

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR
VALÓSZÍNŰSÉGELMÉLETI ÉS STATISZTIKA TANSZÉK

VÉLETLEN GRÁFOK GYÖKERÉNEK IDENTIFIKÁLÁSA ÉS ELREJTÉSE

Szakdolgozat

Témavezető:

GERENCSÉR BALÁZS

Egyetemi adjunktus

Készítette:

DÖBRÖNTEI DÁVID BENCE

matematika BSc



Budapest, 2021

Köszönetnyilvánítás

Elsősorban témavezetőmnek, Gerencsér Balázsnak szeretnék köszönetet mondani. Irányadó útmutatásai, szakmai tanácsai, kérdéseimre adott részletes válaszai nélkül dolgozatom nem jöhetett volna létre. Optimista, pozitív hozzáállásával elérte, hogy a járványhelyzet miatt online térbe költözött konzultációk kellemes hangulatban teljenek.

Köszönet illeti Szőnyi Laurát, Csahók Tímeát és Rancz Adriennét is, akik francia nyelvtudásukkal segítettek munkámat. Nem feledkezhetek meg Kiglics Mátyásról sem, aki számítástechnikával kapcsolatos kérdéseimet idejét nem sajnálva mindig végighallgatta. Hálás vagyok továbbá Pituk Sárának értékes tanácsaiért.

Szeretnék köszönetet mondani középiskolai matematikatanárainknak, közülük is leginkább Spissich László tanár úrnak, hogy elindított ezen a pályán. Köszönettel tartozom továbbá egyetemi oktatóimnak, akik a hagyományos és az online keretek közé szorult oktatás során is arra törekedtek, hogy magas színvonalon adják át a matematika szépségeit.

Végül, de nem utolsó sorban köszönettel tartozom családomnak. Szüleim, akik már általános iskolás koromban felfigyeltek matematika iránti érdeklődésemre, életem minden szakaszában támogattak.

Tartalomjegyzék

1	Bevezetés	1
2	Véletlen növény fák és kapcsolódó fogalmak	3
2.1	Véletlen fák	3
2.2	Pólya-urnák	4
2.3	Pólya-urnák a véletlen fák elméletében	6
3	A gyökér hatása	8
4	Gyökérkereső algoritmusok	10
4.1	Jordan-centralitás	10
4.2	Rumor centralitás	15
4.3	Távolságcentralitás	20
4.4	Összehasonlítás	21
5	A gyökér elrejtése	24
6	Terjedés adott hálózaton	27
6.1	Jordan-centralitás	28
6.2	Rumor centralitás	29
6.3	A gyökér elrejtése	31
7	Összegzés és kitekintés	39

1 Bevezetés

A vírusok, ragályos betegségek megértése régóta foglalkoztatja az emberiséget. A kórokozók biológiai vizsgálatán túlmenően további fontos kérdések is felmerülnek például a vírus terjedésével kapcsolatban. A sokat tanulmányozott SIR (Susceptible, Infectious, Recovered) modell már a múlt század első felében megjelent. A kiterjedt energetikai hálózatok, számítógépes rendszerek, az internet megjelenésével újabb alkalmazási területek nyíltak az időben fejlődő hálózatok elmélete előtt.

Az ilyen, fejlődő folyamatok esetén gyakran lényegi kérdés a legelső egyed ismerete, legyen itt szó akár az első vírusedzőről, akár egy álhír terjesztőjéről a közösségi médiában. Jelen dolgozatban erre a kérdésre próbálunk minél körültekintőbben választ adni.

Habár a kérdés alapos matematikai kidolgozását Shah és Zaman [1] 2010-es cikke indította el, az azóta eltelt évtizedben a témának jelentős szakirodalmi keletkezett. Ennek fényében nem lehet célunk bemutatni a felgyülemlett ismereteket teljes részletességükben, csupán betekintést kívánunk nyújtani a téma sokszínű kérdéskörébe.

A dolgozatban kétféle modellt is ismertetjük a vírusok terjedésének. Az egyik lehetőség az ún. preferential vagy uniform attachment modell, melyek lényege, hogy a fertőzöttek egy véletlen fagráf szerint gyűlnek. Minden időpillanatban egy új csúcshoz kapcsolódik a gráfhoz, mégpedig valamelyik korábbi csúcshoz egy adott valószínűségi eloszlás szerint véletlenül. Bubeck, Devroy és Lugosi [2] cikkében képletesen szólva "Ádám megtalálására" tesz kísérletet ebben a modellben. Alapvető jelentőségű eredményeik jelen dolgozat vonatkoztatási pontjával szolgálnak.

A másik bemutatott lehetőség a diffúziós, vagy terjedési modell: egy adott hálózaton terjed a vírus valamilyen véletlen szabály szerint. Mint látni fogjuk, az uniform és preferential attachment modellek esetében látott technikák jól alkalmazhatóak lesznek ebben az esetben is.

Mindkét esetben mutatunk olyan algoritmusokat, melyek nagy valószínűséggel azonosítják a vírus kiindulópontját, vagy legalábbis megadják a csúcsok egy szűk részhalmozát, amely nagy valószínűséggel tartalmazza a gyökeret.

Mindkét modell közös jellemzője, hogy a vírus terjedését valamilyen fagráf struktúra jellemzi. Más gráfosztályokról általában nagyon keveset tudunk, pusztán remélhetjük, hogy az itt kidolgozott elmélet - legalább részben - általánosítható tetszőleges struktúrára.

Miután kiinduló kérdéseinkre, a gyökér meghatározására, sikerült választ kapnunk, természetesen merül fel a - valamilyen értelemben - ellentétes kérdés: el lehet-e rejteni a forrást egy ellenség elől. Szeretnénk tehát olyan terjedési modellt alkotni, vagy a meglévő modelleinket úgy módosítani, hogy a gyökér detektálása ne legyen lehetséges. Ezt a kérdést például közösségi oldalak, mikroblog-szolgáltatók anonimitási problémái motiválják.

Míg a diffúziós modellben ismert ilyen terjedés, az ún. adaptív diffúzió, a preferential és az uniform attachment modell esetében nincs tudomásunk hasonlóról. Erre az esetre mutatunk egy ötletet, amelyről szimuláció útján kiderül, hogy valamelyest sikeresen rejti el a vírus gyökerét.

Előjáróban megjegyezzük, hogy eddig és a továbbiakban mindig a vírusforrás, a vírus gyökere, a vírus kiindulási pontja stb., vagy csak simán a forrás, gyökér, kiindulási pont kifejezéseket szinonimaként használjuk. Amennyiben nem emeljük ki külön, ezen azt az egyetlen csúcst értjük, amelyik a nulladik időpillanatban a fertőzést hordozza.

A dolgozat lényegi része öt fejezetre oszlik. A 2. fejezetben bevezetjük a véletlen fák fogalmát, és így az uniform és a preferential attachment modelleket is. Továbbá a bizonyítások során kulcsszerepet játszó Pólya-urnák elméletébe adunk egy rövid betekintést.

A 3. fejezetben a véletlen fák gyökerének hatását vizsgáljuk. Habár nem ez a dolgozat fő témája, a kérdés megértésében fontos szerepet játszik, mintegy megalapozza azt.

A 4. fejezet tartalmazza a téma legalapvetőbb eredményeit és azok bizonyításait. Ebben precízen definiáljuk a gyökérkereső algoritmust, és különféle centralitási tulajdonságok segítségével meg is adunk hatásos algoritmusokat a gyökér megtalálására. A 4.4 alfejezetben megmutatjuk, hogy a bevezetett algoritmusok gyorsak is, továbbá szimuláció segítségével a hatékonyságukat is összehasonlítjuk.

A 5. fejezet a gyökér elrejtéséről szól, a kérdés motivációján túl bemutatunk egy lehetőséget a preferential attachment modell módosítására. Erről szimulációval demonstráljuk, hogy alkalmazható lehet az információforrás elrejtésére.

A 6. fejezetben a diffúziós modellt vizsgáljuk, ehhez a 4. fejezetben látott módszerek lesznek segítségünkre. Végezetül a 6.3 részben a gyökér elrejtésére alkalmas adaptív diffúziót mutatjuk be.

2 Véletlen növény fák és kapcsolódó fogalmak

2.1 Véletlen fák

Ebben a fejezetben Banerjee és Bhamidi [3] cikke alapján véletlen fák különböző típusait vezetjük be. Ezek közül többnyire csak a legtöbbet vizsgált uniform és preferential attachment modellekkkel foglalkozunk.

Véletlen fákat - az alábbiakban részletezett - rekurzív módon fogunk előállítani. Ehhez először vezessük be a következő jelöléseket: a számozott T fa izomorfiacsoportha T° , erre úgy is tekinthetünk, mintha a T számozását elhagytuk volna. (Kényelmi okokból a T° csúcsaira továbbra is a T szerinti számozásukkal hivatkozunk.) A csúcsokra általában a számukkal hivatkozunk, de ha hangsúlyozni akarjuk, hogy csúcsról - és nem számról - van szó, akkor az i csúcsot v_i -vel is jelölhetjük. A csúcsok halmazára a szokásos $V(T)$ jelölést használjuk, és a v csúcs fokszámát $d_T(v)$, vagy egyszerűen $d(v)$ jelöli. Gyakran a jelölés egyszerűsítése érdekében azonosítjuk a gráfot a csúcsaival, így például többnyire $|T|$ jelenti a T fa csúcsainak számát, azaz $|V(T)|$ -t.

A T fa számozását *növekvőnek* hívjuk, ha az 1 csúcsból kiinduló összes úton a csúcsok számozása nő. *Rekurzív fának* nevezzük az olyan számozott fákat, amelyek növekvően számozottak. Ha a T fa gyökere $u \in V(T)$, akkor ezt gyakran a (T, u) jelöléssel hangsúlyozzuk. Továbbá $(T, u)_{v\downarrow}$ jelöli a v -ből és leszármazottaiból álló részfat (az u -t tekintve a gyökérnek).

Tekintsünk egy - a pozitív egészen értelmezett - rögzített f függvényt, az ún. csatlakozási függvényt.

2.1 Definíció (Véletlen fa). Egy véletlen fa növekedését leírhatjuk fák egy $\{T_f(j) : 2 \leq j \leq n\}$ sorozatával az alábbi módon:

- $n = 2$ esetén $T_f(j)$ két csúcs egyetlen éllel összekötve, jelölje a csúcsokat $\{v_1, v_2\}$.
- Ha $n \geq 2$, akkor $T_f(n-1)$ csúcsait $\{v_1, \dots, v_{n-1}\}$ -gyel jelölve az n -edik lépésben egy új v_n csúcs kerül a rendszerbe, és éllel csatlakozik az eddigi csúcsok valamelyikéhez véletlenszerűen.

Mégpedig úgy, hogy a v_i csúcsához $f(d_{T_f(n-1)}(v_i))$ -vel arányos valószínűséggel csatlakozik. Ekkor azt a v_j -t, amelyikhez v_n csatlakozik, az ősenek, míg v_n -et a v_j leszármazottjának nevezzük. Ezzel a rekurzióval jutunk az n csúcsú $T_f(n)$ fához.

Szerepeljenek itt a leggyakrabban használt f csatlakozási függvények:

- **Uniform attachment:** $f(\cdot) \equiv 1$. Azaz az új csúcs egyforma valószínűséggel választja bármelyik korábbi csúcsot az ősenek. Ebben az esetben a növekvő sorozat elemeit szokás $UA(n)$ -nel jelölni a $T_f(n)$ helyett.

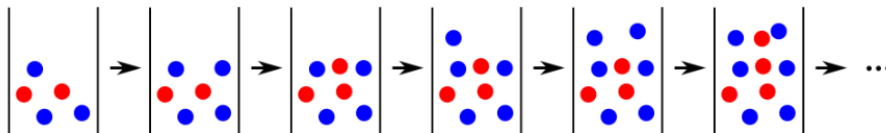
- **"Tiszta" preferential attachment:** $f(k) = k$. Azaz az új csúcs a fokszámmal arányos valószínűséggel csatlakozik a korábbiak valamelyikéhez. Erre a fontos esetre is külön jelölést alkalmazunk: $PA(n)$.
- **Affin preferential attachment:** $f(k) = k + \beta$ valamilyen $\beta \geq 0$ konstanssal. Ez lényegében a hagyományos preferential attachment általánosítása, lehetőséget ad a csatlakozási valószínűség finomhangolására.
- **Szublineáris preferential attachment:** $f(k) = k^\alpha$ valamilyen rögzített $0 < \alpha < 1$ konstansra. Vegyük észre, hogy $\alpha = 0$ esetén az uniform, míg $\alpha = 1$ esetén a preferential attachmentet kapnánk vissza. Ennél általánosabban is használatos a szublineáris attachment elnevezés:

- f nemcsökkenő, nem azonosan 1, $f(k) \geq 1$ minden $k \geq 1$ -re és
- $f(k) \leq k^\alpha$ minden $k \geq 1$ -re, ahol $\alpha \in (0, 1)$.

Vizsgálódásunk középpontjában az uniform és a preferential attachment fák állnak. Ezek fejlődésének megértéséhez kulcsfontosságú fogalom a klasszikus elméletből ismert Pólya-urna. Ezért a következőkben - Rácz és Bubeck [4] összefoglaló művét alapul véve - a Pólya-urnák elméletébe adunk egy rövid betekintést.

2.2 Pólya-urnák

A klasszikus Pólya-urnában kezdetben k kék és p piros golyó van. Minden lépésben kihúzzunk egy golyót véletlenszerűen, és visszateszünk még egy vele azonos színűt, például a 2.1 ábrán¹ az első lépésben kék, míg a másodikban piros golyót húztunk. A kék és piros golyók aránya érdekel minket hosszú idő után. A kék golyók számát X_n -nel, míg a kék golyók arányát $x_n = \frac{X_n}{n}$ -nel jelöljük. Ekkor tehát kezdetben $X_{k+p} = k$.



Ábra 2.1: A klasszikus Pólya-urna egy realizációja

Számoljuk ki a kék golyók számának várható növekedését minden lépésben:

$$\mathbb{E}[X_{n+1}|X_n] = (X_n + 1) \cdot \frac{X_n}{n} + X_n \cdot \left(1 - \frac{X_n}{n}\right) = \left(1 + \frac{1}{n}\right)X_n$$

¹Az ábrát Rácz és Bubeck [4] cikkéből kölcsönöztük.

A fenti egyenlőséget $(n + 1)$ -gyel osztva kapjuk, hogy:

$$\mathbb{E}[x_{n+1}|\mathcal{F}_n] = x_n$$

Azért térhettünk át az X_n -ről az \mathcal{F}_n filtrációra, mert X_n Markov-folyamat. Tehát x_n martingál. Sőt, mivel $x_n \in [0, 1]$, korlátos martingál. A martingál konvergencia tétel szerint tehát egy valószínűséggel konvergál egy valószínűségi változóhoz.

A határeloszlás megtalálásához írjuk fel annak a valószínűségét, hogy az első n húzásból m volt kék. Ennek kiszámításában segít az a megfigyelés, hogy annak a valószínűsége, hogy n húzásból m kék és $n - m$ piros volt nem függ a sorrendtől.

$$\mathbb{P}(X_{n+k+p} = k + m) = \binom{n}{m} \frac{k(k+1) \dots (k+m-1) \times p(p+1) \dots (p+n-m-1)}{(k+p)(k+p+1) \dots (k+p+n-1)}$$

Következésképpen $X_{n+k+p} - k$ eloszlása béta-binomiális. A béta-binomiális eloszlásra úgy is tekinthetünk, hogy először q -t kisorsoljuk béta-eloszlás szerint, melynek paraméterei k és p , majd egy q paraméterű binomiális eloszlást tekintünk. Ezután felírva a nagy számok erős törvényét a binomiális eloszlásra, q feltétel mellett azt kapjuk, hogy $(X_{n+k+p} - k)/n$ egy valószínűséggel q -hoz tart. Mivel $x_n = (X_{n+k+p} - k)/n + o(1)$, ezért x_n is q -hoz tart egy valószínűséggel. Tehát igaz a következő tétel.

