begin{cases} \frac{d_s^{t-1}}{2^{t-1}} & \text{ha } s \neq t \\ \frac{1}{2^{t-1}} & \text{ha } s = t \end{cases}$$

$G_{n,m}$, $m > 1$ definiálásához hozzuk létre a $G_{nm,1}$ gráfot a fenti módon, majd ezek után

$v_i := \{(i-1)m + 1, im + 2, \dots, im\} \forall i = 1, \dots, n$ -re legyenek a gráfunk új csúcsai. Az összevonás hatására az egy halmazon belül futó élek hurokélekké alakulnak, illetve párhuzamos élek is megjelenhetnek.

A tárgyalt modell nevét a Linearized Chord Diagram rövidítéséből kapta. Ebben a gráfképzési formában felvesszünk az x-tengelyen $2n$ pontot, majd köztük egy teljes párosításnak megfelelően hurokat húzunk be. Ezek után az origóból indulva elmegyünk egészen az első jobb végpontig, és az addig szereplő pontokból létrehozuk a gráfunk első csúcsát, majd ezt folytatva haladunk $2n$ -ig. Az új csúcsok közötti éleket az eredeti hurokból képezzük.

Amennyiben a teljes párosításnál a pontokból egyenletes eloszlás szerint választottunk, belátható, hogy az LCD-ből képzett gráfunk fokszámoszlása meg fog egyezni az LCD-modellbe tartozó $G_{n,1}$ gráf fokszámoszlásával.



5. ábra. Gráfképzés LCD-diagramból

4.2.1. Tétel. [2] Az LCD-modellbe tartozó $G_{n,m}(V, E)$ gráf globális klaszterizettségének várható értéke n növelésével a 0-hoz tart, ugyanis

$$\mathbb{E}(C_1(G)) \sim \frac{m-1}{8} \frac{\log^2 n}{n},$$

ha $n \rightarrow \infty$.

Bizonyítás. A továbbiakban [2]-re támaszkodva közlöm a bizonyítás főbb gondolatait.

$$\mathbb{E}(C_1(G)) = \mathbb{E} \left(\frac{3T(G)}{P_2(G)} \right) \sim \frac{3\mathbb{E}(T(G))}{\mathbb{E}(P_2(G))}$$

fennállása miatt meg kell határoznunk a gráfban szereplő háromszögek és tripletek várható értékét.

Első lépésként vizsgáljuk $G_{n,1}$ LCD-modellbeli gráf esetében, hogy mekkora a valószínűsége annak, hogy egy adott szubgráf szerepel benne. Technikai okokból irányítsuk meg az egyes éleit a később hozzáadott végükből a korábban hozzáadott végük felé.

Ehhez vegyünk egy $S(V, E)$ – nem feltétlenül egyszerű – irányított gráfot $\{1, \dots, n\}$ -en, és vezessük be a következő jelöléseket: $V^+(S)$ legyen a csúcsok azon halmaza, ahonnan kilépnek élek, $V^-(S)$ pedig azon halmaza, ahová élek érkeznek. $\forall i \in V(S)$ -re $d_S^{be}(i)$ legyen az i csúcs befoka, $d_S^{ki}(i)$ pedig a kifoka S -ben, $C_S(t)$ pedig azon élek száma S -ben, melyek egyik vége t előtti, másik vége pedig utáni – azaz azoké, amelyek átlépik a t csúcsot.

4.2.2. Lemma. *A fenti jelölésekkel ha S $G_{n,1}$ -nek egy lehetséges részgráfja – azaz $\forall i \in V$ -re $d_S^{ki}(i) \leq 1$ fennáll –, akkor annak a valószínűsége, hogy G valóban tartalmazza S -et:*

$$P_S = \prod_{i \in V^-(S)} d_S^{be}(i)! \prod_{i \in V^+(S)} \frac{1}{2i-3} \prod_{t \notin V^+(S)} \left(1 + \frac{C_S(t)}{2t-3}\right),$$

ami így becsülhető:

$$P_S = \prod_{i \in V^-(S)} d_S^{be}(i)! \prod_{ij \in E(S)} \frac{1}{2\sqrt{ij}} \exp \left(O \left(\sum_{i \in V(S)} \frac{C_S^2(i)}{i} \right) \right). \quad (1)$$

(1) felhasználásával térjünk át $G_{n,m}$ -re, ahol $m \geq 1$ rögzített, és számláljuk meg az ebben részgráfként megjelenő háromszögek számát.

4.2.3. Tétel. *Legyen $G_{n,m}$ egy LCD-beli véletlen gráf, ekkor benne a háromszögek várható értékére ($\mathbb{E}(T)$), a következő teljesül:*

$$\mathbb{E}(T) \sim \frac{m(m-1)(m+1)}{48} \log^3 n,$$

ha $n \rightarrow \infty$.

Bizonyítás. Térjünk vissza az LCD-modell bevezetésénél látott módszerre és gondoljunk $G_{n,m}$ -re, mint $G_{nm,1}$ -ből, a csúcsok csoportosításával származtatott gráfra. Vizsgáljuk meg, hogy egy konkrét csúcshármas mikor alkot háromszöget $G_{n,m}$ -ben.

Az analógia alapján $1 \leq a \leq b \leq c \leq n \in G_{n,m}(V)$ pontosan akkor alkot háromszöget, ha $\exists i, i', j, j', k, k' \in V(G_{nm,1})$, $m(a-1) \leq i, i' < ma$, $m(b-1) < j, j' \leq mb$, $m(c-1) < k, k' \leq mc$ úgy, hogy $ij', jk', i'k \in E(G_{nm,1})$.

Legyen S $G_{nm,1}$ egy olyan részgráfja, amelyre $V(S) = \{i, i', j, j', k, k'\}$ és $E(S) = ij', j'k, i'k$ és alkalmazzuk erre a 4.2.2 lemmát:

$$\begin{aligned}
P_S &= \prod_{x \in V^-(S)} d_S^{be}(x)! \prod_{xy \in E(S)} \frac{1}{2\sqrt{xy}} \exp \left(O \left(\sum_{x \in V(S)} \frac{C_S^2(x)}{x} \right) \right) = \\
&= K_1 \prod_{x \in V^-(S)} d_S^{be}(x)! \prod_{xy \in E(S)} \frac{1}{2\sqrt{xy}} = \\
&= K_2 \prod_{x \in V^-(S)} d_S^{be}(x)! \frac{1}{2\sqrt{mamb}} \frac{1}{2\sqrt{mbmc}} \frac{1}{2\sqrt{mamc}} = K_2 \prod_{x \in V^-(S)} d_S^{be}(x)! \frac{1}{8m^3abc},
\end{aligned}$$

ahol K_1, K_2 korrekciós tagok korlátosak és határértékben az 1-hez tartanak, ha $a \rightarrow \infty$.

