Súlyozott gráfok paramétereinek becslése

TDK dolgozat, 2009. jan. 4.

Készítette: Kói Tamás

BMGE, TTK, V. évfolyam, Matematikus szak

Konzulens: Bolla Marianna docens, Sztochasztika Tanszék

1. Bevezető

Manapság számos probléma vezet nagy és összetett gráfok tulajdonságainak vizsgálatához. Gráfparaméternek nevezzük az olyan gráfokon értelmezett függvényeket, amelyek bármely két izomorf gráfon ugyanazt az értéket veszik fel. Azt mondjuk, hogy tesztelhető egy gráfparaméter, ha igaz az, hogy egy nagy gráf paraméterét jól tudjuk becsülni a paraméter gráfból véletlenül kivett mintákon felvett értékével. A gyakorlatban nagyon előnyös, ha egy paraméterről kiderül, hogy tesztelhető: ahelyett, hogy rengeteg számítási időt felhasználva megpróbálnánk meghatározni a paraméter értékét, megtehetjük, hogy kis mintákat veszünk, és a kis mintán gyorsan számolt értékekkel közelítjük a nagy gráf paraméterét. Lovász László és munkatársai az 1990-es évek kezdeményeire támaszkodva nagyon impozáns elméletet dolgoztak ki annak eldöntésére, hogy tesztelhető-e egy gráfparaméter. A gráfok terén bevezettek egy metrikát, és ezt a teret beágyazták egy kompakt metrikus térbe, az egységnégyzeten értelmezett mérhető, szimmetrikus függvények terébe, az úgynevezett grafonok terébe; ezzel általánosítva a gráf fogalmát. Ennek az elméletnek a felhasználásával több ekvivalens jellemzést adtak gráfok paramétereinek tesztelhetőségére. Különösen figyelemreméltó: akkor tesztelhető egy gráfparaméter, ha ki lehet terjeszteni a grafonok terére úgy, hogy a kiterjesztett paraméter folytonos egy bizonyos normában. Ezzel újabb szál került a kombinatorika és az analízis tudománya közé.

A TDK dolgozatom erről a témáról fog szólni. A témával úgy kerültem kapcsolatba, hogy a konzulensem érdeklődött röviden mincutnak nevezhető paraméterek tesztelhetősége iránt. Gondot okozott az a tény, hogy bár a meglévő elmélet általános volt, a tesztelés definíciója, és a kapcsolódó tételek csak egyszerű gráfokra lettek kimondva. Így konzulensemmel elkezdtünk a tesztelési tétel súlyozott gráfokra vett általánosításán, illetve röviden mincutnak nevezhető paraméterek tesztelhetőségén dolgozni. A munkába Friedl Katalin docens (BME VIK), és Krámli András professzor (Szegedi Egyetem) is bekapcsolódott.

A dolgozat második fejezetében konzulensem, Dr. Bolla Marianna és társszerzői által [1], [2]-ben elért eredményekből mutatok be néhányat. Ezek az eredmények jelentették az inspirációt a súlyozott gráfparaméterek tesztelhetőségének vizsgálatához.

A harmadik fejezet [3]-ban ismertetett definíciókból és tételekből tartalmaz válogatást különös tekintettel gráfok paramétereinek tesztelhetőségére.

A negyedik fejezetben már új gondolatok és eredmények találhatók. A fejezet első felében nagyítóval megvizsgálom az egyszerű gráfok tesztelhetőségének fogalmát. Ezt követően leírom a tesztelhetőség fogalmának súlyozott gráfokra vett általánosítását, továbbá a [3]-ban szereplő tesztelési tétel általánosítását. Az elvégzett munka összességében nem más, mint a dolgok mély átgondolása, a már meglévő fogalmak összeillesztése , és az egyszerű tesztelési tétel bizonyításának érvényben maradásáról való meggyőződés a kisebb hiányosságok pótlásával. A kulcsa mindennek a randomizálási procedúrák, és a hozzájuk kapcsolódó fogalmak pontos megértése volt.

Az ötödik fejezetben néhány gráfparaméter tesztelhetőségét vizsgálom.

A hatodik fejezet kezdetleges számítógépes szimulációk leírását, eredményét és értékelését tartalmazza; továbbá komolyabb szimulációk lehetséges alapgondolatát.

A hetedik fejezetben röviden összefoglalom a leírtakat, továbbá ismertetem, véleményem szerint milyen kérdések megválaszolása lenne fontos a jövőben.

A negyedik fejezet tartalmát a témáról folytatott közös beszélgetések, útmutatások segítségével én dolgoztam ki. Az ötödik fejezet tartalmának kidolgozásában csak kis mértékben vettem részt. Emiatt a fejezet vázlatos, csak az összefüggések megértéséhez mindenképpen szükséges dolgokat tartalmazza. A hatodik fejezetben taglalt teszteket én vé-

geztem.

2. Gráfok klaszterezése a Laplace mátrix spektrumának segítségével

Gráfok klaszterezésén azt értjük, hogy igyekszünk minél jobban csoportokba osztani a csúcsokat úgy, hogy az egyes csoportokon belül futó élek összsúlya nagy, míg a csoportok között futó élek összsúlya kicsi legyen. Nagyon érdekes, hogy gráfok klaszteresíthetőségi tulajdonságaira vonatkozólag képesek vagyunk információt nyerni a gráf úgynevezett Laplace-mátrixának sajátértékeiből. Valahogy egy nagyon szétszórt dologról van információ egy nagyon rendezett dologban.

1. Definíció. *Legyen $P_k = (V_1, \dots, V_k)$ a csúcsok egy k partíciója. Ennek sűrűsége*

$$\rho(P_k) := \sum_{i=1}^{k-1} \sum_{j=1}^{k} (\frac{1}{\alpha_{V_i}} + \frac{1}{\alpha_{V_j}}) \beta(V_i, V_j),$$

ahol α_{V_i} a V_i-beli csúcsok összsúlya, $\beta(V_i, V_j)$ pedig a V_i es V_j csúcshalmaz közt átmetsző élek összsúlya. Legyen

$$\rho_k := \min_{P_k} \rho(P_k).$$

Az imént definiált fogalom a gráf jól klaszteresíthetőségét méri bizonyos szemszögből. $\rho(P_k)$ a partícióhoz hozzárendeli a partíciók között futó élek összsúlyát megsúlyozva egy kicsit. Ez a súly akkor kicsi, ha a vizsgált két partíció mérete közel azonos. Így még pontosabban ennek a minimuma, ρ_k, a gráfnak az olyan jól klaszteresíthetőségét méri, ahol a partícióméretek közel azonossága is fontos tényező. A következő konstans két partíció esetén ennek egy legfeljebb 2 -es szorzóban eltérő változata.

2. Definíció. *Súlyozott gráf Cheeger-konstansának nevezzük a következő kifejezést:*

$$h = \min_{U \subset V(G), \alpha_U \leq \frac{1}{2}} \frac{\beta(U, \overline{U})}{\alpha_U},$$

ahol feltettük, hogy az összes csúcssúly összege 1.

Ezekkel a definíciókkal készen állunk arra, hogy kimondjunk néhány fontos tételt.

1. Tétel. *[2] 3.2. Tétel: Legyen G egy olyan súlyozott gráf, amelyre igaz, hogy*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} = 1.$$

Továbbá a csúcsok súlya legyen a belőlük kiinduló élek összsúlya. A D mátrix az a diagonális mátrix, melynek főátlójában a csúcssúlyok szerepelnek, a B mátrix pedig tartalmazza az élsúlyokat. Legyen λ_1 a $C_D = D^{-\frac{1}{2}}(D - B)D^{-\frac{1}{2}}$ súlyozott Laplace-mátrix legkisebb pozitív sajátértéke. Tegyük fel, hogy $\lambda_1 \leq 1$. Ekkor igaz:

$$\frac{\lambda_1}{2} \leq h \leq \sqrt{\lambda_1 \cdot (2 - \lambda_1)}.$$

Ebből azonnal leolvasható, hogy ha a súlyozott Laplace mátrix legkisebb sajátértéke nagy (1-hez közeli), akkor a gráf nem osztható jól két klaszterbe.

2. Tétel. *[1] tételeiből kivonat: Az 1. Tétel gráfra vonatkozó feltételei mellett igaz:*

$$\sum_{i=1}^{k-1} \lambda_i \leq \rho_k,$$

ahol λ_i a C_D mátrix i-dik nem nulla sajátértékét jelöli (nagyság szerint növekvő sorrendben). ρ_k-ra felső becslés is adható a $k-1$ legkisebb pozitív sajátérték és a hozzájuk tartozó sajátvektorok alapján konstruált optimális $k-1$-dimenziós reprezentánsok úgynevezett k-szórása segítségével.

Emiatt az eddigi kijelentéseinknek bizonyos értelemben a megfordítása is igaz. Ha az optimális $k-1$ dimenziós euklideszi reprezentációt jól lehet k osztályba klaszterezni, illetve ha a súlyozott Laplace-mátrix első $k-1$ sajátértékének összege kicsi, akkor lehet találni a gráfnak is jó k klaszterezését. Nincs garancia arra, hogy a jó euklideszi partíció lesz a gráfot is jól klaszterező partíció, így elsődlegesen annak eldöntésére szolgálhat a tétel, hogy hány osztályos klaszterekre lehet bontani a gráfot. Mindazonáltal a gyakorlatban gyakran jó eredményt ad a gráfon is az euklideszi reprezentáció optimális klaszterezése.

A fenti eredmények jelentették az inspirációt gráfok tesztelhetőségének vizsgálatához. A fent bevezetésre került mennyiségek statisztikus tulajdonsággal bírnak, a gráfok élsúlyainak kicsi változására nem érzékenyek. Ezzel a tulajdonsággal a tesztelhető paraméterek is rendelkeznek. Így természetesen merül fel az elmélet teszteléssel való kapcsolata.

3. Néhány definíció

Ez a fejezet a [3] cikk tartalmába nyújt rövid betekintést. A jelölések nagyrészt megegyeznek a [3] jelöléseivel. Ezen kívül a dolgozatban az összes limeszt, úgy kell érteni, hogy $n \to \infty$.

