MIT TANULHATUNK A BIG DATÁBÓL, AVAGY HOGYAN VÁLASZTUNK KOMMUNIKÁCIÓS CSATORNÁT?

Török János,[1,2] Kertész János[1,3]
[1]BME, Elméleti Fizika Tanszék
[2]MTA–BME Morfodinamika kutatócsoport
[3]Department of Network and Data Science, CEU

Az utóbbi időben a szociális viselkedés vizsgálata egyre inkább multidiszciplináris tudományággá vált. Az emberi kommunikáció és kapcsolat mind nagyobb hányada zajlik digitális rendszereken (telefon, internet), amelyekről rengeteg adat áll rendelkezésre. Az adatok analízisében a szociológusok mellett úttörő szerepet játszottak fizikusok, matematikusok és informatikusok is [1, 2], hiszen számos technikára volt szükség a gráfelmélettől az adatbányászaton és a sokügynökös modellezésen keresztül a statisztikus megközelítésig.

A nagy adatokban rejlő lehetőségek alapos tudományos igényű és nyilvánosságot kapó megértése azért is fontos, hogy minél inkább elkerülhető legyen az ilyen jellegű tudás álságos felhasználása, amire a nagy nemzetközi céges sajnos hajlamosaknak látszanak (a közelmúltban lásd a Facebook és a Cambridge Analytica botrányát [3]). Azonban nagy kérdés, hogy egyes digitális szolgáltatások adataiból mennyire lehet megbecsülni a valódi társadalmi tulajdonságokat. Az alábbiakban ezzel a kérdéssel foglalkozunk.

Mielőtt konkrétabban felvázolnánk a problémát, tekintsük át, hogyan lehet reprezentálni és modellezni a digitális szolgáltatásokon megjelenő emberi kapcsolatokat.