Most nézzük meg, hogy hányféleképpen választhatjuk ki a vizsgált i, i', j, j', k, k' csúcsokat. Az megállapítható, hogy k -nak és k' -nek különbözőnek kell lennie, különben az ő kifoka 2 lenne, ez pedig ellentmondana annak, hogy ő $G_{nm,1}$ -nek egy lehetséges részgráfja. A j -re és j' csúcsokra nincs ilyen kikötés, hiszen „középső” csúcsokként az egyik él mindeképpen a d^{be} , másik a d^{ki} értéküket növeli. Az i és i' választásakor két esetet különíthetünk el, lehetséges $i = i'$, m lehetőséggel és $i \neq i'$, $m(m-1)$ lehetőséggel. Ezeket összegezve az összes lehetőségünk: $mm^2m(m-1) + m(m-1)m^2m(m-1) = m^4(m-1) + m^4(m-1)^2$

Az adott a, b, c csúcsok által megadott háromszögek számának várható értékét $G_{n,m}$ -ben az előző két eredmény alapján becsülhetjük. Az i, i' csúcs választása alapján két eset lesz, vagy $i = i'$, ekkor $d_i^{be} = 2$ vagy $i \neq i'$, ekkor $d_i^{be} \leq 1 \forall i \in V(S)$. Emiatt:

$$\mathbb{E}(T_{abc}) = C_3 \left(m^4(m-1) \frac{2}{8m^3abc} + m^4(m-1)^2 \frac{1}{8m^3abc} \right) = C_3 \frac{m(m-1)(m+1)}{8abc}.$$

Ezután még a, b, c választása a feladatunk, ezeket összegezzük ki az előző, egy konkrét hármásra vonatkozó eredményt:

$$\mathbb{E}(T) = (1 + o(1)) \sum_{1 \leq a \leq b \leq c \leq n} \frac{m(m-1)(m+1)}{8abc} \sim \frac{m(m-1)(m+1)}{48} \log^3 n.$$

□

A továbbiakban a *whp* jelölés a *with high probability*-t rövidíti, melynek jelentése egy A_n eseménysorozatra a következő: $P(A_n) \rightarrow 1$, ha $n \rightarrow \infty$.

4.2.4. Tétel. *(bizonyítás nélkül) Legyen $G_{n,m}$ LCD-modellbeli, $m \geq 1, \epsilon > 0$ rögzített. Ekkor a gráfban szereplő tripletek számára teljesül az alábbi whp:*

$$(1 - \epsilon) \frac{m(m+1)}{2} n \log n \leq P_2(G_n, m) \leq (1 + \epsilon) \frac{m(m+1)}{2} n \log n$$

ha $n \rightarrow \infty$.

Ezzel elértünk a 4.2.1. tétel bizonyításának zárásához a globális klaszterettségi együttható aszimptotikus viselkedéséről. Vessük össze a 4.2.3. tételt és a 4.2.4. tételt:

$$\mathbb{E}(C_1(G)) = \mathbb{E} \left(\frac{3T(G)}{P_2(G)} \right) \sim \frac{m-1}{8} \frac{\log^2 n}{n}.$$

□

4.2.2. Buckley-Osthus-modell/Móri modell

Az LCD-modell esetében felmerülhet az a kérdés, hogy egy új él behúzásakor minden addigi csúcsnak lesz-e esélye az él végpontjává válni. Azt, hogy a válasz igen, az garantálja, hogy amikor egy csúcsot a gráfhoz adunk, vele együtt egy élet is, így fokszáma egy, vagy hurokél esetén kettő lesz. Ezek után a következő csúcs hozzáadásánál már pozitív valószínűséggel szerepel a versenyben.

A valós hálózatok modellezésekor mégis szükség lehet rá, hogy a fokszámok által létrejövő különbséget valamennyire szabályozni tudjuk. Ennek érdekében meghatározhatjuk a csúcsok kezdeti attraktivitását β , amely fokszámtól függetlenül biztosít számukra versenyképességet. Ezt az értéket minél nagyobbra választva a csúcsok fokszámának különbsége egyre kevésbé lesz jelentős befolyással a gráf alakulására.

4.2.2. Definíció (Buckley-Osthus-modell). [3] Azt mondjuk, hogy a $G_{n,m}^\beta(V, E)$ véletlen gráfunk BO-modellbeli, ha ugyanúgy konstruáljuk, mint

az LCD-modellben, azzal a különbséggel, hogy a t . lépésben az új él behúzásakor az i csúcs kiválasztásánál s csúcs esélye a következőképpen alakul:

$$P(i = s) = \begin{cases} \frac{d_s^{t-1} + \beta}{(\beta+2)t-1} & \text{ha } s \neq t \\ \frac{\beta+1}{(\beta+2)t-1} & \text{ha } s = t \end{cases}$$

A BO-féle definícióban szereplő kezdeti attraktivitásra $\beta \geq 0$ egész. A modellt Móri általánosította tetszőleges $\beta > -1$ valósra. Kellően alacsony fokszámok esetén ($\forall i$ -re $0 \leq d^i \leq n^{1/100(\beta+1)}$) mindkét modell fokszám-eloszlása $3 + \beta$ paraméterű hatványeloszlást követ.

4.2.5. Tétel. [4] *A Móri-modell globális klaszterezettségi együttthatójának várható értéke is a 0-hoz tart n növelésével, ugyanis*

$$\mathbb{E}(C_1(G)) \sim K_1 \frac{\log n}{n}$$

teljesül, ha $\beta > 0$ és $n \rightarrow \infty$, ahol

$$K_1 = \frac{3 \left(m(m-1) \frac{(1+\beta)^2}{\beta^2} + m(m-1)^2 \frac{(1+\beta)^3}{\beta^2(2+\beta)} \right)}{\left(\frac{2+5\beta}{2\beta} m^2 + \frac{2-\beta}{2\beta} m \right)}$$

Bizonyítás. A tételt Eggemann és Noble bizonyította [4]-ben $\beta > 0$ feltétel mellett. Mivel a modell az LCD-modell egy általánosítása, így az ahhoz tartozó bizonyításra támaszkodhatunk ebben az esetben is. Az eltérő eredmény miatt azonban ennek főbb lépéseit is végigvesszük.

$$\mathbb{E}(C_1(G)) = \mathbb{E} \left(\frac{3T(G)}{P_2(G)} \right) \sim \frac{3\mathbb{E}(T(G))}{\mathbb{E}(P_2(G))}$$

fennállása miatt meg kell határoznunk a gráfban szereplő háromszögek és tripletek várható értékét.