3. Definíció. *Legyen F egyszerű gráf, míg G él- és csúcssúlyozott gráf, ekkor:*

$$Hom(F,G) := \sum_{\Phi:V(F)\to V(G)} \prod_{i\in V(F)} \alpha_{\Phi(i)}(G) \prod_{ij\in E(F)} \beta_{\Phi(i),\Phi(j)}(G),$$

ahol α_i az i-dik csúcs súlyát jelöli G-ben, míg β_{ij} az i-dik és j-dik élet összekötő él súlya. Ha pedig α_G-vel jelöljük a G csúcsainak összsúlyát, akkor legyen az $F \to G$ homomorfizmussűrűség:

$$t(F,G) := \frac{Hom(F,G)}{{\alpha_G}^k}, \text{ ahol } k \text{ az } F \text{ csúcsainak a száma.}$$

A jelölés onnan származik, hogy ha G egyszerű gráfra úgy tekintünk, mint 1 csúcssúlyokkal, és összeköttetéstől függően 1 vagy 0 élsúlyokkal rendelkező gráfra, akkor $Hom(F,G)$ éppen az F-ből G-be menő homomorfizmusokat számolja össze, vagyis az olyan leképezéseket, amik élet élbe visznek. Ebben az esetben a sűrűség kiszámolásakor ezt osztjuk az összes lehetséges leképezés számával, így annak a valószínűségét kapjuk meg, hogy egyenletesen választva a leképezések közül, mi a homomorfizmus valószínűsége. Lássuk az egyik kulcsdefiníciót:

4. Definíció. *Legyen (G_n) egy súlyozott gráfsorozat, aminek az élsúlyai egyenletesen korlátosak. Azt mondjuk, hogy (G_n) balról konvergens, ha minden F egyszerű gráfra $t(F, G_n)$ konvergens.*

Most megmutatjuk, hogy ez a konvergencia milyen metrikából származik. Ehhez:

5. Definíció.

$$Ha\ S,T \subset V(G)\ akkor\ e_G(S,T) = \sum_{i\in S, j\in T} \alpha_i(G)\alpha_j(G)\beta_{ij}(G).$$

Fontos, hogy S és T nem feltétlenül diszjunkt. Ennek segítségével ha G és $G^{'}$ két ugyanazon címkézett csúcshalmazon értelmezett súlyozott gráf, akkor a vágástávolságuk:

6. Definíció.

$$d_{\square}(G, G^{'}) := \max_{S,T\subset V(G)} \frac{1}{|\alpha_G|^2}|e_G(S,T) - e_{G^{'}}(S,T)|.$$

Kicsit elemezzük a képletet. Megkeressük azt a két csúcshalmazt, amik között futó élek tekintetében a lehető legjobban különbözik a két gráf. Ezek nem feltétlenül nagy halmazok, nagyobb halmazokon lehet, hogy kiegyenlítődnek a különbségek, a halmazok egyik részén az egyik, másikon a másik gráf lehet sűrűbb.

Ez a távolság nagyon megszorított feltételekkel értelmes, szeretnénk ebből egy olyan távolságot csinálni, ami címkézetlen két gráfra is értelmezve van, úgy is, hogy a csúcsszám és a csúcssúlyok nem egyeznek.

7. Definíció. *Legyen G és $G^{'}$ két olyan súlyozott gráf, hogy mindkettő csúcssúlyainak összege 1. Azt mondjuk, hogy az $n \cdot n^{'}$ -es X mátrix frakcionális fedése a két gráfnak, ha igazak a következők:*

$$\sum_{u=1}^{n^{'}} X_{iu} = \alpha_i(G)\ \acute{e}s\ \sum_{i=1}^{n} X_{iu} = \alpha_u(G^{'}).$$

Ezt követően definiálunk két másik gráfot:

8. Definíció. *Legyenek $G[X]$ és $G^{'}[X^{\intercal}]$ az $[n] \times [n^{'}]$-on, mint csúcsokon értelmezett súlyozott gráfok. A közös címkézett csúcshalmazon egy (i,u) csúcsnak a súlya legyen X_{iu}. $G[X]$-ben az $((i,u),(j,v))$ él súlya legyen a G gráf i és j csúcsát összekötő él súlya, míg $G^{'}[X^{\intercal}]$-ban ugyanezen él súlya legyen a $G^{'}$-beli u és v csúcsot összekötő él súlya.*

Így $d_{\square}(G[X], G^{'}[X^{\intercal}])$ már jól definiált. Ennek segítségével:

9. Definíció. *Két súlyozott gráfnak, amelyek csúcssúlyainak összege 1, a távolsága legyen a következő:*

$$\delta_{\square}(G, G^{'}) := \min_{X frak.fed} d_{\square}(G[X], G^{'}[X^{\intercal}])$$

Két tetszőleges gráf távolságát pedig úgy kapjuk meg, hogy annak a két gráfnak a távolságát határozzuk meg, amit úgy nyerünk, hogy a csúcsok súlyát elosztjuk az összcsúcssúllyal.

Be lehet látni, hogy ha a G gráfot felfújjuk k-szorosára, vagyis minden csúcsot k csúccsal helyettesítünk, és minden egyes ilyen csúcsot pontosan azokkal a csúcsokkal kötünk össze (persze ugyanakkora súllyal), amik az eredeti csúccsal összekötött csúcsok helyettesítői, akkor az így kapott gráfot $G[k]$-val jelölve igaz, hogy $\delta_{\square}(G, G[k]) = 0$. Vagyis ez csak egy premetrika (amit metrikának neveznek továbbra is a cikkben egyszerűség végett). Ebből következő további észrevétel, hogy két nagyon különböző csúcsszámú gráf is lehet közel egymáshoz ebben a metrikában, így ez a távolság bizonyos értelemben a gráfok információtartalma közötti különbséget paraméterezi.

3. Tétel. *Legyen (G_n) egy egyenletesen korlátos élsúlyokkal rendelkező súlyozott gráfsorozat. Ekkor (G_n) akkor és csak akkor balról konvergens, ha Cauchy-sorozat $\delta_\square$-metrikában.*

Ezen a ponton érkeztünk el a grafonokhoz. A grafonok azok, amik teljessé teszik a súlyozott gráfok terét ebben a metrikában. Lehetne egyszerűen csak a határobjektumokat grafonoknak hívni, de meg is lehet őket konstruálni:

10. Definíció. *Egy $W : [0,1]^2 \to R$ szimmetrikus, mérhető függvényt grafonnak nevezünk. Azon grafonok halmazát, melyek értékkészlete egy adott I intervallumba esik, $\mathbb{W}_I$-vel jelöljük.*

Minden G súlyozott gráfnak (speciálisan egyszerű gráfnak is) természetesen meg tudunk feleltetni egy W_G grafont. Lenormáljuk a csúcssúlyokat úgy, hogy az összegük egy legyen. A következő intervallumrendszert tekintjük: $I_1 = [0, \alpha_1(G)], I_2 = (\alpha_1(G), \alpha_1(G) + \alpha_2(G)], \ldots$ értelemszerűen. A partíció felé lépcsős grafont építünk: $I_i \times I_j$ felett legyen β_{ij} az értéke (G élsúlyai). Általánosítsuk a homomorfizmussűrűséget grafonokra:

11. Definíció.

$$t(F, W) := \int_{[0,1]^k} \prod_{ij \in E(F)} W(x_i, x_j) dx.$$

Erre a kiterjesztésre triviálisan teljesül, hogy $t(F, G) = t(F, W_G)$. Most általánosítjuk a $\delta_\square$-metrikát grafonokra:

12. Definíció. *Legyen m a $[0,1]^2$-en értelmezett olyan valószínűségmértékek halmaza, amelynek mindkét marginálisa a Lebesgue mérték. Ezt felhasználva:*

$$\delta_\square(W, W') := \inf_{\mu \in m} \sup_{S,T \subseteq [0,1]^2} \left| \int_{(x,u) \in S, (y,v) \in T} (W(x,y) - W'(u,v)) d\mu(x,u) d\mu(y,v) \right|.$$

Látható, hogy $\delta_\square(G, G') = \delta_\square(W_G, W_{G'})$. Így általánosítottuk a gráfokat a vizsgált két konvergenciára való tekintettel.

Vegyük észre, hogy az eddigi definíciók teljesen érzéketlenek voltak a csúcsok skálájára, vagyis a csúcssúlyok összegére. Így az elmélet a gráfok skálainvariáns tulajdonságainak vizsgálatára lesz használható.

Igazak a következő tételek:

4. Tétel. *Minden $W \in \mathbb{W}_I$ grafonhoz létezik egy (G_n) balról konvergens I-beli élsúlyú gráfsorozat, hogy $t(F, G_n) \to t(F, W)$ minden F egyszerű gráfra.*

5. Tétel. *Legyen I egy zárt intervallum, és legyen (W_n) $\mathbb{W}_I$-beli grafonsorozat, ekkor a következők ekvivalensek:*

1. *$t(F, W_n)$ konvergens minden F egyszerű gráfra*
2. *W_n Cauchy-sorozat a $\delta_\square$-metrikában*
3. *létezik egy $W \in \mathbb{W}_I$ úgy, hogy $t(F, W_n) \to t(F, W)$ minden egyszerű F gráfra;*

továbbá $t(F, W_n) \to t(F, W)$ minden F egyszerű gráfra akkor és csak akkor, ha $\delta_\square(W_n, W) \to 0$.

Ezek a tételek összetett okfejtések eredményei. De a végeredmény világos. Ha I zárt intervallum, akkor ha $\mathbb{W}_I$-ben azonosítjuk a $\delta_\square$ távolságban egymástól 0 távolságra levő grafonokat, akkor épp a szintén azonosítás utáni, I-beli élsúllyal rendelkező súlyozott gráfok terének teljes lezárását kapjuk. Ezenkívül ekkor $\mathbb{W}_I$-ben a kétféle konvergencia ekvivalens egymással. Végezetül egy fontos tétel:

6. Tétel. *Ha I zárt intervallum, akkor $(\mathbb{W}_I, \delta_\square)$ kompakt metrikus tér (azonosítás után).*

Most térjünk rá későbbi témánkra, gráfok paramétereire:

13. Definíció. *Egy gráfokon értelmezett valós értékű függvényt paraméternek nevezünk, ha invariáns az izometriára, vagyis két izomorf gráfon ugyanazt az értéket veszi fel.*

14. Definíció. *Egy egyszerű gráfokon értelmezett paraméter tesztelhető, ha minden $\varepsilon > 0$-hoz létezik k, hogy minden, legalább k csúcsú G egyszerű gráfra $P(|f(G) - f(g(k,G))| > \varepsilon) \leq \varepsilon$, ahol $g(k,G)$ a G k csúcsú részgráfjai közül egyenletesen választott véletlen részgráfot jelöli.*

Kevésbé formálisan: tesztelhető egy gráfparaméter ha igaz az, hogy egy nagy gráfból kellő számú, de mindenképpen kis mintát véve a kis gráf paramétere közel lesz a nagy gráf paraméteréhez. Lovász László és kollégái [3] -ban több ekvivalens megfogalmazást adtak egyszerű gráfok tesztelhetőségére. Az egyik közülük különösen figyelemfelkeltő, nagy vonalakban azt mondja, hogy tesztelhető egy paraméter, ha ki lehet terjeszteni grafonokra úgy, hogy a kiterjesztett paraméter a következő normában folytonos:

15. Definíció. *Egy $W(x,y)$ grafon vágási-normája alatt a következőt értjük:*

$$||W||_\square := \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} W(x,y) dx dy \right|.$$

Ez az ekvivalens megfogalmazás voltaképpen azt jelenti, hogy gráfok kombinatorikus tulajdonságait klasszikus analízisbeli eszközökkel lehet vizsgálni.