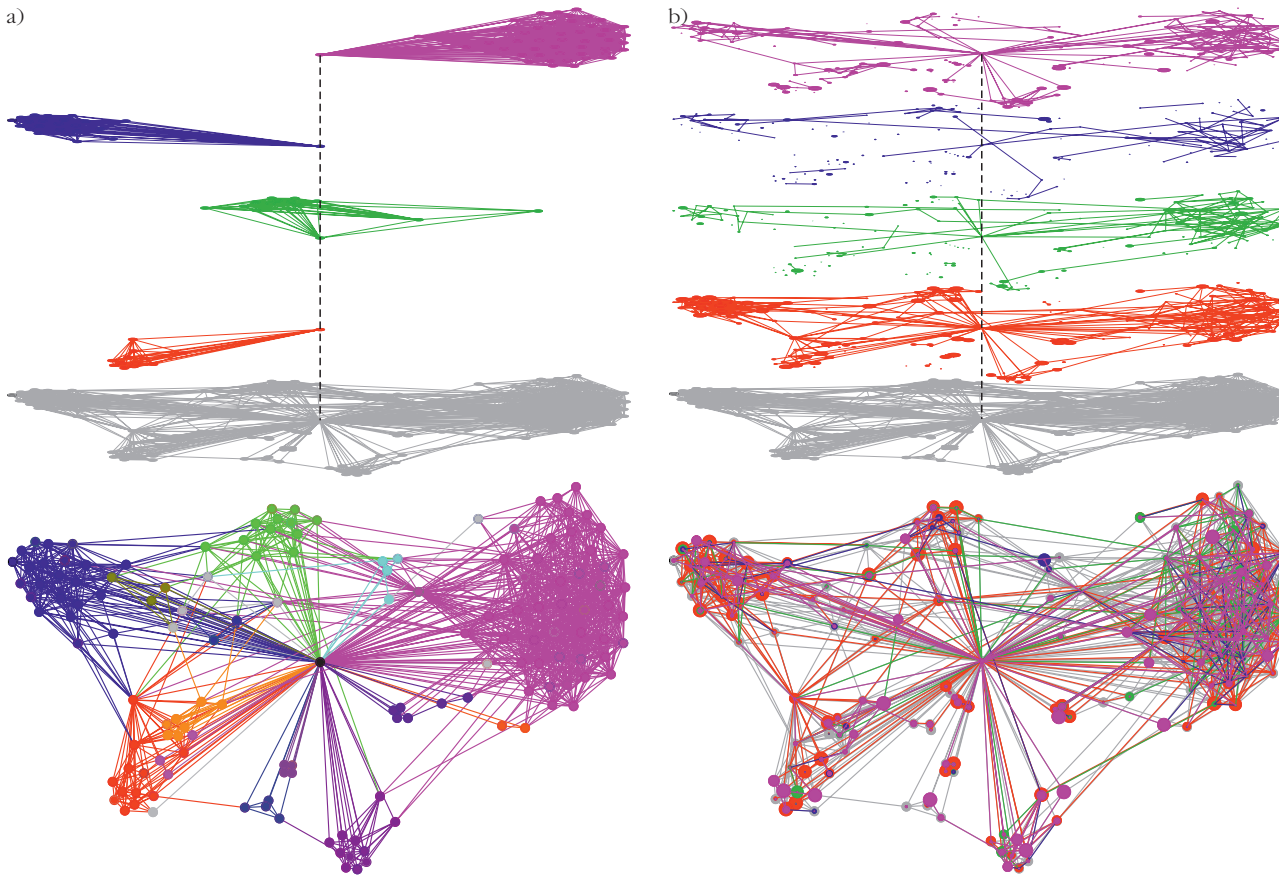
Az emberi kapcsolatok legegyszerűbb modellje egy súlyozott gráf [4], ahol a csúcspontokban az egyének vannak, az élek a szociális kapcsolatot, az élsúlyok pedig a kapcsolatok erősségét jellemzik. A szociális kapcsolatok egyik alapvető ábrázolási módja az úgynevezett egocentrikus hálózat, amely egy olyan gráf, ahol egy központi egyén (az ego), illetve ismerősei (az alterek) a gráf csúcsai, amelyek között az élek a szociális kapcsolatokat jelzik. Az *1.a ábrán* egy tipikus egocentrikus hálózatot mutatunk. Az ábrázoló programban az élek vonzó rugóként, a csúcspontok taszító pontonként vannak modellezve és így alakul ki a szemléltetett struktúra. Jól láthatók a jellegzetes közösségek, amelyek megfelelnek a valós életbeli csoportoknak, mint például munkahely, baráti társaság, rokonok stb. Ezért a kapcsolatokat nem lehet pusztán egyetlen számmal jellemezni, hiszen más jellegű egy szakmai barátság, mint egy rokoni kapcsolat. Ha csak egy típusú kapcsolatokat vizsgálunk (*1.a ábra,* felső sorok), akkor kirajzolódnak a megfelelő közösségek. Minden ilyen közösség egy réteget képez a társadalomban (például munkahelyi ismerettségi hálózat, rokonság, sportolási kapcsolatok), és az ilyen rétegek együttese adja az emberek teljes szociális hálózatát. Az ilyen többrétegű gráfot, ahol különböző rétegeken vannak élek, de a csúcsok minden szinten megfeleltethetők egymásnak – hiszen minden rétegben ugyanarról az emberről van szó – multiplex hálózatnak nevezzük.

A szociális kapcsolataink általános értelemben vett kommunikáció révén jutnak kifejezésre. Legközelebbi barátainkkal gyakori kapcsolatban vagyunk. Megfigyelhető, hogy amennyiben a rendszeres kommunikáció nehézségbe ütközik, akkor kapcsolatunk még a legjobb barátunkkal is elhalványodik (ez alól csak a rokonsági kapcsolat kivétel, ahol hosszú idő után is sokkal könnyebb felvenni a kapcsolatot). Belső igényünk tehát, hogy kapcsolatban legyünk egymással és erre ma már a személyes találkozáson kívül rengeteg lehetőségünk van: telefonálhatunk, küldhetünk sms-t, írhatunk e-mailt, tweetet, vagy azonnali üzenetet valamilyen internetes szociális hálózaton stb. Ezeket kommunikációs csatornáknak hívjuk.

