

OTDK DOLGOZAT

Klaszterezettség vizsgálata geoszociális hálózatokon

Rozgonyi Kristóf

Fizika Bsc., Fizikus szakirány

2. Évfolyam



Témavezető:

Kondor Dániel

ELTE, Komplex Rendszerek Fizikája Tanszék

2014

Kivonat

Komplex rendszerek egy reprezentatív példája, az egyénekből vagy csoportokból és a közöttük levő kapcsolatokból alkotott struktúra, az úgynevezett szociális hálózat. A közösségi portálok (Twitter, Facebook...stb) népszerűsödésével lehetőségünk nyílik valós, a felhasználók alkotta hálózatok tanulmányozására.

Ebben a dolgozatban egy ilyen hálózat vizsgálatával foglalkozunk. A gráfunk egy úgynevezett geoszociális hálózat, ahol a felhasználókhoz földrajzi koordinátákat tudunk rendelni. Az adatokat a Twitter rendszer nyilvános API (application programming interface) segítségével gyűjtöttük. Így 2 454 623 felhasználó között azonosítottunk 3 351 813 kapcsolatot, ahol a kapcsolatot a kölcsönös említések jelentették.

Munkánk során főleg arra a kérdésre kerestük a választ, hogy a hálózat klaszterezettsége hogyan függ a felhasználók távolságától, illetve a köztük levő kapcsolatok erősségétől. Az adatok többféle szűrésével különböző részhálózatokat és az ezekből generált véletlen hálózatokat vizsgáltuk, kitérve a legnagyobb összefüggő részhálózat elemzésére is. Az így kapott adatokra számoltunk a szakirodalomban gyakran használt gráftulajdonságokat. Ezeket úgy választottuk meg, hogy a hálózat egészére jellemző statisztikus paraméterekkel jellemezzék a vizsgált hálózatokat.

A dolgozatban megmértük azt a távolságot, ahol a legösszetettebb a hálózat, illetve belátuk azt az elméletet, hogy a gyenge kapcsolatok fontos szerepet játszanak a társadalmunk összetettségében.

Tartalomjegyzék

1. Bevezetés	1
2. Az adatbázis és a vizsgált gráfok	2
2.1. Az adatgyűjtés	2
2.2. Az adatbázis felépítése	3
2.3. A szociális hálózat megválasztása	3
2.4. A vizsgált gráfok alapját képző tábla	4
2.5. A vizsgált gráfok	5
2.6. A gráfok súlyozása	6
3. Vizsgálati módszerek	7
3.1. Klaszterezettséget jellemző mennyiségek	8
3.1.1. Átlagos klaszterezettségi együttható	8
3.1.2. Geográfiai klaszterezettségi együttható	9
3.1.3. Transzitivitás	11
3.2. Véletlen hálózatok	12
3.3. Élek elhagyása	13
3.4. Vizualizáció	15
4. Eredmények	16
4.1. Véletlen hálózatok	18
4.2. Élek elhagyása	21
4.3. A hálózat gyenge kapcsolatoktól való függése	29
5. Összegzés	32
Hivatkozások	33
1. Függelék	35

1. Bevezetés

A számítógépek elterjedésével egyre nagyobb szerep jut az informatikának a legkülönbözőbb tudományterületek kísérleti adatainak gyűjtésében, feldolgozásában, elemzésében.[1] Például a részecskefizika, csillagászat vagy a hálózat kutatás különböző területein az egyre jobb műszereknek és számítógépeknek köszönhetően hatalmas mennyiségű mérési- és szimulációs adat keletkezik.[2][3]

Manapság sok fizikus, szociológus, matematikus vagy informatikus figyelme fordul a szociális hálózatok vizsgálatára felé, hiszen a közösségi portálok (Twitter, Facebook....stb.) népszerűsödésével rengeteg adat halmozódik fel társadalmunkról, amit maguk a felhasználók tesznek közzé az interneten. Ez a két dolog (a komputerizáció és az online szociális hálózatok népszerűsége) teszi lehetővé, hogy társadalmunk struktúrájáról országhatárokat túllépve kontinentális, vagy akár az egész Föld nagyságának szintjén nyerjünk információt. Az egyik ilyen sokat vizsgált hálózat a Twitter egy online mikroblog szolgáltatás és ismerettség hálózat, ahol a felhasználók maximum 140 karakteres bejegyzéseket (Tweeteket) tehetnek közzé, amit a felhasználó követői automatikusan látnak. A tweetek az egyszerű szövegen kívül tartalmazhatnak például linkeket, hashtageket ¹ vagy személymegjelölést. A felhasználók a tweetjeikhez koordinátákat is megadhatnak, ez általában telefonos GPS alkalmazás segítségével történik.

A Twitter lehetőséget nyújt a kutatóknak, hogy szinte valós idejű adatokat töltsenek le a hálózatról. Ezek az adatok véletlen mintavételezéssel vannak kiválasztva, ezért jól reprezentálják az eredeti adathalmazt. Az ELTE-n is folyik egy ilyen Twitter alapú adatgyűjtő projekt, amiből adatbázis épül. [4]

A Twitter adatok sokszínűsége révén lehetőségünk van a szociális hálózat különböző tulajdonságainak a vizsgálatára. Ebben a dolgozatban a hálózat klaszterezettségét vizsgáljuk, figyelembe véve a felhasználók földrajzi elhelyezkedését és a közöttük levő kapcsolatok erősségét. Először az adatgyűjtést, az adatbázist és a dolgozatban használt szociális hálózatokat mutatjuk be. Ezután ismertetjük a vizsgálati módszereket, végül az utolsó fejezetben bemutatjuk az eredményeket.

¹Hashtag egy szó vagy kifejezés, amit a # szimbólum előz meg. Ezzel általában témákat jelölnek, és a témára való keresést könnyíti meg.

2. Az adatbázis és a vizsgált gráfok

A számítástechnika fejlődésének köszönhetően nem csak a gépek számítási- és adattárolási teljesítménye nő, hanem a begyűjtött adatok mennyisége is. Egy bizonyos mennyiségű adat után a hagyományos file-alapú adatkezelés átláthatatlanná és használhatatlanná válik. A modern adatintenzív kutatásokhoz az adatokat olyan adatbázisokba kell szervezni, amelyek hatékonyan oldják meg a tárolási- és adatkezelési feladatokat. A dolgozathoz használt adatok kezelése és tárolása is egy ilyen adatbázissal történik. [4]

2.1. Az adatgyűjtés

Annak ellenére, hogy az összes üzenet csak kis része érhető el bárki számára szabadon, sok kutatáshoz használnak Twitter adatokat. Az adatletöltés a nyilvános API (application programming interface) segítségével történik. Így az üzenetek 0.5%-hoz férhetünk hozzá, ami hasonló letöltésekhez képest nagynak számít. Az ilyen letöltések ingyenesen hozzáférhetőek és reprezentatív mintát adnak az adatsorból, valamint lehetőséget adnak valós idejű adatletöltésre is. Az adatok adatbázisba való betöltése nem a letöltéssel együtt történik, hanem külön lépés, de ennek a részleteit nem tárgyaljuk ebben a dolgozatban. Alapvetően a letöltött adatok a felhasználó bejegyzései, vagyis a tweetek. Ilyenkor nem csak a tweet szövege, hanem a felhasználóra jellemző információ is letöltődik, esetleg más a tweetet jellemző adat is. A felhasználóra jellemző adatok, amik mindig letöltődnek: a felhasználó azonosítója, felhasználóneve, regisztrációjának az ideje, az, hogy az adott pillanatban a felhasználó kiket követett, illetve, hogy mennyi tweetje van az adott időben, valamint a felhasználó regisztrációjakor megadott hely. A tweetre vonatkozó adatok, amik mindig letöltődnek: a tweet szövege, létrehozásának az ideje és a tweet azonosítója.

A tweetek unicode karaktereket tartalmaznak, így lehetőség van bármilyen szöveget letölteni, beleértve az ázsiai nyelvűeket is. A felhasználók beállíthatják, hogy a bejegyzéseiknél megjelenjen-e a helyzetük, vagyis ha lehetőség van rá, például a mobiljuk GPS-ével, akkor a tweetjeikhez földrajzi koordináták is rendelődjenek-e. Ha egy tweet ilyen -úgynevezett *geo-tagged*- , akkor a letöltött adatok között hosszúsági és szélességi adatok is lesznek.[5] A tweetekre lehet reagálni, vagyis *retweetelni*. Ez olyan, mintha egy e-mailre válaszolnánk. Ha a letöltött adatok között van retweet, akkor az eredeti tweet is letöltődik, és az eredeti tweet adatai is. Mivel egy tweetre bármikor lehet retweetelni, ezért elképzelhető, hogy az adatgyűjtés kezdete előtti időkből is lesznek adatok, amik bekerülnek majd az adatbázisba.

A felhasználóknak lehetőségük van megjelölni más felhasználókat a bejegyzéseikben. Ha

egy tweet személymegjelölést tartalmaz, akkor a megjelölt személy azonosítója is letöltődik. Elképzelhető tehát, hogy a letöltött tweet csupán a bejegyzést létrehozó felhasználóról és a létrehozás idejéről hordoz információt, de olyan tweetek is előfordulnak, amik a bejegyzés helyét is tartalmazzák. Az ilyen geo-tagged tweetek a letöltött adatok csak egészen kis részét, körülbelül 1.5 %-át adják, azonban lehetőség van olyan letöltést csinálni amikor főleg expliciten koordinátás tweeteket tölt le a rendszer, ekkor a bejövő adatok nagy része tartalmaz koordinátákat. Ez másodpercenként 50-60 tweetet jelent. A mi adatbázisunkhoz is egy ilyen letöltést használunk, vagyis az adataink nagy része geo-tagged. Azok az adatok, amiket ebben a dolgozatban használunk, 2012. január és december között lettek gyűjtve. A letöltés folyamatosan ment, de az adatbázisba való betöltés körülbelül havonta történt meg. Az adatbázis maga a Microsoft SQL Server 2012 motort használja, így az adatok könnyen elérhetőek és kezelhetőek az SQL programnyelv segítségével.

2.2. Az adatbázis felépítése

Az adatbázis felépítése az átláthatóság kedvéért viszonylag egyszerű. Az egyik fő tábla, a „*tweet*” tartalmazza a letöltött tweeteket. A táblában a tweet azonosítója, a felhasználó azonosítója, a létrehozás ideje, a retweetek száma, hogy ha retweet volt, akkor melyik tweetre volt válasz, ha vannak, akkor a koordináták, és a bejegyzés szövege található. Egy másik tábla a „*user*” tábla, ami a felhasználók adatait tartalmazza: az azonosítóját, a felhasználónevét, a profil létrehozásának idejét, egy rövid leírást (amit a felhasználók adhatnak meg magukról), valamint a követői számát illetve, hogy engedélyezte-e a GPS adatok közzétételét. A személymegjelölést tartalmazó tábla a „*tweet_user_mention*” a megjelölt személy azonosítóját, a tweet azonosítóját (amiben megjelölték), és a megjelölő személy azonosítóját tartalmazza. A „*tweet_retweet*” táblában azok a tweet azonosítók vannak, amire retweeteltek, valamint a retweetek azonosítója, és a létrehozásuk időpontja. Ez a három tábla a „*tweet*”, „*tweet_user_mention*” és a „*tweet_retweet*” képi az alapját a további tábláknak.