4. Tesztelhetőség

Annak érdekében, hogy jól megértsük a fogalmakat, először megismétlem a [3] egyszerű gráfokra kimondott és bizonyított tesztelési tételét.

7. Tétel. *[3] 6.1-es tétele: Legyen f korlátos egyszerű gráfparaméter. Ekkor a következők ekvivalensek:*

(a) f tesztelhető egyszerű gráfparaméter.

(b) Minden $\varepsilon > 0$ esetén van olyan k pozitív egész, hogy minden legalább k csúcsú G egyszerű gráfra

$$|f(G) - \mathbb{E}(f(g(k,G)))| \leq \varepsilon.$$

(c) Tetszőleges (G_n) balról konvergens egyszerű gráfsorozatra, ahol $|V(G_n)| \to \infty$, az $f(G_n)$ sorozat szintén konvergens.

(d) Az f kiterjeszthető $\mathbb{W}_{[0,1]}$-beli grafonokra úgy, hogy az $\tilde{f}(W)$ kiterjesztett folytonos a vágási-normában és $\tilde{f}(W_G) - f(G) \to 0$, ha $|V(G)| \to \infty$.

(e) Minden $\varepsilon > 0$ valós számhoz van olyan $\varepsilon_0 > 0$ valós és n_0 pozitív egész szám, hogy tetszőleges legalább n_0 csúcsú G_1, G_2 egyszerű gráfra, melyekre $\delta_\square(G_1, G_2) < \varepsilon_0$ teljesül, azokra $|f(G_1) - f(G_2)| < \varepsilon$ szintén teljesül.

Elemezzük egy kicsit a tételt. Észrevehető, hogy a (c), (d), (e) ekvivalens jellemzés nem függ attól, hogy pontosan mi is a randomizálási eljárásunk. Mivel a mintavételezés nagyon természetes, ezért ez utóbbi megjegyzés egyszerű gráfokon nem tűnik annyira meglepőnek. Azonban, mint látni fogjuk, súlyozott gráfokat is vizsgálva a kérdés fontossá válik. A teljes megértés és egység miatt jobbnak láttam már itt tisztázni a kérdést. Kezdetnek egy másik randomizálási lehetőség:

16. Definíció. *Legyen G egyszerű gráf, k pedig egy természetes szám. Ekkor $\xi(k, G)$ jelölje azt a k csúcsú egyszerű gráfot, amit a következőképpen kapunk: G csúcsai közül kiválasztunk k darabot úgy, hogy egyesével választjuk a csúcsokat, minden választásnál G minden egyes csúcsát ugyanakkora valószínűséggel választjuk ki. Vagyis visszatevéssel húzunk k darabot G csúcsai közül. Ezt követően a kihúzott k darab csúcs mindegyikének megfeleltetünk egy pontot. Két pontot összekötünk, ha a nekik megfelelő G-beli csúcsok között futott él. Egyszerűség kedvéért ezt visszatevéses randomizálásnak nevezem.*

Vegyük észre, hogy az eredeti $g(k, G)$ randomizálás is felfogható így, azzal a különbséggel, hogy a már kiválasztott csúcsot nem választhatjuk ki újra, a többi közül választunk egyenletes valószínűséggel. Így az eredeti randomizálást nevezhetjük visszatevés nélküli randomizálásnak.

Fontos észrevétel, hogy tesztelésnek akkor van értelme, amikor a kivett minta sokkal kisebb, mint az eredeti gráf. Ekkor a két mintavételezés lényegében nem tér el egymástól, hiszen nagyon kicsi annak a valószínűsége, hogy lesz olyan csúcs, amit többször is kiválasztottunk. De a tesztelés pontos definíciójában szerephez jut az az eset, amikor közel akkora a vizsgálni kívánt gráf, mint a kivett minta. Ekkor a két randomizálás nem teljesen ugyanaz. Így ezt az utóbbi alternatív randomizálási definíciót látva már kicsit meglepőbb az a tény, hogy az ekvivalens jellemzés (c), (d) és (e) pontja nem függ a definiált randomizálástól.

Most vizsgáljuk meg, hogy a tétel mit mond: egy nagyon természetes mintavételezési eljárással definiálva a tesztelhetőséget azt kapjuk, hogy a tesztelhetőség ekvivalens a függvényről tett (c),(d),(e) állításokkal. Ezek a tesztelhető paraméter más és más tulajdonságát világítják meg, de mint említettem, a randomizálás módjának nincs szerepe. Így egy függvény tesztelhetősége beszédes név, azt jelenti, hogy ha kiderül egy függvényről, hogy tesztelhető (c,d,e állítás), akkor van olyan mintavételezés, ami teszteli a paramétert. A bizonyított tételből tudjuk, hogy a visszatevés nélküli mintavétel ezt mindig megteszi, kérdés, hogy milyen más mintavételezések teszik ugyanezt meg.

Utóbbi probléma megértéséhez közelebb visz bennünket az (e) ekvivalens megfogalmazás. Eszerint akkor tesztelhető egy gráfparaméter, ha elég nagy gráfokra folytonos a $\delta_\square$ metrikában. Így ha egy paraméterről tudjuk, hogy (c), (d), (e) egyike teljesül, továbbá egy randomizálási eljárásról tudunk egy a következő tételhez hasonló állítást, akkor állíthatjuk, hogy a paramétert lehet tesztelni az adott randomizálással:

8. Tétel. *[3] 2.9-es tétele: legyen G súlyozott gráf 1 csúcssúlyokkal és $[-1, 1]$-beli élsúlyokkal, továbbá G legyen legalább k csúcsú. Ekkor*

$$\delta_\square(G, Rand(k, G)) \leq \frac{10}{\sqrt{log_2 k}},$$

legalább $1-e^{\frac{-k^2}{2log_2 k}}$ *valószínűséggel, ahol* $Rand(k,G)$ *a* G *gráf* k *elemű csúcshalmazai közül egyenletesen választott halmaz által kifeszített részgráfot jelöli. (Vegyük észre, hogy ez egyszerű gráfokon egybeesik a* $g(k,G)$ *randomizálással.)*

A fenti tétel elég nagy k-ra egyfajta uniform (gráftól független) monotonitást jelent. A randomizált gráf annál közelebb lesz $\delta_\square$-távolságban egyre növekvő valószínűséggel az eredeti gráfhoz, minél nagyobb k.

Ezeken az okfejtéseken keresztül sikerült jobban megértenünk a tesztelhetőséget. A tesztelhetőség igazából a $\delta_\square$-metrikában való, gráfméretre tekintettel levő folytonosság. Tesztelni pedig azokkal a randomizálásokkal lehet, amik szerinti minta és az eredeti gráf $\delta_\square$-távolsága valószínűségben nullához tart gráftól független uniform módon.

Később látni fogjuk, hogy mivel az alternatív visszatevéses randomizálás megegyezik gráfból, mint lépcsős grafonból vett randomizálással, és arra általánosan van valószínűségben nullához tartást biztosító tétel, ezért egy tesztelhető gráfparamétert ezzel a randomizálással is lehet tesztelni.

Itt is látjuk, hogy a tesztelés gyakorlati kivitelezése többféle lehet. Érdekes lenne megvizsgálni, hogy lehet-e őket ügyesen karakterizálni. Azt, hogy melyiket érdemes választani azon múlik, hogy melyiknél gyorsabb a $\delta_\square$-távolság nullához való konvergenciája. Bár az itt szóba került két randomizálás a gyakorlatban fontos esetekben ugyanarra az eredményre vezet, mégis megjegyzek egy intuitív képet. Amikor többször kihúzok egy csúcsot a visszatevéses randomizálással, akkor az olyan, mintha egyszer húztam volna ki, és azt felfújtam volna. Így a visszatevéses mintavételt úgy képzelhetjük, hogy kihúztunk különböző csúcsokat, és azokat különböző mértékben felfújjuk. Egy korábbi megjegyzésből tudjuk, hogy ha mindegyik csúcsot ugyanakkorára fújjuk, akkor az eredeti és a felfújt gráf $\delta_\square$-távolsága 0. Itt ugyan nem az történt, de hihető, hogy nem lesz nagy az eredeti és felfújt $\delta_\square$-távolsága. Így erre a mintavételezésre úgy gondolhatunk, mint kicsit pazarlóbb mintavételre, amikor is k-nál voltaképpen kevesebb csúcsot, így kevesebb információt veszünk a gráfból. Ebből az intuícióból arra lehet következtetni, hogy egyszerű gráfoknál ha ezen két randomizálás közül kell választani, akkor érdemesebb az eredeti, visszatevés nélküli randomizálást használni.

Ezen előkészületek után általánosítom a tesztelési fogalmat súlyozott gráfokra, továbbá ismertetem a súlyozott gráfparaméterekről szóló tételt bizonyítással együtt. A bizonyítás szinte teljes egészében követi a [3]-ban található egyszerű gráfparaméterekről szóló tétel bizonyítását. Az elvégzett munka a tétel pontos kimondása és a súlyozott esetbeli bizonyításban előforduló kisebb nehézségek, hiányosságok pótlása volt.

Súlyozott gráfok tesztelésénél is több lehetőségünk van a randomizálás definíciójára. A tétel megfelelőjét a $\xi(k,G)$ visszatevéses mintavételezés általánosítására sikerült belátni.

17. Definíció. *Legyen* G *olyan* n *csúcsú súlyozott gráf, amelynek az élsúlyai a* $[0,1]$ *intervallumba esnek, csúcssúlyai tetszőlegesek. Ekkor* $\xi(k,G)$ *gráfon azt a* k *csúcsú egyszerű gráfot értjük, amelyet úgy kapunk, hogy a* G *csúcsaiból egymástól függetlenül visszatevéssel húzunk* k *darabot, úgy, hogy annak a valószínűsége, hogy a* G i*-dik csúcsát húzzuk épp,* $\frac{\alpha_i}{\sum_{i=1}^{n}\alpha_i}$ *. Ezt követően pedig az* i*-edik és* j*-edik húzásra kapott* h_i *és* h_j *csúcsok között* $\beta_{h_i h_j}$ *valószínűséggel húzunk be élet egymástól és a korábbiaktól is függetlenül.*

A definíció megengedi, hogy kétszer ugyanazt a csúcsot húzzuk ki. Ha ez megtörténik, közöttük semmiképpen nem húzunk élet. A fenti randomizálás, mint korábban már utaltam rá, azonos azzal, mintha a gráfnak megfelelő W_G grafonon a [3]-ban is szereplő grafonból történő egyszerű gráf randomizálást hajtanánk végre.