4.2.6. Tétel. *Legyen $G_{n,m}^\beta$ egy BO-beli véletlen gráf, ekkor a benne szereplő háromszögek várható értékére a következő teljesül:*

$$\mathbb{E}(T) = \left(m(m-1) \frac{(1+\beta)^2}{\beta^2} + m(m-1)^2 \frac{(1+\beta)^3}{\beta^2(2+\beta)} \right) \log n + O(1).$$

Bizonyítás. Jelöléseink megegyeznek az előzőkben bevezetettekkel és az elindulásunk is megegyezik: tekintsük $G_{n,1}^\beta$ esetét, vagyis amikor minden lépésünkben pontosan egy csúcsot és élet adunk a gráfunkhoz.

4.2.7. Lemma. *Legyen S $G_{n,1}^\beta$ -nak egy lehetséges részgráfja és $\beta > 0$. Ekkor annak a valószínűsége, hogy $G_{n,1}^\beta$ valóban tartalmazza S -et így írható le:*

$$P_S = \frac{\beta}{\beta + d_s^{be}(1)} \prod_{i \in V^-(S)} \frac{\Gamma(1 + \beta + d_S^{be}(i))}{\Gamma(1 + \beta)} \prod_{i \in V^+(S)} \frac{1}{(2 + \beta)(i - 1) - 2} \prod_{t \notin V^+(S)} \left(1 + \frac{C_S(i)}{(2 + \beta)(i - 1) - 2} \right),$$

ahol $C_S(i)$ az $E(S)$ -beli élek száma amelyek az $\{i, \dots, n\}$ halmazból $\{1, \dots, i - 1\}$ -be mennek. Ebből

$$P_S = \frac{\beta}{\beta + d_s^{be}(1)} \prod_{i \in V^-(S)} \frac{\Gamma(1 + \beta + d_S^{be}(i))}{\Gamma(1 + \beta)} \prod_{(ij) \in E(S): i > j} \frac{1}{(2 + \beta)(i + \beta j)^{\frac{1}{2+\beta}}} \exp \left(O \left(\sum_{j \in S} \frac{C_S^2(j)}{j - 1} \right) \right).$$

Az analógia alapján $1 \leq a \leq b \leq c \leq n \in G_{n,m}(V)$ pontosan akkor alkot háromszöget, ha $\exists i, i', j, j', k, k' \in V(G_{nm,1})$, $m(a-1) < i, i' \leq ma$, $m(b-1) < j, j' \leq mb$, $m(c-1) < k, k' \leq mc$ úgy, hogy $ij', jk', i'k \in E(G_{nm,1}^\beta)$.

Legyen S $G_{nm,1}$ egy olyan részgráfja, amelyre $V(S) = \{i, i', j, j', k, k'\}$, $E(S) = ij', j'k, i'k$, és az előző bizonyítástól eltérő módon már itt vegyük figyelembe, hogy ezeket hogyan választhatjuk.

1. eset: $i = i'$ esetén $d_i^{be} = 2$, $d_{j'}^{be} = 1$ és minden más $s \in S$ -re $d_s^{be} = 0$. Alkalmazzuk a 4.2.7. lemmát:

$$P_S = \frac{\Gamma(3 + \beta)\Gamma(2 + \beta)}{\Gamma^2(1 + \beta)} \frac{1}{(2 + \beta)^3} \left(\frac{1}{i^2 j' (j k k')^{1+\beta}} \right)^{\frac{1}{2+\beta}} \exp \left(O \left(\frac{1}{a} \right) \right) = \frac{(1 + \beta)^2}{(2 + \beta)^2} \frac{1}{m^3} \left(\frac{1}{a^2 b^{2+\beta} c^{2+2\beta}} \right)^{\frac{1}{2+\beta}} \left(1 + O \left(\frac{1}{a} \right) \right).$$

Emellett, az előző bizonyításban látott okok miatt $i = i'$ esetén $m^4(m-1)$ módon választhatjuk ki a csúcsainkat.

2. eset: $i \neq i'$:

$$P_S = \frac{(1+\beta)^3}{(2+\beta)^3} \frac{1}{m^3} \left(\frac{1}{a^2 b^{2+\beta} c^{2+2\beta}} \right)^{\frac{1}{2+\beta}} \left(1 + O\left(\frac{1}{a}\right) \right).$$

Itt i, i', j, j', k, k' megfelelő kiválasztására $m^4(m-1)^2$ lehetőségünk van.

1. és 2. esetet összevetve kapjuk, hogy fix a, b, c mellett a háromszögek várható értéke a következőképpen alakul:

$$\begin{aligned} & \mathbb{E}(T_{abc}) = \\ & = \left(m(m-1) \frac{(1+\beta)^2}{(2+\beta)^2} + m(m-1)^2 \frac{(1+\beta)^3}{(2+\beta)^3} \right) \left(\frac{1}{a^2 b^{2+\beta} c^{2+2\beta}} \right)^{\frac{1}{2+\beta}} \left(1 + O\left(\frac{1}{a}\right) \right). \end{aligned}$$

Ha ezt az értéket a, b, c összes lehetséges választása mellett nézzük, megkapjuk a tételben szereplő kifejezést:

$$\mathbb{E}(T) = \left(m(m-1) \frac{(1+\beta)^2}{\beta^2} + m(m-1)^2 \frac{(1+\beta)^3}{\beta^2(2+\beta)} \right) \log n + O(1).$$

□

4.2.8. Tétel. *(bizonyítás nélkül) Legyen $G_{n,m}^\beta$ BO-modellbeli, $\beta > 0$. Ekkor a gráfban szereplő tripletek számára teljesül az alábbi:*

$$\mathbb{E}(P_2) = \left(\frac{2+5\beta}{2\beta} m^2 + \frac{2-\beta}{2\beta} m \right) n + O\left(n^{\frac{2}{2+\beta}}\right).$$

A 4.2.6. és a 4.2.8. Tétel alapján meghatározhatjuk a globális klaszterezettségi együttható várható értékét.

$$\mathbb{E}(C_1(G_{n,m}^\beta)) = \frac{3 \left(m(m-1) \frac{(1+\beta)^2}{\beta^2} + m(m-1)^2 \frac{(1+\beta)^3}{\beta^2(2+\beta)} \right) \log n}{\left(\frac{2+5\beta}{2\beta} m^2 + \frac{2-\beta}{2\beta} m \right) n} + O\left(\frac{1}{n}\right)$$

□

4.2.3. Egyéb modellek

Barabási és Albert már a Preferential Attachment modell megalkotásakor is említenek olyan tényezőket, amelyeket nem építettek bele a kezdeti modellbe, de a valóságban szintén jellemzőek lehetnek a megfigyelt hálózatokra. Tekintsük át ezekből a legegyszerűbbeket!

