EÖTVÖS LORÁND TUDOMÁNYEGYETEM

TERMÉSZETTUDOMÁNYI KAR

Ádám Réka

VÉLETLEN GRÁFOK ÉS JÁRVÁNYTERJEDÉSI FOLYAMATOK

Szakdolgozat

Alkalmazott matematikus MSc, Sztochasztika szakirány

Témavezető:

Backhausz Ágnes, adjunktus

Valószínűségelméleti és Statisztika Tanszék

Budapest, 2017

Köszönetnyilvánítás

Ezúton szeretném kifejezni hálámat témavezetőmnek, Backhausz Ágnesnek a sok segítségért, melyet mind a szakdolgozatomban szereplő cikkek feldolgozásához adott, mind pedig a szimuláció készítésének folyamatában nyújtott. Ötletei és meglátásai hatalmas segítséget jelentettek a munka során, melyek nélkül e szakdolgozat nem jöhetett volna létre.

Tartalomjegyzék

Ábrák jegyzéke

Bevezetés

Szakdolgozatom témája a járványterjedési folyamatok modellezése véletlen gráfmodellekben.

Az első fejezetben egy áttekintő képet adok a napjainkig megalkotott véletlen gráfmodellekről és ismertetem azokat az eredményeket, melyek a járványterjedés modellezése szempontjából meghatározóak. Igyekszem az alapvető és régebbi, ámde a matematikai kutatás irányát meghatározó modellektől kezdve a legújabb modellekig bemutatni a véletlen gráfokat. Napjainkban egyre fontosabb a hatalmas hálózatok, például szociális hálózataink vizsgálata, ezért igen gyakori és hasznos az újabb és újabb matematikai modellek megalkotása és elemzése.

A második fejezetben a járványterjedés folyamatát leíró modelleket tárgyalom, központban a kontakt folyamat modellel. Ebben a fejezetben a főbb eredmények bemutatásán keresztül arra is rávilágítok, hogy milyen fontosak a járványterjedés matematikai modellje szempontjából a véletlen gráfokkal kapcsolatos eredmények.

A harmadik fejezet egy általam készített szimulációról szól. A szimuláció a duplikációs véletlen gráfmodellben sorsolt gráfokon történő járványterjedést modellezi, és központi kérdése a járványterjedési folyamat viselkedésének jellemzése a gráf paramétereinek függvényében. A szimuláció során készített ábrák szemléltetik a kisorsolt gráfok tulajdonságait és a rajtuk terjedő járványterjedési folyamat viselkedését.

1. fejezet

Véletlen gráfmodellek

A véletlen gráfmodellek vizsgálata és elemzése napjainkban nagy jelentőséggel bír, alkalmazásuk szerteágazó lehet a különféle hálózatok körében – például szociális hálózatok feltérképezésére használhatóak. A matematikai szemszögből történő kutatások kezdete az 1950-es és 1960-as évekre tehető és Erdős Pál és Rényi Alfréd nevéhez köthető, akik először fontos eredményeket bizonyítottak a később róluk elnevezett véletlen gráfmodellről ([1] p. 47 , p. 153., [2]). Azóta számos új modell született, hiszen a hálózatok empirikus vizsgálata során számos olyan tulajdonságot figyeltek meg, amelyeket a korábbi véletlen gráfmodellek – így például az Erdős–Rényi véletlen gráfmodell – nem teljesítettek.

A statikus modellek helyett, melyek egy adott, nagy méretű gráf kapcsolódásait modellezik, létrejöttek a dinamikus modellek, melyek már a gráf méretének növekedésével együtt magyarázzák a kapcsolatok alakulását. Az első ilyen később sokat tanulmányozott modell Barabási Albert–László és Albert Réka által javasolt verzióját ([3]) később Bollobás Béla és Oliver Riordan definiálta matematikailag precíz módon ([4]). Egy másik nagy csoportja a dinamikus modelleknek a duplikációs modellek, melyek csúcsok megkettőződésével fejlődnek és alapjuk a kapcsolatok "öröklődése" ezen megkettőződés során.

A véletlen gráfmodellek segítségével valószínűségi módszerekkel következtethetünk az általuk előállított hatalmas méretű gráf tulajdonságaira – ha ezek a tulajdonságok valódi hálózatokban is a megfigyelések alapján tapasztalhatóak, akkor ezeken kísérletezhetünk az információterjedés, járványterjedés modellezésével. Elsőként tehát azokról a tulajdonságokról írok, amelyeket egy véletlen gráfmodell esetében vizsgálhatunk és amelyeknek az eddigi kutatások alapján szerepe van, vagy későbbiekben szerepe lehet a járványterjedés modellezése szempontjából. A szakdolgozatom nagy részében *irányítatlan* gráfokat vizsgálok, ahol mégsem ott ezt külön jelzem.

1.1. Skálafüggetlenség

A skálafüggetlen jelenség egy olyan jellemzője lehet bizonyos valódi hálózatnak, illetve azok gráfreprezentációjának, mely azt mutatja, hogy a fokszámok jelentős változékonyságot mutatnak. Pontosabban fogalmazva az alacsony fokszámátlag ellenére is léteznek igen nagy fokszámú csúcsok a gráfban. ([1], p.7.) A matematikailag precízebb definíció felépítéséhez először szükség van az alábbi jelölésekre és definíciókra ([1] pp.12-13.) :

- Legyen $(G_n)_{n\geq 1}$ egy olyan gráfsorozat, melyre $G_n = (V_n, E_n)$, vagyis G_n egy n csúcsú gráf, $V_n = \{v_1, \ldots, v_n\}$ a csúcsok halmaza, $E_n \subseteq \{(v, w), v \in V_n, w \in V_n, v \neq w\}$ pedig az élek halmaza.
- Jelölje $P_k^{(n)}$ a pontosan k fokú csúcsok *arányát* G_n-ben.

$$P_k^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{d_i^{(n)}=k} \quad , \tag{1.1}$$

  ahol $d_i^{(n)}$ a v_i csúcs fokszáma G_n-ben.

Általában véve egy véletlen gráfmodell egy $(G_n)_{n\geq 0}$ gráfsorozatot állít elő, melynek egy–egy tagja, egy–egy G_n véletlen gráf tulajdonképpen egy–egy gráfértékű valószínűségi változó, az élek és a esetlegesen a csúcsok halmaza is függ a véletlentől.

1.1.1. Definíció. *(Ritka gráfsorozat). [1] p.12.*

$(G_n)_{n\geq 1}$ *gráfsorozat* ritka gráfsorozat, *ha*

$$\lim_{n\to\infty} P_k^{(n)} = p_k, \qquad \forall k \geq 0 \tag{1.2}$$

teljesül, valamilyen determinisztikus $(p_k)_{k\geq 0}$ *valószínűségi eloszlásra.*

Ha (G_n) *véletlen gráfok sorozata, akkor* $P_k^{(n)}$ *valószínúségi változót jelöl, tehát a limesz* sztochasztikus, eloszlásbeli *vagy éppen* 1 valószínűségű *konvergenciát jelöl,* $(p_k)_{k\geq 0}$ *marad determinisztikus.*

Értelmezhetnénk p_k-t nem determinisztikusként is, ez azonban a gyakorlatban nem igazán fordul elő. Valójában p_k azért valószínűségi eloszlás, mert a k fokú csúcsok arányának k-ra összegezve 1-et kell adnia: $\sum_{k=0}^{\infty} p_k = 1$. Mivel p_k a k fokú csúcsok relatív gyakorisága G_n-ben, ezért a fenti definíció helyett a következőt is mondhatnánk: Legyen X egy diszkrét valószínűségi változó $(p_k)_n$ eloszlással, X_n pedig egy véletlenszerűen választott csúcs fokszáma G_n-ben. Ekkor G_n ritka, ha $X_n \overset{n\to\infty}{\longrightarrow} X$ eloszlásban. Ezt a konvergenciát lecserélhetjük sztochasztikus, vagy 1 valószínűségű konvergenciára. A Barabási–Albert-modellben a legerősebb, 1 valószínűségi konvergencia is igazolható az 1.1.1. definícióban szereplő mennyiségekre, erről kicsit bővebben az 1.3. alfejezetben lesz szó.

Mivel $(p_k)_{k\geq 0}$ valószínűségi eloszlás a természetes számokon, ezért $\sum_{k=1}^{\infty} p_k = 1$, tehát szükségszerűen $p_k \overset{k\to\infty}{\longrightarrow} 0$, vagyis tetszőleges elég nagy méretű gráfot vizsgálva a csúcsok nagy része korlátos fokszámmal rendelkezik, ez indokolja az elnevezést. Legyen $F(k) = \sum_{i=1}^{k} p_i$ a fokszámeloszlás határeloszlásának eloszlásfüggvénye.

1.1.2. Definíció (Skálafüggetlen gráfsorozat). *[1] p.13.*

Egy $(G_n)_{n\geq 1}$ gráfsorozat skálafüggetlen, *ha* ritka gráfsorozat, *valamint valamely τ valós számra teljesül, hogy*

a)

$$\lim_{k\to\infty} \frac{\log\left[1 - F(k)\right]}{\log \frac{1}{k}} = \tau - 1,$$

VAGY egy gyengébb értelemben:

b)

$$\lim_{k\to\infty} \frac{\log p_k}{\log \frac{1}{k}} = \tau.$$

Mivel $\sum\limits_k p_k = 1 < \infty$, ezért szükségszerűen $\tau > 1$ teljesül.

Az 1.1.2. definíció értelmében

$$\frac{\log p_k}{\log \frac{1}{k}} = \tau + o(1), \qquad (\ o(1) \overset{k\to\infty}{\longrightarrow} 0\ )$$

melyből átrendezéssel kapjuk, hogy

$$\log p_k = (\tau + o(1)) \cdot \log k^{-1} \quad = \tau \log k^{-1} + o(1) \log k^{-1} = \tau \log k^{-1} + o(1) \log k^{-1} \tag{1.3}$$

$$= -\tau \log k + o(\log k) \tag{1.4}$$

$$p_k = k^{-\tau} \cdot e^{o(logk)}. \tag{1.5}$$

Az 1.5. alapján az alábbi mondható:

$$e^{o(logk)} \overset{k\to\infty}{\longrightarrow} C \Rightarrow p_k \sim C \cdot k^{-\tau}, \tag{1.6}$$

ahol $\sim$ asszimtotikus egyenlőséget jelöl,

$$\frac{p_k}{C \cdot k^{-\tau}} \overset{k\to\infty}{\longrightarrow} 1.$$

Abban a speciális esetben, ha az $o(\log k)$ tag konstans (jelölje ezt a konstanst c), akkor (1.4) az alábbi alakra hozható:

$$\log p_k = -\tau \log k + c. \tag{1.7}$$

$$p_k = C \cdot k^{-\tau} \tag{1.8}$$

Ebből az alakból az látszik, hogy $\log k$ függvényében ábrázolva a $\log p_k$ értékeket (k pozitív egész), az adott pontok egy negatív meredekségű egyenesen helyezkednek el. Az 1.6 egyenletet, vagy szigorúbb értelemben az 1.8 egyenletet is szokás *hatványtörvényként* nevezni.

Valódi hálózatok esetében persze p_k-t és $F(k)$-t nem tudjuk közvetlenül megfigyelni, hiszen tipikusan nagyon nagy, de mégis csak véges méretű hálózatok, vagy gráfok állnak rendelkezésünkre. Ha azonban elég nagy n méretű gráfokat vizsgálunk, akkor a hatványtörvény teljesülését szokás úgy vizsgálni, hogy vajon a megfigyelhető $P_k^{(n)}$ értékekre teljesül-e az (1.7) egyenlet. Ha tudjuk, vagy feltételezzük, hogy a gráf skálafüggetlen, akkor mivel $P_k^{(n)} \overset{n\to\infty}{\to} p_k$, ezért azt várjuk, hogy ha $\log k$ függvényében ábrázolva a megfigyelt $\log P_k^{(n)}$ értékeket, egy egyenest kapunk, akkor a határértékeloszlás $(p_k)_{k\geq 0}$ is hatványrendben cseng le. Ezért szokásos az alábbi, matematikalag nem precíz felírást vizsgálni:

$$\log P_k^{(n)} \approx -\tau \log k + c. \tag{1.9}$$

Fontos megjegyezni azonban, hogy itt $\approx$ egy közelítést jelöl, mely mögött nincs mindig precíz matematikai állítás ([1] p. 7.). Hasonló egyenlet felírható fix n esetén a k fokú csúcsok számára ([1] p. 7.).

Legyen $N_k \overset{def.}{=} |\{i | d_i^{(n)} = k\}| = nP_k^{(n)}$, ekkor:

$$\log N_k^{(n)} \approx -\tau \log k + c_n, \tag{1.10}$$

ahol $c_n = c - \log n$, hiszen

$$\log N_k^{(n)} = \log nP_k^{(n)} = \log n + \log P_k^{(n)}, \tag{1.11}$$

ezért (1.10) egyenletbe behelyettesítve (1.11) eredményét és átrendezve az egyenletet, visszakapjuk (1.9)-et.

Kivételes eset a $k = 0$ eset, ugyanis ekkor $\log k$ nem értelmezhető, vagyis az (1.7), (1.9), (1.10) kifejezések $k = 0$ esetben semmiképpen nem teljesülhetnek. Ettől azonban eltekinthetünk, mivel a skálafüggetlenség aszimptotikus tulajdonság. Az 1.1.2. definíció azt követeli meg, hogy a tapasztalati fokszámeloszlás konvergáljon egy olyan diszkrét valószínűségi eloszláshoz, mely *aszimptotikusan hatványfüggvény lecsengésű* - ezért véges sok kis k értékre elhanyagolható a fokszámeloszlás-függvény viselkedése.

Sok valódi hálózatra, illetve az azok alapján meghatározott gráfokra azonban éppen a kisebb és közepes k értékekre figyelhető meg a hatványtörvényt követő fokszámeloszlás, nagyon nagy k értékekre azonban már nem. Ez rendszerint akkor figyelhető meg, ha a nagyobb fokszámú csúcsok létrejötte nagy idő vagy pénzbeli költséggel is járna – például egy nagy szociális háló fenntartása és működtetése, vagy egy kutató több emberrel közös

publikációja. Ekkor nem számíthatunk olyan mértékű változékonyságra, mint amilyet a hatványtörvény alapján várnánk. Ezért egy másik alkalmazható eloszlásfajta ezekre a hálózatokra a hatványtörvényt követő eloszlás *exponenciális levágásssal* ([1]p.15.) :

$$p_k = const \cdot k^{-\tau} \cdot e^{\frac{k}{A}} \qquad k \geq 1.$$

Itt A tipikusan nagy érték, A-hoz képest kis k értékekre az eloszlás hatványrendben csökken, míg az A-hoz viszonyítva nagy k értékekre az exponenciális tag legyőzi a hatványrendűt, és az eloszlás exponenciális lecsengésű lesz.

A valódi hálózatok körében igen csak gyakorinak tűnnek a hatványtörvényt teljesítő fokszámeloszlások. Példaként hozható exponenciális levágású hatványtörvényt teljesítő fokszámeloszlásra az *Internet Movie Data base* alapján a filmszínészeket modellező gráf tapasztalati fokszámeloszlása, melyben a kapcsolatot a közös filmben való szereplés jelenti.([1] pp.28-30.). Vitathatóan, de egyes feltételezések szerint az Internet is – például ha a routereket tekintjük a gráf csúcsainak – hatványtörvényt teljesítő fokszámeloszlással rendelkezik, itt azonban felmerült, hogy a fokszámmérések módszere (traceroute-mérések) a felelősek a tapasztalt hatványtörvényért, nem pedig az Internet valódi fokszámeloszlása. ([1] pp. 8-10., pp.44-45.)

Érdekesség továbbá, hogy a hatványtörvényt legelső formájában *Zipf* (1929) fogalmazta meg, aki a szavak relatív gyakoriságát vizsgálta, és azt találta, hogy a k. leggyakoribb szó relatív gyakorisága, $f(k) = const \cdot k^{-\tau}$, és ebben az esetben τ értéke 1-hez közeli. Lotka törvénye, mely kémikusok körében vizsgálta, hogy hány olyan tudós van, akinek nevéhez $k = 2, 3, \ldots$ tudományos mű kapcsolódik, és megállapította, hogy ez is hatványtörvényt követ, itt a τ paraméter 2-höz közeli értéket vesz fel([1] p.43.), a Barabási–Albert-gráf esetén pedig $\tau = 3 + \frac{\delta}{m}$ (ld. 1.3.2. tétel).

Miért ilyen gyakori a skálafüggetlen jelenség a tapasztalati megfigyelések alapján? Erre magyarázatot adhat például a Barabási–Albert-modell fejlődési szabálya, melyet az 1.3. alfejezetben tárgyalok.

Skálafüggetlenség véletlen gráfmoddellekre vonatkozóan

Általában véve egy véletlen gráfmodell egy $(G_t)_{t\geq 0}$ gráfsorozatot állít elő, melynek egy–egy tagja, egy–egy G_t véletlen gráf tulajdonképpen egy–egy gráfértékű valószínűségi változó. Itt t pozitív egész számot jelöl, mégis az n helyett t-vel való indexelést az indokolja, hogy a gráf mérete az idő előrehaladtával növekszik, mint ahogyan a valódi hálózatok mérete is az idő előrehaladtával nő, új csúcsok és élek keletkeznek, bizonyosak pedig eltűnhetnek. Vannak természetesen olyan véletlen gráfmodellek is, melyek egy n csúcsú gráf éleinek alakulását közvetlenül, nem induktívan definiálják, a kapcsolatok itt is meghatározott szabályok szerint a véletlentől függenek (ilyen például az Erdős–Rényi vélet-

len gráfmodell). Az általam vizsgált gráfmodellek a gráf növekedését is modellezik (dinamikus modellek), tehát az időben való fejlődésre is utal az index. Szakdolgozatomban csak diszkrét idejű gráfsorozatokra szorítkozom, ahol az idő paraméter tulajdonképpen a gráf csúcsainak a száma.

Legyen $G_t = (V_t, E_t)$ egy a fenti értelemben vett véletlen gráf, ahol $V_t = \{v_1, \dots, v_t\}$ a csúcsok halmaza, $E_t \subseteq \{(v, w), v \in V_t, w \in V_t, v \neq w\}$ pedig az élek halmaza. Ha a gráfmodellre jellemző, hogy $V_{t-1} \subseteq V_t$ és $E_{t-1} \subseteq E_t$ teljesül $\forall t \in \mathbb{Z}^+$-ra, akkor ez azt jelenti, hogy sem a csúcsok, sem az élek nem törlődnek G_{t-1} csúcsai és élei közül a gráf fejlődése során. Ilyen például a Barabási–Albert-modell az általánosításaival, a parciális duplikációs modell és az általánosított duplikációs modell is (ezekről a modellekről bővebben 1.3. és az 1.5. fejezetekben lesz szó), de nem teljesül eme tulajdonság a Thörnblad–féle modell esetén, ahol G_{t-1} éleinek törlésére is sor kerülhet. (A Thörnblad-féle modell leírását ld. [5])

Mint már korábban is szerepelt, az 1.1.2. definícióban a konvergencia véletlen gráfok esetén eloszlásbeli, sztochasztikus vagy 1 valószínűségű konvergenciát jelöl. A véletlen gráfmodelleket tárgyaló irodalomban azonban az 1.1.2. definíció némiképp eltérő változatai kerülnek említésre ([6], [7]), a következőképpen.