2.3. A szociális hálózat megválasztása

A szociális hálózatot egy gráffal reprezentáljuk, ahol a csúcsok a felhasználók, az élek pedig a közöttük levő kapcsolatok. A kapcsolatok definiálása nagyon fontos, hiszen ez alapján választjuk meg azt a szociális hálózatot, amit vizsgálni akarunk. Twitter adatoknál erre több lehetőség is kínálkozik. Azt viszont szem előtt kell tartanunk, hogy mit akarunk majd megvizsgálni a gráfon. Ebben a dolgozatban a szociális hálózatok klaszterezettségét és a

klaszterezettség felhasználók közötti távolságoktól való függését vizsgáljuk, illetve kíváncsiak vagyunk, hogy a kapcsolatok erőssége hogyan befolyásolja a hálózat összetettségét. Ehhez egy olyan gráfot kell felállítani, ahol a felhasználók valószínűleg ténylegesen ismerik egymást, vagy legalábbis online tartják a kapcsolatot. Az egyik lehetőség az, ha egy irányított gráfot állítunk fel, ahol az élek a felhasználók közötti követések és az irányuk a követő felhasználótól a követett felé mutat. Ilyen gráfokat gyakran használnak Twitter adatok elemzésénél [6]. A mi céljainknak egy ilyen gráf nem felel meg, hiszen sok olyan felhasználót követnek, akik nem követnek vissza. Ilyenek például a meteorológiai állomások vagy a hírességek. Csak a kölcsönös követésekből álló gráfot is vizsgálhatnánk, ám valószínűleg ebben gráfban sok olyan kapcsolat lenne, ami nem takar valós kapcsolatot. Egy másik módszer, ha olyan felhasználók között definiálunk kapcsolatot, akik egymás tweetjeire retweeteltek. Ekkor valamilyen ismeretséget feltételezhetünk a felhasználók között, hiszen válaszolnak egymásnak. Az ilyen gráffal az a baj, hogy egyes populáris bejegyzésre nagyon sok retweet érkezik (pl: a hashtaggal ellátott tweetek között van sok ilyen) és ezért olyan felhasználók kerülhetnek kapcsolatba egymással a gráfunkban, akik valójában nem ismerik egymást. Egy harmadik megoldás, ha az éleket úgy definiáljuk, hogy azok között a felhasználók között van él, akik megemlézték egymást valamelyik bejegyzésükben. Ekkor felállíthatunk egy irányított gráfot, vagy nézhetjük csak a kölcsönös említéseket. Mi az utóbbi módszert választottuk, ekkor ugyanis kölcsönös az ismeretség, bár nem feltétlenül fizikailag, lehet hogy csak online barátságról van szó. Egy így definiált gráf éleit többféleképpen tudjuk súlyozni. Lehet a felhasználók közötti távolsággal, illetve a kölcsönös említések számával, ami jelentheti akár a barátság "erősségét". A dolgozatban egy ilyen kölcsönös említésen alapuló gráfot használunk. Azért döntöttünk végül emellett, mert ekkor feltételezhetjük a legbiztosabban hogy két felhasználó között tényleges ismeretség van, illetve a kétféle -a céljainknak megfelelő- súlyozást tudunk bevezetni.

2.4. A vizsgált gráfok alapját képező tábla

Amint azt láttuk, az adatbázis négy fő táblából áll, és ebből három kell ahhoz, hogy a kölcsönös említéseken alapuló gráfot össze tudjuk állítani. Az adatbázis segítségével több kutatás is folyik [7],[8] az egyetemen, ezért az adatbázist használóknak megvan a lehetőségük arra, hogy létrehozzák a saját tábláikat. Mi is létrehoztuk a saját táblánkat, amit két másik alapján raktunk össze. Ezeket saját célra hozták létre az adatbázis felhasználói. Az egyik, az úgynevezett „links2” tartalmazza azoknak a felhasználóknak az azonosítóit, akik említették egymást, illetve azt, hogy melyik hányszor említette a másikat. Itt csak

olyan felhasználók vannak, akik legalább egyszer kölcsönösen megemlítették egymást. Ez a tábla adja meg a gráfunk éleit. A csúcsokat vehetnénk a „user” táblából, de mi egymásikat választottunk, az úgynevezett „user_location” táblát.

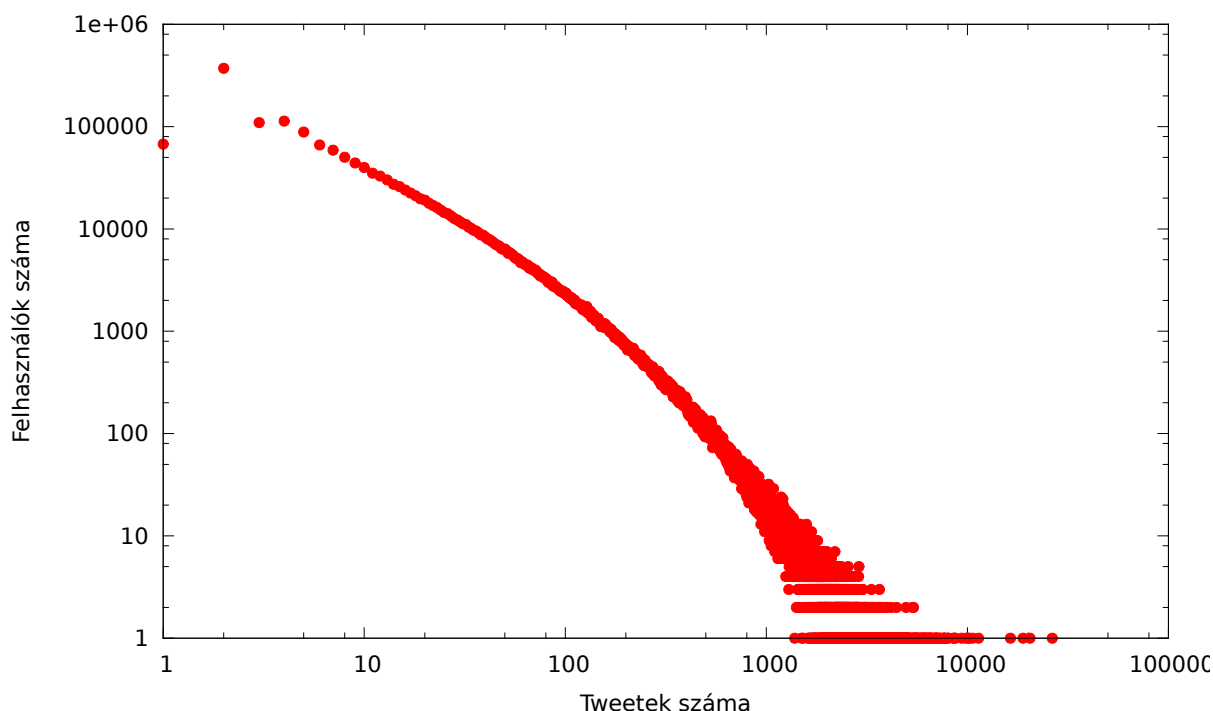
Ez tartalmazza a felhasználók azonosítóját és még sok adatot, mint például, hogy a felhasználók mennyit tweeteltek nappal, illetve éjszaka. A mi szempontunkból három oszlop nagyon fontos; a cluster_id, illetve a lat és lon oszlopok. A lat és lon a felhasználóhoz rendelt földrajzi koordináták, amiket a friend of friend algoritmussal határoztunk meg, [9] 1 km-es küszöbindexet használva. Az így kapott három legnagyobb klaszter középpontjának a koordinátái vannak felhasználónként a „user_location” táblában. Természetesen azokhoz a felhasználókhoz, akikhez nem tudtunk három klasztert azonosítani, kevesebb koordináta-párt rendeltünk. A cluster_id oszlopban három érték lehet; nulla, egy illetve kettő. A nullás a legtöbb tweet alapján meghatározott klaszter, a kettes pedig a legkevesebb tweet alapján. Mi ebben a dolgozatban a nulla cluster_id-val rendelkező adatokat használjuk.

Ebből a két táblából, ami tartalmazza az éleket, valamint a csúcsokat -úgy, hogy azokhoz földrajzi koordináták vannak rendelve- alkottuk meg a „user_mention_graph” táblát. Ez tartalmazza az éleket, azt, hogy hányszor említették egymást a felhasználók, a felhasználók koordinátáit és cluster_id-kat, illetve, hogy melyik felhasználó mennyit tweetelt nappal és éjszaka. Ez a tábla így 14 oszlopos és 22 736 734 soros. Ez nagyobb, mint amennyi szükséges lett volna a dolgozatbeli gráf adatainak a tárolására, de magában hordozza a lehetőséget, hogy a vizsgálatainkat esetlegesen kiterjesszük a jövőben.

2.5. A vizsgált gráfok

A fent említett „user_mention_graph” tábla tartalmazza a vizsgált gráfot. Ez 2 454 623 csúcsot és 3 396 576 élt tartalmaz. A gráf nem összefüggő, 247 288 egymástól független algráfból áll. Ebből a legnagyobb ún. óriáskomponens, vagy legnagyobb összekötött komponens, esetleg legnagyobb algráf 1 830 133 csúccsal, és 3 008 248 éllel rendelkezik. Ez azt jelenti, hogy sok olyan algráf van, ahol két csúcsot köt össze egy él. Ezért érdemes a legnagyobb összekötött komponens is külön megvizsgálni, hiszen az már egy egybefüggő struktúrát alkot. A gráfunkban nagyrészt aktív felhasználók vannak. Ezt onnan tudjuk, hogy a felhasználóinkhoz átlagosan 79 geo-tagged tweet tartozik. Ezek közül minden szűrés után átlagosan 45 tweetet használtunk arra, hogy a nullás cluster_id koordinátákat meghatározzuk, vagyis minden felhasználóhoz átlagosan ennyi koordinátás adatból számoltunk tartózkodási helyeket. Persze sok olyan felhasználó van, akihez csak 1-2 tweet alapján tudtunk tartózkodási helyet megadni.

Ezt a felhasználóhoz tartozó tweetek eloszlása alapján lehet megmondani.



1. ábra. A felhasználók száma azon tweetek számának a függvényében, melyeket a 0 cluster_id-s koordináták meghatározásához használtunk.

A teljes adatbázis alapján megvizsgáltuk az összes felhasználót, akihez tartózkodási helyet tudtunk rendelni. Azt találtuk, hogy itt átlagosan 23.82 tweetet használtunk arra, hogy a 0-ás cluster_id koordinátákat meghatározzuk. Ez azt jelenti, hogy egy sokkal megbízhatóbb adatsort kaptunk, mintha az összes felhasználóból alkotott gráfot vizsgáltuk volna.

2.6. A gráfok súlyozása

A gráfokat többféleképpen súlyozhatjuk. Az egyik lehetőség, ha a kölcsönös említések számával súlyozunk. Ekkor, ha két felhasználó között sok kölcsönös említés van, az erős kapcsolatra utal, hiszen a gráfba csak azok a felhasználók kerülhettek bele, akikhez tartozik legalább egy kölcsönös említés, vagyis az így súlyozott gráfok élsúlya minimum 1.