2.2 Tétel. *Ha az x_n a kék golyók aránya n lépés után egy klasszikus Pólya-urnában, amelyben kezdetben k kék és p piros golyó volt, akkor 1 valószínűséggel:*

$$\lim_{n \rightarrow \infty} x_n = x,$$

ahol $x \sim \text{Béta}(k, p)$

A klasszikus eset természetes általánosítása az, amikor nem csak két színünk van, és nem csak plusz egy golyót helyezünk vissza a kihúzott színből, hanem előre megadott k darabot. Ebben az esetben a fentihez hasonló módon látható be az alábbi tétel.

2.3 Tétel. *Ha az \underline{x}_n az adott színű golyók aránya n lépés után a Pólya-urnában, amelyben kezdetben r_i golyó volt az i színből (összesen m szín van), és k azonos színű golyót teszünk vissza, akkor 1 valószínűséggel*

$$\lim_{n \rightarrow \infty} \underline{x}_n = \underline{x},$$

ahol $\underline{x} \sim \text{Dirichlet}(r_1/k, \dots, r_m/k)$

Legáltalánosabb esetben a visszatevést egy mátrix írja le, ennek sorai a kihúzott golyó színét jelentik, míg oszlopai, hogy az adott színű golyóból mennyit kell visszatenni. Az ilyen ún. visszatevési mátrixok közül a háromszögmátrixok elméletét Janson [5] cikke részletesen tárgyalja. Nekünk a preferential attachment modellel kapcsolatban ebből csak az alábbi eredményre lesz szükségünk.

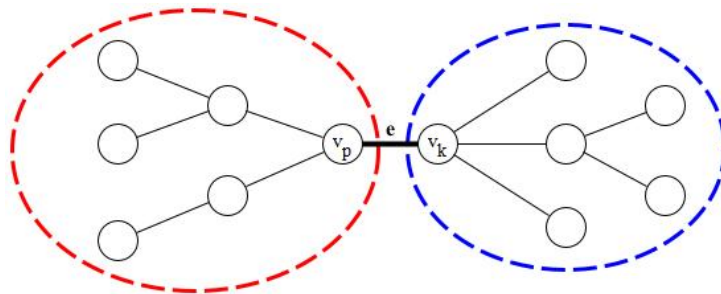
2.4 Tétel. Legyen (X_n, Y_n) rendre a kék és piros golyók száma az n . időpontban egy Pólya urnában, amelynek visszatevési mátrixa: $\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$. Tegyük fel, hogy volt kezdetben néhány piros golyó az urnában. Ekkor Y_n/\sqrt{n} eloszlásban tart egy nemelfajult valószínűségi változóhoz.

A fenti tételben szereplő $\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$ visszatevési mátrix annyiban módosítja a klasszikus Pólya-urnát, hogy minden húzás során még egy második, kék golyót is visszateszünk a húzott golyó színétől függetlenül. Ezért nem meglepő a tétel állítása, miszerint a piros golyók száma n helyett \sqrt{n} -nel arányos.

Mint ahogyan azt előrevetítettük, a Pólya-urnák szoros kapcsolatban állnak az általunk tárgyalt véletlen fa modellekkel. Ezt a kapcsolatot vizsgáljuk meg a következő részben.

2.3 Pólya-urnák a véletlen fák elméletében

Az uniform attachment modellben a klasszikus Pólya-urna az alábbi módon jelenik meg: tekintsük a T fa egy e élét, melynek két végpontja v_k és v_p . Ennek az élnek az elhagyásával a gráf két komponensre esik szét, a 2.2 ábrán² látható módon. Az így kapott két részfa mérete éppen egy klasszikus Pólya-urna szerint növekszik. Tehát teljesül rá a 2.2 tétel.



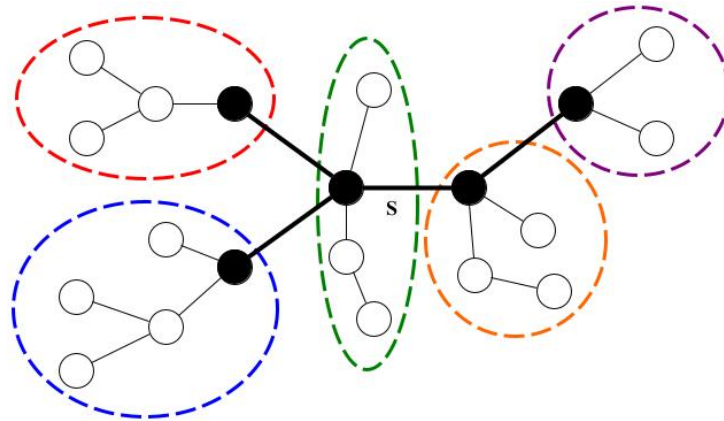
Ábra 2.2: Uniform attachment modellben a két részfa mérete egy klasszikus Pólya-urna szerint növekszik

Amennyiben nem egy élet, hanem egy m csúcsú $S \subset T$ részfa éleit hagyjuk ki a gráfból, úgy m különböző összefüggőségi komponenst kapunk. Ezeknek a részfáknak a mérete olyan Pólya-urna szerint változik, amelyben m különböző színű golyó van. Ebben az esetben tehát a részfák méretének eloszlásáról a 2.3 tétel segítségével mondhatunk valamit.

Olyan Pólya-urnát, ahol két azonos színű golyót teszünk vissza, a preferential attachment fák esetén figyelhetünk meg. A gráfból az előzőhöz hasonlóan egy m csúcsú S részfát kihagyva m komponenst kapunk. Ezek méretét a bennük lévő csúcsok fokszá-

²Ez és a későbbiekben szereplő ábrák a yEd Graph Editor szoftver segítségével készültek.

mainak összegével mérve éppen egy olyan Pólya-urnát kapunk, amelyikbe két, a kihúzottal megegyező színű golyót is visszateszünk.



Ábra 2.3: Uniform és preferential attachment modellben is megjelenik a többszínű Pólya-urna

Preferential attachment fák esetén egy adott csúcs fokszámának változásának megértésében segít a $\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$ visszatevési mátrixú Pólya-urna. Ugyanis rögzítsünk egy v csúcsot, és fokszámát az n -edik lépésben jelöljük Y_n -nel. Jelölje továbbá X_n az összes többi csúcs fokszámának az összegét. Ekkor X_n minden lépésben legalább eggyel nő, de Y_n csak akkor, ha a v lett az új csúcs őse. Tehát (X_n, Y_n) éppen a fenti visszatevési mátrixú Pólya-urna szerint változik, ezért alkalmazható rá a 2.4 tétel.

3 A gyökér hatása

Mielőtt rátérnénk a dolgozat fő témájára, a véletlen fák gyökerének megkeresésére, vizsgáljuk meg, hogy a határeloszlás függ-e egyáltalán a kiinduló gráftól. A válasz megnyugtató: valóban függ, legalábbis valamilyen értelemben. Így reményünk van arra, hogy csak egy későbbi állapot ismeretében következtessünk a kiindulási gráfra, speciálisan a vírusforrás meghatározására is.

Megjegyezzük, hogy ebben a fejezetben a 2.1 definíció egy olyan módosított változatát vizsgáljuk, amelyben a rekurzió egy adott T fából indul, ilyenkor a gráfsorozat n csúcsú tagját (ahol $n \geq |V(T)|$) $T_f(n, T)$ -vel, vagy speciálisan $UA(n, T)$ -vel, $PA(n, T)$ -vel jelöljük.

Szeretnénk eldönteni tehát, hogy van-e összefüggés T és $PA(n, T)$ között. Erre több lehetőségünk van, ahogy Bubeck, Mossel és Rácz [6] utal rá.

Egyrészt tekinthetjük a $PA(\infty, T)$ megszámlálható sok csúcsot tartalmazó fát, amelyet úgy kapunk, hogy a preferential attachment modell rekurzióját minden korláton túl folytatjuk. Kleinberg és Kleinberg [7] megmutatta, hogy ilyen esetben nincs hatása a gyökérnek. Tetszőleges T kiindulási fára a $PA(\infty, T)$ 1 valószínűséggel az egyértelmű izomorfiaosztálya lesz azon megszámlálható csúcsú fáknak, melyekben minden csúcs foka végtelen. Ez a megállapítás tetszőleges növekedési modellre fennáll, amelyben minden rögzített csúcs foka a végtelenbe tart.

A leginkább vizsgált lehetőség azonban képes megmutatni a gyökér hatását, ez az alkalmazások szempontjából is fontos, és talán a legtermészetesebb választás. Ehhez először a hipotézisvizsgálat nyelvére kell lefordítanunk feladatunkat: adott két lehetséges kiindulási gráf, T és S , valamint egy R megfigyelés. Arról szeretnénk dönteni, hogy vajon $R \sim PA(n, T)$ vagy $R \sim PA(n, S)$. Az eredeti kérdést tehát így fogalmazhatjuk át: van-e olyan n -ben aszimptotikusan nem-elhanyagolható erejű teszt, ami eldönti a fenti kérdést? Ez a $PA(n, T)$ és a $PA(n, S)$ eloszlások közötti totális variációs távolság vizsgálatával ekvivalens. Ezért vezessük be az alábbi δ függvényt:

$$\delta(S, T) := \lim_{n \rightarrow \infty} \text{TV}(PA(n, S), PA(n, T)). \quad (3.1)$$

Itt TV jelöli a totális variációs távolságot. Emlékeztetőül, az X és Y valószínűségi változók totális variációs távolsága a véges \mathcal{X} téren vett μ és ν valószínűségi mérték esetén

$$\text{TV}(X, Y) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$$

A δ definíciójában szereplő határérték jóldefiniált, ugyanis $\text{TV}(PA(n, S), PA(n, T))$ nemnövekvő és mindig nemnegatív.

Az alábbi tétel szerint ha az S és T fák - $\vec{d}(S)$ és $\vec{d}(T)$ -vel jelölt - fokszámprofiljai eltérnek, akkor a gyökér hatása érzékelhető ebben az értelemben:

3.1 Tétel (Bubeck, Mossel, Rácz [6]). *Legyen S és T két rögzített, legalább 3 csúcsú fa. Ha $\vec{d}(S) \neq \vec{d}(T)$, akkor $\delta(S, T) > 0$.*

Sőt, néhány speciális esetben több is mondható. Például k csúcsú csillagokra - azaz olyan fákra, amelyekben egy központi csúcsnak minden többi csúcs a szomszédja - igaz az alábbi:

3.2 Tétel (Bubeck, Mossel, Rácz [6]). *Jelölje S_k a k csúcsú csillag gráfot. Ekkor tetszőleges T rögzített fára $\lim_{k \rightarrow \infty} \delta(S_k, T) = 1$.*

Ezek alapján Bubeck, Mossel és Rácz azt sejtették, hogy tetszőleges nem-izomorf kiindulógráfokra $\delta(S, T) > 0$. Sejtésük igaznak bizonyult, erről szól az alábbi:

3.3 Tétel (Curien, Duquesne, Kortchemski, Manolescu [8]). *A (3.1) egyenlőséggel definiált δ függvény metrika a legalább 3 csúcsú fák körében.*

Tehát a preferential attachment modellben van hatása a gyökérnek. Természetesen ugyanezt a kérdést feltehetjük az uniform attachment modell esetében is. A (3.1) egyenlőséghez hasonlóan Uniform attachment fákra is definiálható a delta függvény:

$$\delta(S, T)_{UA} := \lim_{n \rightarrow \infty} \text{TV}(\text{UA}(n, S), \text{UA}(n, T)) \quad (3.2)$$

A 3.3 tétel bizonyításához hasonló gondolatmenettel lehet belátni az alábbi tételt, mely szerint a δ_{UA} függvény a metrika a legalább háromcsúcsú fák izomorfia osztályain.

3.4 Tétel (Bubeck, Eldan, Mossel, Rácz [9]). *Tetszőleges S és T nem-izomorf, legalább 3 csúcsú fákra $\delta_{UA}(S, T) > 0$.*

Megjegyezzük továbbá, hogy az uniform attachment modellben is teljesül a 3.2 tételhez hasonló állítás.

3.5 Tétel (Bubeck, Eldan, Mossel, Rácz [9]). *Jelölje ismét S_k a k csúcsú csillag gráfot. Ekkor tetszőleges T rögzített fára $\lim_{k \rightarrow \infty} \delta_{UA}(S_k, T) = 1$.*

A 3.3 és 3.4 tételek tükrében remélhetjük, hogy a gyökér, vagy legalábbis nagy n esetén a véletlen fa egy korai állapota, nagy valószínűséggel meghatározható. Még akkor is, ha a fejlődésről - a modell szabályától eltekintve - semmilyen információ nem áll rendelkezésünkre. Azt, hogy várapozásunk nem alaptalan, a következő fejezetben bemutatott konkrét algoritmusok támasztják alá.

4 Gyökérkereső algoritmusok

Ebben a fejezetben valamilyen növekedési szabály szerint fejlődő véletlen fát tekintünk, amelyről feltesszük, hogy egyetlen csúcsból fejlődött ki. Célunk csupán az n . lépés utáni gráf - irányítatlan, címkézetlen fa - ismeretében a gyökér felismerése nagy valószínűséggel. Világos, hogy ha egyetlen csúcsot várunk az algoritmustól, akkor nagy valószínűséget nem garantálhatunk, gondolva például az első két csúcs szimmetrikus szerepére.

Ezért inkább a következő értelemben vett algoritmusokat vizsgáljuk:

4.1 Definíció (Gyökérkereső algoritmus). Legyen $0 < \varepsilon < 1$ és $K \geq 1$. A H_K leképezést K költségű, ε hibatűrésű gyökérkereső algoritmusnak nevezzük a $\{T_f(n) : n \geq 2\}$ véletlen fák körében, ha

$$\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_K(T_f(n)^\circ)) \geq 1 - \varepsilon$$

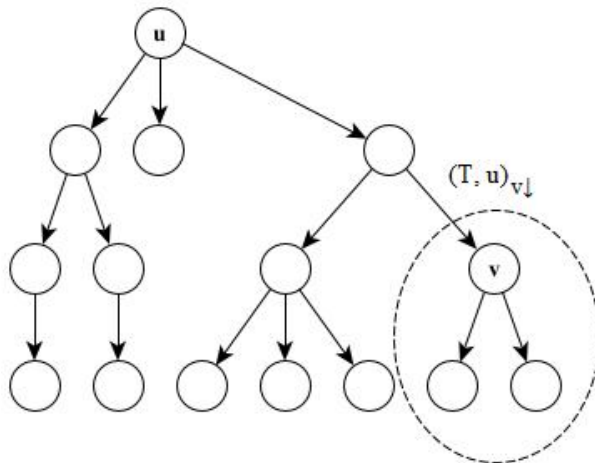
A fenti definícióban H_K egy adott fára K darab csúcsot ad vissza. Általában H_K valamilyen centralitási tulajdonság szerint választ a csúcsok közül. Emögött az a feltételezés áll, hogy a gyökér a gráf "középpontjában" van. A továbbiakban ilyen centralitási tulajdonságokat vezetünk be. Ezek egy részéről ismert, hogy a 4.1 definíció szerinti gyökérkereső algoritmushoz vezetnek.

4.1 Jordan-centralitás

4.2 Definíció. Fákra értelmezhetjük az alábbi $\psi_T : V(T) \rightarrow \mathbb{N}$ függvényt, melyet *Jordan-centralitás*nak nevezünk.

$$\psi_T(u) := \max_{v \in V(T) \setminus \{u\}} |(T, u)_{v\downarrow}|.$$

Azt a csúcsot, amelyik minimalizálja a fenti függvényt, *Jordan-centrum*nak hívjuk.



Ábra 4.4: Példa $(T, u)_{v\downarrow}$ részfára

Emlékeztetőül, $(T, u)_{v\downarrow}$ jelöli a v -ből és leszármazottaiból álló részfat (az u -t tekintve a fa gyökerének). A Jordan-centralitás szemléletesen a következőt jelenti: a T gráfból elhagyjuk az u csúcsot a belőle induló élekkel, a $\psi_T(u)$ a keletkező erdő legnagyobb komponensének méretét adja vissza.

Az elnevezés Camille Jordanra, a híres francia matematikusra utal, akit leginkább a Jordan-mérték, a Jordan-féle normálforma vagy a Jordan-féle görbetétel megalkotójaként tartunk számon. Jordan [10] cikkében adott élszámú gráfok és azok automorfizmusainak osztályozására tesz kísérletet. Vizsgálódásai során jut el a 4.2 definícióban bevezetett centralitási fogalomhoz, melyről be is látja az alábbi állítást.