Az LCD-modell és a Móri-modell esetében is, az eredeti BA-modellhez hasonlóan, az egyes csúcsok előnye kizárólag abból fakad, hogy mikor lettek hozzáadva a gráfhoz, hiszen a később érkező pontoknak kevesebb idejük van foksámuk növelésére, ezért hátrányba kerülnek az új élek behúzásakor a több szomszédal rendelkezőkkel szemben. Emiatt azoknak a csúcsoknak kellene a kész gráfban a legtöbb éllel rendelkezniük, akik a legkorábban kerültek a gráfba, s így a legtöbb ideig gyűjtögettek. Egy, a valós hálózatokban megfigyelt jelenség azonban ellentmond a modell ezen tulajdonságának. Időről időre hallhatunk olyan példáról, amikor egy új érkező gyökeresen felforgatta a rendszerben addig kialakult status quot. Gondoljunk csak például a hirtelen felbukkanó és hatalmas népszerűsége szert tevő honlapokra, alkalmazásokra, vagy a váratlanul berobbanó, óriási ismertségre szert tevő zenekarokra. Egy idő után sikerük növekedése magyarázható a „gazdag egyre gazdagabb lesz” elvvel, az viszont nem, hogy hogyan győzhetik le látványosan a már ismert vetélytársaikat.

Erre magyarázatot az alkalmassági együttható bevezetése adhat, vagyis ha a Móri-modellnél látott általános kezdeti attraktivitás helyett az egyes csúcsok rendelkeznek egy saját β_i alkalmassági mutatóval, amely azt határozza meg, hogy a többiekkel szemben mennyire versenyképesek az új élek bevonzásában. Minél nagyobb egy csúcs alkalmassága, annál nagyobb eséllyel ér el kitörő sikert – nagy foksámot –, leghagyva a régebbi csúcsokat. Ezt a saját alkalmasságot, vagy fitnesszt is figyelembe véve állítható fel az *alkalmassági modell*.

Szintén érdemes lehet foglalkozni a hálózatba már bekerült csúcsok későbbi viselkedésével. Ha újra a valóság ábrázolására gondolunk, könnyen elképzelhető, hogy két már meglévő csúcs között spontán létrejön egy új él, ahogyan az is, hogy a pontjaink idővel öregedhetnek, ami során fitnesszük csökkenhet, vagy akár el is tűnhetnek.

A dolgozat további részében nem fogunk foglalkozni ezekkel a specifikus esetekkel, helyette egy általánosabb matematikai megközelítés kerül bemutatásra.

5. A PA-osztály

5.1. A preferenciális kapcsolódás általánosítása

Az előző fejezetben láthattunk, hogy a Barabási-Albert modell ötletét felhasználva, de különböző módosításokat végrehajtva különböző modellekhez jutunk. Ezek a modellek azonban mind fokszámeloszlásukban, mind a klaszterezettségi együtthatójuk aszimptotikus viselkedésében hasonlítanak egymásra. Érdekes lehet tehát az a kérdés, hogy be tudunk-e vezetni egy olyan paraméteres model családot, amely magába foglalja az egyes változatokat és általános leírását adja a preferenciális kapcsolódás modelljeinek.

Ostroumova, Ryabchenko és Samosvat [8]-ben bevezeti a PA-osztály fogalmát, amely megoldást nyújt a feladatra.

5.1.1. Definíció (PA-osztály). Legyen $G_{n,m}, n > n_0$ véletlen gráf a PA-osztálybeli, ha a kezdeti $G_{n_0,m}$ gráfból kiindulva azt minden lépésben egy csúccsal és ahhoz kapcsolódó m éllel bővítjük a következők szerint:

$$P(d_i^{n+1} = d_i^n | G_{n,m}) = 1 - A \frac{d_i^n}{n} - B \frac{1}{n} + O\left(\frac{(d_i^n)^2}{n^2}\right) \quad (2)$$

$$P(d_i^{n+1} = d_i^n + 1 | G_{n,m}) = A \frac{d_i^n}{n} + B \frac{1}{n} + O\left(\frac{(d_i^n)^2}{n^2}\right) \quad (3)$$

$$P(d_i^{n+1} = d_i^n + j | G_{n,m}) = O\left(\frac{(d_i^n)^2}{n^2}\right), 2 \leq j \leq m \quad (4)$$

$$P(d_{n+1}^{n+1} = m + j | G_{n,m}) = O\left(\frac{1}{n}\right), 1 \leq j \leq m, \quad (5)$$

$\forall i \in V(G_{n,m})$, ahol d_i^n az i csúcs fokszáma $G_{n,m}$ -ben.

A és B paraméterekre fent kell állnia $2mA + B = m$ -nek és $0 \leq A \leq 1$ -nek, hiszen egy lépésben minden élre

$$\begin{aligned} &P(\text{egy behúzendó élnek végpontot választunk}) = 1 = \\ &= \frac{1}{m} \sum_{i \leq n} A \frac{d_i^n}{n} + B \frac{1}{n} = \frac{1}{m} \left(A \frac{2mn}{n} + B \frac{n}{n} \right) = \frac{1}{m} (2Am + B). \end{aligned}$$

Nézzük meg, mit is jelentenek a definíció elvárásai: a 2 - 4 valószínűségek egy tetszőleges $i \leq n$ csúcsra vonatkoznak, az eddig látott preferenciális kapcsolódásos modellekhez hasonlóan az esélyük egy él szerzésére az addigi fokszámukkal arányos, kompenzálva egy uniform taggal. Vegyük észre, hogy az LCD és a BO-modellek definiálásával ellentétben itt nem térünk át a $G_{nm,1}$ gráfra, 4-ben pedig kikötjük, hogy az m él közül egy csúcs kis valószínűséggel szerezzon meg többet is. Az 5 egyenlet már az újonnan hozzáadott $n + 1$ csúcsról szól, aminek fokszáma m lenne, ha csak már meglévő csúcsokba indulnának belőle, így az egyenlet a hurokélek alacsony valószínűségét garantálja.

5.1.1. Lemma. *Az LCD-modell a PA-osztályba tartozik $A = \frac{1}{2}$ és $B = 0$ paraméterekkel. A Móri-modell a PA-osztályba tartozik $A = \frac{1}{\beta+2}$ és $B = \frac{m\beta}{\beta+2}$ paraméterekkel.*

A PA-osztály definiálásával modellek egy kellően tág csoportját határoztuk meg. [8]-ban belátták, hogy ezek mindegyikének fokszámeloszlása hatványeloszlást követ $\gamma = 1 + \frac{1}{A}$ paraméterrel, tehát megfelelnek a valós hálózatok ezen követelményének.

Az előző fejezetben láttuk, hogy az 5.1.1. Lemma alapján ide tartozó LCD- és Móri-modellekhez tartozó véletlen gráfokban a globális klaszterezettségi együttható n növelésével a nullához tart. Kérdés, hogy mit mondhatunk általánosan a PA-osztályba tartozó modellekről.