A másik súlyozási lehetőség, hogy a felhasználók közötti távolságokat adjuk meg súlyként. A távolságot itt nem lehet egyszerűen kiszámolni, hiszen a Földön vagyunk. A dolgozatban bemutatott számítási módszer pontosságát figyelembe véve számunkra megfelelő a Föld gömbbel történő közelítése. Egy gömbön két pont távolságán a közöttük

futó legrövidebb félkörív hosszát értjük. Az erre vonatkozó képletek közül a numerikusan jól kondicionált haversine formulát használjuk. [10]

$$\Delta\omega = 2\arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\Theta}{2}\right) + \cos\Theta_1\cos\Theta_2\sin^2\left(\frac{\Delta\phi}{2}\right)}\right) \quad (1)$$

A képletben polárkoordináták vannak. A Θ polárszög az adott pontba mutató egységvektor z tengellyel bezárt szögét, míg ϕ azimut-szög a vektor xy síkra eső merőleges vetületének, az x tengellyel bezárt szögét jelöli. Mivel távolságokat számolunk, a tengelyek iránya nem számít, mégis jó ha az xy síkot úgy választjuk meg, hogy rajta nyugodjon az egyenlítő, és a z tengelyt a Föld forgástengelyének választjuk úgy, hogy a z irányú egységvektor Észak felé, az x irányú egységvektor pedig a 0 földrajzi hosszúságú fok felé mutasson.

A fenti képlet egy egység sugarú gömbre vonatkozik. A számoláshoz kell a Föld átlagos sugara, mi az irodalomban szokásos $R = 6371$ kilométeres értéket használtuk. Mivel koordinátákat hosszúság és szélesség formátumban kapjuk meg a letöltésnél. A távolságokat egy olyan SQL-ben írt programmal számoltuk, amely bemenetként hosszúság és szélesség adatokat kap és a végén a távolságokat kilométerben mérve adja ki. Ezt minden felhasználópárra, akik között él van kiszámolva megkapjuk a távolság szerinti súlyokat.

3. Vizsgálati módszerek

Gráfok vizsgálatánál releváns kérdés, hogy a gráfban vannak-e jobban összekötött részek, vagy az élek inkább homogénen oszlanak el a csúcsok között. A gráfba jobban beágyazott részeket klikkeknek nevezzük. Azt, hogy mi számít klikknek, többféleképpen is definiálhatjuk. Szigorúan véve azoknak a csúcsoknak a halmaza alkot klikket, amelyekre teljesül, hogy bármely két csúcs között van él, de ennél lehet enyhébb definíciókat is mondani. Az enyhébben definiált csoportokat nevezzük klasztereknek. Mi a lehető legkisebb klikkek, vagyis háromszögek segítségével vizsgáljuk a hálózatok klaszterezettségét.

Mivel mi a hálózat egészére vagyunk kíváncsiak, ezért olyan módszerekkel kell vizsgálnunk a gráfot, amik a globális struktúráról adnak információt, valamint kíváncsiak vagyunk a klaszterezettség különböző paramétereiktől való függésére. Ezért a céljainknak megfelelő módszereket kell választanunk.

A gráfok kezelésére és manipulálására mi a python programnyelvet használtuk (2.7.3-as verzió), ennek is főleg a networkx[11] nevű kiegészítőcsomagját (1.6-2-es verzió). Ez egy

gráfkezelő csomag, ami nagyon leegyszerűsíti a gráfokkal való munkát a beépített függvényei segítségével. A python nyelv előnye az átláthatóság és egyszerűség, a networkx csomag segítségével pedig a gráfokon végzett bonyolult műveletek is leprogramozhatóak néhány sorban. A gráfok könnyű kezelhetősége miatt elég az adatainkat egy él listát tartalmazó adatfile-ban tartani, ami a súlyokat is tartalmazza, mert a program a futása közben generálja le a gráfot, és azt egy változóként kezeli, így téve egyszerűvé a vele való munkát. A networkx csomag használatával kikerültük azt, hogy nekünk kelljen megírni a gráfkezelő programok nagy részét. A python nyelv hátránya viszont a viszonylag hosszú futási idő és a nagy számolási kapacitás igény. Így az általunk használt programok rövidek és könnyen átláthatóak, ellenben hosszú futási idejűek.

3.1. Klaszterezettséget jellemző mennyiségek

Mivel a hálózat globális struktúrájára vagyunk kíváncsiak, ezért érdemes a klaszterezettséget valószínűségi, illetve statisztikus paraméterekkel leírni. A vizsgált mennyiségeket úgy választottuk meg, hogy azok csak a gráf struktúrájától függjenek, vagyis a felhasználók közötti kapcsolatoktól, illetve a háromszögek² számától. Három ilyen mennyiséget mértünk a gráfjainkra, az átlagos klaszterezettségi együtthatót, az átlagos geográfiai klaszterezettségi együtthatót, illetve a tranzitivitást.

3.1.1. Átlagos klaszterezettségi együttható

Az irodalomban gyakran említett globális gráftulajdonság, az átlagos klaszterezettségi együttható. A kiszámolásához előbb definiálnunk kell a klaszterezettségi együtthatót. Egy adott csúcs klaszterezettségi együtthatója a szomszédai között található élek és a szomszédai között maximálisan lehetséges élek számának a hányadosa. Ezt az alábbi formulával számolhatjuk:

$$c_i = \frac{2T_i}{n_i(n_i - 1)} \quad (2)$$

Ahol c_i az i -edik pontra számolt klaszterezettségi együttható, T_i a háromszögek száma amiben az i -edik csúcs részt vesz, n_i pedig az i -edik csúcs fokszáma. Amikor az összes csúcstra kiszámoljuk ezt az együtthatót és kiátlagoljuk, azt nevezzük átlagos klaszterezettségi együtthatónak. Ezt a következőképpen kaphatjuk meg:

²Amikor egy csúcs két szomszédja között is van kapcsolat

$$C = \frac{1}{N} \sum_{i \in G} c_i \quad (3)$$

Ahol C az átlagos klaszterezettségi együttható, N a csúcsok száma, G a gráf csúcsait tartalmazó halmaz, c_i pedig a fent definiált klaszterezettségi együttható az i -edik csúcsra. Ez az érték az $\frac{1}{N}$ -es szorzó miatt nulla és egy közé esik, ami azt jelenti, hogy ez az érték valamilyen valószínűséget jelent. c_i is nulla és egy közé esik, ami akkor egy, ha az adott csúcsra az összes lehetséges háromszög megtalálható, és akkor nulla, ha egy sem. Ezzel az együtthatóval egyfajta tranzitivitást jellemezhetünk a hálózatban, hiszen az átlagos klaszterezettségi együttható megadja, hogy egy véletlenül választott csúcs szomszédai milyen valószínűséggel kötődnek egymáshoz.

A saját ismeretségi körünket megnézve azt láthatjuk, hogy nagyon sok ismerősünk ismeri egymást, ezért egy valós hálózatra nagy valószínűségi értéket mondanánk. A mi gráfunk a társadalomból vett *minta*, amiben vannak olyan felhasználók, akik esetleg csak egy ismeretssel rendelkeznek. Az ilyen felhasználók létezése elképzelhetetlen a valóságban (nincs olyan ember akinek csak egy ismerőse van). Ezek a felhasználók csökkentik az átlagos klaszterezettségi együttható értékét, ezért ennek a változását érdemes vizsgálni.

A szociális hálózatra mért átlagos klaszterezettségi együttható: 0.08177, míg a legnagyobb összefüggő algráfra ez az érték: 0.09879. Ez nem meglepő, hiszen a nagyobb gráfnál az, hogy szeparált algráfok vannak, és hogy ebből sok az olyan csúcspár, amik csak egymással állnak kapcsolatban, lerontja az átlagos klaszterezettségi együttható értékét.

3.1.2. Geográfiai klaszterezettségi együttható

Egy súlyozott hálózaton felmerül a lehetőség, hogy klaszterezettségi együtthatót más módon is számoljunk, úgy, hogy figyelembe vesszük a súlyokat. Szociális hálózatoknál a struktúrát reprezentáló gráf súlyai gyakran a távolságok, ezért már találhatók a távolságokat figyelembe vevő módszerek az átlagos klaszterezettségi együttható kiszámítására az irodalomban. Ebben a dolgozatban bemutatunk egy ilyen módszert, amit a szakirodalomban találtunk. [12] A vizsgált együtthatót az irodalomban irányított gráfokta mondták ki, de mi irányítatlan gráfokra definiáltuk azt.

A geográfiai klaszterezettségi együtthatót tekinthetjük úgy, mint az átlagos klaszterezettségi együttható kiterjesztett változatát. Az előbb definiált klaszterezettségi együttható egy csúcs szomszédaihoz tartozó élek és lehetséges élek arányát mondja meg, míg ennek a geográfiai adaptációja másképp súlyozza az olyan háromszögeket, ahol a pontok földrajzilag közel vannak egymáshoz, és másképp azokat, ahol a pontok távol vannak egymástól.

Az i -edik csúcsra vett geográfiai klaszterezettségi együttható ugyanúgy van definiálva, mint a klaszterezettségi együttható, de minden háromszög, amiben az i -edik csúcs részt vesz, kap egy súlyt:

$$w_{ijk} = e^{-\frac{\Delta_{ijk}}{\beta}} \quad (4)$$

Itt i , j , és k a háromszöget alkotó csúcsok, Δ_{ijk} a háromszög leghosszabb oldala, β pedig egy paraméter. Ez a számolási mód a kis háromszögeket részesíti előnyben; ha az egyik csúcs nincs közel a másik kettőhöz, ez az érték drasztikusan csökken. Az i -edik pont azon szomszédaihoz, akik között nincs él Δ_{ijk} -hoz 0-t rendelünk, hiszen ekkor lesz egy csúcs geográfiai klaszterezettségi együtthatója 0 akkor, ha a szomszédai nem ismerik egymást. Ekkor a geográfiai klaszterezettségi együttható egy csúcsra:

$$gc_i = \frac{2}{n_i(n_i - 1)} \sum_{j,k \in \Gamma_i} w_{ijk} \quad (5)$$

ahol a szumma az összes lehetséges a ponthoz tartozó háromszögre van kiterjesztve. n_i az i -edik pont fokszáma, Γ_i az i -edik pont szomszédainak a halmaza, tehát $n_i = |\Gamma_i|$, Δ_{ijk} pedig a fent definiált súly. Mivel pontosan $n_i(n_i - 1)$ különböző elrendezés van a szomszédok párosítására, és csak a létező háromszögek adnak járulékot a szummába, ezért az eredmény nulla és egy közé lesz normálva. Ez az összefüggés még csak egy adott csúcsra ad értéket, ezért, hogy az egész hálózatra megkapjuk az átlagos geográfiai klaszterezettségi együtthatót, a (3)-as egyenlethez hasonlóan ki kell átlagolnunk a pontokra számolt értékeket:

$$GC = \frac{1}{N} \sum_{i \in G} gc_i \quad (6)$$

Ahol C az átlagos geográfiai klaszterezettségi együttható, N a csúcsok száma, G a gráf csúcsait tartalmazó halmaz, gc_i pedig a fent definiált geográfiai klaszterezettségi együttható az i -edik csúcsra.