A randomizálásunk következménye, hogy tesztelhetőségnek ezen definíció mellett csak olyan súlyozott gráfoknál van értelme, amelyek élsúlyai $[0,1]$-be esnek. Jelöljük az ilyen gráfok halmazát $\mathbb{G}$ -vel. Innentől kezdve ha külön nem említem, súlyozott gráf alatt mindig ebbe az osztályba tartozó gráfot értek.

18. Definíció. *Az f csúcsok skálázására invariáns súlyozott gráfparaméter tesztelhető, ha minden $\varepsilon > 0$ esetén van olyan k pozitív egész, hogy ha $G \in \mathbb{G}$ olyan, hogy*

$$\max_i \frac{\alpha_i(G)}{\alpha_G} \leq \frac{1}{k},$$

akkor

$$\mathbb{P}(|f(G) - f(\xi(k,G))| > \varepsilon) \leq \varepsilon.$$

Vegyük észre, hogy az invariancia a csúcsok skálázására mindenképpen szükséges, hiszen két gráf között mintavétellel nem tudunk különbséget tenni. Ez összhangban van az elmélettel, hiszen - mint már korábban megjegyeztem - a grafonok elmélete invariáns a csúcsok skálázására. A paraméter értelmes egyszerű gráfokon is, miután felfoghatók speciális súlyozott gráfoknak. Emellett, ha egy paraméter ezzel a definícióval tesztelhető a $\mathbb{G}$ gráf halmazon és invariáns az élek skálázására, akkor tesztelhető általában is a súlyozott gráfokon. Mindamellett vegyük észre, hogy az egyszerű gráf tesztelhetőségének fogalmában az szerepelt, hogy minden, legalább k csúcsú gráfra igaznak kell lennie a jó közelítésnek. A fenti feltétel nem az összes, legalább k csúcsú súlyozott gráfon követeli meg a jó közelítést, hanem ezek közül is olyanokon, melyekben nincs domináns csúcssúly. A tétel általános formája:

9. Tétel. *Egy f korlátos súlyozott gráfparaméterre a következők ekvivalensek:*

(a) f tesztelhető súlyozott gráfparaméter.

(b) Minden $\varepsilon > 0$ esetén van olyan k pozitív egész, hogy ha $G \in \mathbb{G}$ olyan, hogy

$$\max_i \frac{\alpha_i(G)}{\alpha_G} \leq \frac{1}{k},$$

akkor $|f(G) - \mathbb{E}(f(\xi(k,G)))| \leq \varepsilon$.

(c) Tetszőleges $(G_n) \subset \mathbb{G}$ balról konvergens súlyozott gráfsorozatra, amelynek nincs domináns csúcssúlya, azaz

$$\max_i \frac{\alpha_i(G_n)}{\alpha_{G_n}} \to 0,$$

az $f(G_n)$ sorozat szintén konvergens.

(d) Az f kiterjeszthető $\mathbb{W}_{[0,1]}$-beli grafonokra úgy, hogy az $\tilde{f}(W)$ kiterjesztett folytonos a vágási-normában és $\tilde{f}(W_G) - f(G) \to 0$, ha

$$\max_i \frac{\alpha_i(G)}{\alpha_G} \to 0.$$

(e) Minden $\varepsilon > 0$ valós számhoz van olyan $\varepsilon_0 > 0$ valós és n_0 pozitív egész szám, hogy tetszőleges G_1, G_2 párra, melyekre

$$\max_i \frac{\alpha_i(G_1)}{\alpha_{G_1}} \leq \frac{1}{n_0}, \qquad \max_i \frac{\alpha_i(G_2)}{\alpha_{G_2}} \leq \frac{1}{n_0},$$

és $\delta_\square(G_1, G_2) < \varepsilon_0$ teljesül, azokra $|f(G_1) - f(G_2)| < \varepsilon$ szintén teljesül.

A csúcsokra vonatkozó skálainvariancia miatt feltehető, hogy $\sum_{i=1}^{n} \alpha_i = 1$. Ezt a tényt nem hangsúlyozom mindig, ahol esetleg hiányát érzi az olvasó, nyugodtan képzelje oda. Néhány jelölés:

$$\alpha_\Phi = \prod_{i=1}^{k} \alpha_{\phi(i)} \tag{1}$$

$$\mathrm{i}nj_\Phi(F,G) = \prod_{ij \in E(F)} \beta_{\phi(i)\phi(j)} \tag{2}$$

$$\mathrm{i}nd_\Phi(F,G) = \prod_{ij \in E(F)} \beta_{\phi(i)\phi(j)} \prod_{ij \in E(\bar{F})} (1 - \beta_{\phi(i)\phi(j)}) \tag{3}$$

Ezekkel a jelölésekkel a korábban bevezetett $t(F,G)$, illetve néhány rokon fogalom:

$$\mathrm{t}(F,G) = \sum_{\Phi} \alpha_\Phi \cdot \mathrm{i}nj_\Phi(F,G) \tag{4}$$

$$\mathrm{t}_{inj}(F,G) = \sum_{\Phi \in \mathrm{Inj}(F,G)} \alpha_\Phi \cdot \mathrm{i}nj_\Phi(F,G) \tag{5}$$

$$\mathrm{t}_{ind}(F,G) = \sum_{\Phi \in \mathrm{Inj}(F,G)} \alpha_\Phi \cdot \mathrm{i}nd_\Phi(F,G) \tag{6}$$

Az általánosítás kulcsa a fenti fogalmak és a randomizálási procedúra kapcsolatának pontos megértése volt. Vegyük észre, hogy $t_{ind}(F,G_n)$ körülbelül annak a valószínűségét adja meg, hogy a G_n-ből való randomizálási procedúra végén F-el izomorf gráfot kapunk. A formula nem más, mint a teljes valószínűség tétele a lehetséges húzássorozatokra. Egy F-ből G-be menő injektív leképezés, ha úgy tekintek az F csúcsaira, mint a húzás helyiértékeire, akkor épp egy húzássorozatot kódol. Tehát α_Φ épp a húzássorozat valószínűsége, ami pedig utána jön, az annak a feltételes valószínűsége, hogy ha adott a kihúzott csúcssorozat, mi annak a valószínűsége, hogy megkapom F-et. Nem véletlenül írtam, hogy csak a körülbelüli valószínűséget kapjuk így meg, hiszen F megkapható úgy is, hogy olyan koordinátákon húzom ki kétszer ugyanazt a csúcsot, amelyekhez tartozó F beli két csúcs között nem fut él, ekkor ezen két koordináta között 1 valószínűséggel nem húzok be élet, és ez rendjén is van. Ezen eseményeknek nem injektív leképezések felelnek meg, ezekre pedig nem szummázunk. Azonban könnyen látható, hogy ha G_n csúcsainak száma végtelenhez tart, miközben a kivett minta csúcsszáma konstans, akkor annak a valószínűsége, hogy többször kihúzom ugyanazt a csúcsot, 0-hoz tart. Ezt a triviális állítást egy fokkal továbbgondolja a következő állítás, mely önmagában is hasznos:

1. Állítás. *Ha G_n olyan gráfsorozat, amelyben a csúcssúlyok összege 1, további az élsúlyok $[0,1]$-beliek, és $\max_i \alpha_i(G_n) \to 0$, akkor*

$$|t(F,G_n) - t_{inj}(F,G_n)| \to 0 \qquad (n \to \infty).$$

Bizonyítás:

A két tag különbsége épp a nem injektív leképezéseken való összegzés. Mivel ezt a különbséget a hátsó 0 és 1 közötti β_{ij} szorzók elhagyása növeli, ezért elég belátni, hogy

$$\sum_{\Phi \notin \mathrm{Inj}(F,G)} \alpha_\Phi \to 0.$$

Ez pedig a kódolásunknak köszönhetően a már korábban is felmerült probléma, miszerint annak a valószínűsége, hogy az összes kihúzott csúcs különböző, 1 hez tart ha $n \to \infty$. Ha n elég nagy, akkor $\max_i \alpha_i(G_n) < c$, így

$$\mathbb{P}(\text{mindegyik különböző}) \geq 1 \cdot (1-c) \dots (1-(k-1)c) \geq (1-(k-1)c)^k$$

Így egyre kisebb c választással látjuk, hogy a kérdéses valószínűség 1-hez tart. ■

Nézzünk egy másik állítást:

2. Állítás. *Tetszőleges $\Phi \in Inj(F,G)$ esetén*

$$inj_\Phi(F,G) = \sum_{F' \supseteq F} ind_\Phi(F',G).$$

Bizonyítás:

Először kis pontosító magyarázat, a jobb oldalon olyan gráfokon szummázunk, amelyeket élbehúzással meg lehet kapni F-ből (magát F-et is ideszámítva). A jobb oldalon minden tagban minden, az F k darab csúcsán elképzelhető potenciális β_{ij} él szerepel szorzóként vagy β_{ij} vagy $(1-\beta_{ij})$ -ként. (Ezek az élsúlyok a G gráf élsúlyai, méghozzá a Φ leképezés által kiválasztott csúcsok között futó élek súlyáról van szó). Mivel csak olyan gráfokra szummázunk, amik F-nél nagyobbak, ezért mindegyik tagban az F összes élének megfelelő G-beli β_{ij} él ugyanilyen formában szerepel szorzótényezőként. Így a minden tagban közös csúcsszorzó ezekkel a szorzótényezőkkel együtt kiemelhető, a kiemelt tag így összesen épp a bal oldal, vagyis $t_{inj}(F,G)$. Így nincs más dolgunk, mint belátni, hogy a kiemelés után megmaradt tagok összege éppen 1. Ezt élszám szerinti indukcióval láthatjuk be könnyen, méghozzá azon élek számára indukciózunk melyek ahhoz hiányoznak, hogy teljes gráfot csináljunk F-ből. Ha egy él hiányzik, és az ezekhez tartozó élsúlyokat rendre $\beta_1, \beta_2, \dots, \beta_k$-val jelöljük, akkor a maradék tag az összes lehetséges következő alakú tagok összege lesz: $a_1(\beta_1) \cdot a_2(\beta_2) \cdot \dots \cdot a_k(\beta_k)$ ahol a megfelelő a_i függvény vagy x vagy $1-x$. Az ilyen tagokon bevezethetünk egy párosítást: azon két tag alkot párt melyek szorzótényezőjében minden a_i függvény ugyanaz, kivéve a_1-t, ami pont különbözik. Egy ilyen párból kiemelve a közös részt, a maradék épp 1 lesz. Ha minden párt összevonunk ily módon, akkor éppen a $(k-1)$ megmaradó éllel kapcsolatos problémához jutunk, ami indukciónk szerint épp 1. ■

Ebből pedig adódik:

$$\mathrm{t}_{inj}(F,G) = \sum_{F' \supseteq F} \mathrm{t}_{ind}(F',G).$$

Ha a fenti eredményt behelyettesítjük a következő formula jobb oldalába, meggyőződhetünk annak azonosság voltáról:

$$\mathrm{t}_{ind}(F,G) = \sum_{F' \supseteq F} (-1)^{|E(F')\setminus(E(F)|} \, \mathrm{t}_{inj}(F',G).$$

Konklúzióként:

3. Állítás. *$t_{inj}(F,G)$ pontosan akkor konvergens minden F-re, ha $t_{ind}(F,G)$ is konvergens minden F-re.*

Persze, hiszen a közöttük levő formulák minden rögzített F-re konstans tag összegét tartalmazzák.