A kommunikációs csatornák kiválasztásánál az emberek sok szempontot mérlegelnek. Először is előfordul, hogy egy adott szolgáltatáshoz egyik ismerősünk nem fér hozzá (például nagyszülőnek nincs Instája), de nagyon fontos, hogy ki melyiket szereti és figyeli állandóan (például egy állandóan lehalkított telefonú gyermek hívogatása eléggé parttalan dolog, amit sok szülő tapasztalatból tud). Tehát rá vagyunk kényszerítve, hogy több kommunikációs csatornát használjunk. Kapcsolataink nem

Török János fizikus egyetemi tanulmányait az ELTE-n végezte, majd 2000-ben PhD fokozatot szerzett fizikából a BME-n. Jelenleg az MTA–BME Morfodinamika Kutatócsoport tagja és a BME Elméleti Fizika Tanszékének docense. Érdeklődési területe a szemcsés anyagok numerikus szimulációja, szemcsék alakváltozásának, továbbá a szociális hálózatok dinamikájának modellezése, illetve kommunikációhoz kapcsolódó „Big Data" elemzése.

Kertész János fizikus, az MTA rendes tagja, egyetemi tanár, a CEU Hálózat- és Adattudományi Tanszékének vezetője, a BME részfoglalkozású professzora. Kutatási területe a statisztikus fizika interdiszciplináris alkalmazásai és a számítógépes társadalomtudomány. Elsősorban a társadalom szerkezetére és működésére jellemző „Big Data" elemzésével és modellezésével foglalkozik.



1. ábra. Egy ember egocentrikus multiplex hálózatának szemléltetése. A felső rész a különböző rétegeket szemlélteti, az alsó ábra pedig ezek összességét. a) A különböző kapcsolattípusok szerint (rokonság, iskolatársak, munkatársak stb.), itt egy jól látható közösség tartozik egy típushoz. b) A különböző kommunikációs csatornákon megjelenő kapcsolatok szemléltetése.

csak az ismeretség típusa, hanem a kommunikáció módja szerint is többrétegű.

Amikor egy-egy digitális szolgáltatás adatait vizsgáljuk, akkor ebből a többrétegű hálózatból csak egy réteget látunk. Mivel, helyesen, az ilyen adatokhoz való hozzáférés még a kutatók számára is csak anonim adatokkal lehetséges, ezért a különböző adatbázisokat nem tudjuk összekötni, hogy megkapjuk az egyének teljes kommunikációs hálózatát.

A szociális kapcsolatok teljes rendszerét csak az összes csatorna együttes elemzése tudná pontosan felfedni. Nagy kérdés tehát, hogy a kommunikációs hálózatok egyetlen, rendelkezésre álló csatorna által lefedett része mennyire adja vissza a teljes szociális hálózat tulajdonságait. Az ezen a részleges, a multiplex egyetlen rétegére vonatkozó gráfon megmért adatok mennyiben és hogyan változnak meg a részlegesség miatt?

Két adatbázissal fogunk foglalkozni: (i) az iWiW ismeretségi hálózattal és (ii) egy mobiltelefon-szolgáltató híváslistájával.

i) Az iWiW-et talán még mindenki ismeri, ez a szociális hálózat, amely 2002-ben indult és 2006-ban üstökösszerű fejlődésnek indult. Élettartamának nagy részében a három legnépszerűbb weboldal közé tartozott Magyarországon, sokak szerint az internet terjedésének egyik hajtóereje volt: 2006-tól 2010-ig az internetelérással rendelkező magyarok 2/3-a a szolgáltatás felhasználója volt. A rendelkezésre álló adatok a következők: a belépés dátuma, a kapcsolatok listája a létesítés dátumával és az utolsó belépés dátuma. Az adatok 2011 januárjáig állnak rendelkezésre, amikor a szolgáltatás népszerűsége már töredékére esett vissza (a felhasználók 2/3-a már nem volt aktív). Az iWiW-szolgáltatást 2014. június 30-án megszüntették.

ii) A másik adat egy nagy európai ország egyik jelentős, 30%-os lefedettségű mobilszolgáltatója. Egy teljes év hívási adatai állnak rendelkezésre (ki, kit, mikor, mennyi ideig hívott). A hívások alapján egy hálózat állítható össze, ahol a csúcsok a hívásokban részt vevő emberek, akik között akkor van él, ha hívták egymást telefonon. Noha a hívásokat mindig az egyik fél kezdeményezi, az élek irányítottságától eltekintünk, hiszen információ mindkét irányban folyik.