4.3 Állítás (Jordan [10]). *Legyen az n csúcsú T fának a Jordan-centruma v^* , ekkor v^* -ot tekintve a fa gyökerének, minden részfára fennáll a következő egyenlőtlenség ($v \neq v^*$):*

$$|(T, v^*)_{v\downarrow}| \leq \frac{n}{2}.$$

Megfordítva, ha egy u csúcsra igaz, hogy minden $v \neq u$ esetén:

$$|(T, u)_{v\downarrow}| \leq \frac{n}{2},$$

akkor u Jordan-centrum. Továbbá egy fának vagy egyetlen Jordan-centruma van, vagy két szomszédos csúcs a Jordan-centrum.

Bizonyítás. Az állításban szereplő egyenlőtlenséget röviden úgy is megfogalmazhatjuk, hogy $\psi(v^*) \leq \frac{n}{2}$. Indirekt bizonyítunk: tegyük fel, hogy $\psi_T(v^*) > \frac{n}{2}$ és mégis v^* a Jordan-centrum.

Legyen v_1 a v^* azon szomszédja, amelynek részgráfja maximális, azaz amelyre

$$|(T, v^*)_{v_1\downarrow}| = \psi_T(v^*).$$

Legyenek továbbá a v_1 szomszédai rendre: v^*, u_1, \dots, u_k , a 4.5 ábrán látható módon. A v_1 -et tekintve gyökernek a megfelelő részgráfméreteket az alábbi módon jelöljük: $|(T, v_1)_{v^*\downarrow}| = r$, és $|(T, v_1)_{u_i\downarrow}| = r_i$ (ahol $i = 1, \dots, k$).

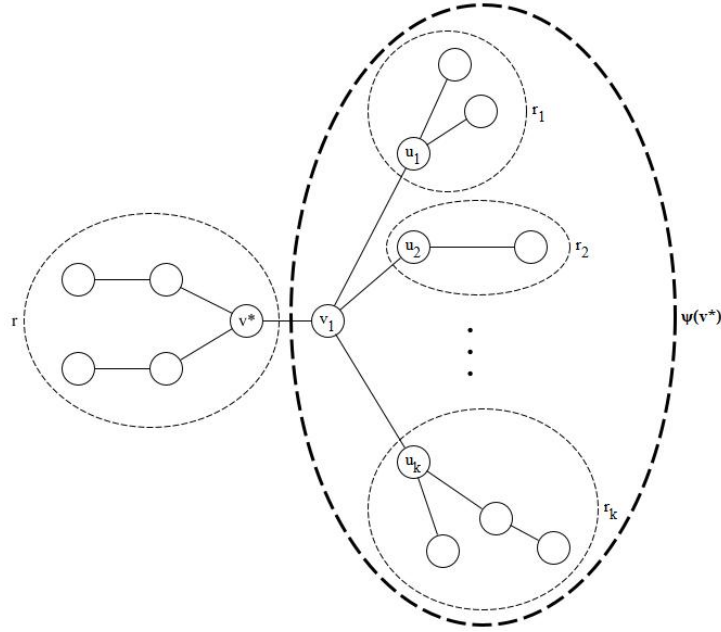
Ekkor fennáll, hogy $\psi_T(v^*) = 1 + \sum_{i=1}^k r_i$, továbbá $n = r + \psi_T(v^*)$. A második egyenlőségéből és indirekt feltevésünkből következik, hogy $r < \frac{n}{2} < \psi_T(v^*)$, míg az elsőből világos, hogy $r_i < \psi_T(v^*)$ minden i -re. Azaz $\psi_T(v_1) < \psi_T(v^*)$ ellentmondva feltevésünknek.

Az állítás második felének igazolásához tekintsük az u csúcsot, amelyre $\psi_T(u) \leq \frac{n}{2}$. Legyen továbbá w az a szomszédja u -nak, amelyen irányú részgráfja maximális, azaz amelyre $|(T, u)_{w\downarrow}| = \psi_T(u)$. Ekkor egy tetszőleges $w_1 \notin (T, u)_{w\downarrow} \cup \{u\}$ csúcsra

$$\psi_T(w_1) \geq |(T, w_1)_{u\downarrow}| > |(T, u)_{w\downarrow}| = \psi_T(u)$$

Továbbá egy tetszőleges $w_2 \in (T, u)_{w\downarrow}$ csúcsra

$$\psi_T(w_2) \geq |(T, w_2)_{u\downarrow}| \geq |(T, w)_{u\downarrow}| = n - |(T, u)_{w\downarrow}| \geq \frac{n}{2} \geq \psi_T(u).$$



Ábra 4.5: Illusztráció a 4.3 tétel bizonyításának első feléhez

Egyenlőség csak abban az esetben teljesülhet, ha $w_2 = w$ és $\psi_T(w) = \psi_T(u) = \frac{n}{2}$.

Ezzel az állítást beláttuk. ■

Most rátérünk annak az igazolására, hogy a Jordan-centralitás a 4.1 definíció szerinti gyökérkereső algoritmust szolgáltat. Jelöljük tehát H_ψ -vel azt a hozzárendelést, amelyik azt a K darab csúcsot adja meg, amelyekre a ψ értéke a legkisebb. Arról, hogy a legkisebb értékeket kell választani, hogy centralitási fogalomhoz jussunk, így győzhetjük meg magunkat: a $\psi_T(u)$ akkor lesz maximális, ha u egy levél volt, ilyenkor ugyanis $\psi_T(u) = n - 1$.

A következő tételek szerint H_ψ jó gyökérkereső algoritmus az uniform és a preferential attachment modellben is.

4.4 Tétel (Bubeck, Devroye, Lugosi [2]). *Legyen $K \geq 2.5 \frac{\log(1/\varepsilon)}{\varepsilon}$. Ekkor*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_\psi(UA(n)^\circ)) \geq 1 - \frac{4\varepsilon}{1 - \varepsilon}.$$

Bizonyítás. A bizonyítás során a csúcsokat időrendben számozzuk. A következőkben $T_{i,k}$ jelöli az i csúcsot tartalmazó komponenst abban a gráfban, amelyet az $UA(n)$ -ből az $\{1, 2, \dots, k\}$ csúcsok közötti élek elhagyásával kapunk. Tehát a $(|T_{1,k}|, \dots, |T_{k,k}|)$ vektor egy k -színű Pólya-urna szerint változik, azaz a 2.3 tétel szerint

$$\frac{1}{n} (|T_{1,k}|, \dots, |T_{k,k}|)$$

egy olyan Dirichlet-eloszláshoz tart eloszlásban, melynek paraméterei $(1, 1, \dots, 1)$

Azt az eseményt vizsgáljuk, amikor az algoritmus nem találja meg a kiinduló csúcsot, azaz amikor $1 \notin H_\psi$. Világos, hogy ilyenkor van egy K -nál későbbi csúcs, amelyikre

$\psi(i) \leq \psi(1)$, tehát ez utóbbi egy bővebb esemény. Ezt tovább bontva kapjuk a következő egyenlőtlenséget:

$$\mathbb{P}(1 \notin H_\psi) \leq \mathbb{P}(\exists i > K : \psi(i) \leq \psi(1)) \leq \mathbb{P}(\psi(1) \geq (1 - \varepsilon)n) + \mathbb{P}(\exists i > K : \psi(i) \leq (1 - \varepsilon)n)$$

A jobb oldalon levő tagokat külön-külön fogjuk megbecsülni.

A ψ definíciójából következik, hogy

$$\psi(1) \leq \max(|T_{1,2}|, |T_{2,2}|),$$

és mivel $|T_{1,2}|/n$ és $|T_{2,2}|/n$ azonos eloszlású, és $[0, 1]$ -en egyenletes valószínűségi változóhoz tart eloszlásban, ezért

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\psi(1) \geq (1 - \varepsilon)n) \leq 2 \lim_{n \rightarrow \infty} \mathbb{P}(|T_{1,2}| \geq (1 - \varepsilon)n) = 2\varepsilon$$

Másrészt minden $i > K$ -ra

$$\psi(i) \geq \min_{1 \leq k \leq K} \sum_{j=1, j \neq k}^K |T_{j,K}|.$$

Mivel ha az $i \in T_{k,K}$, azaz az első $K - 1$ él elhagyása után k és i azonos komponensben van, akkor a $(T, i)_{k \downarrow}$ magában foglalja az összes többi komponenst. Továbbá $\frac{1}{n} \min_{1 \leq k \leq K} \sum_{j=1, j \neq k}^K |T_{j,K}|$ eloszlásban tart a Béta($K - 1, 1$) eloszláshoz. Következésképpen

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbb{P}(\exists i > K : \psi(i) \leq (1 - \varepsilon)n) &\leq \lim_{n \rightarrow \infty} \mathbb{P}\left(\exists 1 \leq k \leq K : \sum_{j=1, j \neq k}^K |T_{j,K}| \leq (1 - \varepsilon)n\right) \leq \\ &\leq K(1 - \varepsilon)^{K-1}. \end{aligned}$$

Összegezve

$$\limsup_{n \rightarrow \infty} \mathbb{P}(1 \notin H_\psi) \leq 2\varepsilon + K(1 - \varepsilon)^{K-1}.$$

Ezzel az adott K választása mellett a tételt beláttuk. ■

4.5 Tétel (Bubeck, Devroye, Lugosi [2]). *Legyen $K \geq C \frac{\log^2(1/\varepsilon)}{\varepsilon^4}$ valamilyen $C > 0$ ismert konstansra. Ekkor*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_\psi(PA(n)^\circ)) \geq 1 - \varepsilon.$$

Bizonyítás. Az előző bizonyításhoz hasonlóan a csúcsokat időrendben számozzuk. A következőkben $T_{i,k}$ jelöli az i csúcsot tartalmazó komponenst abban a gráfban, amelyet a $PA(n)$ -ből az $\{1, 2, \dots, k\}$ csúcsok közötti élek elhagyásával kapunk. Tehát a $2(|T_{1,k}|, \dots, |T_{k,k}|)$ vektor egy olyan k -színű Pólya-urna szerint változik, melynek a viszszevetési mátrixa $2\mathbb{I}_k$ (ahol \mathbb{I}_k a $k \times k$ méretű identitásmátrix). A 2.3 tétel szerint ekkor

$$\frac{1}{n}(|T_{1,k}|, \dots, |T_{k,k}|) \xrightarrow{n \rightarrow \infty} \text{Dir}\left(\frac{d_{PA(k)}(1)}{2}, \frac{d_{PA(k)}(2)}{2}, \dots, \frac{d_{PA(k)}(k)}{2}\right).$$

Legyen $\eta \in (0, 1)$. Ekkor az előzőhöz hasonlóan

$$\mathbb{P}(1 \notin H_\psi) \leq \mathbb{P}(\exists i > K : \psi(i) \leq \psi(1)) \leq \mathbb{P}(\psi(1) \geq (1-\eta)n) + \mathbb{P}(\exists i > K : \psi(i) \leq (1-\eta)n).$$

Ismét felhasználva, hogy

$$\psi(1) \leq \max(|T_{1,2}|, |T_{2,2}|),$$

és mivel $|T_{1,2}|/n, |T_{2,2}|/n$ azonos eloszlású, és Béta(1/2, 1/2) eloszláshoz tartanak, ezért:

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\psi(1) \geq (1-\eta)n) \leq 2 \lim_{n \rightarrow \infty} \mathbb{P}(|T_{1,2}| \geq (1-\eta)n) = \frac{2}{\pi} \arcsin(\sqrt{\eta}) \leq \sqrt{\eta}.$$

Másrész minden $i > K$ -ra a fentiek szerint

$$\psi(i) \geq \min_{1 \leq k \leq K} \sum_{j=1, j \neq k}^K |T_{j,K}|.$$

A $\frac{1}{n} \sum_{j=1, j \neq k}^K |T_{j,K}|$ -et alulról becsli $\frac{1}{n} \sum_{j=2}^K |T_{j,K}|$, ez pedig a 2.3 tétel szerint eloszlásban tart az alábbihoz:

$$\text{Béta} \left(K-1 - \frac{d_{PA(K)}(1)}{2}, \frac{d_{PA(K)}(1)}{2} \right).$$

Tehát

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbb{P}(\exists i > K : \psi(i) \leq (1-\eta)n) &\leq \lim_{n \rightarrow \infty} \mathbb{P} \left(\exists 1 \leq k \leq K : \sum_{j=1, j \neq k}^K |T_{j,K}| \leq (1-\eta)n \right) \leq \\ &\leq K \mathbb{P} \left(\text{Béta} \left(K-1 - \frac{d_{PA(K)}(1)}{2}, \frac{d_{PA(K)}(1)}{2} \right) \leq 1-\eta \right), \end{aligned}$$

ahonnt a béta-eloszlás tulajdonságai és a 2.4 tétel felhasználásával már levezethető a tétel állítása. ■

Bubeck és szerzőtársai belátják továbbá az alábbi két tételt, amelyekben elég kis hibatűrés esetén tetszőleges algoritmus K költségét alulról becslik mindkét modellben.

4.6 Tétel (Bubeck, Devroye, Lugosi [2]). *Létezik olyan $\varepsilon_0 > 0$, hogy minden $\varepsilon \leq \varepsilon_0$ hibatűrés esetén tetszőleges az uniform attachment fa gyökerének megtalálására vonatkozó eljárás K költségére fennáll, hogy:*

$$K \geq \exp \left(\sqrt{\frac{1}{30} \log \frac{1}{2\varepsilon}} \right).$$

Hasonló állítást látnak be a preferential attachment modellben is:

4.7 Tétel (Bubeck, Devroye, Lugosi [2]). *Létezik olyan $c > 0$ konstans, hogy minden $\varepsilon \in (0, 1)$ hibatűrés esetén tetszőleges a preferential attachment fa gyökerének megtalálására vonatkozó eljárás K költségére fennáll, hogy*

$$K \geq \frac{c}{\varepsilon}.$$

A 2.1 definíció után mutatott további speciális esetekben is ismert, hogy a Jordan-centralitás jó gyökérkereső algoritmus. Banerjee és Bhamidi [3] az alábbi módon foglalja össze ezeket a Jordan-centralitáson alapuló algoritmus $K_\psi(\varepsilon)$ költségére vonatkozó ismereteket (C_1, C_2, C'_1 konstasokkal):

- Uniform attachment: $\frac{C'_1}{\varepsilon} \leq K_\psi(\varepsilon) \leq \frac{C_1}{\varepsilon} \exp\left(\sqrt{C_2 \log \frac{1}{\varepsilon}}\right)$
- Preferential attachment: $\frac{C'_1}{\varepsilon^2} \leq K_\psi(\varepsilon) \leq \frac{C_1}{\varepsilon^2} \exp\left(\sqrt{C_2 \log \frac{1}{\varepsilon}}\right)$
- Affin preferential attachment: $\frac{C'_1}{\varepsilon^{\frac{2+\beta}{1+\beta}}} \leq K_\psi(\varepsilon) \leq \frac{C_1}{\varepsilon^{\frac{2+\beta}{1+\beta}}} \exp\left(\sqrt{C_2 \log \frac{1}{\varepsilon}}\right)$
- Szublineáris preferential attachment: $\frac{C'_1}{\varepsilon} \leq K_\psi(\varepsilon) \leq \frac{C_1}{\varepsilon^2} \exp\left(\sqrt{C_2 \log \frac{1}{\varepsilon}}\right)$.

4.2 Rumor centralitás

A Jordan-centralitás mintájára tekinthetjük az alábbi összetettebb függvényt, melytől azt reméljük, hogy szintén jó gyökérkereső algoritmust szolgáltat.

4.8 Definíció. Fákra értelmezhetjük az alábbi $\varphi_T : V(T) \rightarrow \mathbb{N}$ függvényt, melyet *rumor centralitás*nak nevezünk.

$$\varphi_T(u) = \prod_{v \in V(T) \setminus \{u\}} |(T, u)_{v\downarrow}|$$

Azt a csúcsot, amelyik minimalizálja a fenti függvényt, *rumor centrum*nak hívjuk.