Megjegyzem, hogy a fenti állítások a [5] 2.1-es lemmájának inspirációjára készültek. Az állítások megfelelői ki vannak mondva súlyozott gráfokra, de bizonyítva csak egyszerű gráfokra lettek. Az itt leírt egyszerű bizonyítások más alapokra épülnek.

Az eddigi megállapítások birtokában már le tudjuk ellenőrizni, hogy a bizonyítás működik a súlyozott esetre is. A bizonyítás előtt néhány szükséges tételt ismertetek [3]-ból.

10. Tétel. *[3] 4.7-es tételének (ii) pontja: Legyen k pozitív egész. Ha $U \in W_{[0,1]}$, akkor legalább $1 - e^{\frac{-k^2}{2log_2 k}}$ valószínűséggel:*

$$\delta_\square(U, \xi(k, U)) \leq \frac{10}{\sqrt{log_2 k}},$$

ahol $\xi(k, U)$ a súlyozott gráf esetben már ismert randomizálás grafonokra vett általános formája: k darab $[0, 1]$-be eső pontot – $x_1, \ldots, x_k$-t – generálunk függetlenül, egyenletes eloszlással. Ezt a k pontot felrajzoljuk egy papírra, és i, j között $U(x_i, x_j)$ valószínűséggel húzunk élet.

Vegyük észre, hogy a fenti becslés nem függ U-tól, vagyis egyenletesen tudjuk közelíteni véletlen mintával a grafonokat.

1. Lemma. *[3] 5.3-as lemmája: Legyen (G_n) egy súlyozott gráfokból álló sorozat, melynek az élsúlyai korlátosak, továbbá $\max_i \frac{\alpha_i(G_n)}{\alpha_{G_n}} \to 0$. Ekkor ha valamely U grafonra*

$$\delta_\square(U, W_{G_n}) \to 0,$$

akkor (G_n)-nek létezik átcímkézése úgy, hogy az átcímkézett (G'_n)-re igaz:

$$||U - W_{G'_n}||_\square \to 0.$$

A tesztelhetőséget jellemző 9. Tétel bizonyítása:

$(a) \Rightarrow (b)$: (b) feltétel teljesüléséhez tetszőleges ε -hoz keresünk k küszöbszámot. Ehhez válasszunk ε_0 -t úgy, hogy $(1 - \varepsilon_0) \cdot \varepsilon_0 + \varepsilon_0 \cdot 2M$ legyen kisebb ε -nál, ahol M a paraméter korlátja. Ekkor ε_0 -hoz (a) által biztosított küszöbszám jó lesz.

$(b) \Rightarrow (c)$ Legyen G_n a (c) állításnak megfelelően domináns csúcssúllyal nem rendelkező konvergens gráfsorozat. Legyen ε tetszőleges. Ehhez a ε -hoz válasszunk k-t a (b) állítás segítségével. Ekkor elég nagy n-re: $|f(G_n) - \mathbb{E}(f(\xi(k, G_n)))| \leq \varepsilon$. Másrészt a konvergencia definíciójából következik, hogy $t(F, G_n)$ konvergens minden k csúcsú F egyszerű gráfra. Az állításaink miatt így $t_{ind}(F, G_n)$ is konvergens minden k csúcsú F egyszerű gráfra. Jelöljük ezt a limeszt $t_{ind}(F)$-el. Ugyanakkor $t_{ind}(F, G_n)$ n növekedésével egyre inkább valószínűség lesz, annak a valószínűsége, hogy G_n -ből k csúcsú egyszerű gráfot randomizálva a végeredmény F-el izomorf gráf lesz. Így

$$\mathbb{E}(f(\xi(k, G_n))) \to \sum_{\text{F k csucsu graf}} t_{ind}(F) \cdot f(F) := a_k,$$

használva a teljes valószínűség tételét korlátos számú tagot tartalmazó teljes eseményrendszerre. Továbbá

$$|f(G_n) - a_k| \leq |f(G_n) - \mathbb{E}(f(\xi(k, G_n)))| + |\mathbb{E}(f(\xi(k, G_n))) - a_k| \leq 2\varepsilon \text{ ha n nagy.}$$

Így $f(G_n)$ sorozatot vizsgálva mindig találunk olyan számot amelytől egy idő után maximum 2ε -t tér el. Ez pedig implikálja a konvergenciát.

$(c) \Rightarrow (e)$: Tegyük fel, hogy (e) nem áll fenn. Ekkor van olyan ε, G_n és $G_n^{'}$ sorozat, hogy mindkét sorozat domináns csúcssúlya nullához tart, továbbá $\delta_{\square}(G_n, G_n^{'}) \to 0$, és $|f(G_n) - f(G_n^{'}) \geq \varepsilon$. Feltehető, hogy mindkét gráfsorozat konvergens kihasználva, $W_{[0,1]}$ kompaktságát (minden sorozatnak van konvergens részsorozata). Ezért a feltételeink miatt az összefűzött $G_1, G_1^{'}, G_2, \ldots$ sorozat is konvergens. Így (c) miatt ezen sorozat mentén az f értékei is konvergensek. Ez azonban ellentmond annak, hogy $|f(G_n) - f(G_n^{'})| \geq \varepsilon$.

$(e) \Rightarrow (a)$: Tegyük fel, hogy (a) nem áll fenn. Ekkor létezik ε és G_n gráfsorozat, hogy $\max_i \frac{\alpha_i(G_n)}{\alpha_{G_n}} \leq \frac{1}{n}$ és legalább ε valószínűséggel $|f(G_n) - f(\xi(n, G_n))| > \varepsilon$ minden n-re. Most használjuk erre a ε -ra az (e) pontot, így kapjuk a megfelelő tulajdonsággal rendelkező n_0 és ε_0-t. Továbbá [3] 4.7-es tétele alapján (itt használjuk, hogy a randomizálásunk nem más, mint gráfból, mint grafonból vett randomizálás) $\delta_{\square}(G_n, \xi(n, G_n))$ nullához tart valószínűségben. Speciálisan nagy n-re $\delta_{\square}(G_n, \xi(n, G_n)) < \varepsilon_0$ legalább $1 - \frac{\varepsilon}{2}$ valószínűséggel. Használva ε_0 és n_0 választását kapjuk, hogy $|f(G_n) - f(\xi(n, G_n))| < \varepsilon$ legalább $1 - \frac{\varepsilon}{2}$ valószínűséggel. Ez pedig ellentmond annak, hogy legalább ε valószínűséggel az ellenkezője igaz.

Ezt követően belátjuk, hogy (d) is ekvivalens az előzőekkel:

$(e) \Rightarrow (d)$: $W \in \mathbb{W}_{[0,1]}$ grafonra úgy terjesztjük ki f-et, hogy keresünk hozzá konvergáló G_n súlyozott gráfsorozatot, melyben a domináns csúcssúly 0-hoz tart, és a (c) alapján konvergens $f(G_n)$ határértékeként definiáljuk $\tilde{f}(W)$-t. Az (e) miatt a definíció korrekt, nem függ az érték a definiáló sorozattól (két sorozat közel lesz egymáshoz a $\delta_{\square}$ metrikában - háromszögegyenlőtlenség -, így az értékek is közel lesznek egymáshoz). Továbbá ezzel minden $W \in \mathbb{W}_{[0,1]}$ grafonra kiterjesztettük a függvényt, hiszen már egyszerű gráfokkal is meg lehet őket közelíteni.

Először a folytonosságot igazoljuk. Legyen $\varepsilon > 0$. Alkalmazzuk az (e) pontot $\frac{\varepsilon}{3}$-ra, így kapjuk $\varepsilon^{'}$-t és n_0-t. Belátjuk (elég a folytonossághoz), hogy $|\tilde{f}(W) - \tilde{f}(W^{'})| \leq \varepsilon$, ha $||W - W^{'}||_{\square} \leq \frac{\varepsilon^{'}}{3}$. Ehhez nézzünk egy W-hez konvergáló domináns csúcssúly nélküli G_n gráfsorozatot (ha mást nem, egyszerű gráfokból álló sorozatot), ebből a sorozatból vegyünk ki egy olyan G elemet, ami már három dolgot tud (egy idő után mindháromnak egyszerre teljesülni kell): a dominánccs csúcssúly legyen kisebb, mint $\frac{1}{n_0}$, $\delta_{\square}(G, W) < \frac{\varepsilon^{'}}{3}$ (mivel ez a sorozat W közelítése, ez nem gond) és végül $|f(G) - \tilde{f}(W)| \leq \frac{\varepsilon}{3}$. Hasonlóan válasszunk $G^{'}$-t $W^{'}$-höz. Ekkor

$$\delta_{\square}(G, G^{'}) \leq \delta_{\square}(G, W) + \delta_{\square}(W, W^{'}) + \delta_{\square}(W^{'}, G^{'}) \leq \varepsilon^{'}.$$

A középső tag becslésénél felhasználtuk azt is, hogy a $\delta_{\square}$-távolságot két grafon között felülbecsli triviálisan a vágási-norma. Ezért (e) miatt $|f(G) - f(G^{'}| \leq \frac{\varepsilon}{3}$. Így

$$|\tilde{f}(W) - \tilde{f}(W^{'})| \leq |\tilde{f}(W) - f(G)| + |f(G) - f(G^{'})| + |f(G^{'}) - \tilde{f}(W^{'})| \leq \varepsilon.$$

Hátra van annak az igazolása, hogy ez a kiterjesztés bizonyos értelemben tényleg kiterjesztés (limeszben legalábbis). Indirekten tegyük fel, hogy van olyan G_n domináns csúcssúly nélküli sorozat, amely mentén nem igaz, hogy $\tilde{f}(W_{G_n}) - f(G_n) \to 0$. Ez azt jelenti, hogy ennek van olyan részsorozata, amelyre igaz, hogy az előbbi távolság a részsorozat mentén végig nagyobb, mint ε. Kompaktságot kihasználva ennek a részsorozatnak van olyan további részsorozata, ami valamilyen W grafonhoz tart. A jelölést egyszerűsítve tegyük fel, hogy ez maga G_n. Ekkor (c)-t használva $f(G_n) \to \tilde{f}(W)$, továbbá $\tilde{f}$ folytonos, ezért $\tilde{f}(W_{G_n}) \to \tilde{f}(W)$. Ez pedig ellentmondás.