5.1.2. Lemma. [8] *Minden PA-osztálybeli modellben a benne szereplő tripletlek számára (P_2) a következő teljesül whp:*

1. $2A < 1$ esetén $P_2(n) \sim \left(2m(A+B)\frac{m(m-1)}{2}\right) \frac{n}{1-2A}$
2. $2A = 1$ esetén $P_2(n) \sim \left(2m(A+B)\frac{m(m-1)}{2}\right) n \log n$
3. $2A > 1$ esetén $\forall \epsilon > 0$ -ra $n^{2A-\epsilon} < P_2(n) < n^{2A+\epsilon}$

Az 5.1.3. Lemmát összevetve azzal a megállapítással, hogy a megjelenő háromszögek száma a gráfban, $T(n) \leq O(n)$, hiszen minden lépésben legfeljebb $\frac{m(m-1)}{2}$ háromszög jöhetett létre, a következőket kapjuk:

1. $2A < 1 \iff \gamma > 1 + \frac{1}{1/2} = 3$ esetén a $C_1(G_{n,m}) \leq O(1)$
2. $2A = 1 \iff \gamma = 3$ esetén $C_1(G_{n,m}) \leq O\left(\frac{1}{\log n}\right)$
3. $2A > 1 \iff \gamma < 3$ esetén $C_1(G_{n,m}) \leq O(n^{-\lambda})$, $\lambda > 0$

Láthatjuk, hogy abban az esetben is, amikor minden lépésben valóban létre is jön $O(1)$ háromszög, legfeljebb $\gamma > 3$ esetén lehet esélyünk konstans globális klaszterezettségi együttthatót elérni. Emellett az ide tartozó modellek klaszterezettségi együttthatója még rögzített A és m paraméterek mellett is túl különböző viselkedést mutathat. Annak az igénye motiválta tehát a bevezetett, de túl általánosnak tűnő PA-osztály egy megszorítását, hogy ott a klaszterezettségi együtttható már egységesebben meghatározható legyen.

5.2. A T-osztály

5.2.1. Definíció. [9] Azt mondjuk, hogy a $G_{n,m}$ véletlen gráfmodell T-osztálybeli, ha PA-osztálybeli (azaz kielégíti a 2-5 feltételeket) és teljesül rá a következő is:

$$P(d_i^{n+1} = d_i^n + 1, d_j^{n+1} = d_j^n + 1 | G_{n,m}) = e_{ij} \frac{D}{mn} + O\left(\frac{d_i^n d_j^n}{n^2}\right), \quad (6)$$

ahol e_{ij} az i és j csúcsok között futó élek száma, $D > 0$ konstans.

5.2.1. Lemma. *Amennyiben megengedjük a D paraméter 0 -nak választását, akkor az LCD- és a Móri-modell a T-osztályba tartozik $D = 0$ paraméterrel.*

Vizsgáljuk meg először C_1 , azaz a globális klaszterezettségi együtttható viselkedését a T-osztályba tartozó modelleknél.

5.2.2. Lemma. *Legyen $G_{n,m}$ T-osztálybeli, azaz teljesítse 6-t, ekkor whp $T(n) \sim Dn$.*

Bizonyítás. Az állítás bizonyításához először tekintsük az egy lépésben megjelenő háromszögek számának várható értékét. Belátható, hogy 6 teljesülése esetén a hibatag $o(1)$ -ben marad, így ez $D + o(1)$, így n lépés után $ET(n) = Dn + o(n)$. Ezek után belátható a várható érték körüli megfelelő koncentráció is. □

Az előző alfejezetben szereplő 5.1.2. Lemma minden PA-beli modellre, így a T-osztályra is vonatkozik. Ezt, illetve az 5.2.2. Lemmát felhasználva kaphatjuk meg C_1 -et.

5.2.3. Tétel. Minden T -alosztálybeli modellben a globális klaszterezettségi együtthatóra (C_1) a következő teljesül whp:

1. $2A < 1 \iff \gamma > 1 + \frac{1}{1/2} = 3$ esetén

$$C_1(G_{n,m}) \sim \frac{3(1-2A)D}{2m(A+B) + \frac{m(m-1)}{2}} \quad n\text{-től független konstans}$$

2. $2A = 1 \iff \gamma = 3$ esetén

$$C_1(G_{n,m}) \sim \frac{3D}{(2m(A+B) + \frac{m(m-1)}{2}) \log n} \sim O\left(\frac{1}{\log n}\right)$$

3. $2A > 1 \iff \gamma < 3$ esetén $\forall \epsilon > 0$ -ra

$$n^{1-2A-\epsilon} \leq C_1(G_{n,m}) \leq n^{1-2A+\epsilon} \implies C_1(G_{n,m}) \sim O(n^{-\lambda}), \text{ ahol } \lambda > 0$$

Azt kaptuk tehát, hogy a T -alosztálybeli modelleknél is van olyan eset, hogy a globális klaszterezettségi együttható a 0-hoz tart: $D = 0$ vagy $2A \geq 1$. Megfigyelhetjük, hogy mivel az LCD- és Móri-modellek esetén fennáll $D = 0$, így ez alapján a tétel alapján is azt kapjuk, hogy bennük a globális klaszterezettségi együtthatónak 0-hoz kell tartania.

Nézzük meg most a T -alosztályban az átlagos lokális klaszterezettségi együttható, C_2 viselkedését!

Belátható, hogy $G_{n,m}$ -ben az m fokszámú csúcsok száma whp legalább kn , ahol k egy pozitív konstans. Ebből a következő alsó korlátot adhatjuk C_2 -re:

$$C_2(G_{n,m}) = \frac{1}{n} \sum_{i \in V(G)} c(i) \geq \frac{1}{n} \sum_{i \in V(G), d_i=m} c(i) \geq \frac{1}{n} kn \frac{D}{m(m-1)/2} = \frac{2kD}{m(m-1)}$$

Látható, hogy 6 és $D \neq 0$ teljesülése mellett C_2 nem tart a 0-hoz n növelésével, hiszen mindig egy pozitív konstanssal becsülhető alulról. Ezek után definiáljunk hát egy olyan T -alosztálybeli modellt, amelyre $D \neq 0$!

5.3. A Polinomiális modell

A PA-osztály, majd a T -alosztály vizsgálata után felállíthatunk egy konkrét modellt [8] alapján, amely megfelel a (2-6) kritériumainknak.