Vezessünk be néhány új jelölést: legyen $F_k^*(G_t) = F_k^*(t) = \sum_{i=1}^{t} \mathbb{I}_{d_i^{(t)}=k}$, ahol ahol $d_i^{(t)}$ a v_i csúcs fokszáma G_t-ben. Jelölje $\mathbb{F}_k(t)$ ezek várható értékét, $f_k(t)$ pedig a k fokú csúcsok arányának a várható értékét. ($F_k^*(t)$ tehát a korábbi $N_k^{(n)}$ megfelelője, $f_k(t)$ pedig $\mathbb{E}(P_k^{(n)})$ megfelelője.)

$$\mathbb{F}_k(t) = \mathbb{E}(F_k(t)) = \mathbb{E}\left(\sum_{i=1}^{t} \mathbb{I}_{d_i^{(t)}=k}\right), \qquad f_k(t) = \frac{\mathbb{F}_k(t)}{t}.$$

Vegyük most a várható értékek limeszét $t \to \infty$ esetén: $f_k \stackrel{def}{=} \lim_{t\to\infty} f_k(t)$. Azt mondjuk, hogy egy véletlen gráfmodellben igaz a hatványtörvény, ha

$$f_k = \left(1 + O\left(\frac{1}{k}\right)\right) \cdot c \cdot k^{-b}$$

valamely b kitevővel és c konstanssal ([6] p. 242.). Ennél gyengébb definícióként használatos ([7] p. 6.):

$$\lim_{k\to\infty} k^b f_k = c.$$

A fenti definíciók azért kerültek külön megemlítésre, mert nem egyeznek meg azzal, hogy az 1.1.2. definícióban L^1-beli határértéket veszünk, csak abban az esetben, ha a határérték és a várhatóérték képzése felcserélhető.

1.2. Klaszteresedés

A gráfokban a csoportosulások megfigyelésére szolgáló mennyiség az ún. klaszteresedési együttható. Ez a mennyiség azt méri egy gráfban, hogy a csúcsok szomszédai milyen mértékben szomszédai egymásnak is.

Szociális hálózatainkra például nagy mértékben jellemző, hiszen ismerőseink többnyire egymást is ismerik. Ha a hálózatban igaz, hogy egy csúcs két szomszédja nagy valószínűséggel egymásnak is szomszédja, akkor az azt jelenti, hogy a gráfra jellemzőek a csoportosulások.

1.2.1. Definíció. *Klaszteresedési együttható ([1] p.17-18., [8])*

Legyen G egy n csúcsú gráf.

Legyen

$$\Delta_G = \sum_{1\le i,j,k\le n} \mathbb{I}_{((v_i,v_j),(v_i,v_k),(v_j,v_k)\in E)} = 6 \sum_{\substack{1\le i,j,k\le n\\ i<j<k}} \mathbb{I}_{((v_i,v_j),(v_i,v_k),(v_j,v_k)\in E)} = 6h,$$

ahol h a háromszögek száma G-ben.

Legyen

$$W_G = \sum_{1\le i,j,k\le n} \mathbb{I}_{((v_i,v_j),(v_i,v_k)\in E)} = 2 \sum_{\substack{1\le i,j,k\le n\\ j<k}} \mathbb{I}_{((v_i,v_j),(v_i,v_k)\in E)} = 2w,$$

ahol w a G-ben található cseresznyék, másnéven 2 hosszú utak száma.

Ekkor a klaszteresedési együttható

$$CC_G = \frac{\Delta_G}{W_G} = \frac{3h}{w}.$$

Az 1.2.1. definíció nem értelmezett azokra a gráfokra, ahol minden csúcs foka legfeljebb 1 (hiszen ekkor a nevező értéke 0 lenne). Mivel ezek a gráfok pontosan azok a gráfok, melyek izolált csúcsok és élek *diszjunkt* uniói, és ez a gyakorlatban a valós hálózatoknál meglehetősen ritkán fordul elő, ezért ez a megszorítás nem jelent túlzott korlátozást. ([8]).

Ez a CC_G szám felfogható úgy, mint annak a valószínűsége, hogy G-ben az illeszkedő élpárok közül egyenletesen kiválasztva egy élpárt, ők egy háromszöget alkotnak.

Fontos azonban, hogy ez nem egyezik meg azzal a valószínűséggel, hogy egy véletlenszerűen kiválasztott csúcs szomszédai közül szintén egyenletesen kiválasztunk két különböző szomszédot, akkor azok mekkora valószínűséggel ismerik egymást. Ez a valószínűség a klaszteresedési együttható egy másik definíciójához vezet, mely korábbról származik, mint az általam elsőként ismertetett, 1.2.1. definíció.

A klaszteresedési együtthatót lokálisan, egy csúcsra nézve is definiálhatjuk.

1.2.2. Definíció. *Lokális klaszteresedési együttható ([1] p.17-18.), [8])*

Rögzített i-re, v_i csúcsra tegyük fel,hogy $d_i \geq 1$.

$$CC_G(i) = \frac{\sum\limits_{1 \leq j,k \leq n} \mathbb{I}_{((v_i,v_j),(v_i,v_k),(v_j,v_k) \in E)}}{\sum\limits_{1 \leq j,k \leq n} \mathbb{I}_{((v_i,v_j),(v_i,v_k) \in E)}} = \frac{1}{d_i \cdot (d_i - 1)} \sum_{1 \leq i,j,k \leq n} \mathbb{I}_{((v_i,v_j),(v_i,v_k),(v_j,v_k) \in E)}$$

Ha $d_i \leq 1$, akkor $CC_G(i) \stackrel{def}{=} 0$.

$CC_G(i)$ tehát az a valószínűség, hogy v_i szomszédai közül véletlenszerűen kiválasztva két különbözőt, őket is él köti össze.

1.2.3. Definíció. *Klaszteresedési együttható 2. verzió*

$$CC_G^{(2)} = \frac{1}{n} \sum_{1 \leq i \leq n} CC_G(i)$$

Ez pontosan annak az eseménynek a valószínűsége, hogy egy véletlenszerűen kiválasztott csúcs további két, véletlenszerűen kiválasztott szomszédja között is él fut.

Létezik azonban olyan definíció is, mely csak a legalább 2 fokú csúcsokra végzi a lokális klaszteresedési együtthatók átlagképzését. Kérdéses, mennyire jogos a definícióban, hogy 1 fokú csúcsok 0 súllyal járulnak hozzá a klaszteresedési együtthatóhoz, ezért érdemes a globális klaszteresedési együttható elsőként ismertetett, 1.2.1 definíciójára támaszkodni ([8]). A szakdolgozat további fejezeteiben tehát én is ezt a definíciót használom.

1.3. Barabási–Albert-modell

Egy lehetséges magyarázat a skálafüggetlen tulajdonság kialakulására szociális és egyéb hálózatok esetén az először Barabási Albert-László és Albert Réka által megfogalmazott *preferencia-kapcsolódási szabály* (preferential attachment) [3], melynek lényege, hogy a hálózathoz csatlakozó új csúcs hajlamos a korábbi csúcsok közül a nagy fokszámmal rendelkezőkkel kapcsolatba lépni. Ez sok hálózat esetében megfigyelhető jelenség, hiszen például egy szociálisan aktív személllyel, akinek sok barátja van, könnyebben megismerkedhet egy új személy.

A Barabási–Albert-gráf áttörő újdonsága a korábbi gráfmodellekhez képest, hogy azt próbálja megragadni, *miként* jött létre egy hatalmas méretű gráf, mely teljesíti a skálafüggetlenséget, vagyis milyen fejlődési szabály lehet a felelős ezen tulajdonság kialakulásáért. Az Erdős–Rényi véletlen gráfmodell (ezen modellt és tulajdonságait e dolgozatban nem tárgyalom) egy fix n csúcsú gráfról mondja meg, mekkora valószínűséggel álljanak az egyes elemek kapcsolatban – bármely két elemet, egymástól függetlenül p valószínűséggel él köt össze ([1], p. 47.). Az így létrejövő, $(ER_n(p))_{n\geq 2}$ gráfsorozat *nem* skálafüggetlen gráfsorozatot állít elő, bár *skálafüggetlenné tehető*. Más modellek is születtek, melyek nagy n-re fokszámeloszlás tekintetében jól követik az empirikusan megfigyelt tulajdonságait az egyes hálózatoknak. Annak megértését azonban, hogy egy gráf miként változik egy új csúcs felvételével, milyen szabályok szerint fejlődik, a Barabási–Albert-modell, illetve a duplikációs modellek segítik, ezért dolgozatomban leginkább ezen modelleket tárgyalom.

A továbbiakban a már korábban bevezetett $(G_t)_{t\geq 0}$ jelölést használom az adott véletlen gráfmodell által előállított gráfsorozatra, ahol G_t egy t csúcsú gráfot jelöl, t nemnegatív egész szám. Hálózatok esetében az idő előrehaladtával új elem/személy kapcsolódik be a hálózatba, mely a régiekkel új kapcsolatokat épít ki, ez indokolja a méretben egyre növekedő gráfsorozatok vizsgálatát.

A Barabási–Albert-modell a preferencia kapcsolódási szabályt alapul véve azt modellezi, hogy G_{t-1}-ből milyen módon jön létre G_t, az új csúcs mekkora valószínűséggel kerül kapcsolatba a korábbi csúcsokkal. A preferencia-kapcsolódási szabály lényege, hogy az új t. csúcsból nagyobb valószínűséggel húzunk élt olyan, már meglévő, G_{t-1}-beli csúcshoz, melynek már sok szomszédja van – még pontosabban fogalmazva fokszám-arányos valószínűséggel húzunk éleket a már meglévő csúcsokhoz. Ez a kapcsolódási szabály bizonyos értelemben tényleg valósághű, hiszen ha valódi hálózatokra gondolunk, akkor egy új pont (új felhasználó, egy új személy, vagy akár egy új színész) nagyobb valószínűséggel kerül összeköttetésbe olyan régi tagokkal, akinek már sok kapcsolatuk van. Itt érdemes megjegyezni, hogy a Barabási–Albert-modellben a régebbi csúcsok közötti kapcsolat *nem változik*. Ez azért is reális feltételezés, mert számos hálózatban, mint például a

tudományos hivatkozások kapcsolatrendszere, vagy a szociális hálózatok, emberi ismeretségek hálózata során például elmondható, hogy ezek ha egyszer létrejöttek, nem szűnnek meg. Az új csúcsok speciális szabályok alapján kapcsolódnak a régiekhez. Ez persze számos hálózatra – mint például a facebook szociális hálójára – nem teljesül, hiszen ott lehetőség van a korábbi kapcsolatok vagy akár a korábbi csúcsok törlésére is.

1.3.1. Modell (Barabási–Albert-modell). *[4]*

A G_m^t gráfot induktív módon definiáljuk. A paraméterek jelentése: t a gráf csúcsainak száma, m pedig egy a gráf fejlődése során rögzített, pozitív egész, minden új csúcs keletkezése során a belőle újonnan húzott élek száma m.

A G_1^t gráfot akövetkező módon definiáljuk: Kezdetben kiindulhatunk G_1^0-ból, az üres gráfból, vagy G_1^1-ből, amely egy darab v_1 csúcsot és egy hurokélt tartalmaz. (Az ezután következő rekurzió alapján ugyanis az előbbiből 1 valószínűséggel létrejön az utóbbi.) Minden t. lépésben G_1^{t-1}-hez hozzáveszünk egy új, v_t csúcsot, és egy új (v_t, v_i) élt, ahol i-t az alábbi eloszlás szerint választjuk:

$$\mathbb{P}(i = s) = \begin{cases} \frac{d_{G_1^{t-1}}(v_s)}{2t-1} & s \leq t-1 \\ \frac{1}{2t-1} & s = t, \end{cases} \tag{1.12}$$

ahol $d_G(v_i)$ a v_i csúcs G-beli fokszámát jelöli.

Tehát a korábbi csúcsokhoz fokszámukkal arányos valószínűséggel húzunk új élt. Az új csúcs, v_t fokszámát már az új él behúzása előtt is 1-nek tekintjük, hiszen az új él biztosan ebből a csúcsból indul ki, Így az összfokszám, (ha az új él kiindulását még beleszámítjuk, de az él "másik végét" még nem), $2t - 1$ hiszen minden t. lépés után t darab éle van a gráfnak. Nyilvánvaló, hogy G_1^t-ben létrejöhetnek hurokélek, de nem jöhetnek létre többszörös élek.

Ezen rekurzió alapján kisorsolunk egy mt csúcsú, G_1^{mt} gráfot, melynek így mt darab éle is van, majd ezeket m-es csoportokban összehúzzuk egy–egy ponttá, így egy t csúcsú gráfot kapunk. Tehát ha G_1^{mt} csúcsai $v_1^1, v_2^1, \ldots, v_{mt}^1$, akkor az új, G_m^t-vel jelölt gráf csúcsait $(v_1^m, v_2^m, \ldots, v_t^m)$ pedig úgy kapjuk, hogy v_j^m csúcsba fut az összes, $v_{(j-1)m+1}^1, \ldots, v_{jm}^1$ csúcsba futó él.

Direkt módon is definiálhatjuk G_t^m-et. Induljunk ki egy csúcsból és m darab hurokélből. Az új, v_t csúcsból m darab élt húzunk a korábbi $v_1, \ldots, v_{t-1}$ csúcsokhoz. Fontos azonban, hogy egyesével *húzzuk ezeket az éleket, és minden él behúzása után* frissítjük a fokszámokat. *Ez azt jelenti, hogy ha a k-adik él behúzása következik, és az eddig behúzott élek egyik végpontja v_t, a másik pedig $w_1 \ldots w_k$, akkor* az 1.12.-*ben $d_{G_1^{t-1}}(v_s)$-et helyettesítse*

$$\tilde{d}_{G_1^{t-1}}(v_s) = d_{G_1^{t-1}}(v_s) + \#\{1 \leq i \leq k-1 \,|w_i = w_k\}.$$

1.3.2. Megjegyzés. *Az* (1.12) *egyenletben megadott eloszlás valóban egy diszkrét valószínűségi változó eloszlása.*

Jelölje $d_{G_1^{t-1}}(v_s)$*-et az egyszerűség kedvéért* d_s*.* G_1^{t-1}*-ben az élek száma* $t-1$*.*

$$\frac{\sum_{s=1}^{t-1} d_s + 1}{2t-1} = \frac{2(t-1)+1}{2t-1} = \frac{2t-1}{2t-1} = 1.$$

Az előbbi modellnek kétféle általánosítását fogom a továbbiakban bemutatni.

1.3.3. Modell (Általánosított Barabási–Albert-modell). *[1] pp.279-282.*

Az 1.3.1. modellhez hozzáveszünk egy $\delta \geq -m$ *paramétert. Ismét, először a* G_1^t *gráfot konstruáljuk meg, vagyis az* $m = 1$ *esetet nézzük. Ekkor feltehető, hogy* $\delta \geq -1$*.*

Az új él másik végpontjának eloszlását ((1.12) *egyenlet) az alábbiak szerint módosítjuk:*

$$\mathbb{P}(i = s) = \begin{cases} \frac{d_{G_1^{t-1}}(v_s)+\delta}{(2+\delta)(t-1)+(1+\delta)} & s \leq t-1 \\ \frac{1+\delta}{(2+\delta)(t-1)+(1+\delta)} & s = t, \end{cases} \tag{1.13}$$

Ebből az is látszik, hogy az "új" csúcsból minden "régi", v_s *csúcshoz* $d_{G_1^{t-1}(v_s)} + \delta$ *számmal arányos valószínűséggel húzunk élt.*

Jelölje az így kisorsolt gráfot $G_1^t(\delta)$*. Tekintsük most az* $m > 1$ *esetet. Az eredeti* δ *paraméterre, mely a modellt jellemzi, csak* $\delta \geq -m$ *áll fenn. Konstruáljuk meg az előbb ismertetett szabály alapján* $G_1^{mt}(\frac{\delta}{m})$ *gráfot, ez megtehető, hiszen* $\frac{\delta}{m} \geq -1$*. Ez azt is jelenti, hogy minden régi csúcshoz a régi fokszáma +* $\frac{\delta}{m}$ *-mel arányos valószínűséggel húzunk élt.*

Ezután hasonlóan az 1.3.1. modellben leírtakhoz, m-es csoportokat képezve a csúcsokból, húzzuk össze $G_1^{mt}(\frac{\delta}{m})$*-et egy t csúcsú gráffá, így kisorsoltuk a* $G_m^t(\delta)$ *gráfot. A* $\delta = 0$ *esetben visszakapjuk a Barabási–Albert-modellt (1.3.1).*

Ugyanúgy, mint a Barabási–Albert-modellben, direkt módon is definiálhatnánk a G_m^t *gráfot, nem pedig visszavezetve az* $m = 1$ *esetre. Ekkor kezdetben* 1 *csúcsból és m darab hurokélből kell kiindulnunk. Az új, t. csúcsból m darab új élt húzunk egyesével, és mindig frissítjük a fokszámokat. Jelölje a k. él behúzása előtt* d_i *az aktuális fokszámát* v_i*-nek. Minden k. új él behúzásánál minden korábbi,* v_i *csúcshoz* $d_i + \delta$*-val arányos valószínűséggel élt húzunk, kivéve az* $i = t$ *esetet, mert* v_t*-hez* $d_t + 1 + \frac{k\delta}{m}$ *-mel arányos valószínűséggel húzunk élt. Fontos, hogy az m él behúzása során, minden egyes él behúzása után* frissítjük a fokszámokat, *vagyis az éppen aktuális fokszámokhoz alakítjuk a valószínűségeket.*

1.3.4. Megjegyzés. *Az 1.13.-ben valóban egy diszkrét valószínűségi változó eloszlása szerepel.*

Jelölje ismét $d_{G_1^{t-1}}(v_s)$*-et* d_s*. Mivel* G_1^t *konstruálása során egy csúcsból és egy hurokélből indulunk ki, ezért kezdetben igaz az a feltevés, miszerint minden fokszám legalább* 1*. Minden lépésben* 1 *élt veszünk a gráfba, mely az új csúcsból indul ki, tehát minden új csúcs foka legalább* 1 *lesz.*

Éleket nem törlünk. Teljes indukciót használva így látható, hogy minden fokszám legalább 1 *lesz. Az összehúzás során a fokszámok nem csökkenhetnek, tehát minden* $1 \leq s \leq t-1$*-re* $d_s \geq 1$*, vagyis*

$$d_{G_1^{t-1}}(v_s) + \delta \geq 0,$$

ha $\delta > -1$*, valamint az 1.13.-ben tényleg valószínűségi eloszlást definiáltunk:*

$$\frac{\sum_{s=1}^{t-1}(d_s+\delta)+1+\delta}{(2+\delta)(t-1)+1+\delta} = \frac{2(t-1)+(t-1)\delta+1+\delta}{(2+\delta)(t-1)+1+\delta} = \frac{2(t-1)+t\delta+1}{2(t-1)+(t-1)\delta+1+\delta} = 1.$$

Az 1.13. modellnek még további két módosított változata is ismert ([1] p.281-282. modell (b), modell (c)), azonban arra, hogy teljesen kizárjuk a hurokélek megjelenését, a következő definíció alkalmas [9].