Két egyszerű, szociális szempontból is fontos mennyiséget vizsgálunk: a $P(k)$ fokszámeloszlást és a k fokszámú csúcsok szomszédainak $k_{nn}(k)$ várható fokszámát, ami az úgynevezett fokszám-asszortativitással van kapcsolatban. A fokszám egy adott csúcs esetén a belőle kiinduló élek számát méri. A fokszámeloszlás módusa azt mutatja meg, hogy leggyakrabban hány kapcsolata van egy embernek. Meglepő módon mind az iWiW-es, mind a telefonos adat esetében (és gyakorlatilag szinte minden internetes kommunikációs adat esetében is, például twitter, facebook stb.) a fokszámeloszlás módusa k = 1-nél van

(*2. ábra* felső sora), ami ellentmond a hétköznapi tapasztalatnak. Ez ugyanis azt jelentené, hogy az a legvalószínűbb, hogy valakinek csak egyetlen egy szociális kapcsolata van. Ezt az effektust az egyetlen kommunikációs csatorna felhasználásából származó torzításnak kell betudnunk.

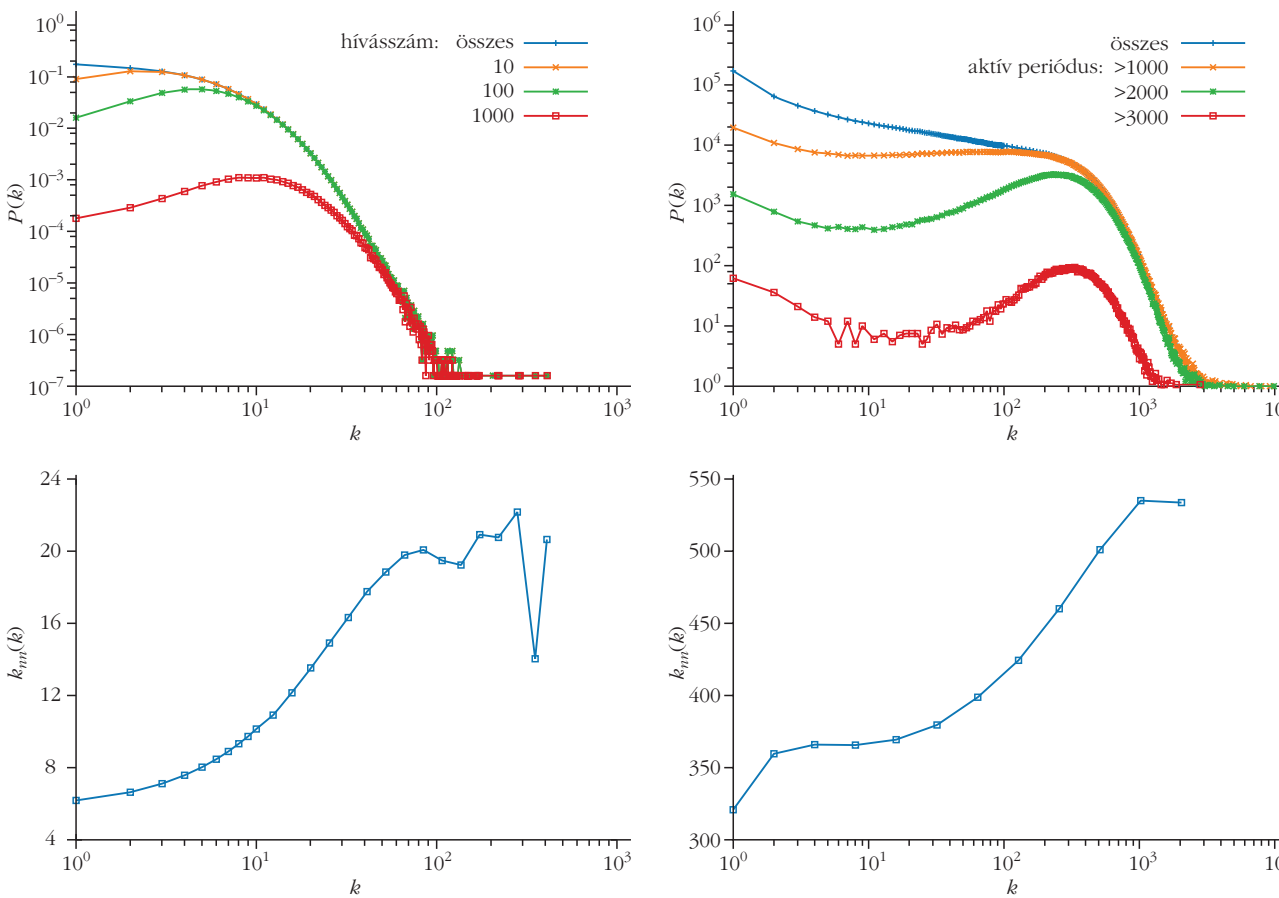
Megvizsgáltuk, hogy mi történik, ha csak a tapasztalt felhasználókat vizsgáljuk. Őket az iWiW esetén úgy definiáltuk, hogy bizonyos időnél többet töltöttek a szolgáltatással, a telefonos adatoknál pedig azokat választottuk ki, aki egy bizonyos számnál többet telefonáltak az adott időszakban. A *2. ábrán* látható, hogy ekkor már 1-nél nagyobb helyen lesz a görbék maximuma. Úgy gondoljuk, hogy azok – adott csatornán mért – kapcsolati hálózata, akik nagy energiát fektetnek e kommunikációs csatornába (hosszú idő alatt alaposan összeválogatják barátaikat egy szociális hálón, sokat telefonálva tartják a kapcsolatot ismerőseikkel) jobban hasonlít a valódi szociális hálózatukhoz, tehát a rajtuk mért mennyiségek jellemzői jobban hasonlítanak a valódi szociális hálózat adataihoz, mint ha minden felhasználót figyelembe vennénk. Az utóbbiak között sok olyan van, aki csak egyszer-egyszer használja az adott csatornát és társadalmi kommunikációja nagy részét máshol bonyolítja le.