5.3.1. Definíció (Polinomiális modell). [8] Legyen $G_{n+1,m}$ egy véletlen irányított gráf úgy, hogy minden él a kisebb csúcsba mutasson. Azt mondjuk, hogy $G_{n+1,m}$ gráf a Polinomiális modellbe tartozik, ha egy kezdeti $G_{n_0,m}$

gráfból kiindulva azt minden lépésben egy csúccsal és m éllel bővítjük a következők szerint:

Legyen $P_{k,l}$ pozitív paraméter $\forall 0 \leq k \leq \frac{m}{2}$ és $\forall 0 \leq l \leq (m - 2k)$ egészre és $\sum_{k,l} P_{k,l} := 1$.

Az $(n + 1)$. lépés előtt $P_{k,l}$ valószínűségekkel választunk egy k_{n+1} és l_{n+1} számot, ez mutatja meg, hogy az m behúzandó él közül mennyit húzzunk be az egyes módok szerint.

Az élbehúzás lehetséges módjai:

1. TF: a már szereplő élek közül választunk egyet egyenletes eloszlás szerint és húzzunk egy-egy új élet a kezdő- és végpontjába is, háromszöget kialakítva.

2. PA: a már szereplő élek közül választunk egyet egyenletes eloszlás szerint és húzzunk egy új élet csak a végpontjába.

3. U: a már szereplő élek közül választunk egyet egyenletes eloszlás szerint és húzzunk egy új élet csak a kezdőpontjába.

Az $(n + 1)$. lépésben behúzzunk $2k_{n+1}$ élet a TF-módszerrel, l_{n+1} élet PA-módszerrel és $m - 2k_{n+1} - l_{n+1}$ élet U-módszerrel.

Az élek behúzásánál alkalmazott három módszer igényelhet némi magyarázatot. A TF a Triangle Formation rövidítése, ebben a lépésben ugyanis biztosan háromszögeket fogunk létrehozni, hiszen egy már létező él két végpontjába húzzunk új éleket. A PA a Preferential Attachment, preferenciális kapcsolódásra utal, hiszen egy csúcs fokszáma akkor fog növekedni ebben a lépésben, ha az a kiválasztott csúcs végpontja volt, azaz ebben a lépésben előnyben részesülnek a magasabb befokkal rendelkező csúcsok. A PA-lépéssel szemben az U Uniform lépést jelent, hiszen itt a kimenő élek után növekedhet az egyes csúcsok fokszáma, a kifoka pedig mindegyiküknek egységesen m .

5.3.1. Tétel. *A Polinomiális modell a PA-osztályba és a T-alsztályba tartozik.*

Bizonyítás. [8] alapján vezessük be $M_{n,m}^{k,l}(i_1, \dots, i_m)$ -et, ami megmutatja, hogy rögzített $k_{n+1} := k$ és $l_{n+1} := l$ értékekre mekkora valószínűséggel kapnak pont az i_1, \dots, i_m csúcsok új éleket az $(n + 1)$. lépésben.

$M_{n,m}^{k,l}(i_1, \dots, i_m) := P(i_1, \dots, i_{2k}$ csúcsok TF-lépésből kapnak új élt,

$i_{2k+1}, \dots, i_{2k+l}$ csúcsok PA-lépésből kapnak új élt, i_{2k+l+1}, \dots, i_m csúcsok

$$\text{U-lépésből kapnak új élt) = } \frac{1}{n^{m-2k-l}} \prod_{x=1}^k \frac{e_{i_{2x}i_{2x-1}}}{2mn} \prod_{y=2k+1}^{2k+l} \frac{d_{i_y}^{be}}{mn},$$

ahol e_{xy} a két csúcs között futó élek száma $G_{n,m}$ -ben, $|E(G_{n,m})| = mn$ és d_y^{be} a csúcs fokszáma $G_{n,m}$ -ben.

Az $(n+1)$. lépésünk vizsgálatához figyelembe kell vennünk k_{n+1} és l_{n+1} kiválasztásának összes lehetséges esetét. Ezt összevetve az előző eredménnyel, a következőt kapjuk:

$$P(i_1, \dots, i_m \text{ csúcsok kapnak új éleket az } (n+1) \text{ lépésben}) =$$

$$= \sum_{k=0}^{m/2} \sum_{l=0}^{m-2k} P_{k,l} M_{n,m}^{k,l}(i_1, \dots, i_m)$$

Az eddigi megalapozás után beláthatjuk a 2-6 feltételek teljesülését a Polinomiális modellre, ezzel pedig a T-alosztályba tartozását (ez az állítás a [8] cikkben bizonyítás nélkül szerepel).

Tekintsük először a 3 feltételt, d_i^n -et az egyszerűség kedvéért jelöljük d_i -vel, d_i^{n+1} -et pedig d_i^* -gal.

$$P_{k,l}(d_i^* = d_i + 1 | G_{n,m}) = P(\text{i élet kap U, TF vagy PA-módon, rögz. } k, l \text{ mellett}) =$$

$$= \frac{1}{n}(m-2k-l) + \frac{d_i}{2mn} 2k + \frac{d_i^{be}}{mn} l = \frac{m}{n} + \frac{(d_i - 2m)k}{mn} + \frac{(d_i^{be} - m)l}{mn} = \frac{m}{n} + \frac{(k+l)(d_i - 2m)}{mn}$$

Ha figyelembe vesszük k, l választását is:

$$\begin{aligned} P(d_i^* = d_i + 1 | G_{n,m}) &= \sum_{k,l} P_{k,l} \left(\frac{m}{n} \right) + \frac{(k+l)(d_i - 2m)}{mn} = \\ &= \frac{d_i}{n} \sum_{k,l} P_{k,l} \frac{k+l}{m} + \frac{m}{n} - \frac{2m}{n} \sum_{k,l} P_{k,l} \frac{k+l}{m} \end{aligned}$$

Legyen $A := \sum_{k,l} P_{k,l} \frac{k+l}{m}$, ekkor, kihasználva a PA-osztály definiálásánál feltett $B = m - 2mA$ állítást is:

$$P(d_i^* = d_i + 1 | G_{n,m}) = \frac{d_i}{n} A + \frac{1}{n} (m - 2mA) =$$

$$\frac{d_i}{n} A + \frac{1}{n} B.$$

A 3 feltételből következik 2 is:

$$P(d_i^* = d_i | G_{n,m}) = 1 - A \frac{d_i}{n} - B \frac{1}{n}.$$

A 4 feltétel arra az esetre vonatkozik, amikor i többször is bekerülne a kiválasztott csúcsok közé. Ekkor az előző érvelést módosítva már $j = 2$ -re látható, hogy minden tag $O\left(\frac{(d_i^n)^2}{n^2}\right)$ vagy $O\left(\frac{1}{n^2}\right)$ lesz, tehát teljesül

$$P(d_i^* = d_i + j | G_{n,m}) = O\left(\frac{(d_i)^2}{n^2}\right), 2 \leq j \leq m.$$

5

$$P(d_{n+1}^{n+1} = m + j | G_{n,m}) = O\left(\frac{1}{n}\right), 1 \leq j \leq m$$

A T-alosztályba tartozáshoz 6-nak is teljesülnie kell. Ezesetben két csúcs együttes beválasztását kell vizsgálnunk.