Ez az együttható figyelembe veszi az élek hosszát is, így nem egy klasszikus értelemben vett valószínűség ad. Akkor lesz nagy ez az érték, ha a hálózatban sok háromszög van, amelyeknél a csúcsok közel vannak egymáshoz, és akkor lesz alacsony, ha kevés háromszög van, vagy a háromszögekben a pontok távol vannak egymástól.

Még nem volt szó β szerepéről, ami egy szabadon megválasztható paraméter. Ennek a megválasztásától függ a geográfiai klaszterezettségi együttható értéke, de ha két hálózatot hasonlítunk össze, akkor mindegy, hogy mit választunk β -nak, mindig ugyanarra a

hálózatra kapjuk a nagyobb értéket. Tehát, ha β -t változtatjuk, akkor a különböző hálózatokra mért értékek aránya is változik, de ha nagyság szerint állítjuk sorba az értékeket, akkor minden β mellett ugyanúgy lennének sorba állítva. Felmerül a kérdés, hogy milyen β -t válasszunk egy-egy hálózathoz. Úgy érdemes választani, hogy β valami a hálózatra jellemző méret legyen, mert így sok β -nál nagyobb és kisebb él is lesz, így a közöttük levő különbség szignifikáns lesz. Jó választás lehet az összes pont átlagos távolsága (ami nem egyezik meg az élek hosszának az átlagával). A dolgozatban mi is így határoztuk meg β értékét, ami a teljes gráfra 8282.9371, vagyis ennyi a teljes gráfban a felhasználók közötti átlagos távolság km-ben.

A teljes szociális hálózatunkra 0.0757-nak mértük ezt az értéket, míg az óriáskomponensre ez az érték: 0.09136. Itt is hasonló okok miatt nagyobb a legnagyobb algráf geográfiai klaszterezettségi együtthatója, mint amit az átlagos klaszterezettségi együtthatónál leírtunk.

3.1.3. Transzitivitás

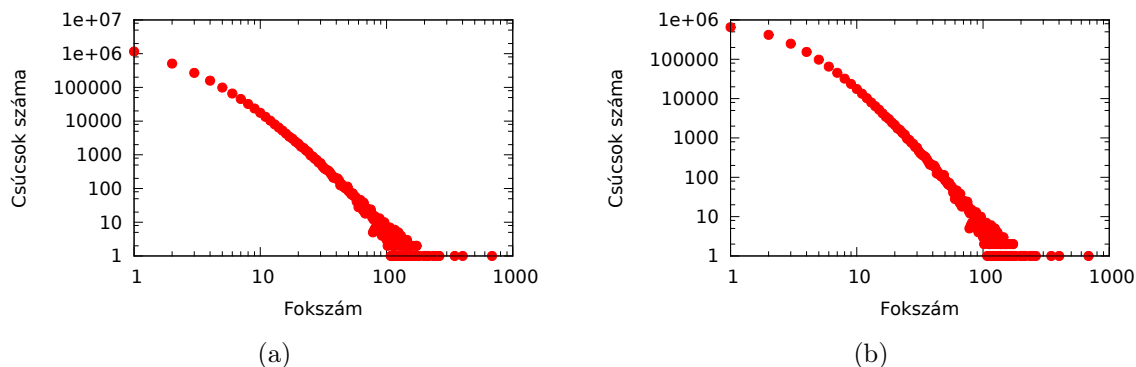
Tegyük fel, hogy egy szociális hálózatban A ismeri B-t és C-t, ekkor ha C is ismeri B-t akkor azt mondjuk, hogy a kapcsolat tranzitív, a három személy egy háromszöget alkot. Az ilyen háromszögek, vagy klikkek szerepe fontos a szociális hálózatokban, mert a háromszögek relatív számából lehet következtetni arra, hogy a hálózat mennyire "sűrű szövésű". Azt, hogy az összes lehetséges háromszög közül mennyi van a gráfban, a transzitivitás mondja meg. Ez az érték nulla és egy közé van normálva és az alábbi módon számolható:

$$T = 3 \frac{\text{A gráfban található háromszögek száma}}{\text{A lehetséges háromszögek száma}} \quad (7)$$

Meg kell jegyeznünk, hogy a transzitivitás és az átlagos klaszterezettségi együttható nem független egymástól, mindkettő a hálózatban levő háromszögek számától függ. Az általunk használt szociális hálóra ez az érték: 0.09101, míg a legnagyobb algráfra: 0.09037. Itt a két érték nem nagyon különbözik, ami azzal magyarázható, hogy az óriáskomponens csúcsai az eredeti gráf csúcsainak körülbelül a 75%-t, az élek száma pedig az eredeti gráf éleinek a 88%-t teszik ki. Így a legnagyobb algráfban kevesebb a lehetséges háromszögek száma, és a pontok számához relatíve több él tartozik, ezért magas ez az érték, és ehhez képest alig alacsonyabb az algráf esetén, mint az eredeti szociális hálónál.

3.2. Véletlen hálózatok

Az egyik vizsgálati módszerünk az, hogy az eredeti gráfokra mért adatokat vizsgáljuk véletlen hálózatokra. Ilyen gráfokat az eredetiekből állítottunk elő úgy, hogy bizonyos tulajdonságaik megegyezzenek, mint például a csúcsok vagy az élek száma. Egy másik számunkra fontos gráftulajdonság, amit szeretnénk megtartani a véletlen gráfok generálásánál, az a fokszámeloszlás.



2. ábra. A gráfok fokszámeloszlása: az (a) a teljes gráfé, míg a (b) ábra a lenagyobb algráfé. A randomizáció után a véletlen hálózatoknak is pont ugyanilyen a fokszámeloszlása.

Ez egy, a hálózatra globálisan jellemző dolog, ami azt adja meg, hogy milyen a felhasználók kapcsolatai számának az eloszlása. Mi azt szeretnénk tehát elérni a fokszámeloszlás megtartásával, hogy ugyanannyi felhasználó legyen akinek sok, és ugyanannyi, akinek kevés kapcsolata van mind az eredeti mind a belőle alkotott véletlen gráfokban is.

Ezeket a véletlen hálózatokat tekintem majd referenciamodellnek. A véletlen gráfokat nem az Erdős-Rényi modell [13], vagy a Barabási–Albert modell [14] segítségével alkottuk meg, hanem az eredeti hálózatból generáltuk. Egy gráfból könnyen előállítható egy másik ugyanakkora gráf, ha a csúcsok közötti éleket összekeverjük. Ezt úgy lehet megtenni, hogy kiválasztunk véletlenszerűen két élt és az őket alkotó csúcspárokból kiválasztunk véletlenszerűen kettőt-kettőt, ami megadja az új éleket. Ha ezt elégszer meg tesszük, akkor egy teljesen új, véletlen hálózatot kapunk.

Az ilyen randomizáció jól működik, ha matematikai hálózatokat vizsgálunk, de itt az éleknek fizikai jelentésük van, ezért ki kell zárunk az olyan lehetőségeket, mint a többszörös élek, vagy a hurokélek. Az ilyen lehetőségeket a randomizáló algoritmus írása közben szem előtt tartottuk és kiküszöböltük. Mi tehát olyan véletlen hálózatokat generálunk, amik megtartják a csúcsok és élek számát, valamint a fokszámeloszlásuk is ugyanolyan, mint az eredeti gráfnak. Az ilyen hálózatok lehetnének akár valós szociális hálózatok is, ahol ugyanúgy vannak népszerű felhasználók sok kapcsolattal és kevésbé népszerűek kevesebb

kapcsolattal.

Az alap gráf, amiből a véletlen hálózatokat generáljuk súlyozott, felmerül tehát a kérdés, hogy a kapcsolatok kicserélése után a súlyok hogyan változnak. Ha a súlyok a felhasználók közötti távolságok voltak, akkor nem kell mást tennünk, mint összekeverni az éleket és az új gráfra kiszámolni a kapcsolatban álló felhasználók közötti távolságokat és azokkal súlyozni a létrehozott véletlen hálózatot. Kölcsönös említésekkel súlyozott gráfoknál nem áll fenn ez a lehetőség, ezért csak a távolságokkal súlyozott gráfokból alkottunk véletlen hálózatokat.

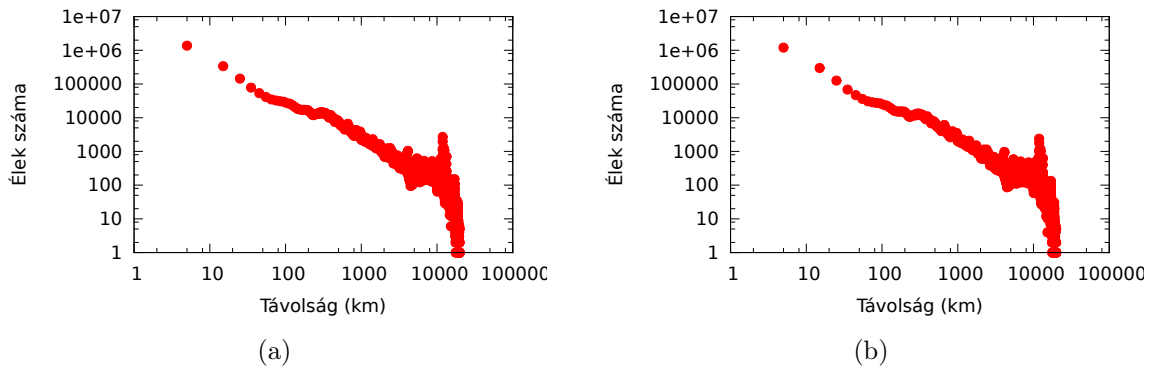
A fent említett módszerekkel létrehoztunk a különböző esetekre egyenként ezer véletlen hálózatot, és az azokra mért gráftulajdonságokból statisztikákat állítottunk fel, majd megvizsgáltuk az eredeti gráfok helyét a statisztikában és ezekből következtettünk annak a tulajdonságaira.

Az átlagos klaszterezettségi együttható és a tranzitivitás mérésekor azt várjuk, hogy a szociális hálózatra ezek az értékek nagyobbak lesznek, mint az eredet gráfból létrehozott véletlen hálózatokra, hiszen a társadalmunkban levő kapcsolatok nem véletlenszerűen jönnek létre. Az ilyen eredmény nem lenne meglepő, inkább az a kérdés, hogy a valós hálózatra mért átlagos klaszterezettségi és átlagos geográfiai klaszterezettségi együttható valamint a tranzitivitás belesimul-e a statisztikákba, vagy inkább outlier-ként viselkedik. A geográfiai együttható mérésekor is azt várjuk, hogy a valós hálózatra mért érték nagyobb lesz, mint a véletlen hálózatra, viszont az nem triviális, hogy hogyan befolyásolják a súlyok a véletlen hálózatokból mért értékeket.