A fokszám-asszortativitás azt méri, hogy mennyire hajlamosak a nagy fokszámú csúcsok nagy fokszámú csúcsokkal kapcsolódni. Ezt a legegyszerűbb a $k_{nn}(k)$ mennyiséggel, a k fokszámú csúcsok szomszédainak átlagos fokszámával mérni: ha monoton növekvő, akkor a hálózat asszortatív, ha csökkenő, akkor disszasszortatív. Szociális hálózatoktól azt várjuk, hogy asszortatívak, ami azt jelenti, hogy a nyitott, sok baráttal rendelkező emberek előszeretettel kapcsolódnak hasonló, nyitott személyekhez. A *2. ábra* alsó sorában látható adatok ezt a feltételezést támasztják alá, de kérdés, hogy mennyire bizonyító erejű egy ilyen ábra. Hiszen előfordulhat, hogy ez az effektus is, mint a fokszámeloszlás monotonitása, csak a mintavétel torzító hatásának következménye.

A probléma vizsgálatához modellezzük a csatornaválasztás mechanizmusát. Ezután – kiindulva a teljes multiplexre jellemző tulajdonságokból – megvizsgáljuk, hogy azok mennyiben torzulnak, ha csak egyetlen csatornát látunk.

Legyen adott egy egyén (i) és jellemezze $0 \le f_i \le 1$ affinitás azt, hogy i mennyire szereti a vizsgált kommunikációs csatornát. Az $f_i = 0$ azt jelenti, hogy nem fér hozzá a csatornához, az $f_i = 1$ pedig azt, hogy ha teheti, csak ezen kommunikál. Legyen i ismerőse j. Ha i és j kapcsolatba szeretne lépni egymással, akkor ki kell választaniuk egy kommunikációs csatornát. Azonban figyelembe kell venniük egymás affinitását is. Mert hiába szeret valaki e-mailt írni, ha a másik azt

2. ábra. Empirikus eredmények mobiltelefon- (bal oszlop) és az iWiW-adatokon (jobb oszlop). A felső sor a fokszámeloszlást mutatja. Az eredeti adat eloszlása kékkel, a tapasztaltabb felhasználókat (telefon esetén hívásszám, iWiW-nél aktivitásiperiódus-limit) rendre sárga, zöld, piros színnel jelöltük. Az alsó sorban az asszortativitás látható a teljes adatra.