A 4 feltétel bizonyításánál látott gondolatmenet itt annyiban módosul, hogy külön kell vizsgálnunk a TF esetet, ahonnan

$$P(d_i^* = d_i + 1, d_j^* = d_j + 1 | G_{n,m}) = \sum_{k,l} P_{k,l} \frac{e_{ij}}{mn} k + O\left(\frac{d_i^n d_j^n}{n^2}\right)$$

Ekkor $D := \sum_{k,l} k P_{k,l}$ választás mellett:

$$P(d_i^* = d_i + 1, d_j^* = d_j + 1 | G_{n,m}) = e_{ij} \frac{D}{mn} + O\left(\frac{d_i^n d_j^n}{n^2}\right)$$

□

5.4. Empirikus eredmények a Polinomiális modellről

A Polinomiális modell empirikus vizsgálatát $m = 6$ választása mellett végeztem. A konkrét modellemben a TF lépéshez k -t ($0 \leq k \leq \frac{m}{2}$) mindig binomiális eloszlás szerint választottam. Ezután a PA lépéshez l ($0 \leq l \leq m - 2k$) az *a*) esetben egyenletes eloszlás szerint, a *b*) esetben szintén binomiális eloszlás szerint került kiválasztásra.

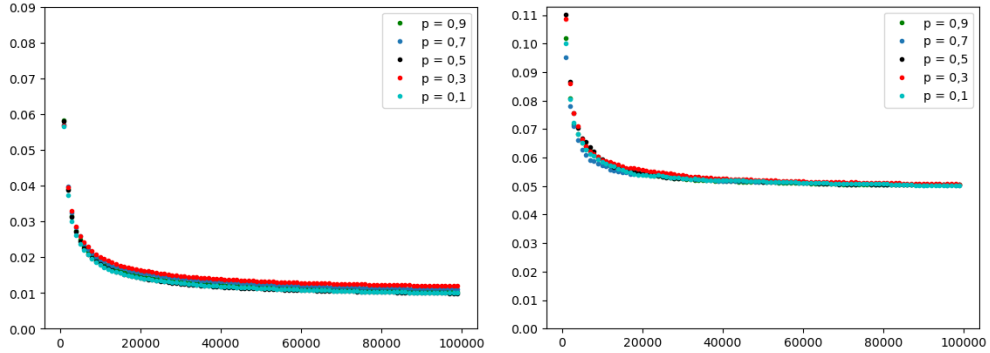
a) A TF-lépés binomiális eloszlásánál alkalmazott p paraméter függvényében a $P_{k,l}$ valószínűségeket a lenti táblázat foglalja össze.

l,k	0	1	2	3
0	$\binom{3}{0}(1-p)^{3\frac{1}{7}}$	$\binom{3}{1}(1-p)^2p^{\frac{1}{5}}$	$\binom{3}{2}(1-p)p^{2\frac{1}{3}}$	$\binom{3}{3}p^3$
1	$\binom{3}{0}(1-p)^{3\frac{1}{7}}$	$\binom{3}{1}(1-p)^2p^{\frac{1}{5}}$	$\binom{3}{2}(1-p)p^{2\frac{1}{3}}$	0
2	$\binom{3}{0}(1-p)^{3\frac{1}{7}}$	$\binom{3}{1}(1-p)^2p^{\frac{1}{5}}$	$\binom{3}{2}(1-p)p^{2\frac{1}{3}}$	0
3	$\binom{3}{0}(1-p)^{3\frac{1}{7}}$	$\binom{3}{1}(1-p)^2p^{\frac{1}{5}}$	0	0
4	$\binom{3}{0}(1-p)^{3\frac{1}{7}}$	$\binom{3}{1}(1-p)^2p^{\frac{1}{5}}$	0	0
5	$\binom{3}{0}(1-p)^{3\frac{1}{7}}$	0	0	0
6	$\binom{3}{0}(1-p)^{3\frac{1}{7}}$	0	0	0

$A = \sum_{k,l} P_{k,l} \frac{k+l}{m}$ összefüggés alapján $\forall p$ -re $A = \frac{1}{2}$ adódik, így az 5.2.3. Tétel alapján C_1 értékére az az elvárásunk, hogy n növelésével ne túl gyorsan, de a 0-hoz tartson.

A következő 6. ábrán p különböző választása mellett 5-5 véletlen gráf megfigyelt klaszterezettségi együtthatóit szemléltetem, az első ábrán a globális (C_1), a másodikon a lokális klaszterezettségi együttható (C_2) viselkedését.

Ahogy láthatjuk, C_1 értéke az elvárásnak megfelelően n növelésével a 0-hoz közelít, $C_2(G)$ viszont alulról korlátosnak tűnik.



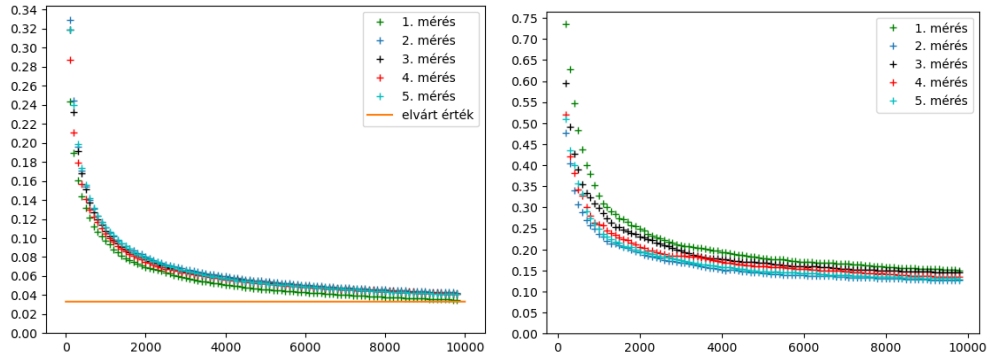
6. ábra. p különböző választása mellett $C_1(G)$ és $C_2(G)$ alakulása

b) Amennyiben az l paramétert is binomiális eloszlás szerint választjuk, már változtatható lesz az A együttható értéke. Válasszuk most p -t és q -t úgy, hogy $A < \frac{1}{2}$ teljesüljön! Ehhez $p := 0,5$, $q := 0,2$.