Megjegyezzük, hogy a rumor centralitást Shah és Zaman vezette be (lásd például: [1]) az ún. megengedett permutációk segítségével, melyek számáról belátták az alábbi összefüggést

$$R(u, T) = n! \prod_{v \in V(T)} \frac{1}{|(T, u)_{v\downarrow}|}.$$

Mi itt inkább a 4.8 definíciót követjük Bubeck, Devroye és Lugosi [2] cikke alapján. Ennek előnye, hogy nem függ a modelltől, tetszőleges fára értelmezett. Hátránya, hogy kevésbé általánosítható tetszőleges gráfra, de erre itt amúgy sem teszünk kísérletet.

A két felírás majdnem egymás reciproka, ugyanis T ismeretében $|T| = n$ is adott, és így $n!$ konstans (minden $u \in V(T)$ -re megegyezik). Továbbá az egyik produktumból hiányzó $|(T, u)_{u\downarrow}|$ éppen n -nel egyenlő, így

$$R(u, T) = (n-1)! \frac{1}{\varphi_T(u)}. \quad (4.3)$$

A rumor centralításra is teljesül a Jordan-centrum egyértelműségéről szóló 4.3 állítás megfelelője:

4.9 Állítás (Shah, Zaman [1]). *Legyen az n csúcsú T fának a rumor centruma v^* , ekkor v^* -ot tekintve a fa gyökerének, minden részfára fennáll a következő egyenlőtlenség:*

$$|(T, v^*)_{v\downarrow}| \leq \frac{n}{2}.$$

Megfordítva, ha egy u csúcsra igaz, hogy minden $v \neq u$ esetén

$$|(T, u)_{v\downarrow}| \leq \frac{n}{2},$$

akkor u rumor centrum. Továbbá egy fának legfeljebb 2 rumor centruma lehet.

Ez a Jordan-centralitásnál látott érveléshez hasonló módon, könnyen igazolható. A két állítást összevetve adódik az alábbi:

4.10 Következmény. *Az n csúcsú T fa Jordan-centruma és a rumor centruma megegyezik, továbbá ez egyetlen csúcsot vagy egy élszomszédos csúcspárt jelent.*

Felhívjuk a figyelmet, hogy a 4.10 következmény nem a két centralitási függvény azonosságát, csupán a minimumhelyük egybeesését állítja. Ezért nem érdektelen kérdés a rumor centralitás gyökerkeresési tulajdonságainak vizsgálata. Jelölje tehát a továbbiakban H_φ azt a hozzárendelést, amelyik egy adott gráfoz azt a K darab csúcsot mondja meg, amelyekre a φ értéke a legkisebb. Az (4.3) egyenlőség miatt ez megegyezik az R szerinti legnagyobb csúcsok kiválasztásával. Célunk megmutatni, hogy a rumor centralitás jó gyökerkereső algoritmushoz vezet a 4.1 definíció értelmében.

Mivel a Jordan-centralitáshoz képest olyan, finomabb képletet alkalmazunk, amelyik több információt tárol a gráf struktúrájáról, ezért azt várjuk, hogy a rumor centralitás K költsége kevesebb lesz ugyanakkora ε hibatűrés mellett. Ezt az intuíciónkat támasztja alá uniform attachment fák esetében az alábbi tétel.

4.11 Tétel (Bubeck, Devroye, Lugosi [2]). *Léteznek olyan $a, b > 0$ univerzális konstansok, hogy ha $K \geq a \exp\left(b \frac{\log(1/\varepsilon)}{\log \log(1/\varepsilon)}\right)$, akkor*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_\varphi(UA(n)^\circ)) \geq 1 - \varepsilon.$$

A tétel bizonyításához szükségünk lesz az alábbi technikai jellegű lemmákra:

4.12 Lemma (Bubeck, Devroye, Lugosi [2]). *Legyen $j_1, \dots, j_l \in \mathbb{N}$ és $s = \sum_{k=1}^l k j_k$, továbbá $k \in [l]$ -re legyen $X_k \sim \Gamma(j_k, k)$, ahol X_1, \dots, X_l függetlenek. Ekkor minden $t \in (0, s)$ -re*

$$\mathbb{P}\left(\sum_{k=1}^l X_k < t\right) \leq \exp\left(-\sqrt{\frac{s}{2}} \log\left(\frac{s}{et}\right)\right).$$

Bizonyítás. Ismert, hogy $\lambda \geq 0$ -ra:

$$\mathbb{E}_{X \sim \Gamma(a,b)} \exp(-\lambda X) = \frac{1}{(1 + \lambda b)^a}$$

Ebből a Chernoff-módszerrel a következőt kapjuk, kihasználva, hogy $x \mapsto \frac{\log(1+x)}{x}$ nem növekvő:

$$\begin{aligned} \mathbb{P}\left(\sum_{k=1}^l X_k < t\right) &\leq \exp(\lambda t) \mathbb{E} \exp\left(-\lambda \sum_{k=1}^l X_k\right) = \exp\left(\lambda t - \sum_{k=1}^l j_k \log(1 + \lambda k)\right) \leq \\ &\leq \exp\left(\lambda t - \sum_{k=1}^l \lambda k j_k \frac{\log(1 + \lambda l)}{\lambda l}\right) = \exp\left(\lambda t - \frac{s}{l} \log(1 + \lambda l)\right) \end{aligned}$$

Legyen $\lambda = \frac{s-1}{t} > 0$, ekkor:

$$\mathbb{P}\left(\sum_{k=1}^l X_k < t\right) \leq \exp\left(-\frac{s \log(\frac{s}{et})}{l}\right)$$

A bizonyítás befejezéséhez elég belátni, hogy $s \geq l^2/2$. ■

4.13 Lemma (Bubeck, Devroye, Lugosi [2]). *Legyen E_1, E_2, \dots független azonos eloszlású 1 paraméterű exponenciális valószínűségi változók sorozata, és legyen*

$$X = \prod_{i=1}^{\infty} \min\left(\sum_{k=1}^i E_k, 1\right).$$

Ekkor minden $t > 0$ esetén fennáll, hogy

$$\mathbb{P}(X \leq t) \leq 6t^{1/4}.$$

Ezek után térjünk rá a 4.11 tétel bizonyítására.

Bizonyítás. Négy lépésben fogunk bizonyítani. Az alábbi jelölést a bizonyítás erejéig vezetjük be: $T := \text{UA}(n)^\circ$. Az $S \geq 1$ értéket pedig a későbbiekben választjuk meg, tetszőlegesen.

1. lépés: Az $\text{UA}(n)$ fa csúcsaira a korábbiaktól eltérő módon hivatkozunk. Indukcióval vezessük be a következő számozását a csúcsoknak: legyen a gyökér jele a \emptyset , a $(j_1, \dots, j_l) \in \mathbb{N}^l$ -nel jelölt csúcs - ahol $l \in \mathbb{N}$ - a (j_1, \dots, j_l) csúcs j_l -edik leszármazottja (születési sorrendben). Tehát a csúcsokat nem természetes számokkal, hanem az $\mathbb{N}^* := \cup_{l=0}^{\infty} \mathbb{N}^l$ halmaz elemeivel címkézzük. Minden $v \in \mathbb{N}^*$ csúcsra legyen $l(v)$ az a szám, amelyre $v \in \mathbb{N}^{l(v)}$ (azaz legyen $l(v)$ a v csúcs mélysége a gráfban). Továbbá legyen

$$s(v) := \sum_{k=1}^{l(v)} (l(v) + 1 - k) j_k.$$

Figyeljük meg az alábbi fontos tulajdonságot: tetszőleges v csúcsra, amelyre $s(v) > 3S$ az alábbiak közül az egyik mindenképpen teljesül:

1. létezik olyan u csúcs, hogy $s(u) \in (S, 3S]$ és $v \in (T, \emptyset)_{u\downarrow}$
2. létezik olyan u csúcs, hogy $s(u) \leq S$ és $v \in (T, \emptyset)_{(u,j)\downarrow}$ valamilyen $j > S$ -re.

A fenti tulajdonságot mélység, azaz $l(v)$ szerinti indukcióval bizonyítjuk. $l(v) = 1$ esetén a második pont $u = \emptyset$ -tel valóban teljesül. Az $l(v) > 1$ esetben jelölje u a v csúcs őst. Ekkor három eset lehetséges:

1. Ha $s(u) > 3S$, akkor az indukciós feltevés alkalmazható u -ra.
2. Ha $s(u) \in (S, 3S]$, akkor az 1. teljesül.
3. Végül ha $s(u) \leq S$, akkor v az u csúcs j -edik leszármazottja valamely $j > S$ -re, mert $s((u, j)) \leq j + 2S$, és $s(v) > 3S$. Ekkor tehát 2. teljesül.

Ezzel beláttuk, hogy valóban 1. vagy 2. valamelyike mindig igaz. Figyeljük meg továbbá, hogy a φ -re teljesül, hogy tetszőleges w csúcsra a v és u csúcsokat összekötő úton: $\varphi(w) \leq \max(\varphi(u), \varphi(v))$. Ezekből már következik az alábbi egyenlőtlenség:

$$\begin{aligned} & \mathbb{P}(\exists v : s(v) > 3S \text{ és } \varphi(v) \leq \varphi(1)) \leq \\ & \leq \mathbb{P}(\exists v : s(v) \in (S, 3S] \text{ és } \varphi(v) \leq \varphi(1)) + \mathbb{P}(\exists v, j : s(v) < S, j > S \text{ és } \varphi((v, j)) \leq \varphi(1)). \end{aligned}$$

2. lépés: Az előző egyenlőtlenség jobb oldalán álló második tagot becsüljük felülről az első taggal a Boole-egyenlőtlen felhasználásával, azaz egyszerűen a valószínűségek összegével. Először is tegyük fel, hogy $v = (j_1, \dots, j_l)$ ekkor:

$$\varphi(v) \leq \varphi(1) \Leftrightarrow \prod_{i=1}^{l(v)} |(T, \emptyset)_{(j_1, \dots, j_i)\downarrow}| \geq \prod_{i=0}^{l(v)-1} |(T, v)_{(j_1, \dots, j_i)\downarrow}| = \prod_{i=1}^{l(v)} (n - |(T, \emptyset)_{(j_1, \dots, j_i)\downarrow}|).$$

Pontosabban ebből következik, hogy:

$$\begin{aligned} & \exists j > S : \varphi((v, j)) \leq \varphi(1) \implies \\ & \prod_{i=1}^{l(v)} |(T, \emptyset)_{(j_1, \dots, j_i)\downarrow}| \cdot \left(\sum_{j=S+1}^{\infty} |(T, \emptyset)_{(v, j)\downarrow}| \right) \geq \prod_{i=1}^{l(v)} (n - |(T, \emptyset)_{(j_1, \dots, j_i)\downarrow}|) \cdot \left(n - \sum_{j=S+1}^{\infty} |(T, \emptyset)_{(v, j)\downarrow}| \right) \end{aligned}$$

Nem nehéz belátni, hogy a $|(T, \emptyset)_{(v, S)\downarrow}|$ sztochasztikusan dominálja $\sum_{j=S+1}^{\infty} |(T, \emptyset)_{(v, j)\downarrow}|$ összeget. Ezek alapján látható, hogy:

$$\mathbb{P}(\exists j > S : \varphi((v, j)) \leq \varphi(1)) \leq \mathbb{P}(\varphi((v, S)) \leq \varphi(1)).$$

Ezt és a Boole-egyenlőtleniséget felhasználva kapjuk, hogy

$$\mathbb{P}(\exists v : s(v) > 3S \text{ és } \varphi(v) \leq \varphi(1)) \leq 2 \sum_{v: s(v) \in (S, 3S]} \mathbb{P}(\varphi((v, S)) \leq \varphi(1)). \quad (4.4)$$

3. lépés: Elérkeztünk a bizonyítás fő részéhez, amelyben belátjuk, hogy tetszőleges v csúcsra, melyre $s(v) \geq 10^{10}$:

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\varphi(v) \leq \varphi(1)) \leq 7 \exp\left(-0.21\sqrt{s(v)} \log(s(v))\right). \quad (4.5)$$

Megfigyelhetjük, hogy az $\left(\frac{1}{n}|(T, \emptyset)_{(j_1, \dots, j_i) \downarrow}|)\right)_{i=1, \dots, l(v)}$ véletlen vektor úgy változik, mint Pólya-urnák egy rendszere, így tehát eloszlásban tart az alábbi valószínűségi változóhoz: $\left(\prod_{k=1}^i U_{j_k, k}\right)_{i=1, \dots, l(v)}$, ahol $U_{j,1}, U_{j,2}, \dots$ mindegyike j darab független $[0, 1]$ -en egyenletes eloszlású valószínűségi változó szorzata. Ebből következik, hogy

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \mathbb{P}(\varphi(v) \leq \varphi(1)) = \\ &= \mathbb{P}\left(\prod_{i=1}^{l(v)} \prod_{k=1}^i U_{j_k, k} \geq \prod_{i=1}^{l(v)} \left(1 - \prod_{k=1}^i U_{j_k, k}\right)\right) = \mathbb{P}\left(\prod_{i=1}^{l(v)} U_{j_i, i}^{l(v)+1-i} \geq \prod_{i=1}^{l(v)} \left(1 - \prod_{k=1}^i U_{j_k, k}\right)\right) \leq \\ &\leq \mathbb{P}\left(\prod_{i=1}^{l(v)} U_{j_i, i}^{l(v)+1-i} \geq \exp\left(-\frac{s(v)}{R}\right)\right) + \mathbb{P}\left(\prod_{i=1}^{l(v)} \left(1 - \prod_{k=1}^i U_{j_k, k}\right) \leq \exp\left(-\frac{s(v)}{R}\right)\right). \end{aligned}$$

Ahol $R > 0$ megfelelően választott szám értékét később határozzuk meg. A 4.12 lemmát alkalmazva:

$$\mathbb{P}\left(\prod_{i=1}^{l(v)} U_{j_i, i}^{l(v)+1-i} \geq \exp\left(-\frac{s(v)}{R}\right)\right) \leq \exp\left(-\sqrt{\frac{s(v)}{2}} \log\left(\frac{R}{e}\right)\right).$$

Továbbá, felhasználva, hogy $1 - \exp(-x) \geq \frac{1}{2} \min(x, 1)$ minden nemnegatív x -re

$$\prod_{i=1}^{l(v)} \left(1 - \prod_{k=1}^i U_{j_k, k}\right) \geq \frac{1}{2^{l(v)}} \prod_{i=1}^{l(v)} \min\left(\sum_{k=1}^i \log(1/U_{j_k, k}), 1\right) \geq \frac{1}{2^{l(v)}} X.$$

Ahol X eloszlásban egyenlő a $\prod_{i=1}^{\infty} \min\left(\sum_{k=1}^i \log(1/U_{1,k}), 1\right)$ szorzattal. Ebből a 4.13 lemmát alkalmazva:

$$\mathbb{P}\left(\prod_{i=1}^{l(v)} \left(1 - \prod_{k=1}^i U_{j_k, k}\right) \leq \exp\left(-\frac{s(v)}{R}\right)\right) \leq 6 \cdot 2^{\frac{l(v)}{4}} \exp\left(-\frac{s(v)}{4R}\right) \leq 6 \exp\left(\frac{\sqrt{s(v)}}{4} - \frac{s(v)}{4R}\right)$$

Ahol felhasználtuk, hogy $s(v) \geq l(v)/2$ a második egyenlőtlenségben. Ebből R -et $e \cdot s(v)^{0.3}$ -nak választva következik a kívánt egyenlőtlenség.