csak nagyon ritkán olvassa. A kommunikációs csatornát tehát valamilyen $p(f_i, f_j)$ szimmetrikus valószínűséggel fogják választani. Ebből az is következik, hogy adott idő alatt egy kapcsolat $p(f_i, f_j)$ mennyiséggel arányos valószínűséggel jelenik meg a kommunikációs csatornán. Tehát a kommunikációs csatorna a valós kapcsolatokból mintavételez, ami azonban nem egyenletes, hanem az egyének affinitásától függ.

Modellünk tehát a következő: adott egy kommunikációs csatorna, az embereket egy affinitás jellemzi a csatorna irányában és a köztük lévő kapcsolatok $p(f_i, f_j)$-vel arányos valószínűséggel jelennek meg a csatornán.

Megvizsgáljuk, hogy a fenti modell – a már említett két mennyiség tekintetében – miként változtatja meg az eredeti hálózat tulajdonságait.

Először is egy emberek közötti szociális hálózatot kell vennünk. Mivel ez nyilván nem ismert ezért kétféle modellt választottunk: (a) véletlen reguláris gráf, ahol minden csúcs fokszáma k_0, de az élek véletlenül vannak behúzva a csúcsok között, (b) Erdős–Rényi-gráf, véletlenszerűen addig húzunk éleket az adott csúcsszámú gráfba, amíg az átlagos fokszám $\langle k \rangle$ nem lesz.

A vizsgált két hálózat minden tulajdonsága jól ismert. A véletlen reguláris gráf fokszámeloszlása k_0-nál Dirac-delta, $k_{nn}(k)$ pedig egyetlen pont. Az Erdős–Rényi-gráf (ER-gráf) fokszám eloszlása Poisson-eloszlás, $k_{nn}(k)$ pedig konstans (nem asszortatív, de nem is diszasszortatív).

Emellett feltesszük, hogy az affinitások egy általános $P(f)$ eloszlásból származnak. Nyilvánvaló $P(f)$ monoton csökkenő, ami azt jelenti, hogy viszonylag kevés ember mutat túláradó lelkesedést egy kommunikációs csatorna iránt. Mi az exponenciális eloszlást választottuk:

$$P(f) = \frac{1}{f_0} \exp\left(-\frac{f}{f_0}\right), \qquad (1)$$

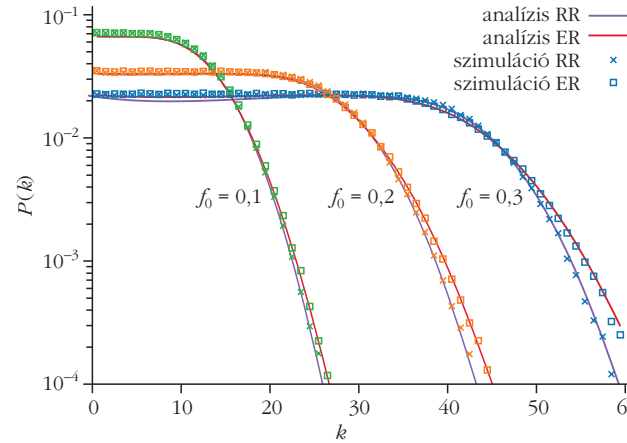
ahol f_0 az átlagos affinitás, ami egy kontrollparaméter. Az egyszerűség kedvéért, minden embernek az ismerőseitől független affinitást adunk a fenti eloszlásból.

Az utolsó kérdés a $p(f_i, f_j)$ függvény megválasztása. A kapcsolat létrejöttéhez – mint már említettük – az szükséges, hogy mindkét fél valamennyire használja ezt a csatornát. Bármelyik is idegenkedik tőle, inkább másik csatornát fognak választani, ezért a két affinitás közül a kisebbik bír fontosabb szereppel. Ezt legegyszerűbben egy minimumfüggvénnyel modellezhetjük:

$$p_{ij} \equiv p(f_i, f_j) = \min\{f_i, f_j, 1\}. \qquad (2)$$

Az 1-et azért vettük a kifejezéshez, hogy a p_{ij} mennyiséget valószínűségként tudjuk használni, amelynek értéke nem lehet nagyobb, mint 1.