1.3.5. Modell (Hurokél nélküli Barabási–Albert-modell). *[9].*

Legyen $m \geq 2$ *egész és* $0 \leq \alpha \leq 1$ *rögzített. Kiindulunk egyetlen csúcsból és* 0 *darab élből, vagyis* $G_1 = (\{v_1\}, \emptyset)$*, vagy* két *csúcsból, melyeket m darab él köt össze,* $G_2 = (\{v_1, v_2\}, E_2)$*, ahol* E_2 *tartalmazza m-szer a* $\{v_1, v_2\}$ *párt. (A gráf növekedési szabálya miatt* G_1*-ből csak* G_2 *jöhet létre.)*

A definíció ismét induktív, G_{t-1} *segítségével definiáljuk* G_t*-t.* G_{t-1}*-hez hozzáveszünk egy új* v_t *csúcsot és belőle kiinduló m darab új élt. Az új* v_t *csúcs m darab szomszédját,* $w_1, w_2, \ldots, w_n$ *csúcsokat egymást követően sorsoljuk. Lehetnek köztük megegyező csúcsok (hiszen többszörös éleket meg akarunk engedni a gráfban).*

Az első szomszéd, w_1 *sorsolásának szabálya:*

- α *valószínűséggel egyenletesen választjuk* w_1*-et a* $\{v_1, \ldots, v_{t-1}\}$ *halmazból;*
- $1-\alpha$ *valószínűséggel a preferenciakapcsolódás szerint:* $\forall (1 \leq s \leq t-1)$*-re*

$$\mathbb{P}(w_1 = v_s) = \frac{d_{G_{t-1}(v_s)}}{Z}, \qquad Z = \sum_{i=1}^{t-1}(d_{G_{t-1}(v_i)}) = 2m(t-2).$$

A k. szomszéd, w_k *sorsolásának szabálya::*

- α *valószínűséggel egyenletesen választjuk* w_k*-t a* $\{v_1, \ldots, v_{t-1}\}$ *halmazból*
- $1-\alpha$ *valószínűséggel a preferenciakapcsolódás szerint:* $\forall (1 \leq s \leq t-1)$*-re*

$$\mathbb{P}(w_1 = v_s) = \frac{\tilde{d}_{G_{t-1}(v_s)}}{Z}, \qquad Z = \sum_{i=1}^{t-1}(\tilde{d}_{G_{t-1}(v_i)}) = 2m(t-2)+k-1,$$

ahol $\tilde{d}_{G_{t-1}(v_s)} = d_{G_{t-1}}(v_s) + \#\{1 \leq i \leq k-1 \,|w_i = w_k\}$.

Az általánosított Barabási–Albert-modell *skálafüggetlen* véletlen gráfsorozatot állít elő. Ennek bizonyításához először is be kell látni, hogy fokszámeloszlásának van határeloszlása, vagyis a kisorsolt gráfsorozot *ritka* gráfsorozat.

1.3.1. Állítás. *[1] p.284. (8.2.4.)*

Az álalánosított Barabási–Albert-modellre (1.3.3. modell)

$$p_k = \begin{cases} 0 & 0 \leq k \leq m-1 \\ (2+\frac{\delta}{m})\frac{\Gamma(k+\delta)\Gamma(m+2+\delta+\frac{\delta}{m})}{\Gamma(m+\delta)\Gamma(k+3+\delta+\frac{\delta}{m})} & k \geq m, \end{cases} \tag{1.14}$$

ahol p_k azonos az 1.1.2. definícióban szereplő határeloszlással, vagyis a k fokú csúcsok arányának sztochasztikus értelemben vett határértéke, Γ pedig a szokásos Γ-függvény.

Az 1.3.1. állítás bionyítását ld. [1] pp.286-303. Mint már az 1.1.2. definíciónál is említettem, a Barabási–Albert-modell esetén a k fokú csúcsok arányára nem csak a sztochasztikus, hanem az erősebb, 1 valószínűségű konvergencia is teljesül, ez következik [10] fő tételéből. Tehát a Barabási–Albert-modell *ritka gráfsorozatot* állít elő.

A speciális $\delta = 0$ esetben, $k \geq m$ -re a határeloszlás az alábbi alakra hozható:
[1] (8.4.4.)

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)},$$

$m = 1$ esetben pedig $p_k = \frac{4}{k(k+1)(k+2)}$, nem más, mint a Barabási–Albert-fa esetén a határeloszlás [11]. A Barabási–Albert-fa az 1.3.5. modell $m = 1, \alpha = 0$ esete. Az tehát, hogy a gráfban megengedünk-e hurokéleket, vagy sem, a határeloszlást nem befolyásolja.

Mivel tudjuk, hogy

$$\frac{\Gamma(k+\delta)}{\Gamma(k+\delta+3+\frac{\delta}{m})} = k^{-(3+\frac{\delta}{m})}\Big(1+O\Big(\frac{1}{k}\Big)\Big),$$

ezért (1.14) egyenletből adódik az alábbi tétel. ([1] pp. 284-286.)

1.3.2. Tétel. *Az általánosított Barabási–Albert-modellben (1.3.3. modell) a fokszámeloszlás határértékére teljesül:*

$$p_k = \left(1+O\Big(\frac{1}{k}\Big)\right)\cdot c_{m,\delta}\cdot k^{-\tau}, \qquad \tau = 3+\delta > 2, \qquad c_{m,\delta} = \frac{\Gamma(m+2+\delta+\frac{\delta}{m})}{\Gamma(m+\delta)}.$$

Tehát az általánosított Barabási–Albert-modell valóban skálafüggetlen gráfsorozatot állít elő, $\tau = 3+\delta > 2$ kitevővel.

1.4. Kitekintés továbbfejlesztett preferencia–kapcsolódáson alapuló modellekre

A preferencia-kapcsolódású modelleknek rendkívül sok továbbgondolása, elemzése és alkalmazása jelent meg, melyek közül csak párat igyekszem bemutatni, példákat adva az ebből induló bonyolultabb modellek megalkotására.

1.4.1. Versengés és preferencia-kapcsolódás

Az 1.3.5. modellt veszi alapul például az a modell [12], melynek az általam eddig bemutatott modellekhez képest új eleme, hogy a csúcsoknak többféle típusa is lehet, és a típusok fejlődését a gráf fejlődésével *együtt* modellezi. A típusokat a matematikai modellben színekkel is azonosíthatjuk, minden csúcs kap egy színt. A modell szemléletes jelentése lehet például az, hogy a piacon felbukkanó új termékről, vagy egy adott nézetről eltérő vélemények születnek, és az alapján, amit az ismerőseinktől hallunk, mi is egy-egy álláspontra helyezkedünk. Ez már azon újdonságok közé tartozik, melyet a szerzők hangsúlyoznak is, hogy a típusok kialakulását a gráf fejlődésével párhuzamosan, vagyis dinamikusan modellezzék. Hasonlóan egy járványterjedési folyamatnak a gráf fejlődésével párhuzamos modellezése is a legfrissebb modellek közé tartozik, erről bővebben a 2.3. alfejezetben írok.

A most következő modellben a csúcsoknak m féle típusa van, minden csúcshoz pontosan egy típus tartozik. A gráf fejlődésével *párhuzamosan* zajlik a típusfejlődés is. Tehát a hurokél nélküli Barabási–Albert-modellbe beépítjük a típusok alakulását is, szintén mint egy véletlen folyamatot. A továbbiakban tegyük fel, hogy 2 típus van (kék és piros). Kiindulunk egy G_0 gráfból, amely csúcsainak típusa is adott. A G_0 gráfból az 1.3.5. modellben leírtak szerint, a típusszámmal megyegyező, m paraméterrel, valamint egy másik, rögzített α paraméterrel fejlődik a gráf. Minden új csúcs hozzávételekor azonban a típusát is sorsoljuk. Az m él behúzása után, ha a kisorsolt m szomszéd közül k darab piros $(1 \leq k \leq m)$, akkor p_k valószínűséggel az új csúcs piros, $1 - p_k$ valószínűséggel pedig kék, ahol $p_k \in (0, 1)$ a modell előre megadott paramétere minden k-ra $(1 \leq k \leq m)$.

A modell *lineáris változatában* $p_k = \frac{k}{m}$ $\forall k$-ra, tehát a valószínűségek megegyeznek azzal, hogy a kisorsolt szomszédok között milyen arányban van jelen a kék és piros típus. Ellenkező esetben a modell *nemlineáris*.

A legfőbb kérdés mindkét modellben $n \to \infty$ esetén a piros és a kék csúcsok arányának határértéke a gráfban. Legyenek ezek az arányok egy konkrét n csúcsú gráfban a_n és b_n.

A *lineáris modellben*, ha a kiindulási gráfban piros és kék csúcsok is vannak, akkor a_n 1 valószínűséggel konvergens, legyen a határértéke a. Továbbá a eloszlásának tartója a

teljes $[0,1]$ intervallum, nem atomos (minden z-re $\mathbb{P}(a = z) = 0$), ahol $a = \lim_{n\to\infty} a_n$ és csak m-től és a kezdeti piros–kék aránytól függ. ([12] Theorem 1.1)

A *nemlineáris* esetben a_n pozitív, vagy nulla valószínűséggel tart az alábbi P polinom $[0,1]$-beli gyökeihez attól függően, hogy P hogyan viselkedik a gyökhelyek környezetében. ([12] Theorem 1.2 – 1.5.)

$$P(z) = \frac{1}{2}\sum_{k=0}^{m}\binom{m}{k} z^k (1-z)^{n-k}\left(p_k - \frac{k}{m}\right)$$

A P polinom a lineáris esetben a nullpolinom, tehát $[0,1] \subseteq Z_p = \{z|P(z) = 0\}$, míg a nemlineáris esetben Z_p atomos, ebből adódik a két modell közti különbség. A nemlineáris modellre vonatkozó állítás bizonyításához [12] szerzői ún. sztochasztikus közelítő folyamatokat használnak, mely bizonyos értelemben jól közelíti a közönséges $\frac{dz(t)}{dt} = P(z(t))$ differenciálegyenlet $(z_t)_{t\geq 0}$ megoldás trajektóriáit.

A modell m típusra is általánosítható, erről és az arra vonatkozó eredményekről bővebben ld. [12]. A modell előnye, mely a fő eredményekből, $m = 2$ esetben is látható, hogy számos paraméterbeállítás mellett, az egyes típusok aránya a gráfban nem tart 1-hez, és nem is hal ki egyik típus sem. Ez egy valósághű jelenséget tükröz.

1.4.2. Fitness

A Barabási–Albert-modellcsaládra jellemző, hogy tipikusan a régebbi csúcsok nagyobb fokszámmal rendelkeznek. Bizonyított, hogy a maximális fokszámmal rendelkező csúcs a fejlődés egy bizonyos pontján túl maximális marad [13], [14]. Ez azonban nem mindig felel meg a valóságnak azokban a hálózatokban, melyeket modellezni kívánunk, ezért a preferencia-kapcsolódáson alapuló modelleknek egy új ága született. A legkorábbi és legegyszerűbb modell a Barabási–Albert-modellen végez bizonyos módosításokat, a következőképpen:

Rögzítsünk egy μ valószínűségi eloszlást, és minden már meglevő vagy valamikor létrejövő v_i csúcshoz tartozzon egy F_i valószínűségi változó. Minden i-re F_i egymástól független és μ eloszlású. Ezek a gráf fejlődése során már *nem változnak*. Minden új csúcs létrejöttekor 1 élt húzunk a már meglevő csúcsokhoz, $F_i \cdot d_{G^{t-1}}$-vel arányos valószínűségekkel. Ebben a modellben ismert, hogy ha a csúcsokat a fokszámukkal arányosan választjuk ki a G_t gráfból, akkor a hozzájuk tartozó fitness értékek határeloszlása $t \to \infty$ esetén vagy abszolút folytonos μ-re ("fit get richer phase"), vagy van egy szinguláris komponense, mely $esssupp\ \mu$-re koncentrálódik ("condensation/Bose-Einstein phase")[13], [15]. Általában feltehetjük, hogy μ kopmakt tartójú, mert az ellenkező eset valójában [15] szerint érdektelen.

1.5. Duplikációs modellek

A duplikációs modellek többségében a biológiai hálózatok, például fehérjék közti kölcsönhatások leírásának céljából jöttek létre. Erre a modellcsaládra is jellemző, mint a Barabási–Albert-gráfra és általánosításaira, hogy a gráfot, mint egy bonyolult sztochasztikus folyamatot írják le, tehát a gráf növekedését magyarázzák.

Ezen modellcsalád alapja, hogy az idő előrehaladtával egy véletlenszerűen választott csúcs (esetleg csak bizonyos valószínűséggel) megkettőződik és az így létrejött, új csúcs a lemásolt csúcs szomszédságát is örökli, ám az így "örökölt" szomszédságban véletlenszerű módosításokat végzünk. Éppen ezért a duplikációs modellben is igaz, hogy a nagyobb fokú csúcsok egy új csúcs keletkezésekor nagyobb valószínűséggel kapnak élt, hiszen több szomszédjuk van, ezért nagyobb az esély rá, hogy valamelyik szomszédjuk lemásolásával keletkezett az új csúcs, és így jó eséllyel a szomszédság az új csúcsra is öröklődött.

1.5.1. Modell (Parciális duplikációs véletlen gráf). *[7] p.3.*

Rögzített egy $0 \leq q \leq 1$ valószínűség. Kiindulunk egy tetszőleges $G_{t_0} = (V_{t_0}, E_{t_0})$ gráfból, $V_{t_0} = \{v_1, \ldots, v_{t_0}\}$ a csúcsok halmaza, $E_{t_0} \subseteq \{(v, w), v \in V_{t_0}, w \in V_{t_0}, v \neq w\}$. Tehát ebben a modellben nem engedünk meg hurokéleket.

Definiáljuk a véletlen gráf fejlődési folyamatát a következőképpen:

Legyen $t \geq t_0$. Egyenletes eloszlás szerint az $\{1, 2, \ldots t-1\}$ halmazból kisorsolva i-t, a $v_i \in V_{t-1}$ csúcsot lemásoljuk, vagyis keletkezik egy új $v_t \notin V_{t-1}$ csúcs. $V_t = V_{t-1} \cup \{v_t\}$. Ami az élek halmazát illeti, $E_{t-1} \subseteq E_t$. Az új csúcs a lemásolt csúcs szomszédait örökli, majd egymástól függetlenül q valószínűséggel mégis töröljük az új csúcsból hozzájuk húzott éleket. Vagyis ha $(v_i, v_j) \in E_{t-1}$, akkor $1 - q$ valószínűséggel $(v_t, v_j) \in E_t$. (Tehát tulajdonképpen csak $1 - q$ valószínűséggel másoljuk le az éleket.)

Így induktív módon definiáltuk a G_t véletlen gráfot G_{t-1} segítségével: $G_t \overset{def.}{=} (V_t, E_t), ha (t > t_0)$.

Általában feltesszük, hogy G_{n_0} összefüggő véges gráf. Abban a speciális esetben, ha $q = 0$, vagyis minden élet biztosan átmásolunk, akkor összefüggő is marad és egy-egy csúcs pont olyan valószínűséggel (a fokszámával arányos valószínűséggel) kapcsolódik az új csúcshoz, mint a preferencia-kapcsolódás esetén. Az együttes eloszlás azonban mégis különböző lesz a két modellben, hiszen a Barabási–Albert-gráfban mindig m darab élet húzunk be, itt viszont a lemásolt csúcs fokszámával azonos számú élet húzunk be.

A parciális duplikációs modell (PD-modell) által definiált véletlengráf egy olyan gráfot eredményez, melyben lehetnek izolált csúcsok és összefüggő komponensek. Ha azonban feltesszük, hogy nem keletkeznek izolált csúcsok, akkor $q \to 1$ esetén határértékként

a preferencia-kapcsolódáson alapuló gráfmodellt kapjuk vissza. Ennek oka, hogy ha nagyon nagy valószínűséggel töröljük a lemásolt éleket, akkor várhatóan legfeljebb 1 élet másolunk le. Ha feltesszük, hogy nem keletkeznek izolált csúcsok, akkor pontosan 1 élet másolunk le. Ebben az esetben minden csúcshoz a fokszámaival arányos valószínűséggel kapja meg az új élet, hiszen erre akkor van esélye, ha az egyik szomszédját másoljuk le. ([7] p.4. Remark 2.5.(2))

Fontos megjegyezni továbbá, hogy itt a kisorsolt v_i csúcs, valamint a "másolata", vagyis v_n között *nem* fut él. Ennek köszönhető, hogy ha egy m osztályú gráfból indulunk ki, akkor a duplikációs modellben a gráfsorozat bármely tagja továbbra is m osztályú gráf marad. ([7]. pp. 4-5. Remark 2.4,2.5.)

1.5.2. Modell (Általános duplikációs modell, módosításokkal). *[6]*

Részben visszanyúlunk az 1.5.1. modell definíciójához, tehát kiindulunk egy $G_{t_0} = (V_{t_0}, E_{t_0})$ gráfból. Rögzítjük $0 \leq q \leq 1$ és $0 \leq r \leq 1$ valószínűségeket. Ezután a gráf növekedése a következők szerint alakul:

A t. lépésben a már meglevő G_{t-1} gráf egy egyenletesen választott csúcsát megkettőzzük, így jön létre a t. csúcs. Az új csúcs a lemásolt csúcs szomszédait is örökli G_{t-1} -ből, majd az ide vezető éleket mégis q valószínűséggel töröljük, $1 - q$ valószínűséggel megtartjuk. (Tehát eddig ugyanazt a fejlődési szabályt definiáltuk, csak más jelölésekkel, $q = 1 - p$-vel, mint a PD-modellben.)

Ezután két verzió szerint járhatunk el. Attól függően, hogy az alábbi két módosítás közül melyiket választjuk, a duplikációs modell kétféle módosított változatát kapjuk.

Módosítások:

1. *Rögzítünk egy a_1 számot, és ha az élek törlése után az új csúcs 0 fokú lett, akkor a_1 darab csúccsal összekötjük, melyeket visszatevés nélkül, egyenletesen választunk a $t - 1$ csúcs közül.*

2. *Rögzítünk egy a_2 számot, és az új csúcsot mindenképpen összekötjük a_2 darab visszatevés nélkül, egyenletesen választott csúccsal.*

Ezt követően a lemásolt csúcs G_{t-1}-beli nem-szomszédaihoz külön-külön, egymástól függetlenül $\frac{r}{t}$ valószínűséggel élt húzunk.

A [6] által megadott definícióban nincs kimondva, hogy a módosítások során visszatevés nélkül választjuk a csúcsokat. Az általam definiált modell, a visszatevés nélküli sorsolás következtében nem tartalmaz többszörös éleket. Fontos továbbá kiemelni, hogy a módosításokban ismertetett lépésekre a véletlenszerűen kiválasztott csúcs lemásolása, valamint "örökölt" éleinek esetleges törlése *után*, azonban a nem-szomszédok közüli sorsolás *előtt* kerül sor, így biztosítható, hogy összefüggő gráfot kapjunk. Ha $r = 0$ és egyik módosítást sem alkalmazzuk, akkor *tiszta duplikációs modellről* beszélünk, ha vagy az r

paraméter pozitív, vagy valamelyik módosítást beépítjük a modellbe (akár mindkettőt is beépíthetjük), akkor nevezzük [6] alapján a modellt *általánosított duplikációs modellnek.* A módosítások nélküli változatát a modellnek már [7] is ismerteti.