4. lépés: A (4.4) és (4.5) egyenlőtlenségeket felhasználva, és feltéve, hogy $S \geq 10^{10}$:

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\exists v : s(v) > 3S \text{ és } \varphi(v) \leq \varphi(1)) \leq 14 \cdot |\{v : s(v) \leq 3S\}| \cdot \exp\left(-0.21\sqrt{S} \log(S)\right)$$

A Hardy-Ramanujan formula Erdős-féle nemaszimptotikus verziója [11] a következő egyenlőtlenség:

$$\left| \left\{ (j_1, \dots, j_l) \in \mathbb{N}^l : l \in \mathbb{N}, \sum_{k=1}^l k j_k = s \right\} \right| \leq \exp \left(\pi \sqrt{\frac{2}{3} s} \right).$$

Ez alapján: $|\{v : s(v) \leq 3S\}| \leq 3S \exp(\pi\sqrt{2S})$. És így $S \geq 10^{10}$ esetén

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\exists v : s(v) > 3S \text{ és } \varphi(v) \leq \varphi(1)) \leq \exp \left(-\frac{1}{100} \sqrt{S} \log(S) \right).$$

Ahonnét már következik a tétel állítása. ■

4.3 Távolságcentralitás

A Jordan-centralitásról és a rumor centralitásról is kiderült, hogy jó gyökérkereső algoritmust szolgáltatnak, ily módon meggyőződünk centralitási tulajdonságukról. Ennek ellenére lehet némi hiányérzetünk, hiszen amikor a "középső" csúcsra gondolunk, akkor a fentieknél természetesebb definíciókat várunk. Egyik ilyen lehetőség a lentebb definiált távolságcentralitás. Annak oka, hogy mégsem ennek vizsgálatával kezdtük az, hogy a távolságcentralításra a fent belátott tételekhez hasonló eredmény nem ismert.

4.14 Definíció (Távolságcentralitás). Egy tetszőleges $G(V, E)$ gráfra a $v \in V(G)$ csúcs távolságcentralitása az alábbi $D(v, G)$ érték:

$$D(v, G) = \sum_{j \in V(G)} d(v, j)$$

ahol $d(v, j)$ a v és j csúcsokat összekötő legrövidebb út hossza. Azt az $u \in V(G)$ csúcsot, amelyikre a távolságcentralitás minimális, távolságcentrumnak nevezzük - amennyiben több ilyen van, az összeset.

A 4.14 definíció azt az intuíciónkat tükrözi, hogy a "központ" az a csúcs, amelyik a többi csúcshoz a legközelebb van.

A távolságcentralításra teljesül az alábbi

4.15 Állítás (Shah, Zaman [1]). *Legyen az n csúcsú T fa távolságcentruma a v_D csúcs, ekkor minden $v \neq v_D$ csúcsra:*

$$|(T, v_D)_{v\downarrow}| \leq \frac{n}{2}$$

A 4.10 következményt és a 4.15 állítást összevetve adódik az alábbi:

4.16 Következmény. *Az n csúcsú T fa Jordan-, rumor és távolságcentruma megegyezik, továbbá ez egyetlen csúcsot vagy egy élszomszédos csúcspárt jelent.*

Ahogy azt már korábban is megjegyeztük, a 4.16 következmény nem állítja a három centralitási tulajdonság, csupán minimumhelyeik egybeesését. Ezek alapján érdeklődésre adhat okot ezen centralitási tulajdonságok hatékonyságának összehasonlítása. A következő szakaszban szimuláció segítségével erre teszünk kísérletet.

4.4 Összehasonlítás

Eddig az algoritmusokat a 4.1 definíció alapján K költségük és ε hibatűrésük alapján vizsgáltuk. Gyakorlati szempontból azonban fontos megvizsgálni az algoritmusok gyorsaságát, azaz a H_K kiszámíthatóságát. Világos, hogy ehhez a Jordan- és a rumor centralitást akarjuk minél gyorsabban meghatározni.

A Jordan-centralitásról könnyen meggondolható, hogy $O(n)$ időben számolható. Ennél meglepőbb talán, hogy a rumor centralitás is lineáris időben kiszámítható. Ez az alábbi észrevételen múlik:

4.17 Állítás (Shah, Zaman [1]). *Ha u és v szomszédos csúcsok az n csúcsú T fában, akkor*

$$R(u, T) = R(v, T) \cdot \frac{|(T, v)_{u\downarrow}|}{n - |(T, v)_{u\downarrow}|}.$$

Avagy

$$\varphi_T(v) = \varphi_T(u) \cdot \frac{|(T, v)_{u\downarrow}|}{n - |(T, v)_{u\downarrow}|}.$$

Bizonyítás. Az $\varphi_T(u)$ és $\varphi_T(v)$ rumor centralitást definiáló szorzatok majdnem minden tényezőben megegyeznek, kivéve az u és v csúcsokhoz tartozókban. Ezekre viszont fennáll, hogy: $|(T, u)_{v\downarrow}| = n - |(T, v)_{u\downarrow}|$, hiszen a $(T, u)_{v\downarrow}$ és a $(T, v)_{u\downarrow}$ két diszjunkt részfa, amelyek az összes csúcsot tartalmazzák. Az állítás első fele a korábban meggondolt (4.3) egyenlőségből következik. ■

Ez alapján Shah és Zaman [1] megadott egy ún. üzenetküldő algoritmust, melynek lényege a következő: először kiválasztunk egy tetszőleges $v \in V(T)$ csúcsot, ezt tekintjük gyökérnek. Ezután kiszámítjuk az összes részfa méretét: $|(T, v)_{u\downarrow}|$, így megkapjuk $R(v, T)$ -ét. Ezt megtehetjük úgy, hogy az összes u csúcs két üzenetet küld az őségnek. Az első üzenet a $|(T, v)_{u\downarrow}|$ részfaméret, amelyet $t_{u\uparrow}$ -nek nevezünk. A második üzenet az u részfájában, azaz $(T, v)_{u\downarrow}$ -ban található csúcsokhoz tartozó részfaméret szorzata, ezt $p_{u\uparrow}$ -vel jelöljük. Ezután az ős csúcs összeadja utódainak $t_{u\uparrow}$ üzeneteit és hozzáad egyet, így megkapja saját részfájának méretét. Majd ezt a gyermekeitől kapott második üzenetek - $p_{u\uparrow}$ értékek - szorzatával megszorozva kapja a saját második üzenetét. Ezeket az üzeneteket aztán felfelé - azaz a v csúcs felé - továbbküldi. Ezek alapján a v csúcs már ki tudja számolni a saját rumor centralitását.

Az $R(v, T)$ ismeretében a 4.17 állítás alapján a v utódainak rumor centralitása már kiszámítható. Minden csúcs üzenetet küld az utódainak saját rumor centralitásáról, ezt $r_{u\downarrow}$ -vel jelöljük. Ezután minden csúcs kiszámítja a saját rumor centralitását az ősetől kapott üzenet és a saját részfaméretének ismeretében.

Az előző szakaszokban beláttuk, hogy a Jordan és a rumor centralitás is jó algoritmus, továbbá a távolságcentralitásról is kiderült, hogy a ezekhez szorosan kötődik. Preferen-

Algorithm 1 Üzenetküldő algoritmus - a rumor centralitás kiszámítására

```
1: Legyen a gyökér  $v \in V(T)$  tetszőleges.
2: for  $u \in V(T)$  do
3:   if  $u$  levél then
4:      $t_{u\uparrow} = 1$ 
5:      $p_{u\uparrow} = 1$ 
6:   else
7:     if  $u = v$  then
8:       
$$r_{v\downarrow} = \frac{n!}{n \prod_{j \in \text{utód}(v)} p_{j\uparrow}}$$

9:     else
10:      
$$t_{u\uparrow} = \sum_{j \in \text{utód}(u)} t_{j\uparrow} + 1$$

11:      
$$p_{u\uparrow} = t_{u\uparrow} \cdot \prod_{j \in \text{utód}(u)} p_{j\uparrow}$$

12:      
$$r_{u\downarrow} = r_{\text{ős}(u)\downarrow} \cdot \frac{t_{u\uparrow}}{n - t_{u\uparrow}}$$

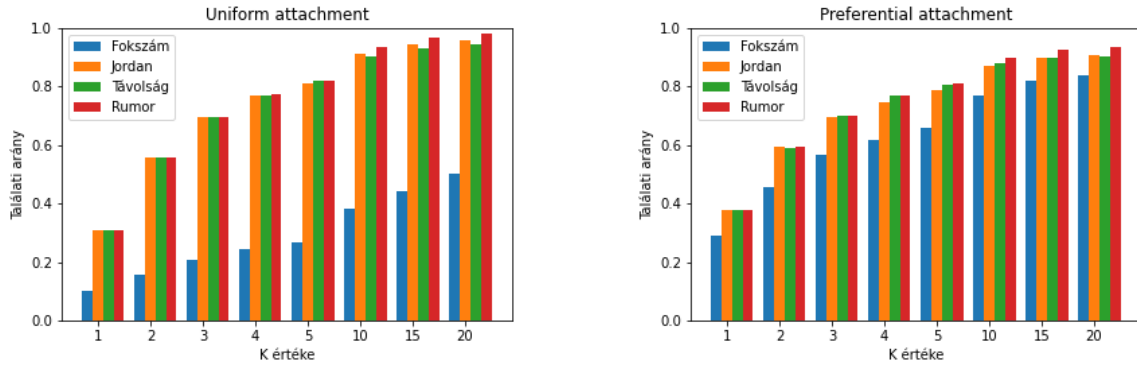
13:    end if
14:  end if
15: end for
```

tial attachment modell esetén érdeklődésre adhat okot a legnagyobb fokú csúcs is, hiszen ekkor érvényesül a "gazdag még gazdagabb lesz" elv, vagyis várhatóan a gyökérnek nagy fokszáma lesz. Vizsgáljuk meg most a különböző centralitási tulajdonságokat a gyakorlatban.

Szimulációnk során 1000-1000 fát generáltunk véletlenszerűen uniform és preferential attachment modell szerint, mindkét esetben 5000 csúcsú fákat vizsgáltunk. A kapott eredményeket a 4.6 ábra szemlélteti. A vízszintes tengelyen a K értékeket, míg a függőlegesen a találati arányt tüntettük fel. Tehát például a $K = 5$ -höz tartozó 0,817 érték az Uniform attachment modellben rumor centralitás esetén azt jelenti, hogy a valódi vírusforrás az esetek 81,7%-ában az 5 leginkább centrális csúcs között volt a rumor centralitás szerint.

Várakozásunknak megfelelően az uniform attachment modellben a fokszám nem garantál magas találati valószínűséget, ennél meglepőbb, hogy a preferential attachment modellben is jelentősen elmarad a másik három centralitási tulajdonságtól. A $K = 1$ -hez tartozó értékek alapján látszik, hogy a biztos találat valószínűsége egyik esetben sem túl magas, de a preferential attachment modellben valamivel nagyobb. Meggyőződhetünk továbbá a 4.16 következményről is: a $K = 1$ -hez tartozó értékek a Jordan-, rumor és a távolságcentralitás esetén megegyeznek, nagyobb K -ra azonban a rumor centralitás valamivel hatékonyabb.

Összességében azt mondhatjuk, hogy kis K értékekre mindhárom centralitási tula-



Ábra 4.6: Különböző centralitási algoritmusok hatékonyságának összehasonlítása.

Mindkét diagram 1000-1000 darab 5000 csúcús fa adatait tartalmazza.

jdonság hasonló találási arányt biztosít. Nagyobb K értékek esetén viszont a rumor centralitás valamivel hatékonyabb. Az előbbi megfigyelés talán kissé ellentmond annak az intuíciónknak, hogy a rumor centralitás egy összetettebb, a gráf struktúrájáról több információt hordozó fogalom, mint a Jordan-centralitás.

A két modellt összevetve azt láthatjuk, hogy a preferential attachment modell esetén kis K -ákra a találási arány nagyobb, mint az uniform attachment modellben, azaz például a pontos detektálás valószínűsége nagyobb. Viszont nagyobb K értékekre a preferential attachment modellben a találási arány lassabban nő, így kis ε hibatűrés esetén az uniform attachment modellben várunk kisebb K szükséges költséget.

Az előbbi megfigyelés megegyezik várakozásunkkal, hiszen a preferential attachment modell esetén az az intuíciónk, hogy a vírusforrás korai megjelenése és a gráf fejlődési szabálya miatt kitüntetettebb szerephez jut, így a pontos felismerésben hatékonyabb. Míg az utóbbi megfigyelés a 4.4 és 4.5 tételekben látott korlátokkal is összhangba hozható.

5 A gyökér elrejtése

Az előző fejezetben különféle módjait ismertettük a gyökér detektálásának. Figyeljük meg, hogy ehhez elég volt a folyamat egy későbbi állapotának ismerete, a fejlődésről szinte semmilyen információnk nem volt.

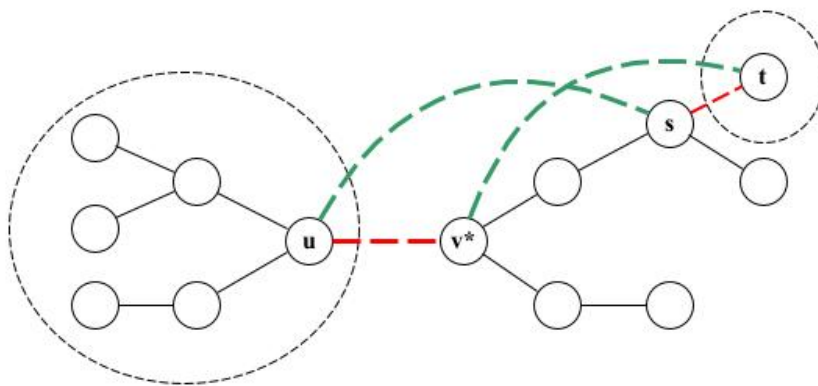
Gyakorlati szempontból azonban sok esetben pontosan a detektálhatóság elkerülése a cél. Gondolva például a különféle népszerű mikroblog-szolgáltatókra, melyek kapcsán gyakran felmerülnek adatvédelmi problémák. Az információ terjesztője részéről természetes igényként jelentkezik az anonimitás. A fentiek alapján azonban kiderült, hogy ehhez általában nem elég az információforrás kilétének titokban tartása.

Ez motiválja a következő kérdést: megadható-e olyan terjedési modell, amelyben a gyökér detektálása nem lehetséges. Ahogy azt a 6.3 fejezetben látni fogjuk, van olyan modell, amelynek ismert megfelelő módosítása a gyökér elrejtésére. Ebben a fejezetben a gyökér elrejtését a 2.1 definícióban bevezetett véletlen fák esetében vizsgáljuk. Megadjuk a preferential attachment modell egy módosított változatát, amelyet a 4.4 fejezetben látott szimulációval hasonlítunk össze, így megtéve az első lépést a kérdés megválaszolása felé.

A következőkben feltesszük tehát, hogy egy ellenség elől akarjuk titokban tartani az információforrás kilétét. Azt is feltesszük továbbá, hogy az ellenség nem ismeri ezt a szándékunkat, így azt feltételezi, hogy hagyományos preferential attachment szerint fejlődik a gráf.

Célunk a 4. fejezetben bevezetett Jordan- és rumor centralitás elrontása lesz. Ezt a következő módosítással próbáljuk elérni: a véletlen fa fejlődésének egy korai, de nem túl korai szakaszában kissé megváltoztatjuk a gráf struktúráját.

A véletlen növe fát az n -edik állapotában módosítjuk, azaz a $PA(n)$ fából egy $PA(n)'$ fát készítünk, majd ebből az új fából folytatjuk a preferential attachment növekedést. A módosítást az alábbi definícióban bevezetett rontó csere jelenti:



Ábra 5.7: A 5.1 definícióban bevezetett rontó csere: a piros éleket a zöldekre cserélve növeljük a v^* Jordan-centralitását.

5.1 Definíció (Rontó csere). Tekintsük tehát a $T = PA(n)$ fát, melynek v^* a gyökere. Ebben két élel fogunk másik kettővel helyettesíteni az alábbiak szerint. A v^* -nak legyen u azon szomszédja, amelyre $|(T, v^*)_{u\downarrow}|$ a legnagyobb. Legyen S az a gráf amelyet a T -ből a v^* csúcs, a belőle induló élek és a $(T, v^*)_{u\downarrow}$ részgráf elhagyásával kapunk. Az S legnagyobb fokú csúcsa legyen s , továbbá t az a szomszédja s -nek amelyre $(T, s)_{t\downarrow}$ nem tartalmazza a v^* csúcsot és minimális méretű. Ezután "cseréljük ki" a $(T, v^*)_{u\downarrow}$ és a $(T, s)_{t\downarrow}$ részgráfokat, azaz töröljük T -ből a $\{v^*, u\}$ és $\{s, t\}$ éleket, és adjuk hozzá a $\{v^*, t\}$ és $\{s, u\}$ éleket. Ezt az eljárást az n . lépésben vett *rontó cserének* nevezzük. (Az u , s és t csúcsok kiválasztása esetén is tetszőlegesen döntünk a döntetlenek esetén.)