3.3. Élek elhagyása

A klaszterezettség vizsgálatára még egy módszert találtunk ki a véletlen hálózatok generálásán kívül. Ez a módszer az élek elhagyása. Súlyozott gráfoknál felmerülhet a kérdés, hogy hogyan változnak a gráf tulajdonságai, ha a nagy, vagy éppen kis súllyal rendelkező éleket kivesszük a hálózatból. Mi is ezt a két módszert követtük, elhagytuk a gráfból a bizonyos értéknél nagyobb súllyal rendelkező éleket, így generáltunk kisebb hálózatokat, illetve úgy, hogy a kisebb súlyú éleket hanyagoltuk el. Ezt más-más módszerrel tettük meg a távolságokkal és a kölcsönös említésekkel súlyozott gráfoknál.

Az élek távolság szerinti súlyait 10 km-es binékbe osztottuk, és úgy ábrázoltuk az élek számát a távolságok függvényében. Az így kapott görbék nagyjából hatványfüggvényt követnek, méghozzá úgy, hogy a kumulált eloszlás (CDF) nagyjából logaritmikus lesz.

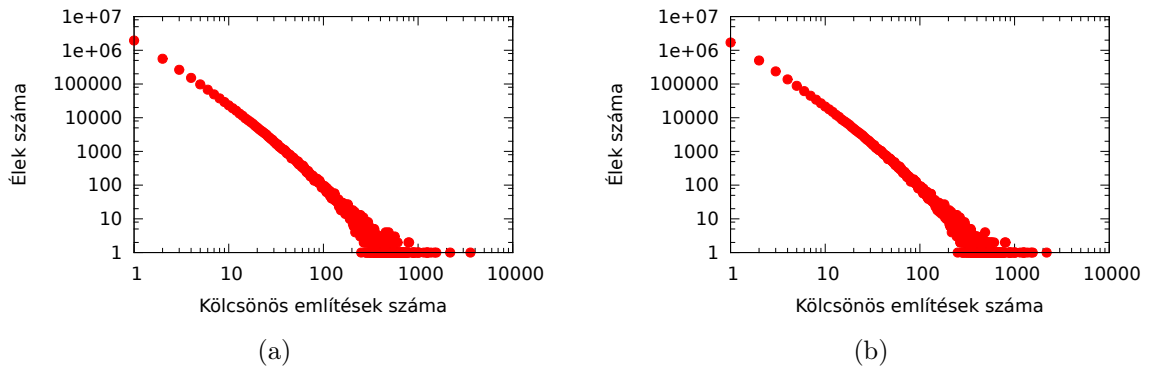


3. ábra. Az élek távolsággal súlyozott eloszlása. Az (a) ábra a teljes gráfé, a (b) ábra pedig a legnagyobb összekötött komponensé. (Itt az adatpontok a binok közepére mutatnak)

Mivel az eloszlások több nagyságrendet is tartalmaznak, a küszöbtávolságokat egy olyan algoritmussal határoztuk meg, ami logaritmikus sorban adja ki a számokat úgy, hogy a felhasználók közötti átlagos távolságot is megadja. A küszöbadatsor végére még önkényesen tettünk egy adatpontot, hogy megfelelően nagy küszöbindex is legyen. Az adatsort logaritmikus skálán ábrázolva egy lineáris jellegű pontsorozatot kapunk. Ekkor a küszöbindexek sűrűsége nagyjából arányos lesz a különböző hosszú élek számával. Így több mérési adatunk lesz a kis távolságokkal súlyozott élek tartományából, míg a nagy súllyal rendelkező élek tartományában ritkábban mérünk, de ez nem baj, mert abban a tartományban sokkal kevesebb az él.

Az így meghatározott küszöbindexek alapján generáltunk "csonka gráfokat". Amikor a nagy távolsággal súlyozott éleket hagytuk el, akkor egyre kisebb szeparált csoportokra bontottuk a gráfot úgy, hogy a csoportok tagjai közel vannak egymáshoz. Ellenkező esetben az egymáshoz földrajzilag közel levő felhasználók között szüntettük meg az éleket, így is szétesik a hálózat, de a megmaradó csoportok tagjai távol vannak egymástól. Az ilyen hálózatokra számoltunk klaszterezettségi és geográfiai klaszterezettségi együtthatókat. A különböző küszöbértékekhez tartozó együtthatók változását vizsgáltuk.

A másik lehetőség hogy a kölcsönös említések szerint súlyozott gráfból hagyunk el éleket. Ekkor a küszöbindexeket lineárisan növeltük (1,2,3...stb), mert a kölcsönös említések száma a barátság erősségére utal. Így olyan hálózatokat hoztunk létre, ahol csak az "erős" barátságok, illetve csak a "gyenge" barátságok tartják össze a gráfot. Ezekre mértünk klaszterezettségi együtthatót, és vizsgáltuk ennek a változását a különböző hálózatokra.



4. ábra. Súlyok eloszlása kölcsönös említések szerint az (a) ábra a teljes gráf, míg a (b) ábra a legnagyobb algráf adatai alapján készült.

Csonka gráfok generálásánál az új hálózatokból kihagytuk az izolált pontokat, hogy csak az adott erősségű kapcsolatokról ténylegesen függő felhasználók maradjanak a gráfban. Az ilyen fajta vizsgálatokat elvégeztük az óriáskomponensre is, de mivel ez egy összefüggő gráf, ezért erre egy másik tulajdonságot is mértünk. Ez pedig az a tulajdonság, hogy adott küszöbindex mellett mennyi algráfra esik szét az eredeti struktúra. Ezt csak a barátságok erőssége szerint vizsgáltuk.

Egy ilyen vizsgálatnál az eredmények nem triviálisak, de általánosságban a tulajdonságok csökkenését/növekedését várjuk, illetve azt, hogy a hálózat nem a kölcsönös említések küszöbindexének a növelésével arányosan esik szét több darabra. Ezzel a módszerrel megvizsgáljuk azt a tézist is, hogy a szociális hálózatokat a gyenge kapcsolatok tartják össze.[15]

3.4. Vizualizáció

A vizualizáció egy hasznos eszköze szinte minden kutatásnak. Nem csak a megértést segíti jobban, hanem többletinformációt is szolgáltat. A mi esetünkben az adataink megbízhatóságát is lehet vele szemléltetni, hiszen ha ábrázoljuk a gráfunkat a Földön, akkor nagyobb csomópontoknak, vagyis városoknak kell kirajzolódniuk a megfelelő helyeken. Az ábrázolástól elsősorban ezt várjuk, de érdekes lehet azt is megfigyelni, hogy ha kiveszünk különböző felhasználókat a gráfból, akkor hogyan változik például a felhasználók sűrűségeloszlása. A Twitter projekt keretében több vizualizációs programot is kifejlesztettek a csoport tagjai. Én a dolgozatomhoz ezek közül a programok közül használok kettőt. Az egyikkel a gráfot, a másikkal pedig a felhasználók sűrűségeloszlását tudjuk ábrázolni a Földön.

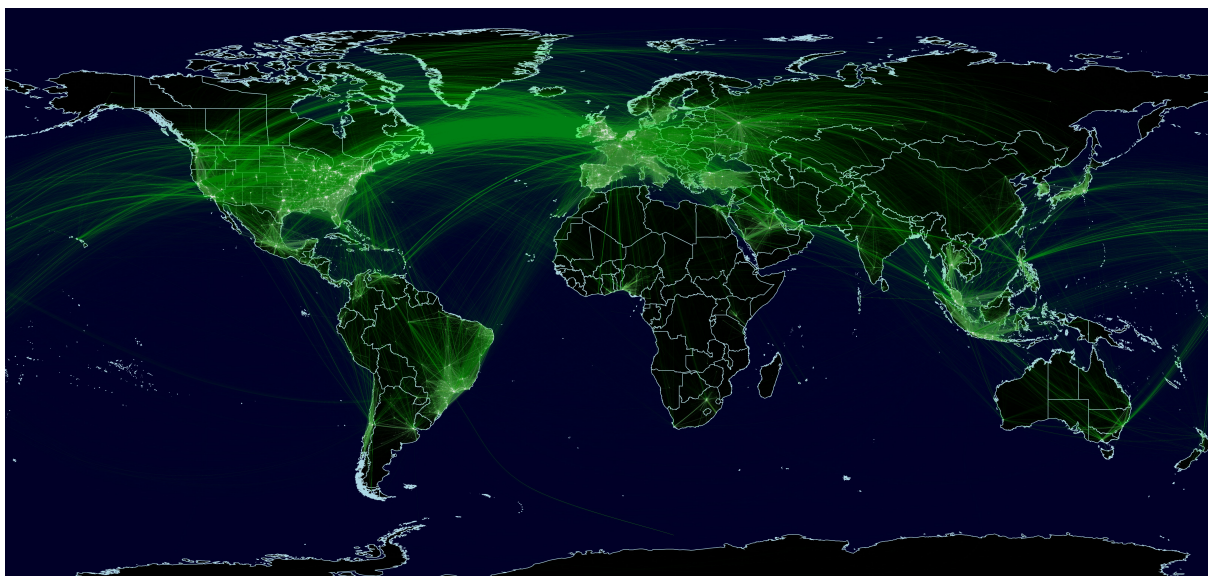
4. Eredmények

Munkánk során egy szociális hálózatot és a legnagyobb összetett komponenst vizsgáltuk. Ezek adatait az alábbi táblázat tartalmazza.

	Teljes gráf	Óriáskomponens
Csúcsok száma	2 454 623	1 830 133
Élek száma	3 396 576	3 008 248
Algráfok száma	247 288	1
Átlagos klaszterezettségi együttható	0.08177	0.09879
Geográfiai klaszterezettségi együttható	0.0757	0.09136
Tranzitivitás	0.09101	0.09037

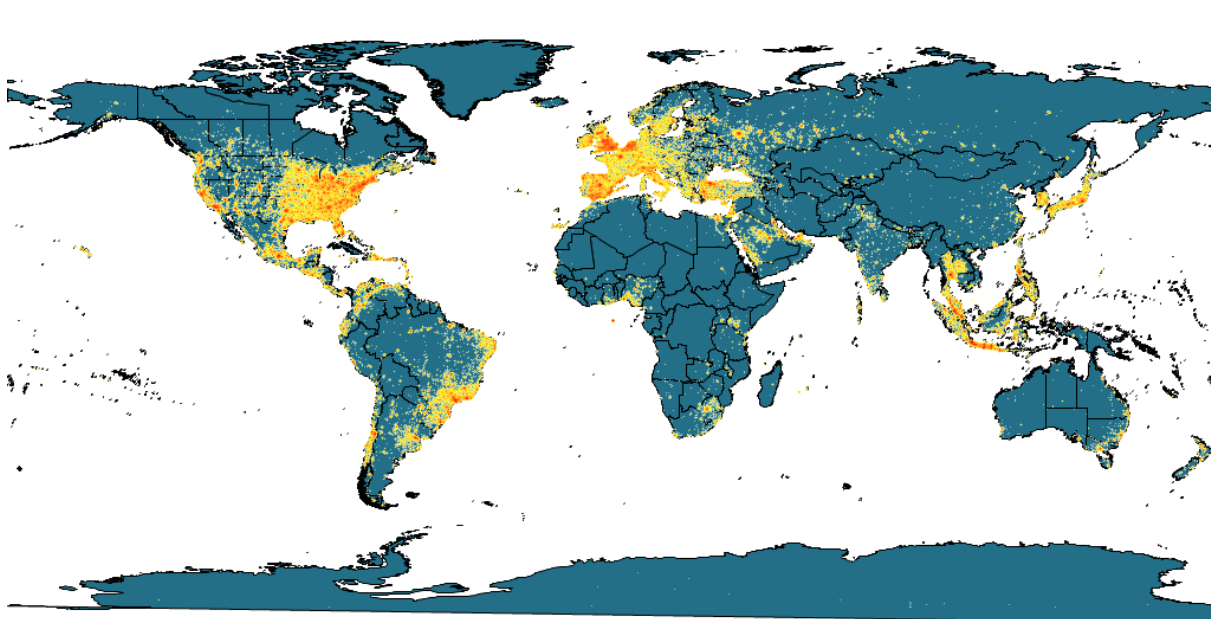
1. táblázat. *A vizsgált gráfok adatai*

A gráfokat először érdemes ábrázolni és csak utána dolgozni velük, hiszen már az ábrázolásból látszik, hogy jó-e az adatszettünk. Mivel ebben a dolgozatban a vizualizációt kiegészítő módszerként alkalmazzuk, csak pár példát hozunk ábrázolásokra. A legtriviálisabb ábrázolása egy ilyen adatsornak az, ha gráfként ábrázoljuk a Földön.