Az általános duplikációs modellben sem fut a kiválasztott csúcs és másolata között él. Következőnek egy olyan modellt mutatok be, melyben a másolat összeköttetésbe kerül a lemásolt csúccsal.

Az 1.5.1. és az 1.5.2. modellek továbbgondolásaként dolgozhatunk olyan modellben is, ahol a megkettőzött csúcsot nem egyenletesen választjuk G_{t-1} csúcsai közül, hanem például a preferencia–kapcsolódási szabály mintájára a fokszámukkal arányos valószínűséggel. Ez tehát valamiféleképpen a duplikációs modellek és a Barabási–Albert-modell ötvözése.

1.5.3. Modell (Preferenciális duplikációs modell). *[16]*

Kiindulunk egy összefüggő, véges G_{n_0} gráfból.

A G_t gráfot $(t \in \mathbb{N}, t > n_0)$ induktívan G_{t-1} segítségével definiáljuk:

A régi csúcsok közül fokszámarányos valószínűséggel kisorsolunk egyet: $\forall$ $(1 \leq s \leq t-1)$-re

$$\mathbb{P}(i = s) = \frac{d_{G_{t-1}(v_s)}}{Z}, \qquad Z = \sum_{j=1}^{t-1} \left(d_{G_{t-1}(v_j)} \right)$$

Az új, v_t csúcs v_i lemásolásával keletkezik. Az éleket egymástól függetlenül, p valószínűséggel megtartjuk, úgy, mint a PD-modellben (1.5.1.): ha $(v_i, v_j) \in E_{t-1}$, akkor p valószínűséggel $(v_t, v_j) \in E_t$. $G_t \overset{def}{=} (V_t, E_t)$

***Valódi iker–változat**: 1 valószínűséggel $(v_t, v_i) \in E_t$.*

***Hamis iker–változat**: $(v_t, v_i) \notin E_t$.*

A csúcsmásolás nem csak egyesével történhet. Egy olyan modellt is bemutatok, melyben egyszerre sok duplikációs lépés egy időben történik. Minden csúcs egyidejűleg megkettőződik, vagyis "gyereke születik".

1.5.4. Modell. *Szimultán duplikációs modell [17]*

Kiindulunk egy G_{t_0} gráfból és minden lépésben minden csúcs megkettőződik, a lemásolt csúcsokat szülőknek, a másolatukat a gyerekeiknek nevezzük. (A továbbiakban ezt a szóhasználatot követem, a szülő és gyerek alatt csúcsokat értek). Tehát G_t-ből G_{2t} fejlődik egy lépés alatt, a megkettőződések egy időben történnek $(t \geq t_0)$ pozitív egész).

- *G_t csúcsait és éleit változatlanul lemásoljuk. A szülők között új élek nem keletkeznek, a régi élek megőrződnek.*
- *Minden gyereket egymástól függetlenül β valószínűséggel él köt össze a saját szülőjével.*

- *Minden gyereket egymástól függetlenül γ valószínűséggel él köt össze más szülővel, ha az a saját szülőjének szomszédja volt G_t-ben.*

- *Bármely két gyereket is a többiektől függetlenül α valószínűséggel él köt össze, ha a szüleik is szomszédosak voltak G_t-ben.*

Az 1.5.4. modellnek tehát három paramétere van: α, β, γ. A paraméterek függvényében a kisorsolt gráfok fokszámeloszlásáról az alábbi tételek ismertek [17]:

1.5.1. Tétel. *Legyen $\beta = 0$.*

- *Ha $(1+\gamma)(\alpha+\gamma) \leq 1$, akkor annak a valószínűsége, hogy egy egyenletesen kiválasztott csúcs G_t-ben izolált csúcs, tart 1-hez, amidőn $n \to \infty$, a 0 fokú csúcsok aránya G_t-ben tart 1-hez, 1 valószínűséggel.*

- *Ha $(1+\gamma)(\alpha+\gamma) > 1$, akkor annak a valószínűsége, hogy egy egyenletesen kiválasztott csúcs G_t-ben izolált csúcs, tart q-hoz, amidőn $n \to \infty$, ahol $q < 1$.*

1.5.2. Tétel. *Legyen $\beta > 0$*

(a) Ha $(1+\gamma)(\alpha+\gamma) < 1$, akkor az 1.5.4. modell által előállított véletlen gráfsorozat ritka gráfsorozat, ahol az 1.1.1. definícióban (1.2) *egyenlőségben* 1 *valószínűségű határértéket veszünk.*
A határeloszlás,$(p_k)_{k\geq 0}$ egy olyan X valószínűségi változó eloszlása (vagyis $\mathbb{P}(X = k) = p_k$), melynek véges p-edik momentuma van, ha $(1+\alpha)^p + (\alpha+\gamma)^p < 2$ és nem véges a p-edik momentuma, ha $(1+\alpha)^p + (\alpha+\gamma)^p > 2$.

(a) Ha $(1+\gamma)(\alpha+\gamma) > 1$, akkor (1.2) *egyenlőségben szintén* 1 *valószínűségű határértéket véve $p_k \equiv 0$ minden k-ra.*

Fontos következménye az 1.5.2. (a) pontjának, hogy az 1.5.4. modell véletlen gráfsorozata *skálafüggetlen* gráfsorozatot állít elő az 1.1.2. definíció értelmében, $\tau = p - 1$ kitevővel, ahol p teljesíti a $(1+\alpha)^p + (\alpha+\gamma)^p = 2$ egyenlőséget. Az 1.5.2. (a) pontja alapján tudjuk erre a p-re, hogy minden $q < p$-re létezik a q-adik momentum, $q > p$-re viszont nem létezik. Ehhez az kell, hogy $\sum_{k\geq 0} p_k k^q < \infty \Leftrightarrow (q < p)$ teljesüljön, ehhez pedig pont az kell, hogy $p_k \sim Ck^{-(p+1)}$ igaz legyen.

2. fejezet

A járványterjedés modellezése

A járványterjedés, egy adott gráfon, vagy egy éppen fejlődő gráfon végbemenő sztochasztikus folyamat. Alkalmazása egy konkrét betegségtől kezdve az internetes vírusokon át akármilyen információterjedésként értelmezhető folyamatra történhet.

2.1. Kontakt folyamat és járványterjedési küszöb

A kontakt folyamat a SIS (susceptible-infected-susceptible) modellek családjába tartozik. Ez azt jelenti, hogy a gráf csúcsainak két állapota van, fertőzött és egészséges állapot. Egy adott t időpillanatban két csoportra oszthatóak a csúcsok állapotaik szerint; beteg és egészséges csúcsok halmaza. Az idő változásával az egészséges csúcsok halmazából először vagy újra a beteg csúcsok halmazába kerülhetnek a csúcsok, a fertőzöttek pedig meggyógyulhatnak, vagyis átkerülhetnek az egészségesek halmazába.

A kontakt folyamat egy olyan járványterjedési modell, melyben a megbetegedett csúcsok meggyógyulásáig eltelt idő exponenciális eloszlású, 1 paraméterrel, tehát átlagosan 1 időegység múlva bekövetkezik a gyógyulás. Ekkor azt mondjuk, hogy a meggyógyulás *rátája* 1. Az egészséges csúcsok ugyancsak exponenciális eloszlás szerinti idő elteltével fertőződnek meg, de a beteg szomszédaik számával arányos paraméterű exponenciális eloszlást követve. Ez az arányossági tényező egy rögzített λ valós szám, mely a járványterjedés egészére jellemző paraméter, minden csúcsra minden t időpillanatban állandó. Tehát a megfertőződés rátája $\lambda|N(v)|$, ahol $N(v)$ a v csúcs beteg szomszédainak halmaza a gráfban.

A legkönnyebben elképzelhető példa egy kontakt folyamattal leírt terjedésre egy internetes vírus terjedése. A számítógépeken levő vírusirtó szofterek időről időre kiszűrik a vírust, azonban nem óvják meg a számítógépet attól, hogy újra megtámadja ugyanaz a vírus, és a következő ellenőrzésig megint fertőzötté válhat.[9]

2.1.1. Definíció (Kontakt folyamat). *[9] p.7.*

Adott egy $G = (V, E)$ gráf, V a csúcsok, E az élek halmaza, valamint egy λ paraméter. A kontakt folyamat egy η_t folytonos paraméterű Markov-folyamat, mely $\forall v \in V$ csúcshoz a t időpillanatban 0 vagy 1 értéket rendel. (Így az egész gráfhoz hozzárendel egy $|V|$ dimenziós $0/1$ vektort.) Ha v csúcs egészséges, akkor $\eta_t(v) = 0$, ha fertőzött, akkor $\eta_t(v) = 1$. Így η_t tulajdonképpen az $A = \{v \in V | \eta_t(v) = 1\}$ halmaz által egyértelműen meghatározott. Az A halmaz tehát a fertőzött csúcsok halmaza, $V \setminus A$ az egészséges csúcsokat tartalmazza.

Az átmeneti ráták:
$$\begin{cases} A \to A \setminus \{v\} & \forall v \in A \qquad 1 \text{ rátával} \\ A \to A \cup \{v\} & \forall v \notin A \qquad \lambda \cdot |N(v)| \text{ rátával} \end{cases}$$

A kontakt folyamat tehát értelmezhető egy vektorértékű, vagy egy halmazértékű (A_t) sztochasztikus folyamatként is.

Belátható, hogy véges gráfokban 1 valószínűséggel kihal a kontakt folyamat által modellezett fertőzés. Ennek az az oka, hogy egy rögzített N csúcsból álló gráfban annak a valószínűsége, hogy minden csúcs meggyógyul, és egy élen sem terjed tovább a fertőzés, akármilyen kis Δt intervallumra pozitív, hiszen külön-külön egy csúcsra és egy élre a gyógyulásnak, illetve a nem fertőzésnek a valószínűsége pozitív, valamint a csúcsok egymástól függetlenül gyógyulnak, az élek egymástól függetlenül fertőznek. Ezért egy fix, N csúcsú gráfra a Δt intervallumon bleüli kihalás valószínűsége alulról becsülhető egy csak N-től függő konstanssal. Ezért 1 valószínűséggel valamely elég nagy t-re bekövetkezik a kihalás. Vagyis ha T elég nagy, akkor $A_T = \emptyset$, a T időpillanatban a fertőzött csúcsok halmaza üres. A kihalás után már nincs él, amely fertőzött és nem fertőzött csúcsok között megy, tehát a fertőzés nem tud tovább terjedni, azt mondjuk, hogy a fertőzés *kihal*. (Ez fordítva nem igaz, ha az összes csúcs fertőzött, akkor is 1 rátával meggyógyulnak.) Ezért véges gráfok esetében a betegség kihalásához szükséges időintervallum hosszát szokás vizsgálni. Ha egy adott betegség terjed a gráfban és a kihalásáig eltelt idő valamilyen *exponenciális függvénye* a csúcsszámnak, akkor azt mondjuk, hogy a betegség exponenciális sebeséggel halt ki. Mivel a járványterjedés egy véletlentől függő folyamat, ezért ha egy konkrét kimenetele esetén a kihalásig eltelt idő exponenciális a csúcsok számában (vagy esetleg még ennél nagyobb nagyságrendű), akkor a betegségből járvány alakult ki. Másképp fogalmazva, ha a kihalás szuper-polinomiális függvénye a csúcsok számának, akkor beszélhetünk járványról ([9] p.2.,7). Ennek alapján vizsgálhatjuk azt az egyes gráfmodellekben, mennyi annak valószínűsége annak, hogy egy betegség járvánnyá alakul.

2.1.2. Definíció (Járványküszöb véges gráfokra, kontakt folyamatra). *A járványterjedési küszöb az a λ_c szám, melyre teljesül, hogy minden $\lambda > \lambda_c$ paraméterű kontakt folyamatra igaz, hogy pozitív valószínűséggel a kihalás a szuperpolinomiális sebességű a csúcsok számában.*

Ha a fertőződés rátája nagyobb, akkor a megfertőződéshez szükséges várható időtar-

tam csökken, hamarabb fertőződik meg az egészséges csúcs, vagyis nagyobb a fertőzés valószínűsége. Tehát minél *nagyobb* a λ paraméter értéke, annál hamarabb fertőződnek meg az egészséges csúcsok, Tehát minél nagyobb a λ értéke, annál több időt vesz igénybe, hogy az összes csúcs meggyógyuljon és kihaljon a járvány. Egyrészt ez indokolja azt, hogy *legalább* exponenciális gyorsasággal nőjön a kihaláshoz szükséges idő, tehát minél gyorsabban nő ez a függvény a csúcsok számában, annál lassabban hal ki a fertőzés. Másrészt ezek alapján kereshetjük azt a kritikus λ értéket, melynél nagyobb paraméterű kontakt folyamatra legalább exponenciális nagyságrendben nő a csúcsok számával a kihaláshoz szükséges idő. Ebből látható, hogy a 2.1.2. definíció valóban értelmes.

Mivel az sok valódi hálózatra jellemző a skálafüggetlenség (1.1.2.), ezért természetes kérdés az, hogy vajon hogyan viselkedik a kontakt folyamat skálafüggetlen gráfokra. [9] fő eredménye szerint, melyet részeletesebben a következő 2.2. részben ismertetek, bármely n-re létezik egy olyan λ_n küszöbszám, hogy nagy valószínűséggel bármely $\lambda > \lambda_n$ fertőzési rátával terjedő kontakt folyamat konstans, pozitív valószínűséggel járvánnyá alakul és $\lambda_n \overset{n\to\infty}{\longrightarrow} 0$.

Mivel a skálafüggetlen gráfok egyben ritka gráfok is (1.1.1), vagyis a csúcsok nagy része korlátos fokszámmal rendelkezik, ha n elég nagy, ezért először érdemes megvizsgálni, hogyan viselkedik a kontakt folyamat egy korlátos fokszámú gráfban.

2.1.3. Állítás. *[9] Lemma.5.1.*

Ha G egy olyan gráf, amelyben a csúcsok maximális fokszáma d, akkor annak a valószínűsége, hogy valamikor is k-nál több csúcs lesz fertőzött a gráfban, $\forall$ k-ra felülről becsülhető $(4\lambda d)^k$ értékkel.

Bizonyítás.

Feltehető, hogy $\lambda d < \frac{1}{4}$, hiszen ellekező esetben $(4\lambda d)^k > 1$ és mivel egy valószínűségre adunk felső becslést, triviális az állítás.

Jelölje $X = |A|$ a fertőzött csúcsok számát bármely időpillanatban. A kontakt folyamat esetében 0 annak a valószínűsége, hogy egyszerre több esemény is bekövetkezzen a csúcsok gyógyulása és megfertőződése közül *ugyanabban az időpillanatban*. Ezért X is egy exponenciális eloszlású időközönként változó folyamat, az alábbi átmeneti rátákkal:

$$X \to X - 1 \qquad X \text{ rátával} \tag{2.1}$$

$$X \to X + 1 \qquad \lambda|c(A, \overline{A})| \text{ rátával} \tag{2.2}$$

Minden csúcs gyógyulásának időpontja 1 paraméterű exponenciális, ezért az X darab csúcs egyikének gyógyulása (X csökken 1-gyel) exponenciális eloszlások minimuma,

tehát szintén exponenciális, a paraméterek pedig összeadódnak, ezért X a ráta. Ugyanezzel az érveléssel $\lambda|c(A,\overline{A})|$ paraméterű exponenciális eloszlású annak az eseménynek a bekövetkezési időpontja, hogy az X csúcs közül az egyik megfertődőzik ($X+1$ eset). Itt $c(A,\overline{A})$ a fertőzött és nem fertőzött csúcsok között menő összes él száma.

Mivel minden csúcs foka legfeljebb d, ezért $|c(A,\overline{A})| < Xd$. Annak a valószínűsége, hogy X csökkenése következik be hamarabb,

$$\frac{\lambda|c(A,\overline{A})|}{\lambda|c(A,\overline{A})|+X} = \frac{\lambda d}{1+\lambda d} < \lambda d.$$

Jelölje továbbá $p_{növ}$ azt a valószínűséget, hogy X növekedése következik be hamarabb. (Annak a valószínűsége, hogy egyszerre következik be a két esemény, 0.)

Ahhoz, hogy X legalább k legyen valamelyik időpillanatban, feltéve, hogy *egyetlen egy* fertőzött csúcsból indulunk ki, legalább k növekedés szükséges az első $2k$ lépésben (ellenkező esetben az első $2k$ lépésben kihalt a folyamat, úgy, hogy X értéke sosem volt $k+1$). Jelöljük ezt az eseményt E-vel.

$$P(E) \leq \sum_{l=0}^{k} \binom{2k}{l+k} (\lambda d)^{k+l} p_{\text{növ}}^{k-l} \leq (\lambda d)^k \cdot \sum_{l=0}^{k} \binom{2k}{l+k} \leq (\lambda d)^k \cdot \frac{2^{2k}}{2} < (4\lambda d)^k.$$

□

Skálafüggetlen gráfok esetében a következő állítás alapozta meg a további okfejtéseket [9], mely arról szól, hogyan viselkedik a kontakt folyamat egy csillagban.

2.1.4. Állítás. *([9])*

Legyen G egy csillag, x a csillag középpontja, levelei pedig $y_1, \dots y_k$. Legyen t időpillanatban a fertőzött csúcsok halmaza A_t. Tegyük fel, hogy $A_0 = x$. Ekkor $\exists$ C konstans, hogy

$$\mathbb{P}(A_{exp(Ck\lambda^2)} \neq \emptyset) = 1 - o(k).$$

Tehát egy csillagban nagy valószínűséggel a csúcsok számában exponenciális ideig túlél a fertőzést.

2.2. Kontakt folyamat a Barabási–Albert-modellben

A Barabási–Albert-modellben kisorsolt gráfokra igazolt az a meglepő tény, miszerint még a legkisebb fertőzési rátával rendelkező kontakt folyamat is pozitív valószínűséggel járvánnyá alakulhat [9]. Fontos leszögezni, hogy ez a fejezet a hurokél nélküli Barabási–Albert-modellben (1.3.5) sorsolt gráfon értelmezett kontakt folyamatról szól, de a gráf a modell működési szabálya szerint sorsolt gráfot rögzíti a járványterjedés előtt, és az a kontakt folyamat során már *nem változik*. Feltesszük továbbá, hogy $t = 0$ időpillanatban egyetlen egy csúcs fertőzött, jelölje r, ez a gyökér.