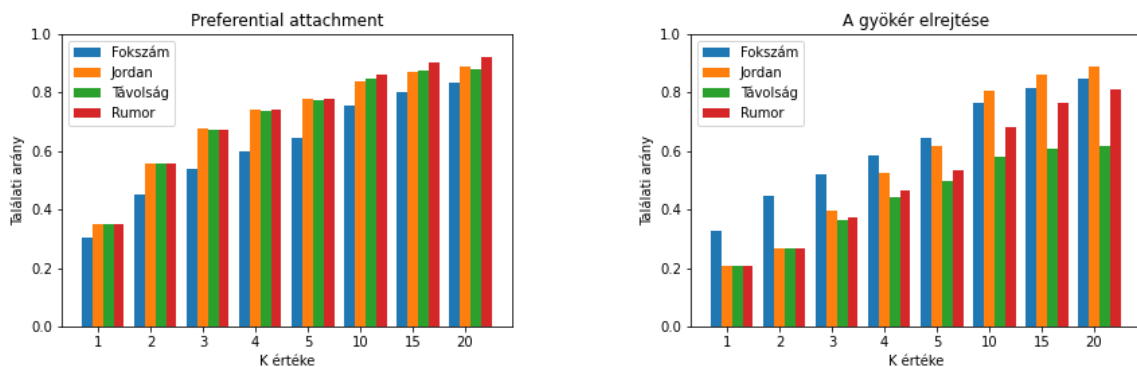
Vegyük észre, hogy a rontó csere növelte a v^* csúcs Jordan-centralitását, azaz a

$$\psi(v^*) = \max_{u \in V(T) \setminus \{v^*\}} |(T, v^*)_{u\downarrow}|$$

értéket, hiszen a legnagyobb részfát áthelyeztük egy másik részfára, így annak mérete nagyobb lett az eredeti maximumnál. Tehát valóban azt várhatjuk tőle, hogy a H_ψ algoritmus hatékonyságát rontani fogja.

Felhívjuk a figyelmet továbbá arra, hogy a rontó csere nem változtatta meg a T fa $\vec{d}(T)$ fokszámprofilját, egyik csúcs fokszáma sem változott. Reméljük tehát, hogy a növekedést nem befolyásoltuk túlságosan, bár a 3.3 tétel szerint - amennyiben nem az eredetivel izomorf gráfot kaptunk - fennáll, hogy $\delta(T, T') > 0$.

Vizsgáljuk meg szimulációval, hogy milyen eredményre vezet egy rontó csere a gyökér elrejtésében. Az 5.8 ábrán látható diagramok mindegyike 2000 csúcsú véletlen gráfokra vonatkozó 1000-1000 teszteredményt összesít. A második esetben azonban $n = 100$ -nál rontó cserét hajtottunk végre, amennyiben erre lehetőség volt.



Ábra 5.8: A hagyományos és a módosított modell összehasonlítása 2000 csúcsú gráfokon, ahol a második esetben rontó cserét hajtottunk végre a 100. lépésben. Mindkét diagramban 1000 méretű mintát összesítettünk.

Megjegyezzük, hogy az esetek nagyjából 85%-ában végrehajtható volt a rontó csere, tehát mind a négy, definícióban szereplő csúcs jóldefiniált volt. Azokban az esetekben,

amikor nem történt csere, a 5.1 definícióban szereplő s vagy t csúcs nincs definiálva. Előbbi akkor fordulhat elő, amikor a v^* egy levél, míg utóbbi abban az esetben, amikor v^* minden u -tól különböző szomszédja levél. Ezekben az esetekben a gráf normál növekedése mellett sem várunk magas találati arányt, ezt a várakozásunkat a szimuláció eredménye alátámasztja. Világos továbbá, hogy minél korábban próbálunk cserélni, várhatóan annál kevesebbszer járunk sikerrel.

Az eredmények azt mutatják, hogy a pontos detektálás jelentősen lecsökkent a rontó csere hatására. Meglepő eredmény azonban, hogy $K = 15$ -nél nagyobb értékekre a Jordan-centralitás hatékonysága nem csökkent. A távolság és a rumor centralitás azonban ezekre az értékekre is elmarad a normál esetben megfigyelt hatékonyságától. A foksám centralitás találati aránya egyik K értékre sem mutat lényeges eltérést, ez megfelel a várakozásunknak, hiszen ez csupán a foksám profiltól függ, és azt nem módosítottuk a rontó csere során.

A fent bemutatott rontó csere tehát érdemes lehet további vizsgálódásra. Az anonim információterjesztéssel kapcsolatban számos további kérdés felmerül, ezekből említünk meg most a teljesség igénye nélkül néhányat.

A rontó csere bevezetése előtt megjegyeztük, hogy az ellenség tudása korlátozott, valójában arról sincs információja, hogy módosítottuk a növekedést. Kérdésként merül fel tehát, hogy az ellenségnek milyen információt adunk, Fanti és szerzőtársai [12] cikkükben három lehetőséget írnak le a diffúziós modellben. Mi itt csak annyit említünk meg, hogy a növekedési szabály, beleértve az esetleges módosítások lehetőségének ismeretét, a vírusgráf, sőt esetleg annak részleges irányítása mind olyan információ, amellyel az ellenség rendelkezhet.

Legáltalánosabb célunk tehát olyan információterjedési modell megadása lehet, amelyben az információforrás megtalálása még az információátadásról való részleges ismeretek esetén sem lehetséges.

6 Terjedés adott hálózaton

Ebben a fejezetben a vírusterjedés egy másik lehetséges modelljét vizsgáljuk meg alaposabban. Bár ez az előző fejezetekben tárgyalt preferential és uniform attachment modellektől lényeges eltéréseket mutat, a korábban kidolgozott eszköztár mégis jól alkalmazható lesz ebben az esetben is.

A következőkben a Shah és Zaman [1] által bevezetett modelljét írjuk le. Csúcsok egy hálózatát a $G(V, E)$ gráffal modellezzük, ahol V csúcsok egy megszámlálhatóan végtelen halmaza, míg E az élek halmaza: $(i, j) \in E$ alakban, ahol $i, j \in V$.

A korábbiakhoz hasonlóan a sokat vizsgált SIR modellnek azt a változatát használjuk, amikor nincs gyógyulás: azaz az SI (*susceptible-infected*) modellt. Tehát ha egy csúcs valamikor megfertőződött, akkor fertőzött is marad. Azt az esetet tekintjük, amikor egyetlen v^* csúcsból indul a fertőzés.

Ha egy $i \in V$ csúcs már fertőzött, akkor megfertőzheti a még nem fertőzött $j \in V$ csúcsot, de csakis akkor, ha a két csúcsot él köti össze, azaz $(i, j) \in E$. Azt az időt, ami alatt az i csúcs megfertőzi a j csúcsot egy λ paraméterű exponenciális valószínűségi változóval modellezzük és $\tau_{i,j}$ -vel jelöljük. Itt az éleket ugyanolyan erősségűnek tekintjük, így az idő skálázásával feltehető, hogy $\lambda = 1$, tehát: $\tau_{i,j} \sim \exp(1)$. Feltesszük továbbá, hogy minden $\tau_{i,j}$ független és azonos elszlású.

A fent bevezetett modellre az alábbi módon is gondolhatunk, mint ahogy arra Shah és Zaman [13] felhívta a figyelmet. A továbbiakban inkább ezt tekintjük, a 2.1 definícióval mutatott hasonlósága miatt.

6.1 Definíció (Terjedés adott hálózaton). Az adott $G(V, E)$ hálózaton való vírusterjedést részfák egy $\{T(j)\}_{j=1}^{\infty}$ sorozatával, ahol a $T(n)$ fa csúcsait $V(T(n)) = \{v_1, \dots, v_n\}$ jelöli érkezési sorrendben. Az $n + 1$ -edik csúcsot a korábbi csúcsok szomszédai közül választjuk ki egyenletesen.

A célunk továbbra is a vírusterjedés $v^* = v_1$ forrásának megtalálása. Ez általános $G(V, E)$ hálózat esetén nehéz feladatnak ígérkezik, ezért az alábbi egyszerűsítéssel élünk: azt az esetet vizsgáljuk, amikor a szóbanforgó hálózat egy reguláris fa, ahol minden csúcs fokszáma megegyezik. A d -reguláris fát a továbbiakban G_d -vel jelöljük.

Mint azt látni fogjuk, a 4. fejezetben bevezetett centralitási tulajdonságok ebben a modellben is hatékony - a 4.1 definíció szerinti - gyökérkeresési algoritmusnak bizonyulnak. Sőt bizonyos speciális esetekre a korrekt detektálás valószínűségéről is tudunk valamilyen pozitív eredményt állítani.

6.1 Jordan-centralitás

Legyen tehát G_d egy d -reguláris fa, amelyen a $v^* = v_1 = T(1)$ csúcsból kiindulva a 6.1 definícióban értelmezett rekurzióval meghatározott $\{T(n)\}_{n=1}^\infty$ gráfsorozat modellezi a terjedést.

Az alábbi tétel szerint a 4.2 definícióban bevezett,

$$\psi_T(u) = \max_{v \in V(T) \setminus \{u\}} |(T, u)_{v\downarrow}|$$

képlettel definiált Jordan-centralitás ebben a modellben is jó gyökerkereső algoritmus.

6.2 Tétel (Khim, Loh [14]). *Tegyük fel, hogy $d \geq 3$. Ha ε elegendően kicsi és $K \geq \frac{C}{\varepsilon}$, akkor*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_\psi(T(n)^\circ)) \geq 1 - \varepsilon,$$

ahol $C > 0$ egy csak a d -től - azaz ε -tól nem - függő konstans.

Ennek igazolásához szükséges az alábbi eredmény.

6.3 Tétel (Khim, Loh [14]). *A G_d $d \geq 3$ -reguláris fán értelmezett $\{T(n)\}_{n=1}^\infty$ folyamatra tetszőleges $\eta \in (0, 1)$ és $K > 3$ esetén*

$$\limsup_{n \rightarrow \infty} \mathbb{P}(1 \notin H_\psi(T(n)^\circ)) \leq C_1 \eta^{1 + \frac{1}{d-2}} + C_2 K^{2 + \frac{1}{d-2}} (1 - \eta)^{K-1 + \frac{1}{d-2}}.$$

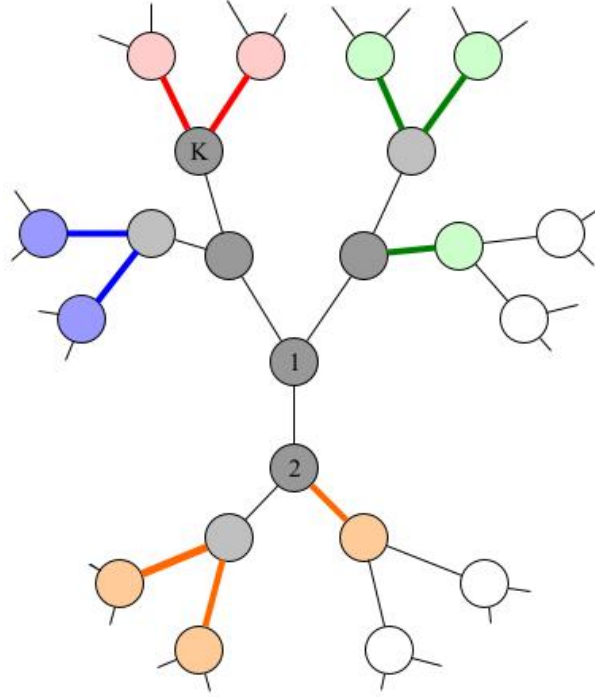
Érdeemes összevetni az eredményeket a 4. fejezetben igazoltakkal, gondolva itt 4.4 és a 4.5 tételekre. Ezek alapján feltűnhet, hogy ha a $d \rightarrow \infty$, akkor a terjedési modell egyre jobban hasonlít az uniform attachmenthoz. Szemléletesen a nagy fokú reguláris fák úgy viselkednek, mint az uniform attachment modell.

A 6.3 tétel bizonyítása lényegében a 4.4 tételben látott Pólya-urnákra épülő bizonyítást követi. Ennek részletezésétől itt most eltekintünk, csupán az alábbi kulcsfontosságú megfigyelést igazoljuk, amely mutatja a Pólya-urnákkal való szoros kapcsolatot.

6.4 Lemma (Khim, Loh [14]). *Az első K csúcs közötti él kihagyásával kapott részfák méreteinek a teljes gráf méretéhez vett aránya eloszlásban Dirichlet-eloszláshoz tart:*

$$\frac{1}{n} (|T_{1,K}|, \dots, |T_{K,K}|) \xrightarrow{d} \text{Dirichlet} \left(\frac{d_{G_d}(1) - d_{T_K}(1)}{d-2}, \dots, \frac{d_{G_d}(K) - d_{T_K}(K)}{d-2} \right)$$

Bizonyítás. Emlékeztetőül a csúcsokat időrendben számozzuk, továbbá $T_{i,K}^n$ jelöli az i csúcsot tartalmazó komponenst abban a gráfban, amelyet az $T(n)$ -ből az $\{1, 2, \dots, K\}$ csúcsok közötti élek elhagyásával kapunk. Jelölje $E_{i,K}^n$ a $T_{i,K}^n$ és a $G_d \setminus T(n)$ között futó élek halmazát, vagy ezzel ekvivalensen azon csúcsok halmazát, amelyeket az adott részfa megfertőzhet.



Ábra 6.9: Illusztráció a 6.4 lemma bizonyításához. A $T(K)$ részfa csúcsait sötétszürkével, míg a további fertőzött csúcsokat világosszürkével jelöltük. Az $E_{i,K}^n$ élhalmazokat a színosztályok jelentik.

Figyeljük meg, hogy ezeknek a halmazoknak a méretei egy Pólya-urna szerint fejlődnek, amelynek visszatevési mátrixa $(d-2)\mathbb{I}_K$. A 2.3 tétel szerint tehát

$$\frac{1}{n(d-2)+2}(|E_{1,K}^n|, \dots, |E_{K,K}^n|) \xrightarrow{d} \text{Dirichlet}\left(\frac{d_{G_d}(1) - d_{T_K}(1)}{d-2}, \dots, \frac{d_{G_d}(K) - d_{T_K}(K)}{d-2}\right),$$

ahol $n \rightarrow \infty$, és kizártuk azon komonenseit a vektornak, amelyben $d_{G_d}(i) - d_{T_K}(i) = 0$. Fennáll továbbá, hogy:

$$|E_{i,K}^n| = (d-2)|T_{i,K}^n| - d_{T_K}(i) + 2.$$

Ahonnét következik, hogy

$$\lim_{n \rightarrow \infty} \frac{1}{n} |T_{i,K}^n| = \lim_{n \rightarrow \infty} \frac{1}{n(d-2)+2} |E_{i,K}^n|.$$

Ebből már következik a lemma állítása. ■

6.2 Rumor centralitás

Az alábbi tétel szerint a 4.8 definícióban bevezett,

$$\varphi_T(u) = \prod_{v \in V(T) \setminus \{u\}} |(T, u)_{v \downarrow}|$$

képlettel definiált rumor centralitás jó gyökérkereső algoritmus ebben a modellben is

6.5 Tétel (Khim, Loh [14]). *Tekintsük a G_d $d \geq 3$ -reguláris fán értelmezett $\{T(n)\}_{n=1}^\infty$ terjedési folyamatot. Jelölje $H_{\varphi,L}^*(T)$ azon u csúcsok halmazát a T fában, melyre a rumor centrum és az adott u csúcs távolsága legfeljebb L . Ekkor $L \geq 2$ esetén*

$$\limsup_{n \rightarrow \infty} \mathbb{P}(1 \notin H_{\psi,L}^*(T(n)^\circ)) \leq 7 \exp \left(-\frac{L}{2} \log \left(\min \left\{ \frac{L(d-2)}{4ed^2 \log(L)}, \frac{L}{2d^2} \right\} \right) \right).$$

Speciálisan, megfelelően kis ε esetén

$$\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_{\psi,L}^*(T(n)^\circ)) \geq 1 - \varepsilon,$$

feltéve, hogy $L \geq \frac{a \log(7/\varepsilon)}{\log(b \log(7/\varepsilon))}$, ahol a és b - csak d -től függő - konstansok.