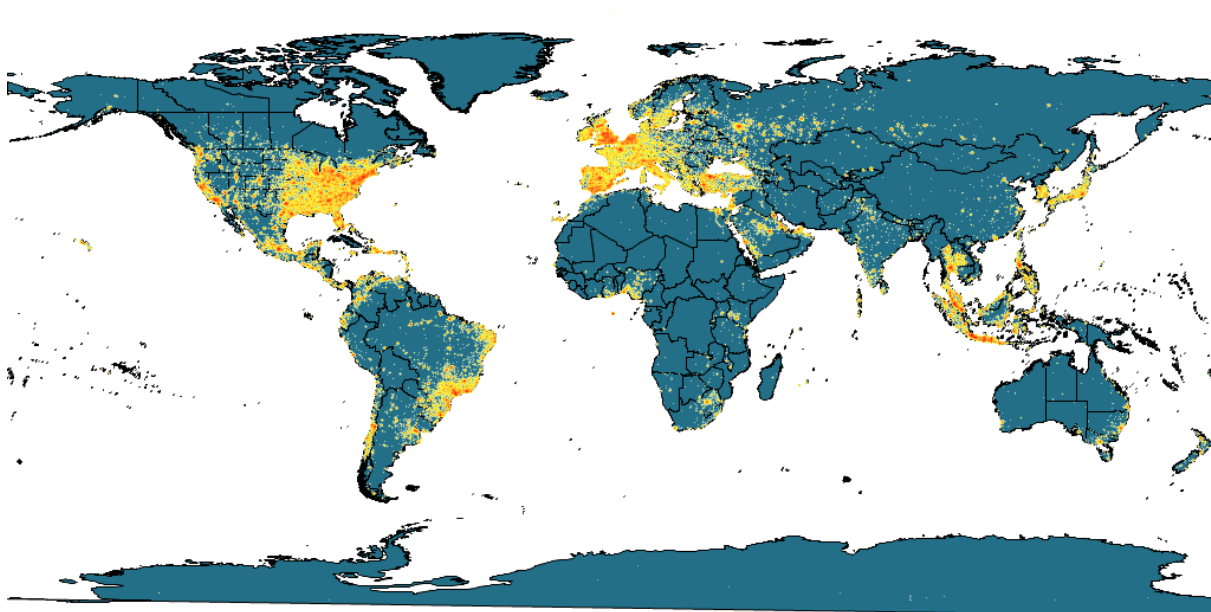
5. ábra. A teljes gráf ábrázolása a Földön

Egy másik lehetséges ábrázolás az, ha a felhasználók sűrűségének az eloszlását ábrázoljuk. Az ábrákon a piros tartományokon vannak a legsűrűbben a felhasználók, és minél sárgább egy tartomány annál ritkábban helyezkednek el az adott régióban.



6. ábra. A teljes gráf felhasználóinak a sűrűségeloszlása a Földön

Ezzel a módszerrel ábrázoltuk a legnagyobb összekötött komponenst is.



7. ábra. Az óriáskomponens felhasználóinak a sűrűségeloszlása a Földön

Látszik, hogy a két gráf földrajzi elhelyezkedése között nincs nagy különbség, illetve ez nem mutatható ki a mi módszereinkkel. Mivel a két gráf ábrázolásai hasonlóak, nem feltétlenül érdemes külön vizsgálni őket.

Az ábrákon megjelentek a nagyvárosok, és a legtöbb felhasználót olyan helyre azonosítottuk, ahol a Twitter-t széles körben használják. A kapcsolatokat megnézve sem találunk

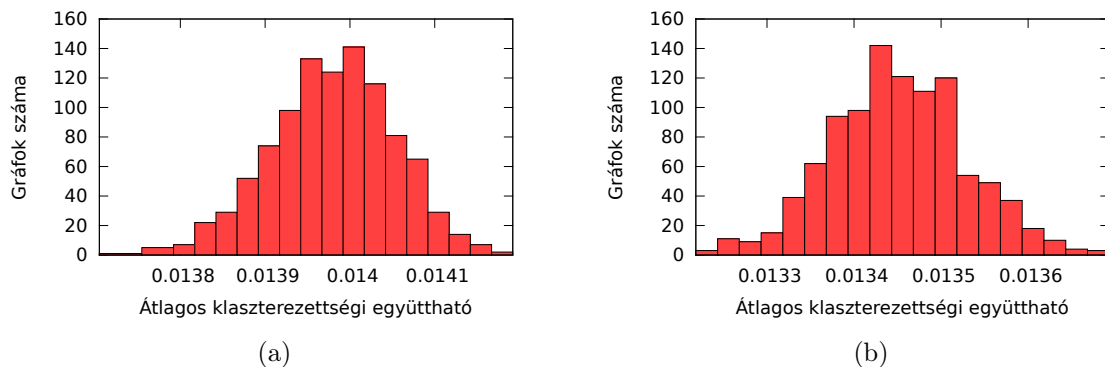
jelentős számú valótlannak tűnő kapcsolatot, vagyis az ábrákból azt állapíthatjuk meg, hogy a vizsgált adataink jól reprezentálják a valóságot.

A további eredményeket a vizsgálati módszerek szerint mutatjuk be. Ez alól kivétel a hálózat gyenge kapcsolatoktól való függése, ami külön kerül tárgyalásra.

4.1. Véletlen hálózatok

A gráfjainkból 1000 véletlen gráfot generáltunk és ezekre mértünk átlagos klaszterezettségi együtthatót, átlagos geográfai klaszterezettségi együtthatót, valamint tranzitivitást. A véletlen gráfokra mért értékeket hisztogramként ábrázoltuk, és ezek alapján vontuk le a következtetéseinket. Ezeknél a méréseknél azt vártuk, hogy a valódi hálózatra mért értékek nagyobbak, mint a véletlen hálózatokra mérték.

A szociális hálózatra mért átlagos klaszterezettségi együttható: 0.08177, míg a legnagyobb összefüggő algráfra ez az érték: 0.09879. Az átlagos klaszterezettség mérésénél

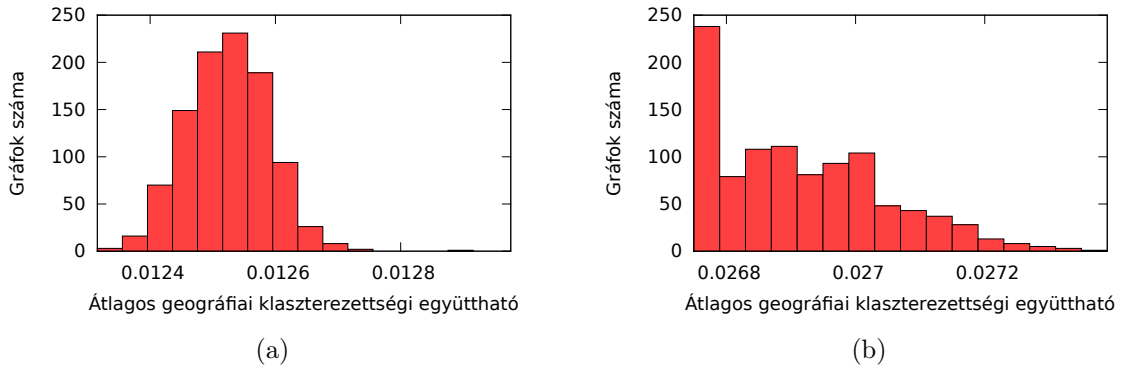


8. ábra. A véletlen hálózatokból alkotott hisztogram 0.000025-ös szélességű binekkel, ahol az (a) ábra a teljes gráfból alkotott, a (b) ábra pedig a legnagyobb algráfból alkotott hálózatok értékeinek az eloszlása

azt találtuk, hogy a kapott hisztogramok jó közelítéssel normáeloszlást alkotnak. Az itt kapott eloszlásoknál a legnagyobb átlagos klaszterezettségi együttható értéke: ~ 0.0142 a teljes gráfra, a legnagyobb részgráfra pedig ~ 0.0137 . A várakozásainknak tehát megfelel az az eredmény, miszerint a valós hálózatok jobban összekötöttek, mint a véletlen hálózatok. Ami viszont meglepő, hogy a valós hálózatok egyáltalán nem illenek bele a statisztikákba. Valós hálózatok esetén körülbelül 6-7-szer nagyobb az átlagos klaszterezettségi együttható, vagyis a valószínűsége annak, hogy egy felhasználó barátai is ismerik egymást.

Az geográfai klaszterezettségi együttható vizsgálata esetén a gráfokból először véletlen hálózatokat generáltunk úgy, hogy nem tartottuk meg az élsúlyokat, hanem a randomizáció

után újrasúlyoztuk a gráfot a felhasználók közötti valós távolságokkal. Ha megtartottuk volna az éleket, akkor valószínűleg más eredményeket kaptunk volna, de a mi célunk nem az ilyen absztrakt struktúrák vizsgálata. A gráfokra számolt értéke a geográfiai klaszterezetségi együtthatónak: 0.0757 a teljes hálózatra és 0.09136 a legnagyobb összekötött komponensre. Itt is azt várjuk, hogy a valós hálózatokra mért értékek nagyobbak lesznek, mint a véletlen hálózatokra mérték.