l, k	0	1	2	3
0	$\binom{3}{0}(1-p)^3\binom{6}{0}(1-q)^6$	$\binom{3}{1}(1-p)^2p\binom{4}{0}(1-q)^4$	$\binom{3}{2}(1-p)p^2\binom{2}{0}(1-q)$	$\binom{3}{3}p^3$
1	$\binom{3}{0}(1-p)^3\binom{6}{1}(1-q)^5q$	$\binom{3}{1}(1-p)^2p\binom{4}{1}(1-q)^3q$	$\binom{3}{2}(1-p)p^2\binom{2}{1}(1-q)q$	0
2	$\binom{3}{0}(1-p)^3\binom{6}{2}(1-q)^4q^2$	$\binom{3}{1}(1-p)^2p\binom{4}{2}(1-q)^2q^2$	$\binom{3}{2}(1-p)p^2\binom{2}{2}q^2$	0
3	$\binom{3}{0}(1-p)^3\binom{6}{3}(1-q)^3q^3$	$\binom{3}{1}(1-p)^2p\binom{4}{3}(1-q)q^3$	0	0
4	$\binom{3}{0}(1-p)^3\binom{6}{4}(1-q)^2q^4$	$\binom{3}{1}(1-p)^2p\binom{4}{4}q^4$	0	0
5	$\binom{3}{0}(1-p)^3\binom{6}{5}(1-q)q^5$	0	0	0
6	$\binom{3}{0}(1-p)^3\binom{6}{6}q^6$	0	0	0

Ekkor $A = \sum_{k,l} P_{k,l} \frac{k+l}{m} = 0,35$, ebből $B = m - 2mA = 0,65$ és $D = \sum_{k,l} kP_{k,l} = 1,5$ adódik. Ezekből az 5.2.3. Tétel alapján C_1 -re a következőt kapjuk:

$$C_1 = \frac{3(1-2A)D}{2m(A+B) + m(m-1)/2} = 0,03309.$$



7. ábra. $p = 0,5$, $q = 0,2$ választása mellett $C_1(G)$ és $C_2(G)$ alakulása

Az 7. ábrán demonstrált gráfok a Polinomiális modellek azon csoportjába tartoznak, amelyek globális klaszterezettségi együtthatójára kellően nagy n -ekre konstans értéket várunk el. Ez került ábrázolásra a mérési eredmények mellett, melyek láthatóan hozzásimulnak az egyenesünkhöz.

Az átlagos lokális együtthatójuk az empirikus eredmények alapján szintén nem tart a 0-hoz.

Összességében a gyakorlati vizsgálat és az elméleti eredmények alapján elmondható, hogy a Polinomiális modell $A < \frac{1}{2}$ és $D > 0$ esetén megfelel a valós hálózatok modelljeivel szemben támasztott klaszterezettségi elvárásoknak, hiszen globális és átlagos lokális klaszterezettségi együtthatójuk is szigorúan pozitív marad n növelésével.

6. Összegzés

Dolgozatomban igyekeztem egy koherens, átfogó képet adni a véletlen hálózatok különböző klaszterezettségi mutatóiról, összefüggéseikről és vizsgálatukról, különös tekintettel a globális és átlagos lokális klaszterezettségi együttthatókra.

Áttekintettük a klaszterezettségi együttthatók fogalmát, a magasabb szerveződési szintet vizsgáló diadikus klaszterezettséget és a konkrétan meghatározott, erősen klaszterezett csoportok megtalálását célzó modularitást.

Ezt követően láttuk a modularitás témakörében definiált konfigurációs modellre épülő Newton-féle algoritmusok működését. Ezek szolgáltatnak nekünk az elvárt klaszterezettségi együttthatót jól közelítő véletlen gráfokat, ehhez azonban külső megszorításokat alkalmaznak, így nem modellezik megfelelően a valós hálózatok önszerveződő létrejöttét.

Ezen igény mentén bemutattuk a dinamikus Barabási-Albert modellt, illetve az erre épülő LCD- és Buckley-Osthus-modelleket. Ezek fokszámeloszlásban és törvényszerűségeikben jól tükrözik a valódi hálózatok viselkedését, azonban beláttuk, hogy globális (és átlagos lokális) klaszterezettségi együttthatójuk a 0-hoz tart, ez pedig ellentmond a tapasztalt konstans értékeknek.

Az általános PA-osztály bevezetése részleges megoldást jelentett erre a problémára. Erről beláttuk, hogy valóban az eddig vizsgált modellek egy általánosítása, majd egy plusz feltétel bevezetése, vagyis a T-alosztály felállítása után elértük, hogy az ide tartozó modellek globális klaszterezettségi együttthatója a paraméterek megfelelő választása esetén egy pozitív konstans értékhez tartson.

A dolgozat zárásaként definiáltuk a konkrétan meghatározott Polinomiális modellt, ami a PA-osztályba, azon belül pedig a T-alosztályba tartozik. A modell empirikus vizsgálata alátámasztotta a korábbi elméleti eredményeket a T-alosztályba tartozó modellek globális klaszterezettségi együttthatójáról, és eredményekkel szolgált az átlagos lokális klaszterezettségi együtttható aszimptotikusan konstans viselkedéséről is.

Hivatkozások

- [1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [2] Béla Bollobás and Oliver M. Riordan. Mathematical results on scale-free random graphs. In *Handbook of graphs and networks*, pages 1–34. Wiley-VCH, Weinheim, 2003.
- [3] Pierce G. Buckley and Deryk Osthus. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Math.*, 282(1-3):53–68, 2004.
- [4] N. Eggemann and S. D. Noble. The clustering coefficient of a scale-free random graph. *Discrete Appl. Math.*, 159(10):953–965, 2011.
- [5] Lenwood S. Heath and Nidhi Parikh. Generating random graphs with tunable clustering coefficients. *Phys. A*, 390(23-24):4577–4587, 2011.
- [6] Attila Jung. Véletlen gráf családok modularitásának aszimptotikus viselkedése, szakdolgozat. 2019.
- [7] M. E. J. Newman. Random graphs with clustering. *Phys. Rev. Lett.*, 103:058701, Jul 2009.
- [8] Liudmila Ostroumova, Alexander Ryabchenko, and Egor Samosvat. Generalized preferential attachment: tunable power-law degree distribution and clustering coefficient. In *Algorithms and models for the web graph*, volume 8305 of *Lecture Notes in Comput. Sci.*, pages 185–202. Springer, Cham, 2013.
- [9] Liudmila Ostroumova Prokhorenkova. General results on preferential attachment and clustering coefficient. *Optim. Lett.*, 11(2):279–298, 2017.
- [10] M. Ángeles Serrano and Marián Boguñá. Clustering in complex networks. I. General formalism. *Phys. Rev. E (3)*, 74(5):No. 056114, 9, 2006.

NYILATKOZAT

Név: Babolcsay Barbara

ELTE Természettudományi Kar, szak: Matematika BSc

NEPTUN azonosító: WJ38E8

Szakdolgozat címe:

Véletlen hálózatok klaszterezettsége

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2023. 06. 06.



a hallgató aláírása