2.2.1. Tétel. *([9] p. 5. Theorem 2.1.) Minden $\lambda > 0$ paraméterre létezik egy olyan N küszöbszám, hogy ha egy $n > N$ méretű Barabási–Albert-gráfot veszünk és egyenletesen kiválasztunk egy v csúcsot ebből a gráfból, akkor $1 - O(\lambda^2)$ valószínűségge v-re igaz, hogy egy v gyökerű kontakt folyamat járvánnyá alakulásának valószínűsége alulról becsülhető a*

$$\lambda^{C_1 \frac{\log\left(\frac{1}{\lambda}\right)}{\log\log\left(\frac{1}{\lambda}\right)}}$$

értékkel és felülről becsülhető a

$$\lambda^{C_2 \frac{\log\left(\frac{1}{\lambda}\right)}{\log\log\left(\frac{1}{\lambda}\right)}}$$

értékkel, ahol a konstansok nem függnek λ és n értékétől.

2.2.2. Tétel. *([9]p. 5. Theorem 2.2.) Minden $\lambda > 0$ paraméterre létezik egy olyan N küszöbszám, hogy ha egy $n > N$ méretű, tipikus skálafüggetlen gráfot veszünk, egyenletesen kiválasztunk egy v csúcsot ebből a gráfból, akkor egy v gyökerű kontakt folyamat járvánnyá alakulásának valószínűsége alulról becsülhető a λ^{C_3}-mal és felülről becsülhető λ^{C_4}-gyel.*

Összefoglalva a két tétel állítását a következőket mondhatjuk: A csúcsok nagy részére, vagyis $O(\lambda^2 n)$ csúcs kivételével igaz, hogy $\mathbb{P}(J_v) \sim \lambda^{\Theta\left(\frac{\log(\lambda^{-1})}{\log\log(\lambda^{-1})}\right)}$, ahol J_v azt az eseményt jelöli, hogy v csúcsból induló kontakt folyamat szerint terjedő fertőzés járvánnyá alakul. A figyelmen kívül hagyott $O(\lambda^2 n)$ számú csúcs hatása az átlagos túlélés valószínűségére jelentős, $\mathbb{P}(J_v)_{\text{átlag}} \sim \lambda^{\Theta(1)}$, ahol $\mathbb{P}(J_v)_{\text{átlag}}$ éppen az a valószínűség, hogy egyenletesen választva egy v csúcsot a betegségből járvány fejlődik.

A két tétel közti jelentős különbség mögött az áll, hogy ha egy λ^{-2}-nél nagyobb fokszámú csúcsból indul a fertőzés, akkor nagy valószínűséggel járvánnyá alakul, viszont ha λ^{-1}-nél jelentősen kisebb fokszámmal rendelkező csúcsból, akkor nagyon gyorsan kihal a fertőzés. ([9], p.3.)

2.2.3. Következmény. *A Barabási–Albert-modellben sorsolt gráfokra a kontakt folyamatra vonatkozó járványküszöb, $\lambda_c = 0$. A 2.2.1. és a 2.2.2. tételek következményeképp a járvány kialakulásának valószínűségére alsó és felső becslés adható λ függényében (λ a szokásos módon a kontakt folyamatra jellemző paraméter). Az alsó becslés mutatja, hogy a járvány kialakulásának valószínűsége* minden $\lambda > 0$ *esetén szigorúan pozitív.*

Az eredmények azért kötődnek a Barabási–Albert-modellhez, mert a Barabási–Albert-gráfra alkalmazható Pólya-urna reprezentáció segítségével nagyságrendileg megállapítható egy csúcs környezetében a maximális fokszám, melyet a 2.2.1. és a 2.2.2. tételek bizonyításához elengedhetetlenek. A bizonyításról, valamint a Pólya-urna reprezentációról bővebben ld.[9].

A Pólya-urna reprezentáció lényege, hogy minden urnában az eddig ott levő golyók száma N_i, u pedig egy előre rögzített paraméter, akkor az új golyó i. urnába való elhelyezésének valószínűsége arányos $N_i + u$-val. Ezzel ismét ekvivalens az, ha minden urnához rendelünk egy p_i paramétert, akkor az új golyó, az előző lépésektől függetlenül p_i valószínűséggel kerül az i. uránba. Pólya azt is meghatározta, hogy u és $\{N_i\}$ függvényében a határeloszlás milyen paraméterű β eloszlás. A Pólya-urna reprezentáció kapcsolata a Barabási–Albert-modellel a következő: minden új él behúzása az utóbbiban megfelel egy új golyó i. urnába való behelyezésével. A Pólya-urna reprezentáció segíségével egy formális definíció is adható, mely ekvivalens az 1.3.5. modell definíciójával és alapjául szolgál a járványterjedésről szóló fő eredmények bizonyításához [9].

2.3. Kontakt folyamat és skálafüggetlen gráf párhuzamos fejlődése

Egy újabb és még fejlődésben lévő területe a járványterjedés matematikai tanulmányozásának azon modellek megalkotása és vizsgálata, melyek a járványterjedési folyamatot és a gráf fejlődésének folyamatát párhuzamosan zajló folyamatként kezelik. Ebben a fejezetben egy ilyen modellt, és egy járványterjedéssel kapcsolatos eredményt ismertetek, [18] alapján.

A gráf fejlődése a most következő modellben gráf éleinek behúzására, illetve törlésére korlátozódik, tehát nem a gráf méretének növekedésével párhuzamosan vizsgálandó a kontakt folyamat, hanem a kapcsolatok alakulásával. Ezt azért fontos hangsúlyozni, mert az 1. fejezetben ismertetett modellekben a kapcsolatok a gráf növekedésével párhuzamosan fejlődtek, a gráf méretének növekedésével párhuzamosan jöttek létre új élek. Most azonban rögzítjük a gráf méretét, és csak az élek változása jelenti a gráf fejlődését.

A modellalkotás során [18] három feltételezéssel él, melyek alapjaiban meghatározzák a modellt.

- Az gráf élhálózatának fejlődése független a kontakt folyamattól
- Az élek fejlődése és a kontakt folyamat *azonos időskálán* mozog
- A csúcsoknak van egy hierarchiája, ún. erőssége, mely az idő során nem változik. Ezt az erősséget határozza meg a csúcsok címkézése; az alacsonyabb címke erősebb csúcsot jelöl, belőle nagyobb valószínűséggel húzunk élt. Tehát a címke bizonyos értelemben hasonlít az 1.4.2. részben szereplő fitness értékhez, hiszen *időben nem változik* és befolyásolja a későbbi élek csatlakozását az adott csúcshoz. (De a fitness érték minél nagyobb, annál nagyobb valószínűséggel csatlakozik él a csúcshoz.)

2.3.1. Modell. *Legyen $(G_t)_{t\geq 0}$ egy olyan gráfsorozat amelyre $G_t = (V, E_t)$, V a csúcsok halmaza, E_t pedig az élek halmaza (t egy folytonos idő paraméter). A csúcsok halmaza tehát fix, $\forall$ t-re azonos, az idő előrehaladtával se nem nő, se nem csökken a csúcsok száma. A csúcsokat cimkékkel látjuk el, $V = \{1, \dots N\}$. A címkézés szintén változatlan a gráf fejlődése során, nem függ t-től. Az élek halmaza, E_t, viszont az idő elteltével a véletlentől függő módon változik. Rögzítsük továbbá $\beta > 0$ és $1 > \gamma > 0$ paramétereket.*

A $t = 0$ esetben $G_0 = (V, E_0)$ véletlen gráf, amelyre az élek eloszlása a következőképpen alakul; $\forall$ $\{x, y\} \subseteq V$ rendezetlen csúcspár között a többi a többi csúcspártól függetlenül $p_{x,y}$ valószínűséggel fut él, ahol

$$p_{x,y} = \min\left\{\frac{\beta N^{2\gamma-1}}{x^\gamma y^\gamma}, 1\right\}.$$

A $t > 0$ esetben a G_t gráf éleinek fejlődése a következő szabály szerint alakul: minden csúcs egymástól függetlenül κ rátával frissíti a szomszédait. Az x cimkével ellátott csúcsból a frissítés esetén $\forall$ $y \in V \setminus \{x\}$ csúcshoz egymástól $p_{x,y}$ valószínűség valószínűséggel húzunk élt; függetlenül egymástól és az addig jelenlevő élektől.

A gráfon az előbb leírt fejlődési folyamattól függetlenül *egy fertőzés a kontakt folyamat a 2.1.1. definíciója szerint zajlik, jelölje ezt η_t.*

Korábbi munkák alapján G_0 fokszámeloszlása hatványrendben cseng le, $\tau = 1 + \frac{1}{\gamma}$ kitevővel, sőt a hatványtörvény $\forall$ G_t gráfra teljesül ugyanazzal a τ-val [18], [19].

Ugyanúgy, mint a nem fejlődő gráfon végbemenő kontakt folyamat esetén, ebben a modellben is elmondható, hogy ha $N < \infty$, akkor a kontakt folyamat 1 valószínűséggel kihal; vagyis 1 valószínűséggel $\forall$ v csúcsra $\eta_t(v) = 0$, ha t elég nagy. Ezért itt is azt érdemes vizsgálni, hogy a kihalás mikor következik be, jelölje ezt az időpontot a továbbiakban T_{ext}.

A fő eredmény, mely a kontakt folyamatról ismeretes a 2.3.1. modellben, azt mondja ki, hogy ha $\gamma > \frac{1}{3}$, akkor bármilyen λ paraméterű kontakt folyamat legalább exponenciális ideig fennmarad a gráfban. Ha azonban $\gamma < \frac{1}{3}$, akkor megfelelően kicsi λ paraméterű kontakt folyamatra a kihalás legfeljebb $N^{\frac{1}{2}}$ nagyságrendű időn belül bekövetkezik. Mivel a $\gamma < \frac{1}{3}$ feltétel ekvivalens azzal, hogy $\tau > 4$, ezért a modell szerint létrejövő skálafüggetlen gráf hatványkitevőjével is megfogalmazható a tétel. [18].

2.3.1. Tétel. *A 2.3.1. modellben tegyük fel, hogy $t = 0$ időpillanatban minden csúcs fertőzött; $\eta_t(x) = 1$ $\forall 1 \leq x \leq N$-re.*

(a) $\gamma > \frac{1}{3} (\Leftrightarrow \tau < 4)$ és tetszőleges β, λ pozitív paraméterek választása esetén $\exists$ C konstans, hogy $\forall N > 0$-ra teljesül:

$$\mathbb{P}(T_{ext} \leq e^{cN}) \leq e^{-cN} \tag{2.3}$$

(b) $\gamma < \frac{1}{3} (\Leftrightarrow \tau > 4)$ *esetén* $\exists\ \lambda_c > 0$, *hogy* $\forall\ \lambda < \lambda_c$ *paraméterű kontakt folyamatra* $\exists C$ *konstans, hogy*

$$\mathbb{E}(T_{ext}) \leq C\sqrt{N}.$$

Ahelyett, hogy konkrét kimenetelekről állapítjuk meg, hogy egy betegség járvánnyá alakult-e vagy sem, érdemesebb ennek az eseménynek a valószínűségét vizsgálni, mint azt már az előző alfejezetben is tettük. Bevezetek ezért egy általános definíciót, mely sztochasztikus értelemben definiálja, mit jelent az, hogy egy eseményrendszer bekövetkezése exponenciális ideig teljesül.

2.3.2. Definíció. *Egy* $(A_t)_{t\geq 0}$ *eseményrendszer bekövetkezése exponenciális idejű, ha* $\exists\ c > 0$ *univerzális konstans, hogy* $\forall\ N$*-re*

$$\mathbb{P}\Big(\bigcap_{s\leq e^{cN}} A_s\Big) \geq 1 - e^{-cN}. \tag{2.4}$$

A (2.4). egyenlettel ekvivalens:

$$1 - \mathbb{P}\Big(\bigcap_{s\leq e^{cN}} A_s\Big) = \mathbb{P}\Big(\overline{\bigcap_{s\leq e^{cN}} A_s}\Big) = \mathbb{P}\Big(\bigcup_{s\leq e^{cN}} \overline{A_s}\Big) \leq e^{-cN} \tag{2.5}$$

Legyen most $A_t = \{\exists v \in V | \eta_t(v) = 1\}$, vagyis az az esemény, hogy a t időpontban még nem halt ki a járvány. Ekkor $\overline{A_t}$ azt az eseményt jelöli, hogy a t. időpontban már kihalt a járvány. Vagyis az (2.5) egyenlet bal oldalán pontosan annak a valószínűsége áll, hogy $T = e^{cN}$ idő előtt kihal a járvány.

2.3.3. Következmény. *2.3.1. Tétel (a) rész* $\Leftrightarrow \gamma > \frac{1}{3} (\Leftrightarrow \tau < 4)$ *és tetszőleges* β, λ *pozitív paraméterek választása esetén a kontakt folyamat a 2.3.1. modellben exponenciális ideig fennmarad (nem hal ki).*

2.4. Kitekintés más járványterjedési modellre

A kontakt folyamat, mint járványterjedési modell meglehetősen bonyolult és a modellezés szempontjából nem kényelmes, mert a járványterjedést folytonos paraméterű folyamatként kezeli. Ezért vagy valamilyen értelemben a diszkretizáltját vesszük ennek a folyamatnak (ld. a 3. fejezetben), vagy egyszerűbb, diszkrét időben fejlődő járványterjedési modelleket vizsgálunk. Diszkrét paraméterű járványterjedési modellre példa a Reed-Frost-modell, illetve annak egy módosított változata nem homogén populációra. Ez a modell azért kapott itt helyet, mert a klaszteresedési együttható és a járványterjedés kapcsolatot első ízben vizsgálták ebben a modellben, mely a 3. fejezetben található szimuláció egyik központi kérdése.

2.4.1. Modell (Reed-Frost inhomogén populációban). *[20] Jelölje $(\xi_t)_{t\geq 0}$ a Reed-Frost-féle* diszkrét idejű *járványterjedési folyamatot, ahol t nemnegatív egész. Adott egy $G = (V, E)$ gráf, V a csúcsok, E az élek halmaza. Hasonlóan a kontakt folyamathoz, $\xi_t(v) = 1$, ha $v \in G$ fertőzött, $\xi_t(v) = 0$, ha v egészséges és legyen $\xi_t(v) = -1$, ha immunis, vagyis többet nem fertőződhet meg. Ha egy csúcs immunissá válik, akkor tulajdonképpen* kilép *a járványterjedési folyamatból.*

A $t + 1$. lépésben minden v beteg csúcs, egymástól függetlenül, p valószínűséggel adja tovább a fertőzést a még nem fertőzött szomszédainak, majd kilép *a járványterjedési folyamatból (immunissá válik.) Vagyis $\forall \{v, u\} \in E$ csúcspárra egymástól függetlenül:*

$$\xi_t(v) = 1, \xi_t(u) = 0 \Rightarrow \mathbb{P}(\xi_{t+1}(u) = 1) = p,$$

valamint $\forall$ v-re, amelyre $\xi_t(v) = 1$, arra $\xi_{t+1}(v) = \xi_{t+2}(v) = \ldots = -1$, vagyis a v csúcs immunissá válik a fertőzésre.

Ez a modell az előzőektől eltérően egy SIR-modell, mert a csúcsoknak három állapota is lehetséges, fertőzött, egészséges és immunis. A folyamatból való kilépés miatt legfeljebb $|G|$ lépésben véget ér a járványterjedési folyamat; vagy minden csúcs meggyógyul, vagy csak fertőzött csúcsok maradnak a folyamatban. Tehát azt mondjuk, a járványterjedési folyamat *véget ér*, ha vagy csak egészséges és immunis, vagy csak fertőzött és immunis csúcsok maradnak. Fontos, hogy míg az előbbi esetben a teljes gráf *meggyógyul*, addig az utóbbiban a gráfnak *egy pozitív hányada* vált fertőzötté, hiszen bizonyos csúcsok immunissá váltak a fertőzésre. Ezért a járványterjedési küszöb 2.1.2. definíciója itt nem értelmes. Ezért ebben a járványterjedési küszöb helyett azt az R_0 értéket szokás vizsgálni, mely a várható értéke annak, hogy egy egyenletesen választott csúcs ha megfertőződik, hány csúcsnak adja át a fertőzést. *Nagy járványkitörésnek* pedig az felel meg, ha $|G| \to \infty$ esetén egy többé-kevésbé determinisztikus pozitív hányada kertőződik meg a gráf csúcsainak, melyre rárakódhat egy kis rendű Gauss-zaj. Tétel mondja ki, hogy nagy járványkitörésnek akkor és csak akkor pozitív a valószínűsége, ha $R_0 > 1$ [21].

Vegyük most a Reed–Frost-modellt. Tegyük fel, hogy *kezdetben* 1 *véletlenszerűen választott csúcs fertőzött*. Rögzítsünk továbbá egy n csúcsú gráfot. Ekkor az első lépésben a megfertőzött csúcsok számának várható értéke kifejezhető a következőképpen:

$$R_0 = \frac{1}{n}\sum_{i=1}^{n} d_i \cdot p = p \cdot \frac{1}{n}\sum_{i=1}^{n} d_i. \tag{2.6}$$

Ha egy adott véletlen gráfmodellben dolgozunk és $\frac{1}{n}\sum_{i=1}^{n} d_i$ csak a modell paramétereinek függvénye, akkor értelmes kérdés lehet, hogy rögzített paraméterek esetén melyik az a legkisebb p valószínűség, melyre $R_0 > 1$. (Vagy ha nincs legkisebb ilyen p érték, akkor ezek infimumát vizsgálhatjuk, vagyis azt, hogy melyik az a legkisebb p érték, melyre már nem teljesül.) A továbbiakban ezt a küszöböt nevezem járványkitörési küszöbnek.

A járványterjedés és a klaszteresedési együttható kapcsolatát [20] szerzői egy olyan véletlen gráfmodellben vizsgálták, melyben egy G_n gráf n darab csúcsának mindegyikére egyformán teljesül, hogy mind az m csoportnak egymástól függetlenül r valószínűséggel tagja. Itt $m = \lfloor \beta n^\alpha \rfloor$ és ahhoz, hogy a klaszteresedési együttható kalibrálható legyen a modellben, $\alpha = 1$-et feltételeznek.

Jelölje $c = \lim_{n\to\infty} c_n$, ahol c_n a fenti véletlengráf modellben egy n méretű gráfban azt a feltételes valószínűséget, hogy két csúcs között él fut, feltéve, hogy van legalább egy közös szomszéduk. Ez tulajdonképpen egyfajta elméleti klaszteresedési együttható, de az 1.2.1. definíciótól abban különbözik, hogy nem a konkrét, kisorsolt n méretű gráfok klaszteresedési együtthatója, hanem a modell alapján könnyen számolható elméleti valószínűség, a kisorsolt gráf figyelembevétele nélkül. A konkrét modellben $\alpha = 1$ esetén $c = \frac{1}{1+\beta\gamma}$.

A [20] által ismertetett modellnek tehát 2 szabad paramétere van, β és γ, Ezekkel R_0 és c kifejezhető. A szerzők egy csúcs fokszámának aszimptotikus várható értékét is rögzítik, $\beta\gamma^2$. Mindezen feltételezések mellett a következő állítást fogalmazzák meg: az aszimptotikus klaszteresedési együttható (c) növelésével R_0 értéke is nő, több különböző p paraméterű Reed–Frost-féle járványterjedés esetén ([20], p.13. felső ábra). Ebből viszont (2.6). alapján az következik, hogy egy rögzített n csúcsú gráfban p küszöbértéke, a járványkitörési küszöb csökken.