Felhívjuk a figyelmet, hogy a fenti tételben nem a szokásos H_φ függvényt használtuk. Könnyen látható azonban, hogy a tétel következtében a 4.1 definíciónak is megfelel az algoritmus: $K = \frac{d(d-1)^L - 2}{d-2}$ a G_d -ben egy adott csúcstól legfeljebb L távolságra levő csúcsok száma. Így az ε hibatűréshez elég $1/\varepsilon$ -ban szublineáris K költség.

A 6.2 tétel bizonyításához szükséges az alábbi tétel, amelyben becslést adunk annak a valószínűségére, hogy a rumor centrum messze van a valódi vírusforrástól.

6.6 Tétel (Khim, Loh [14]). *Tekintsük a G_d $d \geq 3$ -reguláris fán értelmezett $\{T(n)\}_{n=1}^\infty$ terjedési folyamatot. Legyen v egy csúcs G_d -ben, és jelöljük $l(v)$ -vel a v_1 és a v közötti - egyértelmű - út hosszát. Ekkor $l(v) > 1$ esetén*

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\varphi_{T(n)}(v) \leq \varphi_{T(n)}(v_1)) \leq 7 \exp \left(-\frac{l(v)}{2} \log \left(\min \left\{ \frac{l(v)(d-2)}{4e \log(l(v))}, \frac{l(v)}{2} \right\} \right) \right).$$

A fenti két tétel bizonyítása a 4.11 tételnél bemutatott módszerhez hasonlóan történhet, ezért ennek közlésétől eltekintünk.

A rumor centralitásról belátható, hogy bizonyos esetekben nemcsak a 4.1 definíció értelmében jó gyökerkereső algoritmus, hanem gyöker pontos meghatározására is alkalmas lehet. Erről szól az alábbi eredmény:

6.7 Tétel (Shah, Zaman [1]). *A d -reguláris végtelen fán való terjedésre igaz, hogy:*

- $d = 2$ esetén $\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_{\varphi,1} T(n)^\circ) = O(\frac{1}{\sqrt{n}})$
- $d > 2$ esetén $\liminf_{n \rightarrow \infty} \mathbb{P}(1 \in H_{\varphi,1} T(n)^\circ) > 0$.

A tétel első fele azt az intuíciónkat erősíti meg, hogy amennyiben a terjedés egy egyenesen történik, a gyöker pontos meghatározása nem lehetséges. Abban az esetben, amikor $d \geq 3$ a pontos detektálás valószínűsége nullától egyenletesen elhatárolódik.

A 4.16 következmény alapján a 6.7 tétel a Jordan- és a távolságcentrum esetében is érvényben marad.

6.3 A gyökér elrejtése

Láttuk tehát, hogy a 6.1 definícióban bevezetett diffúziós modellben is identifikálható a terjedés kiindulópontja. A 5. fejezetben leírt anonimitási kérdések tehát ezzel a modellel kapcsolatban is felmerülnek. Ebben az esetben azonban jóval többet tudunk: Fanti és szerzőtársai [12] megadtak egy ún. adaptív terjedési eljárást, amellyel elérhető a gyökér tökéletes elrejtése. Ebben a részben ezt az eljárást mutatjuk be.

A hálózat amelyen az információ terjed ebben az esetben is egy G_d d -reguláris végtelen fa lesz. Azonban a 6.1 definícióban látottaktól eltérő módon itt egy időpontban egyszerre több csúcs is megfertőződhet.

6.8 Definíció (Adaptív diffúzió). A G_d végtelen d -reguláris fán való adaptív diffúziónak nevezzük azt a terjedési folyamatot, amelyet az 2. algoritmus ír le.

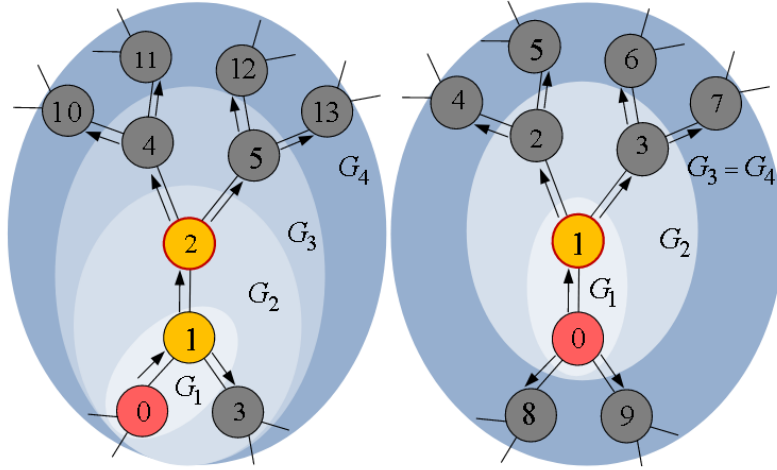
Ennek lényege a következő: minden lépésben biztosítjuk, hogy minden páros pillanatban, azaz $n \in \{2, 4, \dots\}$ esetén a fertőzött $T(n)$ részgráf kiegyensúlyozott fa legyen. Sőt elérjük azt is, hogy az összes levél távolsága a középponttól $n/2$ legyen ilyenkor. A gráf mindenkori középpontját "virtuális (vírus/információ) forrásnak" nevezzük, és v_n -nel jelöljük. Az igazi vírusforrásra a szokásos $v_0 = v^*$ jelölést alkalmazzuk.

Az első lépés mindig ugyanaz: $n = 0$ esetén a v^* vírusforrás egyforma valószínűséggel választ a szomszédai közül egyet, hogy az legyen v_2 , azaz a virtuális forrás a $n = 2$ időpillanatban. Az $n = 1$ pillanatban v^* továbbítja az üzenetet v_2 -nek. Ezután az $n = 2$ pillanatban v_2 megfertőzi a még egészséges szomszédait, így létrehozva a $T(2)$ gráfot. Majd a v_2 csúcs dönt arról, hogy marad virtuális vírusforrás, vagy továbbadja ezt a pozíciót egyik szomszédjának.

Amennyiben v_2 marad a virtuális forrás, tehát $v_4 = v_2$, akkor a levelekhez eljuttatja az ún. infection message-t, ezzel jelezve nekik, hogy fertőzzék meg még nem fertőzött szomszédjaikat. Erről az M_n darab üzenetről feltesszük, hogy egyből végbemegy. Az üzenetek biztosítják, hogy az $n = 3$ lépésben a terjedés minden irányban szimmetrikus. Az $n = 4$ pillanatban nincs további terjedés. Ez látható a 6.10 ábra³ jobb oldalán.

Ha v_2 úgy dönt, hogy lemond a virtuális forrás pozícióról, akkor átadja azt egy véletlenszerűen választott szomszédjának, kivéve a korábbi virtuális forrásokat, esetünkben ez v_0 . Ezzel tehát a virtuális forrás a valódi vírusforrástól egy lépéssel távolabb kerül. Ezután a v_4 kiküldi a leveleknek, hogy tovább fertőzzenek, a v_2 irányába azonban nem küld ilyen üzenetet. Így a fertőzött gráf csak egy részfájában aszimmetrikus. Az $n = 4$ időpillanatban a virtuális forrás változatlanul v_4 marad, és megismétli az előbbi üzenetet a levelek felé. Ezzel a megfertőzött rész ismét szimmetrikus lesz a virtuális forrás körül, ezt az esetet mutatja be a 6.10 ábra bal oldala.

³Ezt az ábrát Fanti és társai [12] cikkéből kölcsönöztük.



Ábra 6.10: Az adaptív diffúzió két lehetséges elindulása. A valódi forrást piros, míg a virtuális forrásokat sárga csúcs jelöli. Az $N = 4$ időpontban a virtuális forrást piros szegéllyel tüntettük fel.

Ezt az eljárást folytatjuk minden lépésben: a v_n virtuális forrás véletlenszerűen megtartja vagy átadja pozícióját, ettől függően a fertőzött részgráf determinisztikusan nő úgy, hogy biztosítsa a szimmetriát a v_{n+2} csúcs körül.

Ezt a véletlen terjedési folyamatot idő-inhomogén (időfüggő) Markov-láncként is felfoghatjuk, ahol az állapot a virtuális forrás aktuális pozíciója $\{v_n\}_{n \in \{0, 2, 4, \dots\}}$. Mivel a vizsgált hálózat egy végtelen d -reguláris fa, és a következő virtuális forrás egyenletesen választódik az előző szomszédai közül, ezért elég a v^* és a v_n közötti távolság Markov-láncként vizsgálni:

$$h_n := \delta_L(v^*, v_n)$$

minden n -re, ahol $\delta_L(v^*, v_n)$ jelöli a v^* és v_n csúcsok közötti távolságot a hálózatban.

Minden időpillanatban a protokoll véletlenszerűen eldönti, hogy a virtuális forrás maradjon-e ugyanaz ($h_{n+2} = h_n$), vagy kerüljön a valódi forrástól eggyel távolabb ($h_{n+2} = h_n + 1$). A keletkező $\{h_n\}_{\{2, 4, 6, \dots\}}$ Markov-lánckot a fenti döntési valószínűség megadásával határozzuk meg. Minden n -re annak a valószínűségét, hogy a virtuális forrás nem változik az alábbi módon jelöljük:

$$\alpha_d(n, h) := \mathbb{P}(h_{n+2} = h_n | h_n = h)$$

Például, ha $\alpha_d(n, h) = 1$ minden n -re és h -ra, akkor a virtuális forrás sohasem változik a $n > 1$ időpillanatokban, azaz v_1 marad. Ilyenkor a terjedés nagyfokú szimmetriája miatt a valódi forrás identifikálása nagy valószínűséggel megtehető. Ellenben ha $\alpha_d(n, h) = 0$ minden n -re és h -ra, akkor a virtuális forrás mindig változik, és így a valódi forrás mindig a fertőzött részfa egy levele lesz. Ez biztosítja, hogy a megtalálása nehéz feladat.

A kihívást $\alpha_d(n, h)$ megfelelő megválasztása jelenti: a terjedés elég gyors és a forrás elrejtésére is alkalmas kell, hogy legyen.

Hogy problémánkat megfelelő kontextusba helyezzük, tekintsük azt a determinisztikus terjedést, amikor minden lépésben minden fertőzött csúcson az összes még egészséges szomszédjának továbbadja a fertőzést, ekkor a fertőzött csúcsonak száma az n időpillanatban fennáll, hogy

$$F_n = 1 + \frac{d((d-1)^n - 1)}{d-2}$$

Ebben az esetben a fertőzött részgráf egy kiegyensúlyozott reguláris fa, ahol a levelek egyenlő távolságra vannak a forrástól. Valamilyen értelemben a fenti mennyiség jellemzi tehát a lehető leggyorsabb terjedést.

Másrészt viszont az előző részben vizsgált modellben a forrás detektálása lehetséges, a 6.2, 6.5 és 6.7 tételek értelmében, ezt szeretnénk itt elkerülni.

Vizsgáljuk most meg az adaptív diffúzió terjedési és anonimitási tulajdonságait. Jelölje $p^{(n)} = [p_h^{(n)}]_{h \in \{1, \dots, n/2\}}$ a Markov-lánc állapotainak eloszlását a n időpillanatban, azaz $p_h^{(n)} = \mathbb{P}(h_n = h)$. Ekkor az állapotváltozást az alábbi $((n/2)+1) \times (n/2)$ -es sztochasztikus mátrix írja le:

$$p^{(t+2)} = \begin{bmatrix} \alpha_d(n, 1) & & & & & \\ 1 - \alpha_d(n, 1) & \alpha_d(n, 2) & & & & \\ & 1 - \alpha_d(n, 2) & \ddots & & & \\ & & \ddots & \alpha_d(n, n/2) & & \\ & & & 1 - \alpha_d(n, n/2) & & \end{bmatrix} p^{(n)}$$

A h_n tekinthető szigorúan pozitívnek, mert a $n = 0$ pillanatban a virtuális forrás mindenképpen megváltozik, így $h_n \geq 1$, ha $n \geq 1$.

Azt szeretnénk elérni, hogy páros n időpillanatokban fennálljon az alábbi egyenlőség:

$$p^{(n)} = \frac{d-2}{(d-1)^{n/2} - 1} \begin{bmatrix} 1 \\ (d-1) \\ \vdots \\ (d-1)^{n/2-1} \end{bmatrix} \in \mathbb{R}^{n/2} \quad (6.6)$$

ha $d > 2$, és $d = 2$ esetén pedig: $p^{(n)} = (2/n)\mathbf{1}_{n/2}$ legyen, ahol $\mathbf{1}_{n/2}$ a csupa egyet tartalmazó vektor $\mathbb{R}^{n/2}$ -ben. A virtuális forrástól h távolságra levő csúcsonak száma:

$$d(d-1)^{h-1}$$

és a szimmetria miatt mindegyik egyforma valószínűséggel volt a vírusforrás:

$$\mathbb{P}(T(n) | v^*, \delta_L(v^*, v_n) = h) = \frac{1}{d(d-1)^{h-1}} p_h^{(n)} = \frac{d-2}{(d-1)^{n/2} - 1}$$

Ez $d > 2$ esetén nem függ a h -tól, így a virtuális forrástól eltekintve az összes csúcs egyforma valószínűséggel volt a valódi forrás.

Ez és a kívánt 6.6 egyenlőség alapján felírható egy rekurzió. Melyet megoldva, és kihasználva, hogy kezdetben $p^{(2)} = 1$ azt kapjuk, hogy az $\alpha_d(n, h)$ valószínűség alábbi megválasztása esetén teljesül a 6.6 egyenlőség:

$$\alpha_d(n, h) = \begin{cases} \frac{(d-1)^{n/2-h+1}-1}{(d-1)^{n/2+1}-1} & \text{ha } d > 2 \\ \frac{n-2h+2}{n+2} & \text{ha } d = 2 \end{cases} \quad (6.7)$$

Látható, hogy ez $d > 2$ -re közelítőleg $(d-1)^{-h}$ -val egyenlő. A paraméterek ilyen megválasztásával az adaptív terjedés viszonylag gyors, nevezetesen $F_n = O((d-1)^{n/2})$ lesz a fertőzött csúcsok számának nagyságrendje a n -edik időpillanatban. Továbbá ilyenkor teljesül az alábbi eredmény.