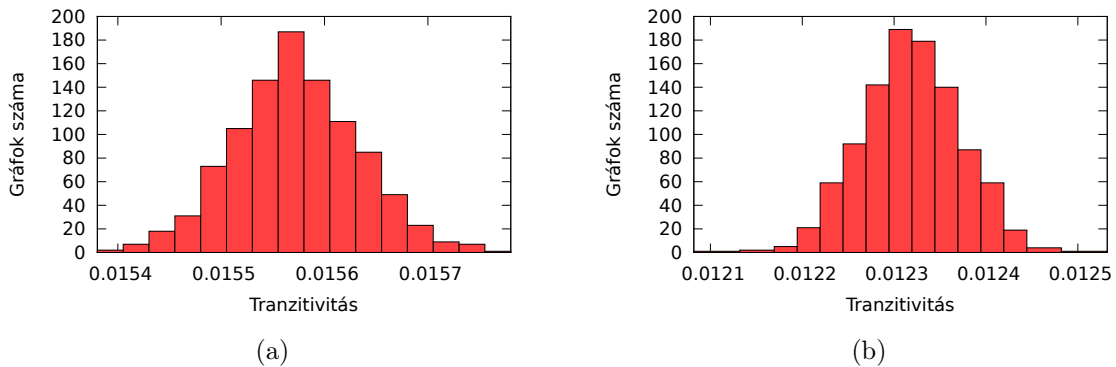
9. ábra. A véletlen hálózatokra mért értékekből alkotott hisztogramok, 0.00004-es méretű binekkel. Az (a) ábra a teljes gráfból alkotott, míg a (b) ábra a legnagyobb algráfból létrehozott véletlen hálózatokra mért értékek eloszlása.

A hisztogramok ábrázolása után azt találtuk, hogy a teljes hálózathoz létrehozott véletlen gráfok jó közelítéssel normáeloszlást adnak, de a legnagyobb algráhból létrehozottak nem. A két esetre mért geográfiai együttható értéke is más, a teljes hálózathoz létrehozott gráfokra ~ 0.0124 - 0.0128 , míg az óriáskomponensből alkotott gráfokra ~ 0.0268 - 0.0272 . Amíg az átlagos klaszterezetségi együttható mérésekor a két gráf esetében a véletlen hálózatokra mért értékek közel voltak egymáshoz, itt azt tapasztaljuk, hogy az óriáskomponensből alkotott véletlen hálózatra mért értékek kétszer akkora, mint a teljes hálózathoz létrehozott véletlen gráfokra számolt értékek. Látszik, hogy a földrajzi elhelyezkedés hogyan hathat ki a társadalom összetettségére. Ennél a módszernél ugye figyelembe vettük a felhasználók egymástól vett távolságait úgy, hogy a geográfiai klaszterezetségi együttható számításakor a kis területű háromszögek nagyobb súlyt kapjanak, mint a nagy területűek. A kapott együttható kiszámolásánál a két talán legfontosabb tényező, adott csúcs-, él- és fokszámeloszlás mellett, hogy a gráf milyen földrajzi kiterjedésű, illetve, hogy milyen a felhasználók földrajzi sűrűsége (a felhasználók átlagos távolsága). Ez alapján azt mondhatjuk, hogy két azonos csúcs-, él- és fokszámeloszlással rendelkező hálózat közül, amelyek ugyanolyan népsűrűségű területen helyezkednek el, az a földrajzilag kisebb kiterjedésű, amelyikre nagyobb a geológiai klaszterezetségi együttható értéke. Ugyanilyen

hálózatok közül, amelyek ugyanakkora földrajzi régióban vannak, annak kisebb a felhasználók közötti átlagos távolsága, amelyikre a geológiai klaszterezettségi együttható értéke nagyobb. Több ilyen hálózatot akár sorba is tudunk rendezni a kiterjedése, vagy a felhasználók közötti átlagos távolság alapján.

Az adatok alapján feltételezhetjük, hogy a legnagyobb algráf kisebb földrajzi kiterjedésű, illetve kisebb átlagos felhasználók közötti távolsággal rendelkezik, mint a teljes hálózat. Ezt nem mondhatjuk biztosra, mert az ábrázolásaink felbontása nem elég nagy, ahhoz, hogy ezt el tudjuk dönteni. Csak feltételezzük, hogy ezért ekkora a differenciál a véletlen hálózatokra mért értékek között. Valós hálózatokra is van egy differenciál, aminek az oka a két hálózat topológiájának a különbségében keresendő. Ezt a különbséget ki lehet nagyítani véletlen hálózatok generálásával, ugyanis itt csak a hálózat topológiája befolyásolja a kapott eredményeket. Ezzel lehet megmagyarázni a véletlen gráfokból alkotott hisztogramok közötti különbséget. Vagyis azért különbözik a két ábra, mert a vizsgált hálózatok topológiája más jellegű.

A harmadik mennyiség, amit véletlen gráfok generálásával vizsgáltunk, az a tranzitivitás. A tranzitivitás azt mondja meg, hogy az összes lehetséges háromszög közül mennyi van a gráfban. Ez a teljes hálózatra: 0.09101, míg a legnagyobb algráfra: 0.09037.



10. ábra. A véletlen hálózatokra mért értékekből alkotott hisztogramok, 0.000025-ös méretű binekkel. Az (a) ábra a teljes gráfokból alkotott, míg a (b) ábra a legnagyobb algráfból létrehozott véletlen hálózatokra mért értékek eloszlása.

A gráfokból létrehozott véletlen hálózatok hisztogram ábráiból az látszik, hogy a véletlen hálózatokra mért értékek jó közelítéssel normáeloszlásból valóak. Azt lehet látni, hogy a valós hálózatokra mért értékek 6-7-szer magasabbak, mint a véletlen hálózatokra számolt értékek. Ugyanilyen eltérést találtunk az adatok között, amikor az átlagos klaszterezettségi együtthatót vizsgáltuk.

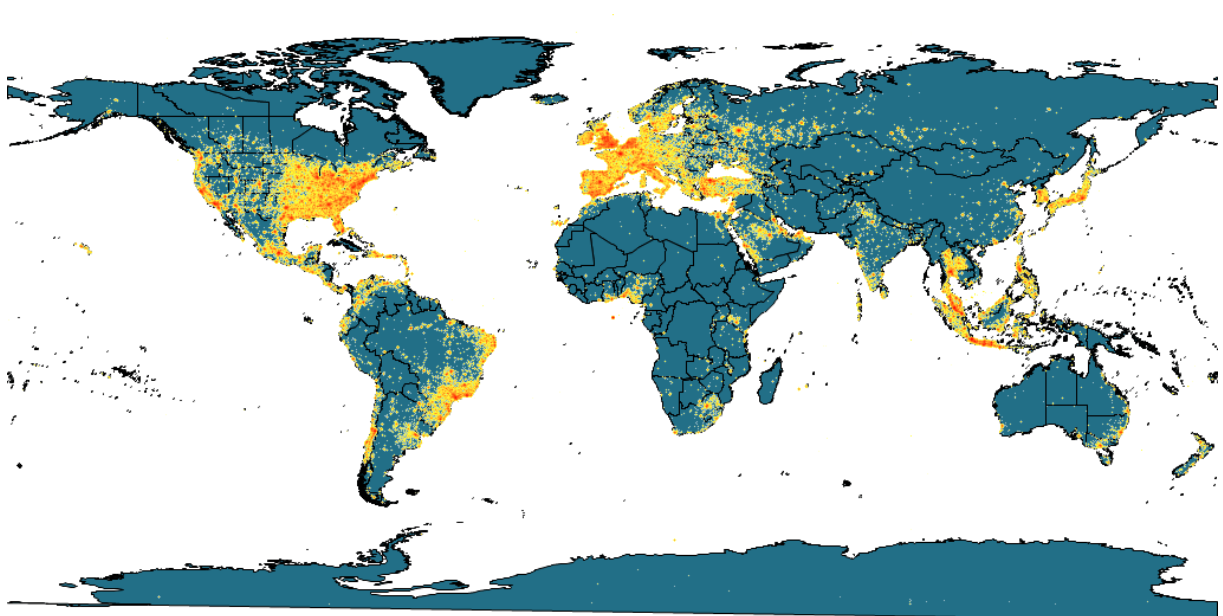
Összességében elmondhatjuk, hogy a valós adatszettünkre mért értékek sokkal nagyobb-

bak, mint a belőlük alkotott véletlen hálózatra számoltak. A felállított statisztikákban a valós adatok outlier-ként viselkedtek. Ez a társadalom önszerveződésével magyarázható. Külön érdekes eredményt adott a geográfiai klaszterezettségi együttható vizsgálata. Ekkor ugyanis megfigyelhettük, hogy a klaszterezettség erősen függ a geoszociális hálózatok topológiájától. Illetve ennek az együtthatónak a segítségével egy lehetséges módszert adtunk arra, hogy ugyanolyan "nagy" geoszociális hálózatok közül kiválasszuk a földrajzilag kisebb területen elhelyezkedőt.

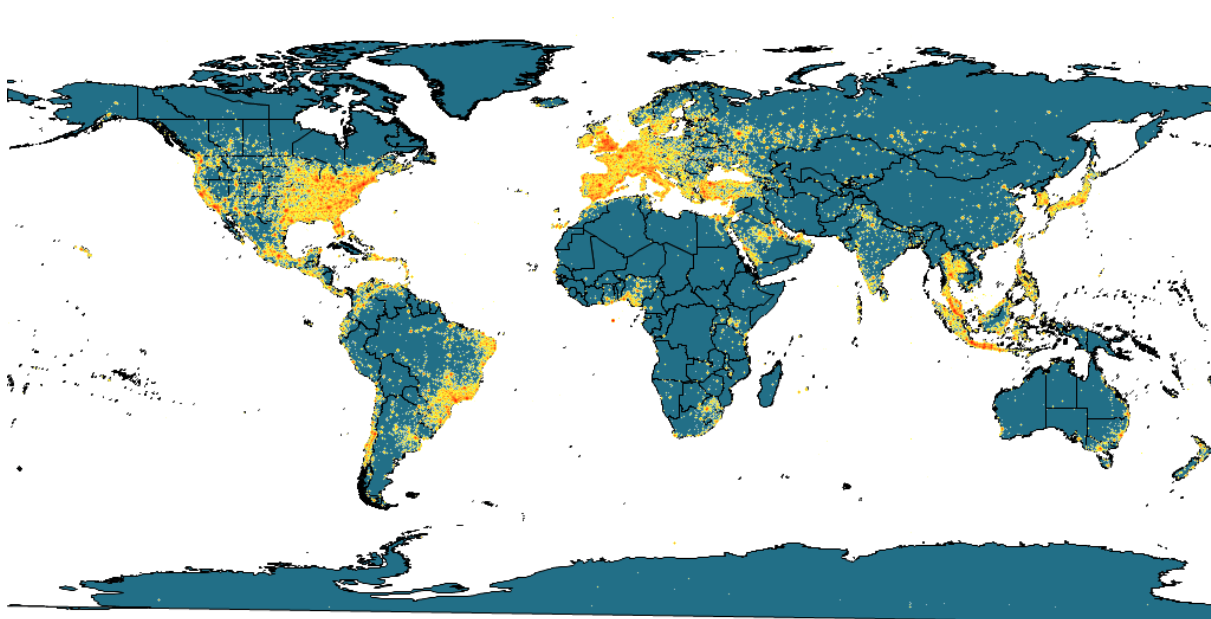
4.2. Élek elhagyása

Ennél a módszernél a vizsgált gráfokból új gráfokat generáltunk a bizonyos küszöbindexnél nagyobb, illetve kisebb súlyú élek kiszűrésével. A gráfjainkat kétféleképpen is súlyoztuk. Tekintsük ezek közül először azt az esetet, amikor a felhasználók közötti távolságok a súlyok!