A jelenség heurisztikus igazolásaképp a következő magyarázat adható: ([20] p. 14.) a klaszteresedési együttható növekedése *ebben a konkrét modellben* azt vonja maga után, hogy a csúcsok kevesebb, de nagyobb csoportok tagjai. Ebből viszont az következik, hogy kisebb valószínűséggel kerülik el a fertőzést, vagyis a járvány könnyebben elterjed [20]. Tehát egy nagyobb paramétertartományra kell teljesülnie, hogy pozitív valószínűséggel járvány tör ki. Következésképpen a járványkitörési küszöb csökken.

Ez utóbbi eredmények alátámasztják, hogy érdemes a klaszteresedési együttható függvényében vizsgálni a járványterjedési folyamatot, mely a következő fejezetben ismertetett szimuláció egyik célja. A 2.6 egyenletben szereplő $\frac{1}{n}\sum_{i=1}^{n} d_i$ mennyiség pont az élsűrűség, melynek rögzítése a szimuláció során is felmerül majd.

3. fejezet

Szimuláció – járványterjedés a duplikációs modellben

A duplikációs modellben sorsolt gráfokon való járványterjedésről nem születtek még elméleti eredmények, ezért itt volt helye a szimuláció elkészítésének, amelynek segítségével azt reméltük, képet kaphatunk arról, miként befolyásolják a duplikációs modell paraméterei, vagy a kisorsolt gráf paraméterei a járványterjedés folyamatát.

A fő cél a gráfban található csoportosulásoktól való függés vizsgálata volt, ezért esett a választás a duplikációs modell szimulációjának elkészítésére, melynek sorsolási szabálya elősegíti a klikkek kialakulását a véletlen gráfban.

3.1. Az általános duplikációs modell szimulációja

Az 1.5.2. modell szimulációja során elsőként ki kellett választani, mi legyen a kiindulási gráf, G_0. Az általam megvalósított szimulációban egy háromszögből indulok ki, tehát az első három csúcs automatikusan össze van kötve a másik kettővel. Tehát $t_0 = 3$, $G_{t_0} = (V_3, \{v_1, v_2\}, E_3)$, ahol $V_3 = \{v_1, v_2, v_3\}, E_3 = \{\{v_1, v_2\}, \{v_1, v_2\}\}$ és $t \geq 4$-re alkalmazom a duplikációs modell fejlődési szabályát.

A másik felmerülő kérdés, hogy az a_1 illetve a_2 véletlenszerűen egyenletes eloszlás szerint választott csúcsot vajon visszatevéssel, vagy visszatevés nélkül választjuk-e. Másképpen fogalmazva, megengedünk-e a gráfban többszörös éleket? Ezt a kérdést tulajdonképpen már az 1.5.2. definíció kimondásával tisztáztam, inkább csak még egyszer hangsúlyozom. Az általam megvalósított szimulációban, mely az első módosítást alkalmazza csak, ezeket a csúcsokat visszatevés nélkül választom, tehát nem engedek meg többszörös éleket. Sőt, az általános duplikációs modell általam ismertetett definíciója alapján hurokélek sem fordulhatnak elő a gráfban, mivel a t. csúcs szomszédai kizárólag az előző, $t - 1$ csúcs közül kerülnek ki.

További problémaként felmerült, hogy $t \leq a_1$ illetve $t \leq a_2$ esetén, visszatevés nélkül nem tudok egyenletes eloszlás szerint elegendő számú csúcsot választani; ekkor az általam készített szimulációban az 1. módosítás beépítésével az esetleg létrejövő izolált csúcs az összes korábbival összeköttetésbe kerül.

A duplikációs modell és az ennek segítségével kisorsolt gráfokon való járványterjedés szimulációját R-ben készítettem.

A program három fő részre tagolódik:

- A véletlen gráf sorsolása a duplikációs modell szerint
- A gráfra jellemző élsűrűség és klaszteresedési együttható kiszámolása
- A járványterjedés szimulációja

(A forráskód megtalálható az A függelékben.)

3.1.1. A véletlen gráf sorsolása – a *duplication függvény*

A gráfot szomszédsági lista segítségével tárolom. Egy G gráfot egy olyan lista jellemez, melynek az i. eleme szintén egy lista, mégpedig az i. csúcs szomszédainak listája. Általában a keletkezett gráfok szomszédsági mátrixa egy jellemzően elég ritka mátrix, ezért volt optimálisabb listákat használni.

Az első három lépést kivéve (legyártom a háromszögnek megfelelő éllistát), a 4. lépéstől kezdve a program minden lépésben meghívja a *duplication* függvényt. Ennek a függvénynek 5 paramétere van: X a gráf szomszédsági listája, q, r és a_1 a duplikációs modell 1.5.2 definícióban szereplő paraméterei, valamint t, a gráf aktuális mérete.

A *duplication* függvény ismételt meghívásával a véletlen gráf fejlődése modellezhető. A forráskódban a *lepes* változó jelöli a gráf maximális méretét, vagyis azt, hogy hány csúcsú gráfot sorsolunk ki a duplikációs modell segítségével.

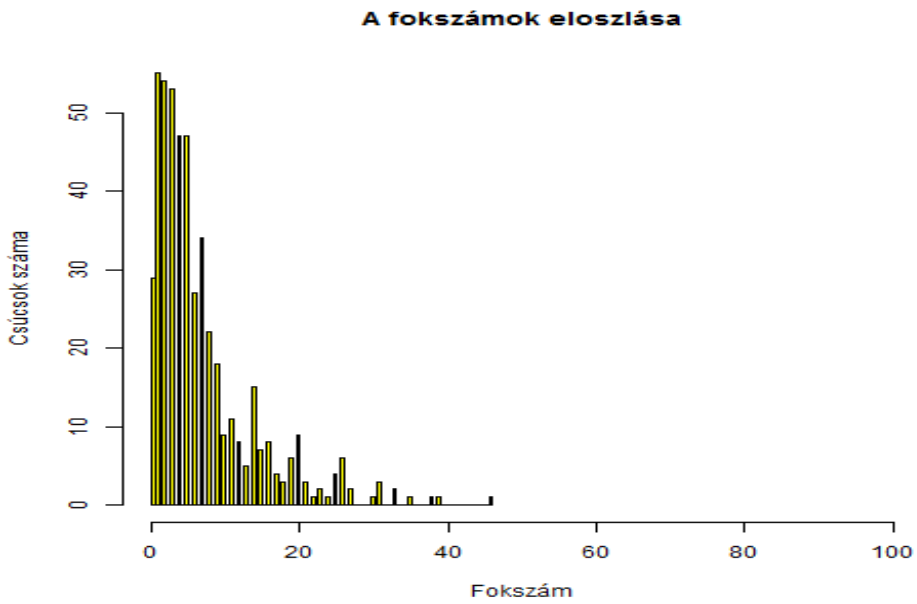
Összefoglalva tehát a végleges, kisorsolt gráf 4 paramétertől függ, melyeket az alábbi módon választhatunk meg:

- a q és r paraméterek, melyek valószínűségeket jelölnek, tehát $q, r \in [0, 1]$ valós értékek
- $N \in \mathbb{N}$ és $N \geq 4$, a gráf végső mérete
- $a_1 \in \mathbb{N}$, érdemes N-hez képest kicsinek választani

$G\ (q, r, a_1, N)$ az általános duplikációs modell 1.5.2. definíciója szerint sorsolt véletlen gráf.

3.1.2. A módosítás szerepe

Annak érdekében, hogy ne keletkezzen túl nagy számban izolált csúcs, szükség volt az 1.5.2 definícióban szereplő 1. módosítás beépítésére a szimulációban.



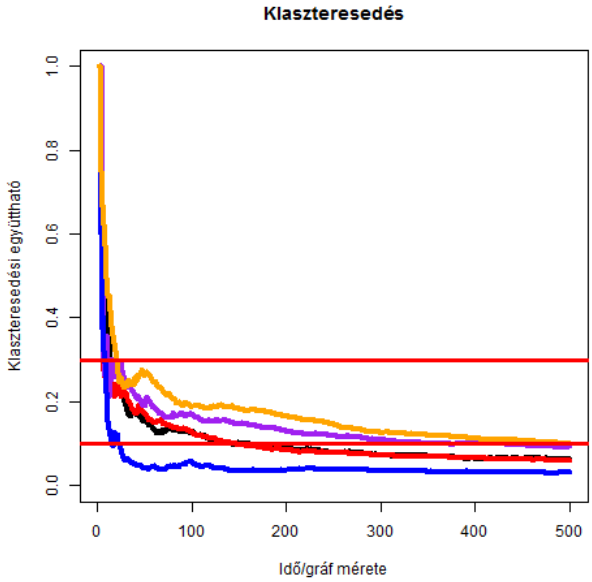
3.1. ábra. Egy módosítás beépítése nélkül sorsolt gráf fokszámeloszlása

A 3.1 oszlopdiagram egy, a program segítségével sorsolt 500 csúcsú gráf fokszámeloszlását mutatja. A vízszintes tengelyen látahatóak az előforduló fokszámok, a függőleges tengelyen pedig az, hogy adott fokszámú csúcsból hány darab található a kisorsolt gráfban.

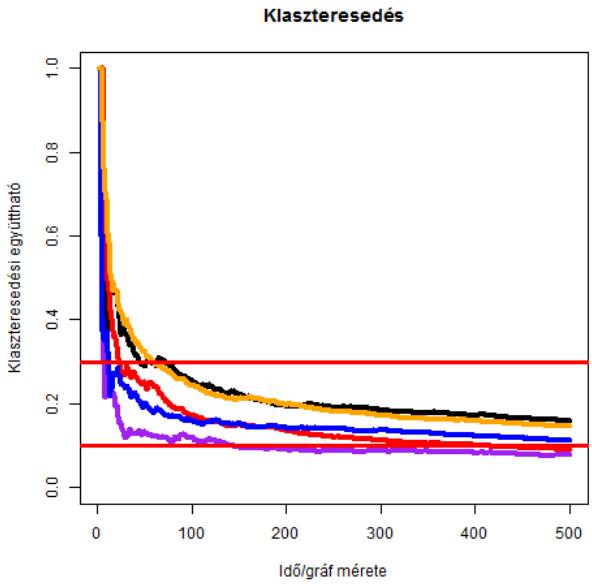
A szimuláció elkészítése során az első kitűzött cél az volt, hogy a program segítségével olyan gráfot tudjunk sorsolni, amelyben létrejönnek csoportosulások a gráfban, hiszen majd később ezek hatását szeretnénk megfigyelni a járványterjedésre vonatkozóan. A csoportosulások megfigyelésére szolgáló mennyiség az ún. klaszteresedési együttható, melynek az 1.2.1. szerinti definíciójával fogok dolgozni. Kiemelendő, hogy a klaszteresedési együttható $c = \frac{3h}{w}$, ahol h a gráfban található háromszögek száma w pedig a cseresznyék száma és ez a c szám annak a valószínűsége, hogy két véletlenszerűen kiválasztott illeszkedő él egy háromszöget alkot.

Valódi hálózatok (pl. biológiai és számítógépes hálózatok, repülőtérhálózatok) klaszteresedési együtthatójáról [8] tartalmaz adatokat. Figyelembe véve ezeket a tapasztalati értékeket a kitűzött cél az volt, hogy a sorsolt gráf klaszteresedési együtthatója 0.1 és 0.3 közé essen és így a valódi hálózatok esetében tapasztalt értékekkel azonos nagyságrendű legyen.

Míg a duplikációs modell módosítás nélküli változatában a nagyon kicsi q és r paraméter sem vezetett célra, a klaszteresedési együttható már egy 500 csúcsú gráfban is jóval 0.1 alá csökkent, és ott is maradt, ez látható a 3.2a ábrán.



(a) Módosítás beépítése nélkül



(b) Módosítás beépítése, $a_1 = 10, q = 0.2, r = 0.8$

3.2. ábra. Klaszteresedési együttható az idő függvényében

Ahhoz, hogy a klaszteresedési együttható 1000 csúcsú gráf esetén és 0.1 és 0.3 közé essen, szükség volt *az* 1.5.2 definícióban szereplő módosítások valamelyikének beépítésé-

re; én az 1. számú módosítást választottam. A módosítás beépítésével már elérhető, hogy a klaszteresedési együttható a kívánt értékek között maradjon, kis q, de nagy r paraméter választása esetén. Az általam kipróbált esetek közül $q = 0.2$ és $r = 0.8$ paraméterek választása esetén a klaszteresedési együttható az 500 csúcsú gráfok vizsgálata során 0.1 felett maradt, 5-ből 4 gráf esetén, és az egy kivételes esetben is csak kicsivel ment a 0.1 szint alá, ezt szemlélteti a 3.2b. ábra.

A fent ábrázolt esetekben mind egy 500 csúcsú gráfot kaptunk a sorsolással. Felmerülhet a kérdés, hogy mivel láthatjuk, hogy a klaszteresedési együttható a gráf méretének növekedésével egyre lassabban, de folyamatosan csökken, vajon megfelelő méretű gráfokat sorsoltunk-e a valódi hálózatok modellezéséhez. [8] alapján az ott szereplő hálózatok élszáma 10^3 és 10^8 nagyságrendek közé esik, ezért célszerű ellenőrizni, hogy a kisorsolt 500 csúcsú gráfok élszáma megfelelő nagyságrendű-e.

A program gyorsítása után ezt ellenőriztem és 1000 darab különböző, egymástól függetlenül véletlenszerűen sorsolt 500 csúcs gráf esetén azt tapasztaltam, hogy az élek száma 10^4 nagyságrendű volt (átlagos élszám a gráfokra: 6033.92), valamint ezekre a végleges klaszteresedési együttható is többségében 0.1 és 0.3 közé esett (klaszteresedési együtthatók átlaga: 0.1393997). A 3.2a. és a 3.2b ábrák alapján az is látszik, hogy a klaszteresedési együttható az idő előrehaladtával egyre kevésbé csökken.

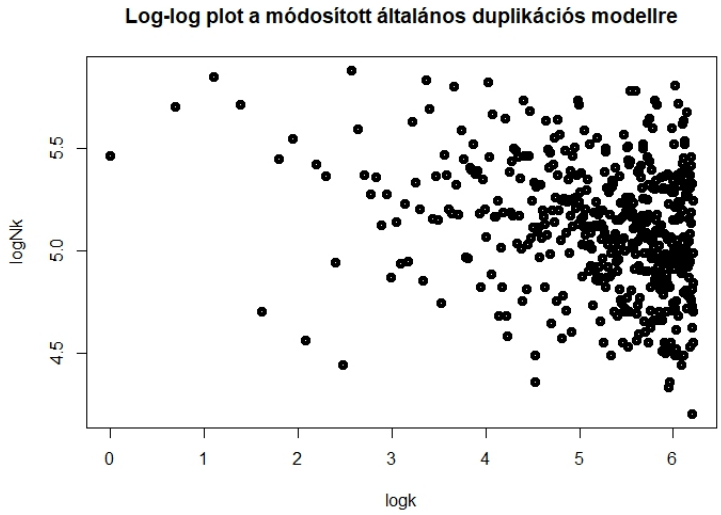
Tehát mondhatjuk azt, hogy a valós hálózatokra [8] alapján jellemző klaszteresedési együtthatót tudunk elérni a sorsolt gráfokban a módosított duplikációs modell szimulációjának segítségével.

A klaszteresedési együttható kiszámításának gyorsításában kulcsszerepet játszott a *mapply* és az *intersect* függvények használata. A *clustering* függvény a háromszögek és a cseresznyék összeszámolásáért felel. A háromszögszámolásnál végigmegy minden csúcsnak minden szomszédján, így ezekre a csúcspárokra külön külön a szomszédlisták metszetével növeli a háromszögek számát tartalmazó változót (itt használjuk az *intersect függvényt*). Így a háromszögek számának háromszorosát kapjuk, pont ez szerepel a klaszteresedési együtthatóban. A cseresznyék száma a tapasztalati fokszámeloszlás-vektorból (minden csúcs szomszédainak számát tartalmazza), melyet a *fokszameloszlas* függvény által megkapunk, már egyszerűbb, e kettőből pedig megkapható a klaszteresedési együttható.

A *mapply* függvény segítségével több, esetemben 1000 gráf klaszteresedési együtthatója könnyen és gyorsan kiszámítható. Az 1000 gráf szomszédsági listáját tárolva egy újabb listában, valamint ezek fokszámeloszlás-vektorait egy másikban, a három függvényt, *clustering*, *elszam*, *elsuruseg* függvényeket ezen listák páronként megfelelő elemeire kell alkalmazni. Ez a *mapply* függvény segítségével igen gyorsan lefut. Az eredményeket egy új, 1000 elemű vektorban kapjuk.

3.1.1. Megjegyzés. *A klaszteresedési együttható időbeni változásának vizsgálatához minden lépésben a program újraszámolta a klaszteresedési együtthatót, ez hosszú futásidőt eredményezett. (Igaz, a dinamikus módon való számítás csökkenthette volna a futásidőt, az* intersecct *függvény alkalmazásával együtt.) A módosítás szerepének vizsgálatakor, valamint a megfelelő paraméterbeállítások kereséséhez tehát, amikor még az időbeni változást is figyelni kellett, csak viszonylag kevés, öt gráfot vizsgáltam egyszerre. A módosítás szükségességére azonban már kevés adatból is fény derült. Ha azonban már csak a gráf végleges klaszteresedési együtthatóját akarom ellenőrizni, azt már* 1000 *gráfra is könnyen elvégzi a program.*

További érdekességként elkészítettem egyetlen, az általánosított duplikációs modell által $q = 0.1$ és $r = 0.9$ paraméterbeállítások mellett sorsolt 500 csúcsú gráfora az 1.1. alfejezetben tárgyalt mennyiségekre a tapasztalati fokszámeloszlás log-log ábráját; $\log k$ függvényében ábrázoltam a $\log N_k$ mennyiséget, ahol N_k a k fokszú csúcsok *száma* a gráfban.



3.3. ábra. Log-log plot a duplikációs modellben

A 3.3. ábra alapján a kisorsolt gráf nem tűnik sem ritka, sem pedig skálafüggetlen gráfnak, hiszen nagy arányban vannak jelen a nagy fokú csúcsok és az ábra nem közelít egy egyeneshez sem.

3.2. A járványterjedés modellezése

A duplikációs modell alapján kisorsolt gráfokon az alábbi járványterjedési modellt szimuláltam.