$(d) \Rightarrow (c)$: Legyen G_n domináns csúcssúly nélküli konvergens súlyozott gráfsorozat. Legyen W a limesze ennek a sorozatnak. Definíció szerint

$$\delta_\square(W_{G_n}, W) \to 0.$$

Így a [3] 5.3-as lemmája alapján létezik a (G_n) súlyozott gráfsorozatnak olyan $(G_n^{'})$ címkézése, hogy $||W_{G_n} - W||_\square \to 0$. Így $\tilde{f}$ folytonossága miatt: $\tilde{f}(W_{G_n^{'}}) - \tilde{f}(W) \to 0$. Ezzel kész vagyunk, hiszen: $f(G_n) - \tilde{f}(W_{G_n^{'}}) = f(G_n) - \tilde{f}(W_{G_n})$. Utóbbi pedig (d) miatt 0-hoz tart. Így $f(G_n)$ valóban konvergens. ■

Most látjuk, hogy egy konkrét randomizálásra igaz a tétel. Az egyszerű esethez hasonlóan itt is elmondható, hogy ha egy f korlátos, súlyozott gráfparaméterre, mely invariáns a csúcsok skálázására (ahogy az egész elmélet) igaz a (c), (d), (e) állítások valamelyike, illetve a választott mintavételezés a $\delta_\square$ metrikában egyenletesen nullához tart, akkor lehet a randomizálással tesztelni a gráfparamétert. Ez a gyakorlatban elég. Mégis a jobb megértés érdekében kiemelem a választott randomizálás azon tulajdonságait, ami miatt igaz a teljes tétel:

A $(b) \Rightarrow (c)$ implikációnál láthatjuk, hogy szükség van arra, hogy a kapott minta egyszerű gráf, illetve arra, hogy rögzített k csúcsú F egyszerű gráf esetén domináns csúcssúly nélküli (G_n) konvergens súlyozott gráfsorozat mentén $t_{ind}(F, G_n)$ az F megkapásának valószínűségéhez tartson. Továbbá a $(c) \Rightarrow (e)$ implikációnál szükségünk van [3]-beli 4.7 Tétel analogonjára, vagyis arra, hogy a gráf és a kivett minta $\delta_\square$-távolsága valószínűségben uniform módon 0-hoz tart (gráftól függetlenül). Egyéb állítást nem használunk ki a randomizálással kapcsolatban.

Mint fent említettem, ha egy paraméterről kiderül, hogy tesztelhető a fenti értelemben, akkor minden olyan randomizálással lehet tesztelni, amire teljesül az uniform nullához tartás. Ez a megjegyzés a súlyozott esetben különösen fontos, hiszen mint láttuk, a tétel bizonyításához szükség volt arra, hogy egyszerű gráfot randomizáljunk. Én intuitíven úgy gondolom, hogy ha randomizálás helyett behúznánk az éleket súlyozottan, akkor egy jobban közelítő (gyorsabb az uniform nullához tartás) gráfsorozatot kapunk. A következő tétel biztosítja, hogy lehet súlyozott gráffal is tesztelni:

11. Tétel. *[3] 4.7-es tétel (i) pontja: Legyen k pozitív egész. Tetszőleges U grafonra legalább $1 - e^{\frac{-k^2}{2log_2 k}}$ valószínűséggel*

$$\delta_\square(U, \mathbb{H}(k, U)) \leq \frac{10}{\sqrt{log_2 k}} ||U||_\infty,$$

ahol $\mathbb{H}(k, U)$ a grafonból történő súlyozott mintavételt jelöli.

Továbbá ha be tudjuk látni a paraméterről, hogy folytonos a $\delta_\square$-metrikában növekedő gráfokon, vagyis az (e) állítást, akkor utóbbi tétel segítségével tesztelési eredményt kapunk az uniform élkorláttal rendelkező gráfosztályokon is.

Kerek a világ! A $\mathbb{G}$ osztálynak része az egyszerű gráfok osztálya. Teljesül amit elvárunk: a $\mathbb{G}$-n való tesztelhetőségből következik az egyszerű gráfokon való tesztelhetőség, hiszen az előbbi (e) ekvivalens jellemzéséből következik utóbbi (e) ekvivalens jellemzése.

Úgy gondolom a $\mathbb{G}$ osztály nem különleges. Mint fent említettem a tesztelési tétel más élsúlykorláttal rendelkező gráftereken (élek korlátossága fontos) a $t(F, G)$ bonyolultabb definíciója miatt problémásabb. Ötletként felmerült bennem, hogy $t(F, G)$-t lehetne úgy módosítani, hogy más gráfosztályokból való módosított egyszerű mintavétellel összhangban legyen (ha az élek $[A, B]$ intervallumba esnek, akkor A vagy B súlyú éleket

randomizálunk). Mindenesetre vagy a bonyultabb definíciók kezelésével, vagy $t(F,G)$ ekvivalens módosításával, vagy ügyes normálással, várható a tesztelési tétel általánosítása általánosabb gráfosztályokra.

5. Tesztelhető paraméterek

Ebben a fejezetben néhány súlyozott gráfparaméter tesztelhetőségét vizsgálom.

19. Definíció. *Klasztertérfogattal súlyozott mincut:*

$$\mu_k(G) = \min_{P_k} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{1}{\alpha_{V_i} \cdot \alpha_{V_j}} \cdot e_G(V_i, V_j),$$

ahol $P_k = (V_1, \ldots, V_k)$ a csúcsok k partíciója.

Vegyük észre, hogy ez a paraméter invariáns a csúcsok skálázására. Korlátosságát úgy láthatjuk be, hogy élei súlyának növelése 1-ig csak növeli a paraméter értékét; az n csúcsú teljes gráfon pedig csúcssúlyoktól függetlenül az értéke $\binom{k}{2}$. Így $\mathbb{G}$ osztályon felmerül a tesztelhetőség. Azonban tudjuk, hogy ha ezen az osztályon tesztelhető, akkor az egyszerű gráfokon is annak kell lennie. De ez a paraméter ott sem tesztelhető.

Tegyük fel indirekten, hogy tesztelhető. Rögzítsünk egy kicsi ε-t. A tesztelhetőség definíciója alapján ekkor van olyan m, hogy minden legalább m csúcsú gráfból m elemű mintát véve a mintán számolt gráfparaméter értéke $1-\varepsilon$ valószínűséggel legfeljebb ε-ra tér el az eredeti gráf paraméterétől. Adok egy m-nél nagyobb csúcsszámmal rendelkező gráfot, amelyre a fenti jól közelíthetőség nem igaz. G n csúcsú gráf álljon egy $n-k+1$ csúcsú teljes gráfból és $k-1$ darab egymással nem összekötött pontból, melyek egy éllel kapcsolódnak a nagyobb komponenshez. Ekkor G-n a paraméter értéke közel 0 (ha n tényleg nagy), de az m elemű mintánk n növelésével 1 hez tetszőleges közeli valószínűséggel m csúcsú teljes gráf lesz, amin a paraméter értéke $\binom{k}{2}$. Ez pedig ellentmond a tesztelhetőségnek.

Nézzünk egy másik példát:

20. Definíció. *Csúcssúlyokkal súlyozott mincutnak nevezzük a következő paramétert:*

$$f_k(G) := \min_{P \in P_k} \frac{1}{\alpha_G^2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} e_G(V_i, V_j)$$

Ez a paraméter is invariáns a csúcsok skálájára, továbbá korlátos, így tudunk tesztelhetőségről beszélni. A paraméter értékét felülbecsülhetjük egy $k-1$ különálló pontból és a maradék halmazból álló partícióval, amit az élek 1-el való felül becslése után tovább becsülhetünk a következő módon:

$$f_k(G) \leq (k-1) \cdot \frac{\alpha_{max}(G)}{\alpha_G} \cdot \frac{\alpha_G}{\alpha_G} + \binom{k-1}{2} \cdot \left(\frac{\alpha_{max}(G)}{\alpha_G} \right)^2.$$

Ebből következik, hogy az olyan (G_n) sorozatok mentén, ahol a domináns csúcssúly 0-hoz tart, a paraméter értéke is 0-hoz tart, ebből pedig az ekvivalens állítások (c) pontja

miatt következik a paraméter tesztelhetősége. Mindazonáltal ez a tesztelhetőség triviális, hiszen a (G_n) sorozatok mentén történő nullához tartásra uniform felső becslésünk van a domináns csúcssúlytól függően. Így egyszerűen csak arról van szó, hogy a 0 értékkel jól becsüljük a paramétert, ha kicsi a domináns csúcssúly.

Az eddigi paraméterek nagyon lokálisak voltak, így nem is nagyon várhattunk tesztelhetőséget (utóbbinál triviális tesztelhetőség ugyan van, de az nem túl érdekes). Ha a paramétereket a partícióra vett megkötésekkel globálisabbá tesszük, akkor jó eséllyel tesztelhető gráfparamétereket kapunk.

Ehhez jelölje P_k^c a csúcsok olyan $(V_1, \ldots, V_k)$ k-partícióit, melyekben $\frac{\alpha_{V_i}}{\alpha_G} \geq c$ $(i = 1, \ldots, k)$, ahol $c \leq \frac{1}{k}$ rögzített konstans. Továbbá legyen $a = (a_1, \ldots, a_k)$ valószínűségi eloszlás az $\{1, \ldots, k\}$ halmazon. P_a álljon a V olyan k-partícióiból, amelyekre:

$$\left(\frac{\alpha_{V_1}}{\alpha_G}, \ldots, \frac{\alpha_{V_k}}{\alpha_G}\right)$$

körülbelül a eloszlású.

Továbbá legyen $f_k^c(G)$, $f_k^a(G)$, illetve $\mu_k^c(G)$, $\mu_k^a(G)$ a megfelelő paraméterek azon módosulatai, amelyeknél a minimumot csak P_k^c-, illetve P_k^a-beli partíciókon tekintjük. Ezek a módosulatok már tesztelhető paraméterek. Ez a [4]-ben kiépített elméletből kijön. Bár bizonyításuk pontosan ki van dolgozva, a dolgozatban csak a bizonyítások alapgondolatát közlöm. A szükséges fogalmakat is csak körülbelül van lehetőségem ismertetni. Minden nem részletezett jelölés megtalálható [4]-ben. Lássuk a szükséges tételeket:

12. Tétel. *[4] 2.14 Tétel: Legyen (G_n) egyenletesen korlátos élsúlyú, domináns csúcssúly nélküli súlyozott gráfsorozat. Ekkor a következők ekvivalensek:*

1. *A (G_n) sorozat balról konvergens.*
2. *(G_n) hányadosai (faktorgráfok) konvergensek a d_1^{Hf} Hausdorff-távolságban.*
3. *A (G_n) sorozat mikrokanonikus alapállapoti energiája konvergens.*

13. Tétel. *[4] 2.15 Tétel, részlet: Legyen (G_n) egyenletesen korlátos élsúlyú domináns csúcssúly nélküli súlyozott gráfsorozat. Ekkor a (G_n) sorozat alapállapoti energiái konvergensek.*

Vegyük észre, hogy $f_k^a(G)$ felfogható egy azonosan 0 mágneses mezőjű, és egy $(k \times k)$-s J kölcsönhatás mátrixú rendszer mikrokanonikus alapállapoti energiájaként, ahol J a következő: $J_{ii} = 0$ $(i = 1, \ldots, k)$-ra és $J_{ij} = -1/2$ ha $(i \neq j)$:

$$f_k^a(G) = \min_{\Phi \in \Omega_a(G)} E_\Phi(G, J, 0),$$

ahol $\Omega_a(G)$ az olyan $V(G) \to \{1, \ldots, k\}$ leképezések halmaza, amelyekre $\forall i$-re

$$\left| \sum_{u \in \Phi^{-1}(i)} \alpha_u(G) - \alpha_G a_i \right| \leq \alpha_{max}(G).$$

Egyszerűbben elmondva, a leképezéseket partíciónak gondolva, azon partíciókat gyűjtjük össze, amelyek közel a eloszlásúak a csúcssúlyozott gráfon.