6.9 Tétel (Fanti, Kairouz, Oh, Ramchandran, Viswanath [12]). *Tekintsük a G_d végtelen csúcsú $d \geq 2$ -reguláris fát, és az ezen vett 6.8 definíció szerinti adaptív diffúziót a v^* kiindulási ponttal, ahol az $\alpha_d(n, h)$ valószínűségeket a (6.7) egyenlőségben leírt módon választjuk. Az ellenség egy $N \geq 0$ időpontban megbecsli a v^* forrás helyét maximum-likelihood becsléssel: \hat{v}_{ML} . Ekkor az alábbi tulajdonságok érvényesek:*

1. a fertőzött csúcsok száma a N időpontban:

$$F_N \geq \begin{cases} \frac{2(d-1)^{(N+1)/2-d}}{(d-2)} + 1 & \text{ha } d > 2 \\ N + 1 & \text{ha } d = 2 \end{cases} \quad (6.8)$$

2. a forrás meghatározásának valószínűsége a maximum-likelihood becsléssel a N időpontban:

$$\mathbb{P}(\hat{v}_{ML} = v^*) \leq \begin{cases} \frac{d-2}{2(d-1)^{(N+1)/2-d}} & \text{ha } d > 2 \\ (1/N) & \text{ha } d = 2 \end{cases} \quad (6.9)$$

3. a maximum-likelihood becslés és a valódi forrás közötti várható távolság alulról becsülhető az alábbi módon:

$$\mathbb{E}[\delta_L(\hat{v}_{ML}, v^*)] \geq \frac{d-1}{d} \frac{N}{2}. \quad (6.10)$$

Bizonyítás. 1. Az adaptív diffúzió során páros N esetén a fertőzött $T(N)$ részgráf olyan fa, amelyben a virtuális forrásnak d , míg a többi nem levél csúcsnak $d-1$ szomszédja van, és a levelek távolsága a virtuális forrástól $N/2$. A csúcsok száma ebben az esetben tehát:

$$F_N = 1 + d(1 + (d-1) + \dots + (d-1)^{N/2-1}) = 1 + d \frac{(d-1)^{N/2} - 1}{(d-1) - 1} = \frac{d(d-1)^{N/2}}{d-2} - \frac{2}{d-2}$$

Páratlan N -re $\alpha_d(N, h)$ valószínűséggel az előbbihez hasonló a $T(N)$ fa, csak a levelek távolsága a virtuális forrástól $(N+1)/2$. Míg $1 - \alpha_d(N, h)$ valószínűséggel $T(N)$ lényegében két $d - 1$ -reguláris fa, melyek mélysége $(N - 1)/2$, a gyökerüknél összekapcsolva. A páros esethez hasonlóan látható, hogy a csúcsok F_N számát $d > 2$ az alábbi egyenlőség adja:

$$F_N = \begin{cases} 1 & , \text{ ha } N = 0 \\ \frac{2(d-1)^{(N+1)/2}}{d-2} - \frac{2}{d-2} & , \text{ ha } N \geq 1, N \text{ páratlan, } 1 - \alpha \text{ valószínűséggel} \\ \frac{d(d-1)^{(N+1)/2}}{d-2} - \frac{2}{d-2} & , \text{ ha } N \geq 1, N \text{ páratlan, } \alpha \text{ valószínűséggel} \\ \frac{d(d-1)^{N/2}}{d-2} - \frac{2}{d-2} & , \text{ ha } N \geq 1, N \text{ páros} \end{cases}$$

Hasonlóképpen $d = 2$ esetén:

$$F_N = \begin{cases} 1 & , \text{ ha } N = 0 \\ N + 1 & , \text{ ha } N \geq 1, N \text{ páratlan, } 1 - \alpha \text{ valószínűséggel} \\ N + 2 & , \text{ ha } N \geq 1, N \text{ páratlan, } \alpha \text{ valószínűséggel} \\ N + 2 & , \text{ ha } N \geq 1, N \text{ páros} \end{cases}$$

Ebből már következik a (6.8) egyenlőtlenség.

2. A virtuális forrásról látható, hogy nem lehetett a valódi vírusforrás, azaz

$$\mathbb{P}(T(N)|v^* = v_N) = 0.$$

A többi csúcsról igazoljuk, hogy egyformán valószínű jelöltek a valódi vírusforrásra. Figyeljük meg, hogy egy fában bármely két csúcs között létezik egy egyértelmű út, így tetszőleges $u \in T(N) \setminus \{v_N\}$, sőt a v_N elhelyezkedése miatt ennek hossza legfeljebb $\lceil N/2 \rceil$. Tehát létezik virtuális források egy $\{v_i\}_{i=0}^N$ sorozata, melyre $v_0 = u$, azt kell belátni, hogy tetszőleges u -ra ennek a sorozatnak a valószínűsége azonos. A G_d hálózat regularitása, és a $T(N)$ szimmetriája miatt páros N esetén tetszőleges u_1, u_2 pontokra, melyekre $\delta_L(u_1, v_N) = \delta_L(u_2, v_N) = h > 0$ fennáll, hogy:

$$\mathbb{P}(T(N)|v^* = u_1) = \mathbb{P}(T(N)|v^* = u_2)$$

Ha még azt is figyelembe vesszük, hogy az $\alpha_d(n, h)$ -kat a (6.7) egyenlőséggel definiáltuk, és hogy a v_N -től h távolságra pontosan $d(d-1)^{h-1}$ csúcs található kapjuk, hogy a

$$\mathbb{P}(T(N)|v^* = w_1) = \mathbb{P}(T(N)|v^* = w_2)$$

egyenlőség tetszőleges a v_N virtuális vírusforrástól különböző csúcsokra fennáll.

Páratlan N -re, ha a virtuális forrás nem változik, a fenti gondolatmenet megismételhető, mert a $T(N)$ szimmetrikus marad. Amikor éppen változik a vírusforrás, akkor a gráf a $\{v_{N-1}, v_N\}$ éle mentén lesz szimmetrikus. Ilyenkor a v_N és v_{N-1} csúcsok egyike

sem lehet a valódi vírusforrás, és ezt az ellenség is meg tudja határozni. A fenti éllel összekött két részfa szimmetriája és a folyamat konstrukciója (6.7) miatt tetszőleges $w_1, w_2 \in T(N) \setminus \{v_N, v_{N-1}\}$ csúcsokra:

$$\mathbb{P}(T(N)|v^* = w_1) = \mathbb{P}(T(N)|v^* = w_2)$$

Tehát páratlan N esetén:

$$\mathbb{P}(v_{ML} = v^*) = \frac{1}{F_N - 2}$$

Ebből már következik a (6.9) egyenlőtlenség.

3. Az előző pont alapján lényegében minden a virtuális forrástól különböző pont egyformán valószínű, így az ettől vett átlagos távolságot kell csak kiszámolni. Természetesen ebben az esetben is külön ügyelve arra az esetre, amikor a N páratlan és a gráf nem teljesen szimmetrikus. Mindhárom esetben egyszerű számolás mutatja, hogy a (6.10) egyenlőtlenség valóban teljesül. ■

Habár a paraméterek fenti megválasztásával sikeresen elrejtettük a forrást, a terjedés üteme elmarad a determinisztikusnál látott $O((d-1)^N)$ -től. Ez a konstans faktor veszteség a kitevőben azonban elkerülhetetlen: ahhoz, hogy a determinisztikus modelltől eltérjünk, megfelelő késleltetéseket kell bevezetni a terjedésben.

A fent leírt optimális paraméterválasztáshoz ismerni kell a d fokszámot. Erre nincs mindig szükség: tekinthetjük például azt az esetet, amikor $\alpha_d(n, h) = 0$ minden d, n és h értékre. Az alábbi 6.10 állítás szerint az optimális esettel összevethető hatékonyságú ez a paraméterezés is, az elrejtés tekintetében. Ezt szemléletesen úgy láthatjuk, hogy a fertőzési gráfban nagyjából annyi belső pont van, mint levél, így elég a levelek között elrejtetni a vírusforrást. Egyetlen kivétel ez alól a $d = 2$ eset, ilyenkor ugyanis a fertőzött részgráf egy út, ennek csupán két levele van, így ilyenkor a detektálás valószínűsége triviálisan $1/2$. A 6.9 tételhez hasonló módon bizonyítható az alábbi:

6.10 Állítás (Fanti, Kairouz, Oh, Ramchandran, Viswanath [12]). *Tekintsük a G_d végtelen csúcsú $d > 2$ -reguláris fát, és az ezen vett 6.8 definíció szerinti adaptív diffúziót a v^* kiindulási ponttal, ahol az $\alpha_d(n, h) = 0$ választással élünk minden d, n és h értékre. Az ellenség egy $N \geq 0$ időpontban megbecsli a v^* forrás helyét maximum-likelihood becsléssel: \hat{v}_{ML} . Ekkor az alábbi tulajdonságok érvényesek:*

1. a fertőzött csúcsok száma a $N \geq 1$ időpontban:

$$F_N \geq \frac{(d-1)^{(N+1)/2}}{d-2} \quad (6.11)$$

2. a forrás meghatározásának valószínűsége a maximum-likelihood becsléssel a N időpontban:

$$\mathbb{P}(\hat{v}_{ML} = v^*) = \frac{d-1}{2 + (d-2)F_N} \quad (6.12)$$

3. a maximum-likelihood becslés és a valódi forrás közötti várható távolság alulról becsülhető az alábbi módon:

$$\mathbb{E}[\delta_L(\hat{v}_{ML}, v^*)] \geq \frac{N}{2}. \quad (6.13)$$

Az 6.8 definícióban látott adaptív diffúzió tehát sikeresen elrejti a valódi vírusforrást d -reguláris fán való terjedés esetén. Fanti és szerzőtársai számos további kérdést vizsgálnak [12] cikkükben. Ezekről jelentős terjedelmük miatt csak az alábbi néhány gondolatot ragadjuk ki.

Természetes kérdésként felmerül, hogy ha az ellenség több időpontban is megfigyelheti a terjedést, mennyivel javul a valódi gyökér megtalálásának valószínűsége. Belátják, hogy az - optimálisan paraméterezett - adaptív diffúzió esetén legfeljebb a fertőzött részgráf méretében logaritmikus szorzóval növekszik a detektálás valószínűsége a megfigyelések számától függetlenül.

Fanti és társai felhívják a figyelmet arra is, hogy nem minden hálózat topológiára optimális az adaptív diffúzió. Ellenben belátják, hogy megfelelően választott $\alpha(n, h)$ esetén az adaptív diffúzió - a szükséges módosítások elvégzése után - négyzet rácson is eléri a gyökér tökéletes elrejtését.

Megjegyzik továbbá, hogy a kérdésnek van játékelméleti vonatkozása is. Két játékos a tervező és az ellenség játssza az alábbi játékot: a tervező tetszőleges stratégiával terjesztheti az üzenetet a v^* forrásból kiindulva feltéve, hogy egy körben csak egy távolságra tudja továbbítani. Az ellenség pedig a forrás egy \hat{v} becslését számítja ki tetszőleges stratégiával a rendelkezésére álló több-kevesebb információ alapján. A tervező a detektálás valószínűségét minimalizálni, míg az ellenség maximalizálni akarja. Bizonyos feltételek teljesülése esetén ekkor az adaptív diffúzió (gyenge) domináns stratégia. A kérdés ilyen irányú vizsgálata nem ismeretlen a szakirodalomban, de jelen dolgozat lehetőségein messze túlmutat.

Algorithm 2 Adaptív diffúzió

Input adott hálózat $G = (V, E)$, forrás $v^* \in V$, időpont N , fokszám d

Output a fertőzött csúcsok halmaza V_N

```
1:  $V_0 \leftarrow \{v^*\}$ ,  $h \leftarrow 0$ ,  $v_0 \leftarrow v^*$ 
2:  $v^*$  kiválasztja egyik  $u$  szomszédját véletlenszerűen
3:  $V_1 \leftarrow V_0 \cup \{u\}$ ,  $v_1 \leftarrow v_u$ 
4: Jelölje  $S(u)$  az  $u$  szomszédainak halmazát
5:  $V_2 \leftarrow V_1 \cup S(u) \setminus \{v^*\}$ ,  $v_2 \leftarrow v_1$ 
6:  $n \leftarrow 3$ 
7: for  $n \leq N$  do
8:    $v_{n-1}$  választ egy valószínűségi változót  $X \sim U(0, 1)$ 
9:   if  $X \leq \alpha_d(n-1, h)$  then
10:     for all  $v \in S(v_{n-1})$  do
11:       Infection Message( $G, v_{n-1}, v, V_n$ )
12:     end for
13:   else
14:      $v_{n-1}$  véletlenszerűen választ egy  $u \in S(v_{n-1}) \setminus \{v_{n-2}\}$ 
15:      $h \leftarrow h + 1$ 
16:      $v_n \leftarrow u$ 
17:     for all  $v \in S(v_n) \setminus \{v_{n-1}\}$  do
18:       Infection Message( $G, v_n, v, V_n$ )
19:     if  $n + 1 > N$  then
20:       break
21:     end if
22:     Infection Message( $G, v_n, v, V_n$ )
23:   end for
24: end if
25:  $n \leftarrow n + 2$ 
26: end for
27: procedure INFECTION MESSAGE( $G, u, v, V_n$ )
28:   if  $v \in V_n$  then
29:     for all  $w \in S(v) \setminus \{u\}$  do
30:       Infection Message( $G, v_n, w, V_n$ )
31:     end for
32:   else
33:      $V_n \leftarrow V_{n-2} \cup \{v\}$ 
34:   end if
35: end procedure
```

7 Összegzés és kitekintés

A dolgozatban véletlen fák gyökerének identifikálását és elrejtését tűztük ki célul. Ezen kérdések kapcsán kétféle vírusterjedési modell is bemutatásra került: a preferential és uniform attachment fák, illetve a végtelen, reguláris fán való diffúzió esetében is láthattunk a gyökér detektálására alkalmas algoritmusokat.

A gyökér elrejtését a reguláris fán való terjedés esetén az adaptív diffúzió segítségével sikerült elérni, míg a preferential attachment modellnek bemutattuk egy ígéretes módosítását.

Bár mint láttuk, az elmúlt időszakban számos ilyen eredmény napvilágot látott, a témakör még sok izgalmas megválaszolatlan problémát tartogat. Zárásként ezek közül említünk meg most néhányat.

Természetesen merül fel a kérdés, hogy általános gráfokra igazolható-e hasonló eredmény? Az itt bemutatott Pólya-urnákra épülő bizonyítások várhatóan nem általánosíthatók, így ennek megválaszolásához új technikák kidolgozására lesz szükség.

A valóságban gyakran a vírusgráfnak csak bizonyos részleteit ismerjük, viszont rendelkezésünkre állhat esetleg több más információ, például néhány él irányítása. Ebbe az irányba tehetünk egy lépést, ha feltesszük, hogy több különböző időpontban is megfigyelhetjük a gráfot.

A gyökér elrejtése esetén is felmerülnek a fenti kérdések. Általánosabb gráfosztályokon is sikerülhet-e a forrás elrejtése? Mi történik, ha az ellenség több információ birtokában van, például több időpontban is megfigyeli a gráfot, vagy egyes csúcsok "kémkednek", és elmondják ki fertőzte meg őket?

Irodalomjegyzék

- [1] Devavrat Shah and Tauhid Zaman. Detecting sources of computer viruses in networks: theory and experiment. In *SIGMETRICS*, pages 203–214, 2010.
- [2] Sébastien Bubeck, Luc Devroye, and Gábor Lugosi. Finding Adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172, 2017.
- [3] Sayan Banerjee and Shankar Bhamidi. Root finding algorithms and persistence of jordan centrality in growing random trees. *arXiv preprint arXiv:2006.15609*, 2020.
- [4] Miklós Z Rácz and Sébastien Bubeck. Basic models and questions in statistical network analysis. *Statistics Surveys*, 11:1–47, 2017.
- [5] Svante Janson. Limit theorems for triangular urn schemes. *Probability Theory and Related Fields*, 134:417–452, 03 2006.
- [6] Sébastien Bubeck, Elchanan Mossel, and Miklós Z Rácz. On the influence of the seed graph in the preferential attachment model. *IEEE Transactions on Network Science and Engineering*, 2(1):30–39, 2015.
- [7] Robert D Kleinberg and Jon M Kleinberg. Isomorphism and embedding problems for infinite limits of scale-free graphs. In *SODA*, pages 277–286, 2005.
- [8] Nicolas Curien, Thomas Duquesne, Igor Kortchemski, and Ioan Manolescu. Scaling limits and influence of the seed graph in preferential attachment trees. *Journal de l'École polytechnique-Mathématiques*, 2:1–34, 2015.
- [9] Sébastien Bubeck, Ronen Eldan, Elchanan Mossel, and Miklós Z Rácz. From trees to seeds: on the inference of the seed from large trees in the uniform attachment model. *Bernoulli*, 23(4A):2887–2916, 2017.
- [10] Camille Jordan. Sur les assemblages de lignes. *Journal für die reine und angewandte Mathematik*, 70:185–190, 1869.
- [11] P. Erdős. On an elementary proof of some asymptotic formulas in the theory of partitions. *Annals of Mathematics*, 43:437–450, 1942.
- [12] Giulia Fanti, Peter Kairouz, Sewoong Oh, Kannan Ramchandran, and Pramod Viswanath. Hiding the rumor source. *IEEE Transactions on Information Theory*, 63(10):6679–6713, 2017.
- [13] Devavrat Shah and Tauhid Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, 2011.
- [14] Justin Khim and Po-Ling Loh. Confidence sets for the source of a diffusion in regular trees. *IEEE Transactions on Network Science and Engineering*, 4(1):27–40, 2017.

NYILATKOZAT

Név: Döbröntei Dávid Bence

ELTE Természettudományi Kar, szak: Matematika Bsc

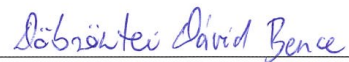
NEPTUN azonosító: CS52QE

Szakedolgozat címe:

Véletlen gráfok gyökerének identifikálása és elrejtése

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

Budapest, 2021. május 31.



a hallgató aláírása