A járványterjedés körönként zajlik, vagyis diszkrét paraméterű sztochasztikus folyamat, jelölje μ_t, ahol t pozitív egész, μ_t pedig a fertőzött csúcsok halmaza által egyértelműen meghatározott diszkrét paraméterű Markov-folyamat (hasonlóan, mint a 2.1.1. definícióban η_t, a 2.4.1. modellben ξ_t). Ha egy v csűcs fertőzött, akkor $\mu_t(v) = 1$, ha v egészséges, akkor $\mu_t(v) = 0$. Minden körben egy fertőzött csúcs a többitől függetlenül, rögzített p valószínűséggel meggyógyul és a gráf minden élén, mely egy fertőzött csúcsot egy egészséges csúccsal köt össze, szintén egymástól függetlenül, s valószínűséggel terjed tovább a fertőzés. Így tehát az, hogy egy fertőzött csúcs melyik körben fog meggyógyulni, geometriai eloszlású lesz p paraméterrel. Egy egészséges csúcs pedig minden körben $1 - (1 - s)^{N_i}$ valószínűséggel betegszik meg, ahol az N_i szám az egészséges csúcs fertőzött szomszédainak számát jelöli. Ez a paraméter tehát körönként módosul, hiszen újabb csúcsok fertőzödnek meg, az egészségesek pedig meggyógyulhatnak, ezáltal körönként változik a csúcsok fertőzött szomszédainak száma, ellentétben a gyógyulásra jellemző p paraméterrel, amely állandó. Formálisan:

$$\mu_t(v) = 1 \Rightarrow \mathbb{P}(\mu_{t+1}(v) = 0) = p \tag{3.1}$$

$$\mu_t(v) = 1, \mu_t(u) = 0 \Rightarrow \mathbb{P}(\mu_{t+1}(u) = 1) = s \tag{3.2}$$

Ebben a járványterjedési modellben a terjedési folyamat egy diszkrét idejű folyamat. Minden élen egymástól függetlenül egy előre rögzített valószínűséggel terjed a fertőzés, ha az él beteg és egészséges csúcs között fut, tehát bizonyos szempontból hasonlít a Reed–Frost modellre (2.4.1). Fontos különbség azonban, hogy ez a modell egy SIS típusú modell, a csúcsoknak két állapota és a csúcsok nem válnak immunissá a gyógyulás után, újra megfertőződhetnek. Másrészt az is megkülönbözteti a 2.4.1. modelltől, hogy a beteg csúcsok minden körben csak egy bizonyos előre rögzített p valószínűséggel gyógyulnak meg, míg a Reed–Frost modellben biztosan meggyógyulnak és ki is lépnek a folyamatból.

A kontakt folyamattal pedig az köti össze ezt a modellt, hogy a geometriai eloszlás az exponenciális eloszlás diszkretizáltjaként is felfogható, tehát bizonyos értelemben a kontakt folyamat diszkretizáltjáról beszélünk. Bár a kontakt folyamat esetében 0 annak a valószínűsége, hogy bármely két esemény (a gyógyulások és a fertőzések közül) egy időben történne, addig itt igen sok esemény történik egy időpillanatban. Azonban a kontakt folyamatra igaz az, hogy egy kis időintervallum alatt már nagyon sok esemény bekövetkezhet.

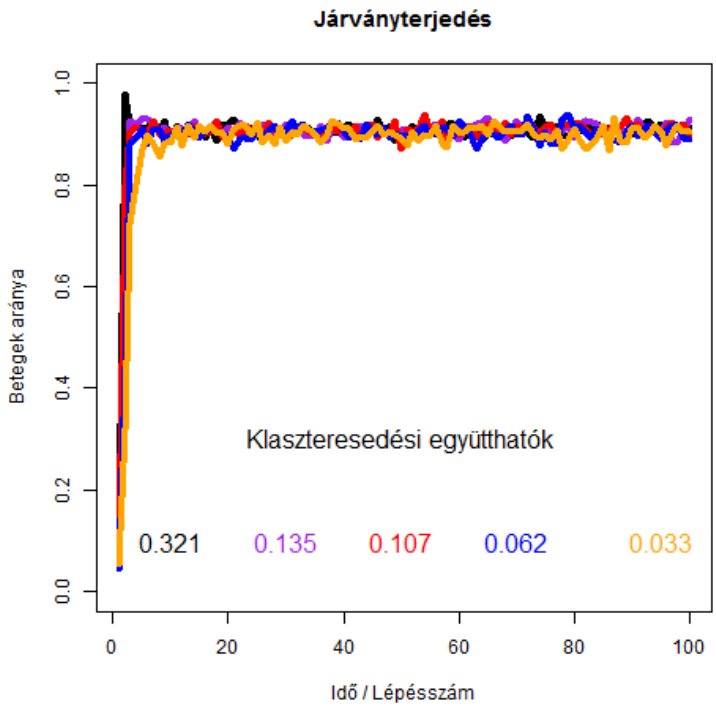
A járványterjedés szimulációját némiképp lassúvá teszi, hogy a fertőzött szomszédok számát minden körben minden csúcsra vonatkozóan újra kell számítania a programnak.

A forráskódban szereplő *jarvanyterjedes_idoben* függvény a megadott p, s paraméterek mellett egy tetszőleges L körből álló járványterjedési folyamatot szimulál egy már meglevő G gráfon, mely szomszédsági listával adott. Ez a függvény a járványterjedés minden körében meghívja a *foksz_fert* függvényt, mert kiszámítja és egy vektorban tárolja, hogy az egyes csúcsoknak hány darab fertőzött szomszédja van. Ehhez egy másik, bináris vektorban körönként azt is számon kell tartani, hogy épp melyik csúcs fertőzött, melyik nem.

3.3. A klaszteresedési együttható és a járványterjedés kapcsolatának vizsgálata

Az első kísérlet során 5 különböző, a duplikációs modell által sorsolt gráfon vizsgáltam a járványterjedési folyamat első 100 körét. Minden egyes körre vonatkozóan eltároltam a beteg csúcsok arányát a gráfban.

Mivel a fő célunk a klaszteresedési együttható szerepének meghatározása, ezért 5 különböző klaszteresedési együtthatójú gráfot választottam ki. A klaszteresedési együtthatók között nagyságrendileg különbözők is voltak, a legnagyobb együttható értéke 0.321 a legkisebb pedig 0.033 volt.



3.4. ábra. Járványterjedés $p = 0,5$ $s = 0.9$ paraméterekkel

A 3.4. ábrán látható ennek az első kísérletnek az eredménye. A vízszintes tengelyen ábrázoltam a járványterjedés köreit (vagy másképpen a járványterjedés időbeni előrehaladását), a függőleges tengelyen pedig az aktuális körben/időpillanatban a betegek arányát a gráfban. A különböző színű görbék különböző gráfokon futtatott járványterjedéseknek felelnek meg. [1]

A 3.4. ábra alapján úgy tűnik, hogy a betegek aránya egy idő után nem függ a klaszteresedési együtthatótól, de a járvány gyorsabban elterjed a nagyobb klaszteresedési együtthatójú gráfokban (a 3.4. ábrán a fekete görbe éri el a legmagasabb szintet a legelső körökben), míg a kisebb együtthatójú gráfban (sárga görbe) mintha lassabban nőne a betegek aránya a járványterjedés során. Az is látható, hogy ezen az öt gráfon a betegek aránya függetlenül a klaszteresedési együtthatótól ugyanarra a szintre áll be.

3.3.1. Az élsűrűség és a klaszteresedési együttható különválasztása

Fő célom arra fényt deríteni, hogy a klaszteresedési együtthatótól hogyan függ a járványterjedés folyamata.

Mivel értelemszerűen nagyobb élsűrűségű gráfokhoz nagyobb klaszteresedési együttható fog tartozni a gráfokban, ezért ahhoz, hogy pusztán a klaszteresedési együttható, és ne az élsűrűség hatását figyeljük meg a járványterjedésre vonatkozóan, először szükség lenne *azonos élsűrűségű*, de *különböző klaszteresedési együtthatójú* gráfokra.

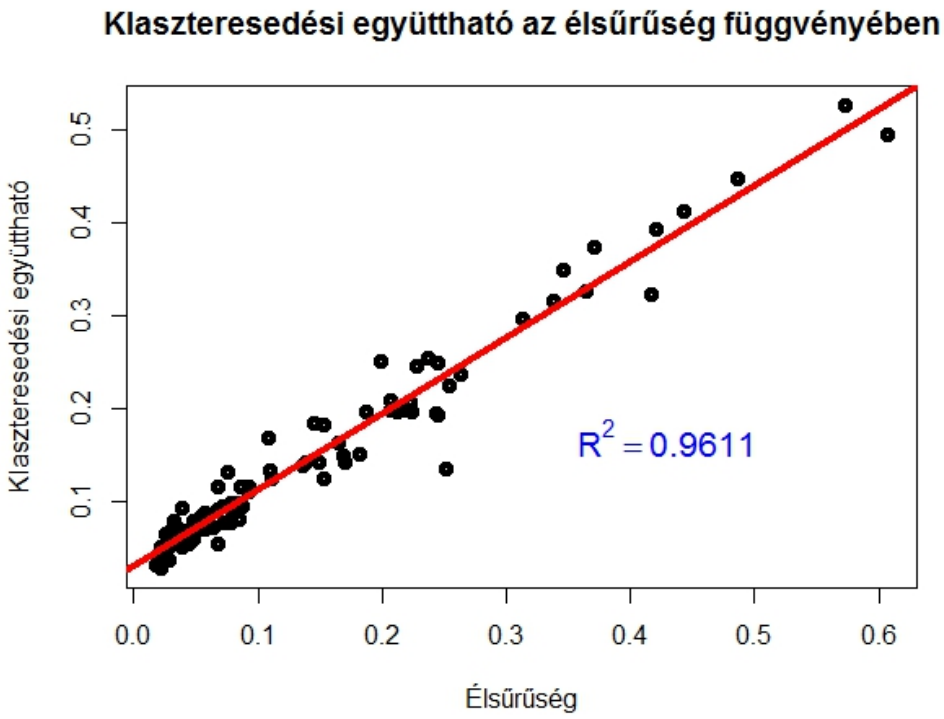
A következő cél tehát azonos élsűrűségű, de különböző klaszteresedési együtthatóval bíró gráfok előállítása volt a duplikációs modell segítségével. Ez azért nehézkes, mert pusztán a duplikációs modell sorsolási szabályával nem lehet azonos élsűrűségi gráfokat sorsolni.

Ennek érdekében a gráf paramétereit is sorsoltam. A q paramétert egyenletes eloszlás szerint a $[0, 0.5]$ intervallumból, az r paramétert pedig szintén egyenletes eloszlás szerint a $[0.5, 1]$ intervallumból. (Azt a motivációt megőrizve, hogy a klaszteresedési együttható ne legyen nagyon alacsony, ez pedig jellemzően a kisebb q paraméter választása mellett teljesült.) Majd a véletlen paraméterekkel a duplikációs modell segítségével összesen 100 különböző, 500 csúcsú gráfot sorsoltam. Ezekre a gráfokra a program segítségével kiszámítottam az élsűrűség és a klaszteresedési együttható értékét. A 3.5. ábra ezen 100 gráf élsűrűségének függvényében ábrázolja a hozzájuk tartozó klaszteresedési együtthatót (tehát minden pont egy gráfnak felel meg, vízszintes koordinátája az élsűrűség, függőleges koordinátája a klaszteresedési együttható). A 3.5 ábrán egy lineáris regresszióval készült egyenes is látható, melyre az $R^2 = 0.9722$ érték igen nagy, tehát erős lineáris kapcsolatra utal. Vagyis a klaszteresedési együttható lineárisan növekszik az élsűrűséggel,

[1] A görbe megnevezés nem teljesen indokolt, hiszen a járványterjedés egy diszkrét időben zajló folyamat, mindazonáltal a jól láthatóság érdekében ábrázoltam folytonos görbékkel a járványterjedést.

sőt, egy-két kivételtől eltekintve nagyjából azonos értékeket vesz fel a két mennyiség.

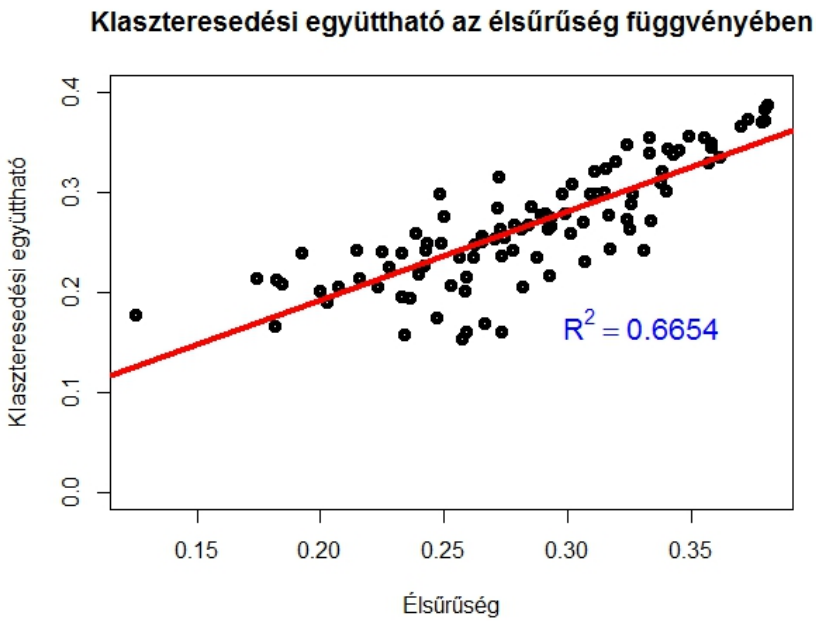
Innen azon természetes ötlet alapján indultam tovább, hogy megnéztem, mely q és r paraméterek esetén volt a klaszteresedési együttható és az élsűrűség különbségének abszolútértéke a legnagyobb (általában az előbbi a kisebb mennyiség). Ehhez a $q \approx 0.04$ és $r \approx 0.92$ paraméterértékek tartoztak. Itt érdemes megjegyezni azt is, hogy talán még alkalmasabb lett volna azt vizsgálni, hogy a két mennyiség hányadosa mikor a legnagyobb, illetve legkisebb. Erre azt találtam, hogy az $q \approx 0.4$ és az $r \approx 0.86$ paraméterek esetén a legnagyobb az eltérés aránya. Ezekkel a paraméterekkel sorsolva a gráfokat azonban a klaszteresedési együtthatók 0.03 és 0.08 körül mozogtak. Mivel ezek az értékek nem felelnek meg a valós hálózatoktól elvárt *kellően nagy* értéknek, ezért nem lenne célszerű ezekkel tovább dolgozni.



3.5. ábra. Az élsűrűség és a klaszteresedési együttható kapcsolata sorsolt véletlen gráfparaméterek (q, r) esetén

Mivel $q \approx 0.04$ és $r \approx 0.92$ paraméterek mellett volt a legnagyobb az abszolút eltérése a két vizsgált mennyiségnek (az élsűrűség és a klaszteresedési együttható), ezért alkalmasnak tűnt ehhez közeli paraméterekkel sorsolni a gráfokat, majd összehasonlítani az élsűrűségüket a klaszteresedési együtthatókkal.

3.3.2. Klaszteresedési együttható és a járványterjedés paraméterei közötti kapcsolat



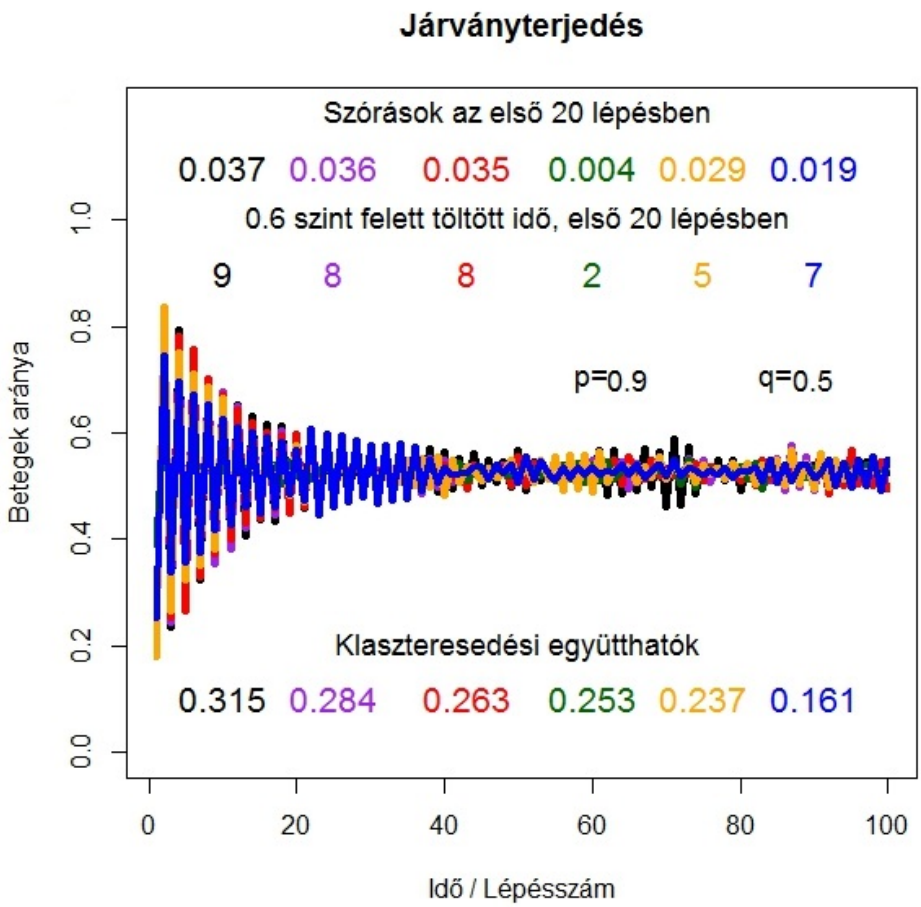
3.6. ábra. Eltérő élsűrűség és klaszteresedés, $q = 0.1, r = 0.9$

A továbbiakban tehát már fix, $q = 0.1$ és $r = 0.9$ paraméterekkel újabb 100 gráf sorsolása után ismét megvizsgáltam a klaszteresedési együttható és az élsűrűség együttes viselkedését. A 3.6. ábrán szintén az élsűrűség függvényében ábrázoltam sorsolt gráfok klaszteresedési együtthatóját. Az ábrán látható egyenes lineáris regresszióval készült, az R^2 értéke azt mutatja, hogy a lineáris kapcsolat nem kifejezetten erős a két mennyiség között (ellentétben az előző esettel, 3.5. ábra).

Tehát míg nagyobb skálán jól látszik a lineáris kapcsolat, addig leszűkítve a paramétereket a $q = 0.1$ és $r = 0.9$ esetre, már csak egy véletlen zaj figyelhető meg. Az a 3.6. ábráról is leolvasható, hogy ezen rögzített paraméterek mellett kapunk azonos élsűrűségű, de igen eltérő klaszteresedési együtthatóval bíró gráfokat, tehát ez a paraméterbeállítás megfelel a célnak.

Ezt követően kiválasztottam egy olyan gráfcsoportot, amelyeknek az élsűrűsége közel azonos. A továbbiakban a kisorsolt 100 gráf közül csak azokat vizsgáltam, melyeknek az élsűrűsége 0.270 és 0.274 közé esett. Ebben a konkrét esetben hat ilyen gráf volt. Ezen a hat gráfon a járványterjedés időbeni alakulását vizsgáltam a klaszteresedési együttható függvényében.