Így a $f_k^a(G)$-re a [4] 2.14 Tételből közvetlenül következik a tesztelhetőség előző fejezetben igazolt (c) ekvivalense. Az $f_k^c(G)$ paraméter tesztelhetősége egy kicsit összetettebb.

A [4] 2.15-ös Tétel grafonokra történő általánosításon keresztül kerül bizonyításra ([4] 3.5 Tétel). Ezt a bizonyítást lehet úgy módosítani, hogy belássuk $f_k^c(G)$ tesztelhetőségét is.

$\mu_k^a(G)$ tesztelhetősége a következő triviális összefüggésen múlik:

$$\mu_k^a(G) = \min_{G/P \in \hat{S}_a(G)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij}(G/P),$$

ahol $\hat{S}_a(G)$ azon k-faktorok halmaza, amelyeknél a faktorgráf csúcssúlyai (az $\frac{\alpha_{V_i}}{\alpha_G}$ mennyiségek) közel a-eloszlásúak.

A $\mu_k^a(G)$ paraméter tesztelhetőségét ezúttal is az ekvivalens állítások (c) pontjával lehet igazolni. Annyit megjegyeznék, hogy domináns csúcssúly nélküli (G_n) sorozatok mentén $\mu_k^a(G_n)$ konvergenciája a fenti azonosság miatt a faktorgráfok élsúlyain múlik. Róluk pedig a [4] 2.14 Tétel alapján tudjuk, hogy d_1^{Hf}-távolságban konvergensek. Ebből jön ki a tesztelhetőség. Ugyanúgy, mint az előző esetben, ebből $\mu_k^c(G_n)$ tesztelhetősége is következik.

6. Kezdetleges tesztek, komolyabb tesztek lehetősége

Mathematica-ban dolgoztam. Beprogramoztam $\mathbb{G}$ elemeiből a két említett randomizálási procedúrát: a visszatevésest és a visszatevés nélkülit. Az első esetben nem kell mást tenni, mint a csúcssúlyok alkotta vektornak elkészíteni a kumulatív vektorát és generálni a $(0, \alpha_G)$ intervallumból k darab véletlen számot, aztán bináris kereséssel a kumulált listában megkeresni a generált számoknak megfelelő súlyt. Végül a kiválasztott csúcsok között a megfelelő valószínűségekkel éleket kell randomizálni. Tegyük fel, hogy a mintagráf csúcsszáma konstans, és a súlyok K bitesek. Ekkor az első fázis $O(n^2)$ művelettel megoldható (ahol n jelöli a gráf csúcsszámát). Továbbá ha feltesszük, hogy a véletlen szám generálása problémamentes, akkor a második fázis is megy $O(n^2)$ művelettel. A harmadik fázis független n-től, így ha véletlen szám generálása problémamentes, akkor konstans művelettel megoldható. Tekintve, hogy az elmélet sűrű gráfokon használható, kijelenthetjük, hogy a randomizálás lineáris az inputban a fenti feltételek mellett. Megjegyzem, hogy a műveleti igény kiszámításánál nem törekedtem a pontosságra, célom megfelelően kicsi felsőbecslés volt. Emellett biztos vagyok benne, hogy a randomizálást ügyesebb algoritmussal gyorsabban is meg lehet oldani. Továbbá a randomizálási algoritmust a konkrét feladatra érdemes szabni.

A második randomizálás majdnem így megy, annyi a különbség, hogy egyesével generálom a véletlen számokat, és megnézem, hogy az épp kisorsolt csúcsot kihúztam-e már. Ha igen, akkor újra sorsolok egészen addig, amíg egy még ki nem húzott csúcsot nem sikerül kivennem. Itt felhasználtam, hogy az egyenletes húzással kapott csúcs azon feltétel mellett, hogy eddig ki nem húzott csúcsot húztam, egyenletes eloszlású a ki nem húzott csúcsokon. Gyakorlati problémánál, amikor a tesztelendő gráf nagy a mintához képest, nincs különbség a két algoritmus között. Azért programoztam két algoritmust, mert összemérhető esetben kíváncsi voltam a randomizálások viszonyára.

Ezen kívül beprogramoztam egy $f_2^c(G)$-t és egy $\mu_2^c(G)$-t számoló algoritmust. Ezeket a lehető legegyszerűbben oldottam meg, a csúcsok összes részhalmazát végignézik, és úgy minimalizálnak. Erre azért volt szükség, mert a csúcsok súlyozása miatt bármelyik csúcsrészhalmaz lehet elég nagy. Ezekkel a kezdetleges algoritmusokkal a programom 20 csúcsú mintagráfon körülbelül 20 percig futott. Ezen a ponton körülnéztem a paraméterek szakirodalmában. Láttam, hogy nehéz problémákról van szó, így nagyságrendi javulást

nem várhatok. Nyerhettem volna időt az algoritmus finomításával, például a csúcssúlyok előrendezésével, továbbá Mathematica helyetti C++ használattal, de úgy gondoltam, hogy nem éri meg az erőbedobást, úgyis csak kezdetleges eredményeket kaphatok. Így 20 csúcsú mintákat használva végeztem kezdetleges teszteket. A kivett minta alacsony csúcsszáma miatt a kapott eredményeket óvatosan kell kezelni. Egy kiegészítő információ; a tesztelési tételben kulcsszerepet játszó $\delta_{\square}$-távolságra adott valószínűségi becslés $k = 2^{20}$ körül lesz releváns. Ha külön nem említem, akkor az $f_2^{\frac{1}{4}}(G)$-t teszteltem.

Az első tesztben négy különböző méretű gráfot generáltam. A nagy gráfban az egyik komponens 2000, a másik 1600 csúcsból állt, komponensen belül 0.9, közöttük 0.6 valószínűséggel húztam be minden egyes élet egymástól függetlenül. A közepes és a kis és pici gráf ugyanígy képződött csak a komponensek mérete volt eltérő: (1000,800), (500,400), (50,40). Minden gráfhoz két csúcssúlyt is hozzárendeltem, az azonosan egy súlyvektort, illetve egy független E(0,1) elemekből álló súlyvektort. Így voltaképpen 8 gráfunk van. Ezek olyan gráfok, melyekről tudjuk, hogy nagyjából a két definiáló komponens állítja be a mincutot, így ezeken kiszámoltam az értékeket és vizsgáltam, hogyan közelíti a gráfokból vett minta. Minden gráfból 20 elemű mintát vettem. Kétféle randomizálásra is lefuttattam a tesztelést: visszatevésre és visszatevés nélkülire. Így összességében 16-féle vizsgálatot végeztem, mindegyiket 10-szer hajtottam végre, rand1 jelöli a visszatevéses randomizálását, rand2 a visszatevés nélküli randomizálást. Nagy elemszámnál a rand1 és rand2 ugyanazt adja, több független kísérletnek érdemes tekinteni, a pici gráfnál már érdekes a különbségüket vizsgálni.

(2000,1600)	nemekvi/rand1	nemekvi/rand2	ekvi/rand1	ekvi/rand2
érték	0.1478	0.1478	0.1482	0.1482
átlag	0.1275	0.1283	0.1235	0.1180
szórás	0.0099	0.0084	0.0113	0.0066
(1000,800)	nemekvi/rand1	nemekvi/rand2	ekvi/rand1	ekvi/rand2
érték	0.1471	0.1471	0.1480	0.1480
átlag	0.1198	0.1233	0.1233	0.1248
szórás	0.0082	0.0051	0.0070	0.0079
(500,400)	nemekvi/rand1	nemekvi/rand2	ekvi/rand1	ekvi/rand2
érték	0.1488	0.1488	0.1483	0.1483
átlag	0.1195	0.1215	0.1183	0.1245
szórás	0.0086	0.0126	0.0086	0.0103
(50,40)	nemekvi/rand1	nemekvi/rand2	ekvi/rand1	ekvi/rand2
érték	0.1389	0.1389	0.1385	0.1385
átlag	0.1185	0.1183	0.1145	0.1120
szórás	0.0087	0.0060	0.0088	0.0073

A teszt az alacsony mintaszámhoz képest elég jó közelítést mutat. Ugyan a minták mindig az igazi érték alatt vannak, de a szórásuk elég kicsi. Megjegyzem, hogy ez az eredmény nem igazán a tesztelhetőséget igazolja, inkább a nagy számok törvényének érvényesüléséről van szó. A generált gráfok szerkezete körülbelül ugyanaz, tekinthetünk úgy rájuk, mint egymás felfújtjaira kis zajjal terhelve. A vett minta pedig azon múlik, hogy a két nagy komponens közül hány darab esik az egyikbe, hány darab esik a másikba. Ezért nem tapasztaltunk a közelítésben különbséget a különböző gráfméreteknél, illetve a különböző csúcssúlyozásnál sem (nagy számok törvénye csúcsokra is érvényesült). Kétféleképpen tekinthetünk erre az eredményre: áltesztelési eredményként, illetve tesztelési

eredményként speciális esetre (felfoghatjuk a tesztelési tételt sok kisebb tétel uniformizálásának). Még egy észrevétel, ha nagyon szeretnénk, akkor beleláthatjuk ahol releváns, vagyis a pici gráfnál, hogy a visszatevés nélküli randomizálásnak kicsit kisebb a szórása, ahogy intuitíven várjuk is. Bár a különbség nem szignifikáns.