A fokszámeloszlást szerencsére analitikusan is ki lehet számolni [5]. A kapott fokszámeloszlások mindig monoton csökkenők, még a véletlen reguláris gráf esetében is, amikor a kezdetben a fokszámeloszlás egy Dirac-delta volt (minden csúcs fokszáma k_0).



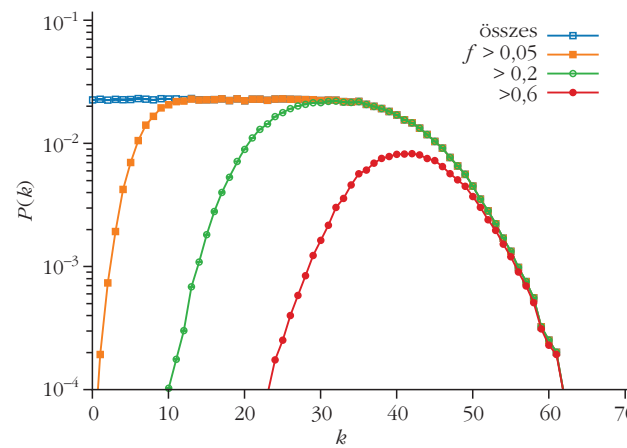
3. ábra. A szimulált kommunikációs csatornák fokszámeloszlása. Az x a reguláris véletlen, a négyzet a ER-gráf kiindulási hálózatot mutatja, a folyamatos vonalak az analitikus eredmények. A három görbecsoport az affinitáseloszlás különböző f_0 paramétereihez tartozik.

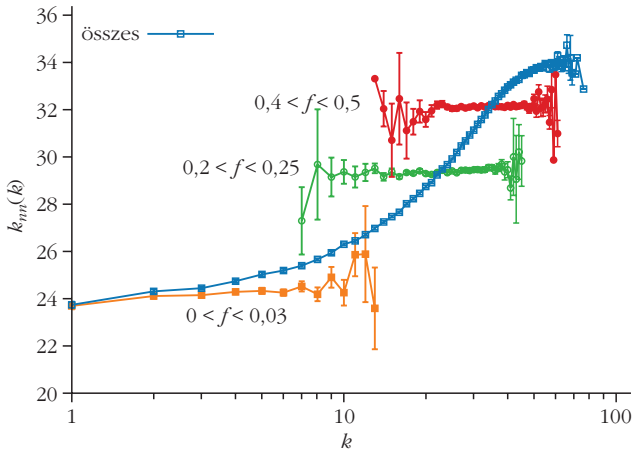
Ezt a *3. ábra* szemlélteti, ahol mindkét kiindulási hálózatot megvizsgáltuk, és a mintavételezett fokszámeloszlások – a kiindulási gráftól függetlenül – minden esetben nagyon hasonlóan néznek ki. Csak az affinitások f_0 paramétere van hatással az eredményekre, de a görbék monoton jellege megmarad. Ez azt jelenti, hogy a mintavételezési folyamatnak sokkal nagyobb hatása van az eredményekre, mint a kiindulási hálózatnak, és az empirikus adatokon látható monoton csökkenő fokszámeloszlást is a mintavételezés okozhatja.

Megvizsgáltuk, hogy mi történik, ha csak a nagy affinitású nódusok fokszámeloszlását vizsgáljuk meg. A *4. ábrán* jól látható, hogy a fokszámeloszlás maximuma újra véges értéknél lesz, mint ahogy az empirikus adatokon is látható. Ez azt sejteti, hogy a nagy affinitású felhasználók hálózata sokkal jobban hasonlít az eredetihez, mint a többié.

Vizsgáljuk meg az asszortativitásra jellemző $k_{nn}(k)$-t! Itt az empirikus adatok is a társadalomról várható viselkedést mutattak. Már említettük, hogy a ER-hálózatnál a $k_{nn}(k)$ konstans, azonban a mintavé-

4. ábra. A szimulált kommunikációs csatornák fokszámeloszlása $f_0 = 0{,}3$ és ER-gráf esetén. A kék görbe a teljes mintavételezett hálózat fokszámeloszlását mutatja, a többi csak azon nódusokét, amelyek affinitása a megadott értéknél nagyobb.