Minden gráfra egy 100 körös járványterjedési folyamatot futtattam $p = 0.9$ és $s = 0.5$ paraméterekkel. (Ezek a paraméterek a továbbiakban szintén rögzítettek, mindegyik jár-



3.7. ábra. Járványterjedés azonos élsűrűségű gráfokon, $p = 0.9, s = 0.5$

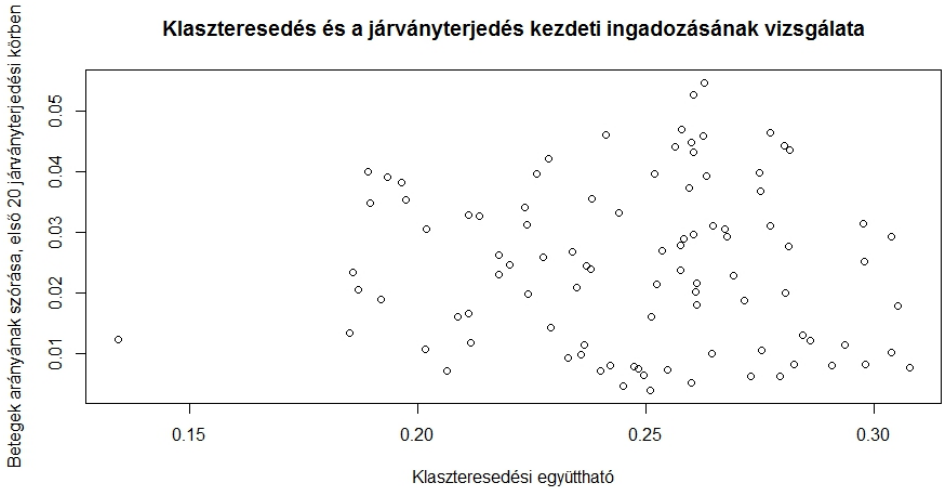
váynterjedési szimulációt ezkkel a paraméterekkel futtattam.) Ezt ábrázoltam a 3.7. ábrán. A járványterjedés során a betegek aránya mind a 6 esetben egy idő után azonos szintre áll be. Erről bővebben a következő részben, a járványterjedés hosszútávú viselkedésénél írok. Az azonban, hogy a járványterjedés első köreiben mekkora ezen mennyiségnek a szórása, igen eltérő. A nagyobb klaszteresedési együtthatójú gráfoknál nagyobb ingadozás figyelhető meg.

Megnéztem, hogy a járványterjedés első 20 körében hányszor esett a betegek aránya 0.6 fölé. Itt az az összefüggés látszik kirajzolódni, hogy a nagyobb klaszteresedési együtthatóval bíró gráfok esetében a beteg aránya többször esik a 0.6 szint fölé. Bár a legalacsonyabb klaszteresedési együttható esetében mégis meglepően sokszor, 7 körben is a 0.6 szint fölé kerülünk, az ábrán látható, hogy a kék görbe ingadozása mégis kisebb a járványterjedés első szakaszában. A betegek arányának szórását is megvizsgáltam, szintén a járványterjedés első 20 körében.

A 3.7. ábrán azt a feltevést fogalmaztam meg, hogy minél nagyobb a klaszteresedési együttható a gráfban, annál nagyobb a szórás a beteg csúcsok arányát illetően a járványterjedés első 20 lépésében (egy gráf kivételével ez teljesül).

Ez a hatás tehát már teljes mértékben a klaszteresedési együttható hatásának tudható be, hiszen közel azonos élsűrűségű gráfokat sorsoltam.

Nagyobb mintára is elvégeztem a kísérletet. Az élsűrűség kiszűrésével 96 olyan gráfot sorsoltam, melyeknek élsűrűsége közel azonos, 0.26 és 0.274 közé esik, és a klaszteresedési együttható függvényében vizsgáltam a járványterjedés során a betegek arányában tapasztalt szórást az első 20 járványterjedési körben. A nagyobb minta viszont már nem igazolta, hogy a klaszteresedési együttható növekedésével nő a szórás. Nyitott kérdés maradt, hogy az ingadozás hogyan függ a klaszteresedési együtthatótól.



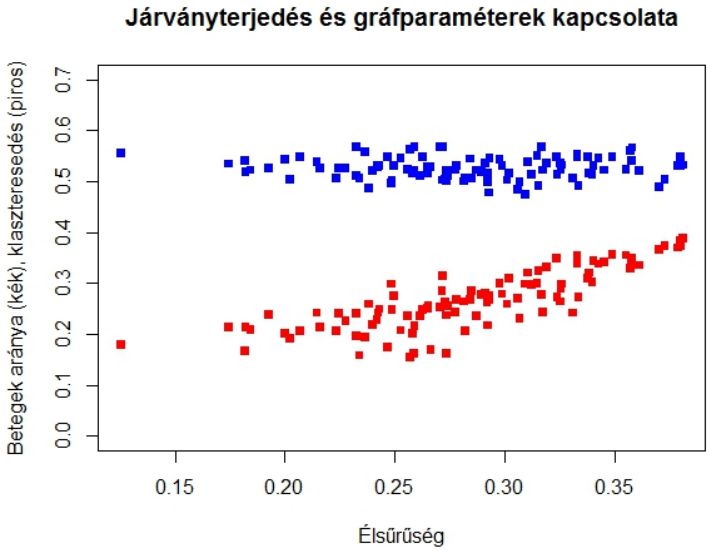
3.8. ábra. Betegek arányának szórása, első 20 lépésben azonos élsűrűségű gráfokon

3.3.3. A járványterjedés hossszútávú viselkedése

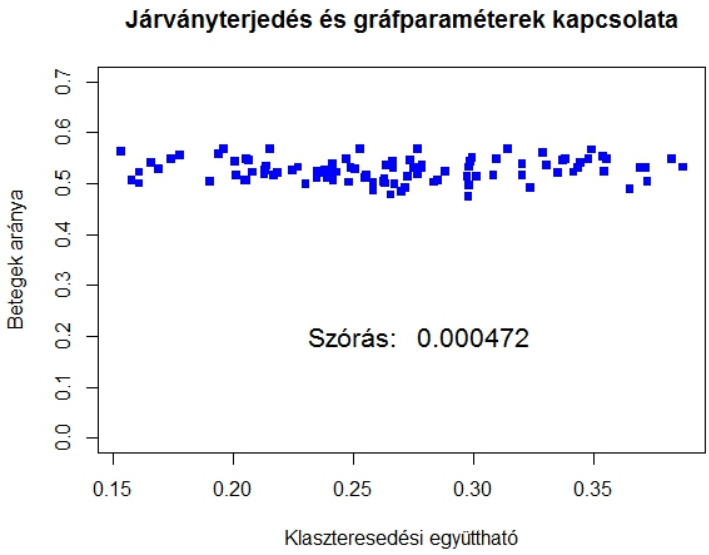
A 3.4. alapján már sejthető, hogy a klaszteresedési együtthatótól nem függ az az érték, amely köré a betegek aránya stabilizálódik, sok lépésben futtatva a járványterjedést. Az élsűrűség kiszűrése után, azonos élsűrűségű gráfokon is ugyanezt tapasztaljuk (a 3.7. ábra).

Az *élsűrűség szerinti szűrés nélkül*, 100 különböző gráfra is megvizsgáltam az élsűrűség függvényében a klaszteresedési együttható és a betegek arányának végső értékét egy 100 körös járvány lezajlása után (3.9. ábra), valamint csak a klaszteresedési együttható függvényében a beteg arányának végső értékét szintén egy 100 körös járvány lezajlása után (3.10), és azt tapasztaltam, hogy a betegek arányának végső alakulása nem függ sem a

klaszteresedési együtthatótól, sem az élsűrűségtől.



3.9. ábra. Betegek végső aránya és klaszteresedés az élsűrűség függvényében $p = 0.9, s = 0.5$



3.10. ábra. Betegek végső aránya a klaszteresedési együttható függvényében $p = 0.9, s = 0.5$

3.4. Eredmények összegzése, további kérdések felvetése

A járványterjedés szimulációja alapján azt tapasztaltam, hogy a járványterjedés paramétereinek rögzítése mellett a gráf élsűrűségétől és a klaszteresedési együtthatótól nem függ a beteg csúcsok aránya a gráfban, ha a járványterjedésre kellően nagy lépésszám után, kellő idő elteltével nézünk rá. A betegek aránya különböző gráfokra azonos egyensúlyi állapot körül stabilizálódik. A járványterjedés kezdeti szakaszában megfigyelhető ingadozás, a járványterjedés sebessége azonban igen eltérő képet mutat.

A tapasztalatok rengeteg kérdést felvetnek, melyeket érdemes lenne tovább vizsgálni a szimuláció segíségével.

Vajon mi az oka annak, hogy a járvány hosszútávú viselkedése nem függ az élsűrűségtől és a klaszteresedéstől? Vajon függhet-e valamilyen módon a duplikációs modell paramétereitől (N, q, r, a_1) a betegek arányában beálló egyensúlyi állapot?

Mi okozhatja a kezdeti ingadozásban tapasztalható változénykony képet? Milyen más mennyiséget érdemes még vizsgálni a gráfra vonatkozóan a klaszteresedési együttható mellett, hogy magyarázatot kapjunk a járványterjedés kezdeti fázisaiban tapasztalható különbségekre? Milyen más mennyiség jellemezheti még a gráf struktúráját, ami befolyásolhatja a járványterjedést?

Érdemes lenne továbbá összehasonlítani más véletlen gráfmodellekben, például a Barabási–Albert-modellben hogyan függ a klaszteresedési együtthatótól a járványterjedés? Vajon ott is igaz marad-e, hogy azonos élsűrűség mellett, eltérő klaszteresedési együtthatójú gráfok sorsolása esetén már nem függ a betegek aránya a gráftól, ha a járványterjedés paramétereit rögzítjük?

A. függelék

A duplikációs modell szimulációjának forráskódja

A.1. A duplikációs modell szimulációjához tartozó kódok

A.1.1. A *duplication* függvény

```
duplication<-function(X,q,r,t,c1){
  w<-sample(1:(t-1),1)
  j<-0
  vector<-c()
  if(!is.character(X[[w]])){   #w nem 0foku, szomszedait sorsoljuk
    n<-length(X[[w]])
    indicator<-rbinom(n,1,1-q) #1-q valoszinuseggel megtartjuk
    vector<-X[[w]]*indicator #sorsolas a szomszedok kozul
    vector<-vector[vector!=0]
    X[vector]<-lapply(X[vector], function(x) append(x, t))

    szomszed<-X[[w]]
    nemszomszed<-rep(1:(t-1))
    nemszomszed[szomszed]<-rep(0,n)
    nemszomszed<-nemszomszed[nemszomszed!=0]

    indicator2<-rbinom(t-1-n,1,r/t)
    vector2<-nemszomszed*indicator2 #sorsolas nemszomszedok kozul
    vector2<-vector2[vector2!=0]
    vector<-append(vector,vector2)
```

```
    X[vector2]<-lapply(X[vector2], function(x) append(x, t))
    }
  else { #0 foku, nincs szomszed
    indicator2<-rbinom(t-1,1,r/t)
    vector2<-rep(1:(t-1))*indicator2 #sorsolas a nemszomszedok kozul
    vector2<-vector2[vector2!=0]
    X[vector2]<-lapply(X[vector2], function(x) append(x, t))
    vector<-vector2
    }
  if(length(vector)==0){
    minimum<-min(length(X),c1)
    vector<-sample(1:(t-1),minimum)
    X[vector]<-lapply(X[vector], function(x) append(x, t))
  }
  X[[t]]<-vector
  return(X)
}
```

A.1.2. A *duplication függvény meghívása*

```
lepes <-500
a1<-10
minta<-1000
q=0.1
r=0.9
G<-list()
a<-c(2,3)
b<-c(1,3)
c<-c(1,2)
for (futas in (1:minta)){
  G[[futas]]<-list()
  G[[futas]][[1]]<-a
  G[[futas]][[2]]<-b
  G[[futas]][[3]]<-c
}
for (u in (4:lepes)){
  for (futas in (1:minta))
    G[[futas]]<-duplication(G[[futas]],q,r,u,a1)
```

```
}
```

A.2. A klaszteresedési együttható és az élsűrűség számolása

A.2.1. A *clustering* függvény

```
clustering<-function(X,c){# X graf, c fokszameloszlasvektor
  haromszogek<-0        #haromszogszamolas
  for (i in 1:length(X)){
    if (!is.character(X[[i]])){
      L<-X[[i]]
      for (j in 1:(length(X[[i]]))){
        k<-X[[i]][j]
        M<-X[[k]]
        haromszogek<-haromszogek + length(intersect(L,M))
      }
    }
  }
  haromszogek<-haromszogek/2

  cs<-0 #cseresznyeszamolas
  for (i in 1:length(c)){
    cs<-cs+choose(c[i],2)
  }

  clust<-0
  clust<-haromszogek/cs
  return(clust)
}
```

A.2.2. A *fokszameloszlas* függvény

```
fokszameloszlas<-function(X){
  c<-c()
  for(i in 1:length(X)){
    if (is.character(X[[i]])){
      c[i]<-0
    }
    else {
```

```
        c[i]<-length(X[[i]])
      }
    }
    return(c)
}
```

A.2.3. Az *elsuruseg* és az *elszam* függvény

```
elsuruseg<-function(X,c){# szomszedlista, fokszameloszlas-vektor
  E<-sum(c)
  n<-length(X)
  S<-0.5*E/choose(n,2)
  return(S)
}
elszam<-function(X,c){# szomszedlista, fokszameloszlas-vektor
  E<-sum(c)
  S<-0.5*E
  return(S)
}
```

A.2.4. *mapply, lapply* függvények alkalmazása

```
g<-list() #g elemei: vektorok, fokszameolaszlasvektorok
g<-lapply(G, fokszameloszlas)

H<-c() #H vektor, elemei a graf vegleges klaszteresedesi egyhoja
ES<-c() #ES vektor, elemei a graf vegleges elsurusege
H<-mapply(clustering, G, g)
ES<-mapply(elsuruseg, G, g)
ELSZAM<-mapply(elszam, G, g)
```

A.3. A járványterjedés szimulációjához tartozó kódok

A.3.1. A *foksz_fert* függvény

```
foksz_fert<-function(X,j){
# vektort ad vissza,
#i. koord: i. csucsnak hany fertozott szomszedja van
#X graf, j: 0/1 vektor, fertozott/nem fertozott
```

```
  c<-c()
  for(i in (1:(length(X)))){
    if (is.character(X[[i]])){
      c[i]<-0
    }
    else {
      s<-j[X[[i]]]
      c[i]<-sum(s)
    }
  }
  return(c)
}
```

A.3.2. A *jarvanyterjedes_idoben* függvény

```
szures<-which(ES > 0.26 & ES<0.274)

jarvanyterjedes_idoben<-function(X,q,p,L){
  n<-length(X)
  Fert<-c(1, rep(0,(n-1)))
  vel<-c()
  B<-c()
  for (l in (1:L)){
    f<-foksz_fert(X,Fert)
    for (i in (1:n)){
      if ((Fert[i]==0)&&(f[i]>0)){
        vel[i]<-rbinom(1,1,1-(1-q)^f[i])
        if(vel[i]==1){
          Fert[i]<-1
        }
      }
      else{
        if (rbinom(1,1,p)==1){
          Fert[i]<-0
        }
      }
    }
    B[l]<-sum(Fert)/n
```

```
  }
 return(B)
 #B:betegek aranya L db lepes mindegyikeben, L elemu vektor
}

L<-20
pj<-0.8
qj<-0.5


Bgido<-list()
futasindex<-0
for (futas in szures){
 futasindex<-futasindex +1
  Bgido[[futasindex]]<-jarvanyterjedes_idoben(G[[futas]],qj, pj, L)

}

#VarB<-sapply(Bgido, var)
```

Irodalomjegyzék

[1] REMCO VAN DER HOFSTAD, *Random graphs and complex networks, Volume I.*, Cambridge Series in Statistical and Probabilistic Mathematics, (2016)

[2] ERDŐS, P., RÉNYI, A., *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 5, (1960), pp. 17-61

[3] BARABÁSI, A.-L., and ALBERT, R. *Emergence of scaling in random networks* Science, 286 (5439), (1999) pp. 509-512.

[4] BOLLOBÁS, B.,RIORDAN, O. *The diameter of a random scale-free graph* Combinatorica, 24 (1), (2004) pp. 5-34.

[5] Backhausz, A., T. F. Móri, *Further properties of a random graph with duplications and deletions*, Stochastic Models 32(1), (2016), 99–120.

[6] BEBEK, G., BERENBRINK, P., COOPER, C., RIEDETZKY, T., NADEAU,J., SAHINALP, S. *The degree distribution of the generalized duplication model.* (`http://web.stanford.edu/ saberi/epidemic.pdf`)

[7] FELIX HERMANN, PETER PFAFFELHUBER, *Large-scale behavior of the partial duplication random graph, Preprint.*, arXiv:1408.0904

[8] `http://konect.uni-koblenz.de/statistics/clusco`

[9] BERGER, N., BORGS, C., CHAYES, J.T., SABERI, A. *On the spread of viruses on the internet* in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 301–310, ACM, New York. (`http://web.stanford.edu/ saberi/epidemic.pdf`)

[10] Ágnes Backhausz, Tamás F. Móri, *Local degree distribution in scale free random graphs* Electronic Journal of Probability, 16 (54), (2011), pp. 1465-1488, *http://ejp.ejpecp.org/article/view/916/1116*

[11] MÓRI, T., *Diszkrét paraméterű martingálok*, Typotex Kft., (2011), pp. 71-73.

[12] TONĆI ANTUNOVIĆ, ELCHANAN MOSSEL, MIKLÓS Z. RÁCZ, *Coexistence in Preferential Attachment Networks. Combinatorics, Probability and Computing*, (2016), Vol. 25, no. 6p. 797–822. DOI 10.1017/S0963548315000383.

[13] DEREICH, S., ORTGIESE, M. *Robust Analysis of Preferential Attachment Models with Fitness* Combinatorics, Probability and Computing, (2014), Vol. 23, no. 3p. 386–411. DOI 10.1017/S0963548314000157.

[14] DEREICH, S., MÖRTERS, P. *Random networks with sublinear preferential attachment: degree evolutions* Electron. J. Probab., (2009), 14:no. 43, 1222–1267.

[15] BORGS, C. CHAYES, J., DASKALAKIS, C., ROCH, S. *First to market is not everything: an analysis of preferential attachment with fitness* In STOC'07—Proceedings of the 39th Annual ACM Symposium on Theory of Computing, (2007), pp. 135–144. ACM, New York,

[16] COHEN, N., JORDAN, J., VOLIOTIS, M. *Preferential duplication graphs*, Journal of Applied Probability, 47(2), (2010), 572-585. doi:10.1017/S0021900200006823

[17] JORDAN, J. *Randomised Reproducing Graphs* Electron. J. Probab. Volume 16 (2011), paper no. 57, 1549-1562.

[18] JACOB, E., MÖRTERS, P., *The spread of infections on evolving scale-free networks*, (2015), arXiv:1512.00832

[19] CHUNG, F., LU L., *Complex Graphs and Networks*, (2006) (Am Math. Society, Providence, RI).

[20] BRITTON, D., DEIJFEN, M., LAGERÅS A. N., LINDHOLM M., *Epidemics on random graphs with tunable clustering*, J. Appl. Prob. 45,(2008) 743–756 (2008)

[21] ANDERSSON, H. Britton, T. *Stochastic Epidemic Models and Their Statistical Analysis* Springer Science and Business Media, (2012)