A második tesztet az inspirálta, hogy nehéz olyan gráfot konstruálni, aminek előre tudjuk a mincut értékét. Az előző tesztben ezt csináltam, de az eredmény a túlzott specialitás miatt nem győzött meg. Így itt azzal próbálkoztam, hogy kisebb, teljesen véletlen gráfból generáltam 10 mintát, és a mintán vett függvényérték szórását vizsgáltam. Az első táblázat 50 csúcsú olyan gráfból generál, aminek élei és csúcsai is egymástól független $E(0,1)$ eloszlásúak. A táblázatban függőleges vonallal van elválasztva a két független kísérlet.

	rand2	rand1	rand2	rand1
szórás	0.0069	0.0123	0.0049	0.0091

A következő táblázat első négy oszlopa 50 csúcsú véletlen egyszerű gráfból kapott eredményeket tartalmazza, míg az utolsó kettő 100 csúcsú súlyozott véletlen gráfból kapott eredményeket.

	rand2	rand1	rand2	rand1	rand2	rand1
szórás	0.0099	0.0085	0.0079	0.0143	0.0053	0.0078

Azt tudom mondani az eredményekről, hogy a minták szórása egyik esetben sem volt túl nagy, amit pozitív eredménynek lehet interpretálni ilyen kis mintaelemszámnál. A különböző kísérletekben nem tértek el egymástól szignifikánsan a szórások. Továbbá itt is megfigyelhetjük, hogy a visszatevés nélküli randomizálás egy árnyalatnyival jobbnak bizonyult.

Ennek a kísérletnek az első felét elvégeztem a μ_2-re is, amiről tudjuk, hogy nem tesztelhető paraméter. A tesztet 50 csúcsú véletlen súlyozott gráfra végeztem el:

	rand1	rand2	rand1	rand2
szórás	0.0510	0.0642	0.0681	0.0387

Ennél a nem tesztelhető paraméternél egy nagyságrenddel magasabb a szórás a tesztelhető $f_2^{\frac{1}{4}}$ szórásához képest.

Végezetül a következő kísérletnél 22 csúcsú véletlen gráfot generáltam. Ezen kiszámoltam a paraméter értékét, és ebből vettem $10, 11, ..., 20$ elemű mintát, mindegyikből 10 darabot. Mindezt azért, hogy megvizsgáljam, javul-e a közelítés a minta elemszámának növelésével. Mindezt elvégeztem négyszer, két súlyozott és két egyszerű véletlen gráfot generálva egymástól függetlenül.

	súly. át.	súly. sz.	súly. át.	súly. sz.	egy. át.	egy. sz.	egy. át.	egy. sz.
igazi	0.0865		0.0788		0.0744		0.0661	
10	0.0670	0.0142	0.0630	0.0164	0.0690	0.0120	0.0740	0.0295
11	0.0694	0.0162	0.0678	0.0150	0.0653	0.0137	0.0620	0.0214
12	0.0847	0.0130	0.0632	0.0120	0.0806	0.0075	0.0722	0.0158
13	0.0769	0.0074	0.0680	0.0102	0.0781	0.0114	0.0722	0.0136
14	0.0735	0.0097	0.0658	0.0091	0.0668	0.0056	0.0607	0.0124
15	0.0716	0.0128	0.0680	0.0113	0.0680	0.0084	0.0578	0.0073
16	0.0750	0.0078	0.0734	0.0080	0.0758	0.0096	0.0691	0.0090
17	0.0758	0.0062	0.0685	0.0073	0.0747	0.0047	0.0702	0.0083
18	0.0735	0.0069	0.0657	0.0087	0.0735	0.0050	0.0642	0.0077
19	0.0740	0.0086	0.0648	0.0105	0.0679	0.0044	0.0640	0.0106
20	0.0823	0.0063	0.0718	0.0075	0.0778	0.0049	0.0698	0.0049

Mindkét esetben észrevehető némi monotonitás. A súlyozott esetben nem erőteljes, az egyszerű esetben viszont szépen látszik.

Összefoglalva az eddig leírtakat; nagyon lényeges, hogy a randomizálási procedúra gyorsan végezhető. Ezen kívül fontos, hogy a végzett tesztek kezdetlegesek a vett minta alacsony csúcsszáma miatt, így a kapott eredmények is óvatosan kezelendők. Mindazonáltal a tesztelési struktúrába kis betekintést lehetővé tettek. Most írok komolyabb tesztek végzéséhez felmerülő ötletekről.

Mint azt a korábbi fejezetben leírtam, kicsit körülnéztem a szakirodalomban mincut számoló algoritmusokat keresve. A tesztelés szempontjából érdektelen balanszírozás nélküli mincut szoros kapcsolatban áll a maximális folyam problémával, és ismertek gyors polinomidejű algoritmusok. Ezzel szemben a balanszírozott változat nehéz probléma. A szakirodalomban rá és különböző változataira számos közelítő algoritmus létezik. Nehézségét a következő probléma illusztrálja.

A (k, v) balanszírozott partíció probléma alatt ($v \geq 1$) azt a feladatot értjük, hogy egyszerű gráfok csúcsait akarjuk k részre osztani, úgy, hogy a partíciók mérete ne haladja meg a $v\frac{n}{k}$ értéket, és a partíciók között futó élek száma minimális legyen. Ez a paraméter nem teljesen a mi balanszírozott mincutunk, hiszen a partíciók mérete felülről van korlátozva, de nagyon hasonlít hozzá. Én úgy képzelem, hogy ez a paraméter kezeli azt is, hogy hány darab partícióra éretik el a minimum, mert lehetséges, hogy néhány partícióbeli halmaz csak egy-egy kevéssé kapcsolódó csúcsból áll. Igaz a következő tétel:

14. Tétel. *[6] Theorem 1: Ha létezik polinom idejű közelítő algoritmus véges közelítő faktorral $k \geq 3$-ra a (k,1) balanszírozott partíció problémára, akkor $P = NP$.*

Ez utóbbi tétel számomra azt mutatja, hogy ha a balanszírozásunk nagyon szigorú, akkor nagyon nehéz problémával állunk szemben. Továbbá megjegyzem, hogy minél gyengébb a balanszírozás, annál nagyobb teret engedünk a közelítésnek, annál könnyebbé válik a probléma.

A mincut különböző változataira létező közelítő algoritmusok mögött nagyon komoly lineáris programozás van. A feldolgozásuk meglehetősen időigényes. Ezért nem végeztem el. Mégis lehetőségként felmerül, hogy közelítő algoritmusok segítségével próbáljuk a balanszírozott mincut tesztelhetőségét vizsgálni. Idea egyszerű, mind a tesztelendő nagy gráfra, mind a mintára közelítőleg számolnánk csak ki a paraméter értékét. Ha a közelítés nem túl rossz, akkor jó eséllyel várunk tesztelést alátámasztó eredményeket. Továbbá ez

a gondolat visszafelé is működhet, tesztelés segítségével lehetne gyorsítani a már meglévő közelítő algoritmusokat.

Ezen kívül komolyabb tesztelési szimulációk készítése céljából felmerül olyan gráfparaméterek keresésének lehetősége, melyek polinomidőben meghatározhatók.

7. Összefoglalás, továbblépési lehetőségek

A dolgozatban bevezettem az olvasót a grafonok és tesztelhetőség elméletébe. Sikerült kicsit mélyebben megérteni a randomizálási procedúra és tesztelhetőségét viszonyát. Ennek segítségével sikerült a [3]-ben kimondott egyszerű tesztelési tételt súlyozott gráfokra általánosítani. Ezt felhasználva tisztáztam néhány paraméter tesztelési viszonyát. Végezetül közöltem kezdetleges teszteredményeket kiegészítve komolyabb tesztek lehetséges alapgondolatával.

Mindeközben maradtak meggondolandó problémák. Gyakorlati szempontból fontos lenne megvizsgálni a súlyozott gráfból vett súlyozott minták közelítő képességét. Emellett a tesztelés pontosabb kidogozása is kívánatos lenne általában $\mathbb{W}_I$-n. Ezen kívül felmerült bennem olyan általános elmélet lehetősége, amellyel különböző gráfosztályokon való tesztelhetőséget lehetne vizsgálni.

Továbbá jó lenne az ismertetett tesztelési elméletet az első fejezetben ismertetett spektrálelméletben hasznosítani. Egy fontos kapcsolódó tétel [4]-ből:

15. Tétel. *[4] 2.15 Tétel (iv) pontja: Ha (G_n) balról konvergens súlyozott gráf sorozat egyenletesen korlátos élsúlyokkal, akkor (G_n) spektruma konvergens, vagyis ha $\lambda_{n,1} \geq \cdots \geq \lambda_{n,|V(G_n)|}$ jelöli G_n adjacencia mátrixának sajátértékeit, akkor $|V(G_n)|^{-1}\lambda_{n,i}$ és $|V(G_n)|^{-1}\lambda_{n,|V(G_n)|+1-i}$ konvergens minden $i > 0$-ra.*

Megjegyzem, hogy a bal-konvergenciánál a sajátértékek konvergenciája ténylegesen gyengébb. Lehetőségként felmerül, hogy a sajátértékek konvergenciája ekvivalens egyes mincut paraméterek konvergenciájával.

Ugyan nem a dolgozathoz szorosan kötődő továbblépési irány, de a területnek fontos problémája, hogy nem biztosított a ritkább gráfokon (például internet) az elmélet haszna. Egy ritka gráfokból álló sorozat az n^2 normáló tényező miatt a $\delta_\square$-metrikában az azonosan nulla grafonhoz tart. Így ezen konvergencia sebessége és a tesztelési tétel által kapott küszöbindexek viszonya határozza meg azt, hogy használható-e az elmélet ritka gráfokra. Ha jól értesültem, az utóbbi időben születtek eredmények a probléma kezelésére.

Hivatkozások

[1] M. Bolla, G. Tusnády, *Spectra and optimal partitions of weighted graphs*, Discrete Mathematics, 1994,Vol.128, 1-20

[2] M. Bolla and G. Molnár-Sáska, *Isoperimetric properties of weighted graphs related to the laplacian spectrum and canonical correlations*, Studia Scientiarum Mathematicarum Hungarica, 2002, Vol. 39, 425-441

[3] C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós, K. Vesztergombi, *Convergent Sequences of Dense Graphs I: Subgraph Frequencies, Metric Properties and Testing*, 2006, http://arxiv.org/PS_cache/math/pdf/0702/0702004v1.pdf

[4] C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós, K. Vesztergombi, *Convergent Sequences of Dense Graphs II: Multiway Cuts and Statistical Physics*, 2007, http://www.cs.elte.hu/∼lovasz/ConvRight.pdf

[5] L. Lovász, B. Szegedy *Limits of dense graph sequences*, 2006, J. Comb. Theory B 96, 933-957., http://www.cs.elte.hu/∼lovasz/limits.pdf

[6] Konstantin Andreev, Harald Räche, *Balanced Graph Partitioning*, 2004, Proceedings of the sixteenth annual ACM symposium on Parallelism in algorithms and architectures, Session Alghorithms, Barcelona, 120-124, http://www.dcs.warwick.ac.uk/∼harry/pdf/balanced_partition_journal.pdf