5. ábra. Asszortativitásra jellemző $k_{nn}(k)$ a mintavételezés után egy ER-gráfból kiindulva. A kék görbe a teljes mintavételezett hálózat asszortativitását mutatja, a többi csak azon nódusokét, amelyek affinitása a megadott tartományban van.

telezés után a hálózat asszortatív lesz (*5. ábra*). Érdekes módon még a görbe alakja is kifejezetten hasonlít az empirikus görbékre. Tehát ismét azt mondhatjuk, hogy az egyetlen kommunikációs csatorna adataiból látható asszortatív viselkedés ténye az eredeti szociális hálózatra nézve nem tekinthető bizonyítéknak. Itt is megvizsgáltuk, vajon mi történik, ha csak a nagy affinitású nódusokat vizsgáljuk, illetve csak azokat, amelyek affinitása egy kis tartományba esik. Ekkor visszakapjuk a ER-gráfra jellemző konstans affinitást.

Összefoglalva, azt vizsgáltuk, hogy mennyiben különbözik az egyetlen kommunikációs csatornán megfigyelhető tulajdonság a teljes szociális hálózatra jellemzőtől. Mivel az utóbbiról nincs elég adatunk, ezért a problémát fordítva vizsgáltuk. Alapvető és egyszerű feltevésekből kiindulva próbáltuk leírni az emberek viselkedését a kommunikációs csatornák használata során, és a teljes szociális hálózatot leegyszerűsítve modelleztük. Azt kaptuk, hogy az egyetlen csatornára szűkített mintavételezés torzítja az eredeti hálózat tulajdonságait, mégpedig a nagyszámú egycsatornás megfigyelés által szolgáltatott eredmények irányába. Ez a jelenség olyan erős lehet, hogy új tulajdonságok jelenhetnek meg (monotonitás a fokszámeloszlásban), illetve erősödhetnek fel (fetételezhetően az asszortativitás esetében). A torzítás arra vezethető vissza, hogy egy csatorna használatánál nagy súllyal szerepelnek olyanok, akiknek az nem fő kommunikációs eszközük, és így az ő hálózatuk csak töredékesen jelentkezik. Ezt igazolták azok az empirikus adatokon és a modellen is végzett mérések, amelyek az „érett” felhasználók esetében a torzítások csökkenését mutatták. Megjegyezzük, hogy itt csak a modell legegyszerűbb változatát ismertettük. Részletesebb vizsgálatok bebizonyították, hogy az említett torzítások függetlenek a modell részleteitől.

Az eredményeknek fontos tanulságuk van a kutatók és a felhasználók számára is. A kutatóknak az, hogy ha csak egyetlen csatorna információi állnak rendelkezésükre, akkor az eredmények a teljes szociális hálózat tulajdonságaira nézve félrevezetők lehetnek, és a felhasználókat jobban közelíti, ha csupán a nagy affinitású („érett”) egyéneket veszik figyelembe. A felhasználóknak pedig tudniuk kell, hogy minél intenzívebben használ valaki egyfajta szolgáltatást, róla annál többet lehet tudni, hiszen kapcsolati hálójának annál nagyobb része válik elérhetővé.

Irodalom

1. Barabási A.-L.: *Villanások: A jövő kiszámítható*. (ford.: Kepes J.) Libri Könyvkiadó, Budapest (2016) ISBN 9789633105139.
2. Staar Gy.: Gráflimesz, könyvek és család. Beszélgetés Lovász László matematikussal. *Természet Világa 145* (2014) 530–535.
3. *A cambridge analytica botrány*. dátum: 2018. május 29., https://index.hu/aktak/a_cambridge_analytica_botrany
4. Barabási A.-L.: *Behálózva – A hálózatok új tudománya*. (ford.: Vicsek M.) Helikon Kiadó, Budapest (2013) ISBN 9789632272580.
5. J. Török, Y. Murase, H.-H. Jo, J. Kertész, K. Kaski: What Big Data tells: Sampling the social network by communication channels. *Physical Review E94* (2016) 052319.