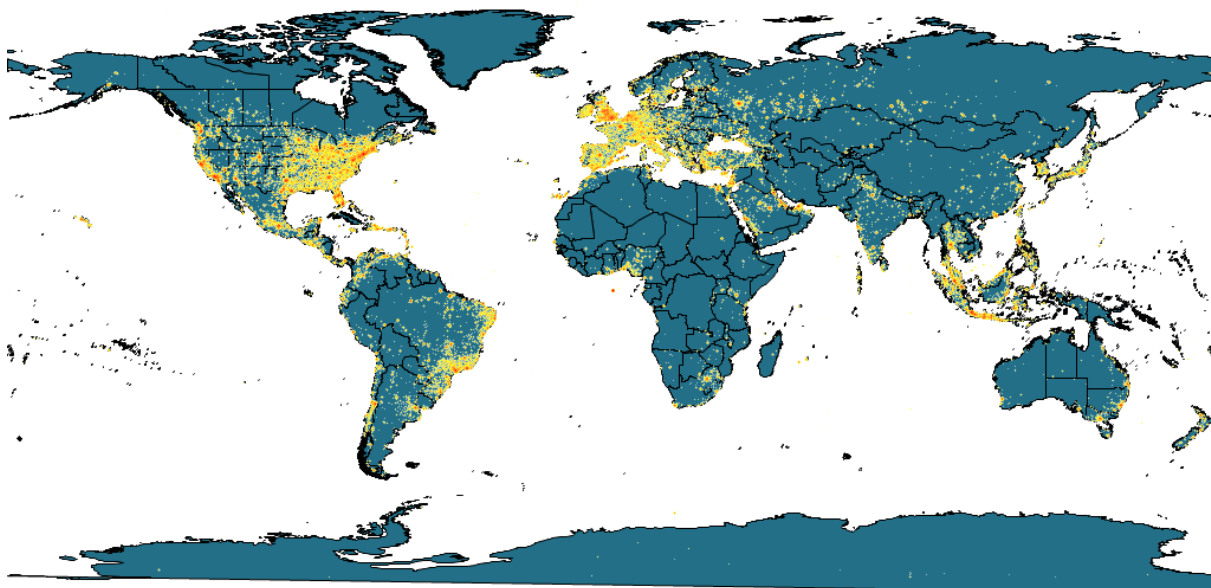
Érdekes kérdés ilyenkor, hogy a felhasználók sűrűségének az eloszlása hogyan változik a küszöbindexek függvényében. Az alábbi ábrákon olyan gráfok felhasználóinak a sűrűségeloszlását ábrázoljuk, ahol adott küszöbindex mellett a küszöbindexnél kisebb éleket, és az így keletkezett izolált pontokat kivettük a teljes gráfból.



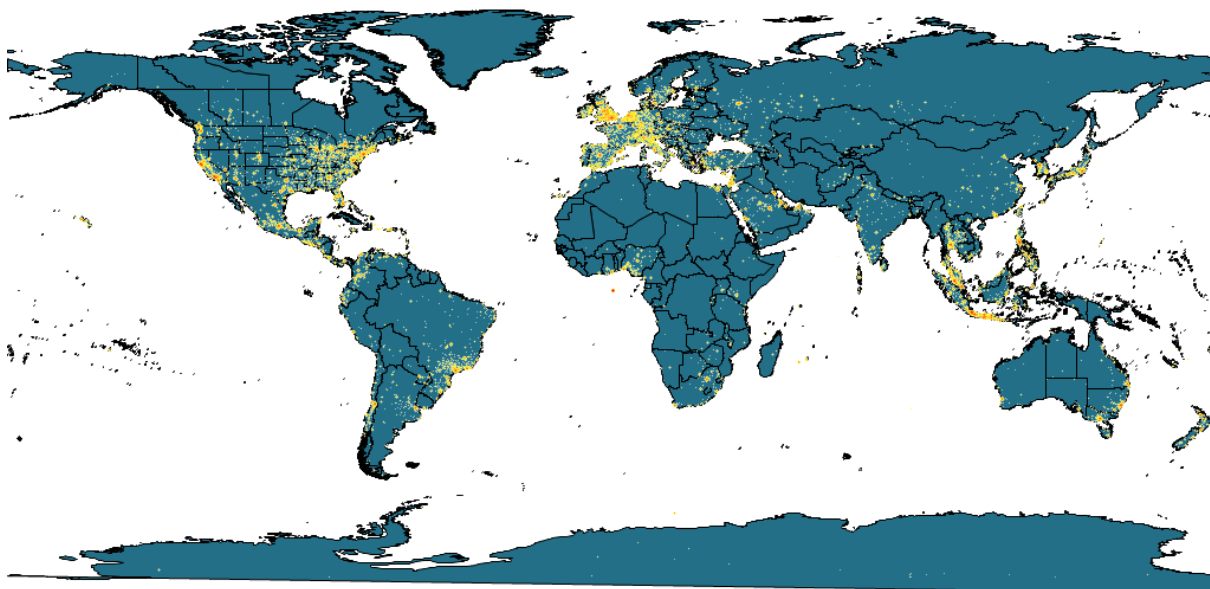
11. ábra. Felhasználók sűrűségeloszlása, amikor a teljes gráfból küszöbindexnél kisebb éleket elhagytuk. A küszöbindex: 11 km



12. ábra. Felhasználók sűrűségeloszlása, amikor a teljes gráfból küszöbindexnél kisebb éleket elhagytuk. A küszöbindex: 158 km



13. ábra. Felhasználók sűrűségeloszlása, amikor a teljes gráfból küszöbindexnél kisebb éleket elhagytuk. A küszöbindex: 1111 km



14. ábra. Felhasználók sűrűségeloszlása, amikor a teljes gráfból küszöbindexnél kisebb éleket elhagytuk. A küszöbindex: 8121 km

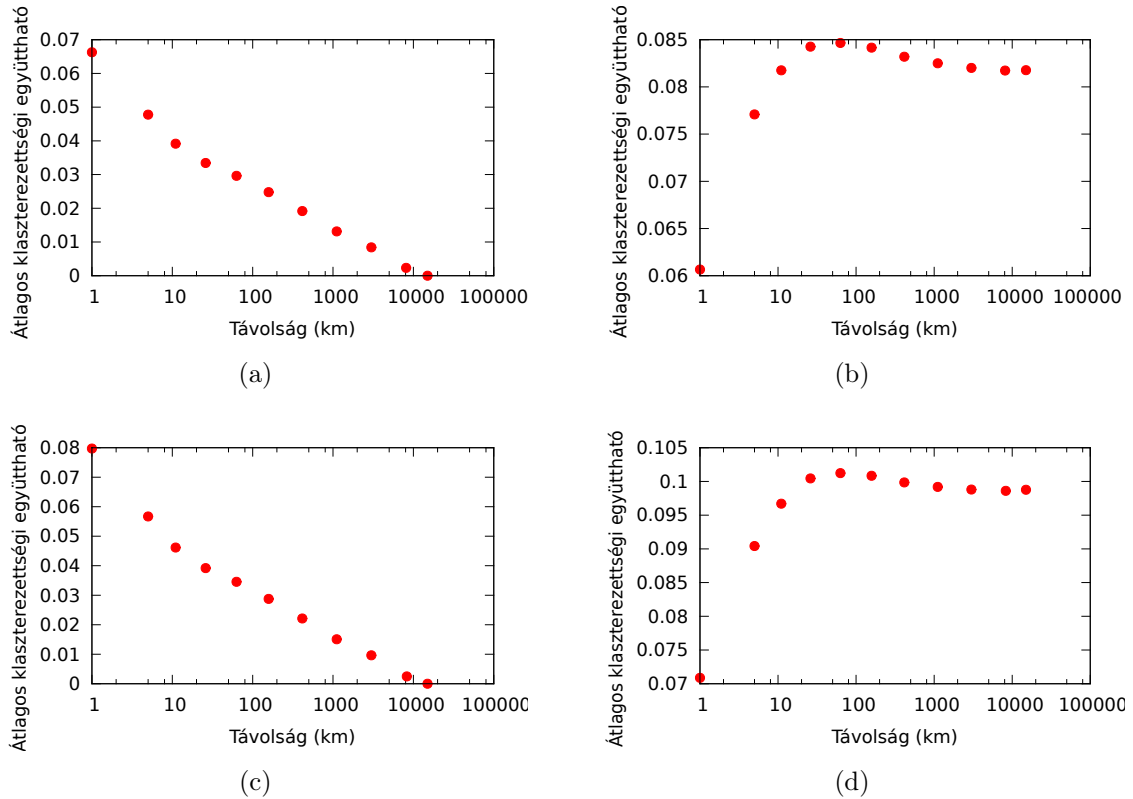
A felhasználók sűrűségét egy olyan programmal ábrázoltuk, ami minden ábrázolásakor a legsűrűbb tatományt jeleníti meg pirossal. Vagyis ha két olyan gráfot ábrázolunk vele ahol a felhasználók sűrűségeloszlásának a jellege ugyanolyan, de az egyik gráf jelentősen több felhasználót tartalmaz mint a másik, akkor -ha mindkét gráf elég nagy- ugyanolyan ábrákat fogunk kapni. Ez tehát nem a legjobb módszer egy ilyen fajta vizsgálatnál.

Az ábrákon jól látszik a program fent említett tulajdonsága, vagyis hogy minden ábrához külön kalibrálja az eloszlás skáláját. 11-es küszöbindex mellett a sűrűségeloszlás szinte ugyanolyan, mint a teljes gráf esetén. A küszöbindexek növelésével azért egy idő után ki tudjuk venni, hogy a Föld mely területén helyezkednek el azok a felhasználók, akik távoli ismerősökkel rendelkeznek. Az adatokra azt lehet mondani, hogy behúzódnak a városokba, a küszöbindexek növelésével. Látszik, hogy pl. Észak-Amerika, vagy Európa felhasználóinak közel homogén eloszlása megszűnik. A klaszterezettség mélyebb megértéséhez azonban szükséges a szociális háló kvantitatív vizsgálata is.

Ennél a vizsgálati módszernél mindhárom gráftulajdonság változását vizsgáltuk a küszöbindexek függvényében. Az ábrázolásokból kihagytuk a nulla küszöbindexhez tartozó értékeket.

Az átlagos klaszterezettségi együtttható vizsgálatánál elvileg ahogy vesszük ki az éleket, ez az érték csökken, viszont ezáltal a csúcsok száma is, hiszen az izolált pontokat nem hagytuk benne a hálózatba. Abban az esetben, ha a csúcsok száma a küszöbindexek

függvényében gyorsan változik, akkor az átlagos klaszterezettségi együttható nőhet is, ha az eredménygráfban az élek relatív száma nagyobb mint a kiindulógráfban.



15. ábra. Az átlagos klaszterezettségi együtthatók értéke a küszöbindexek függvényében: az (a) ábrán az eredeti gráfból a megadott küszöbindexnél kisebb, a (b) esetben a nagyobb éleket vettük ki. A (c) és a (d) ábrákon ugyanezt ábrázoltuk, csak a kiindulógráf a legnagyobb algráf volt.

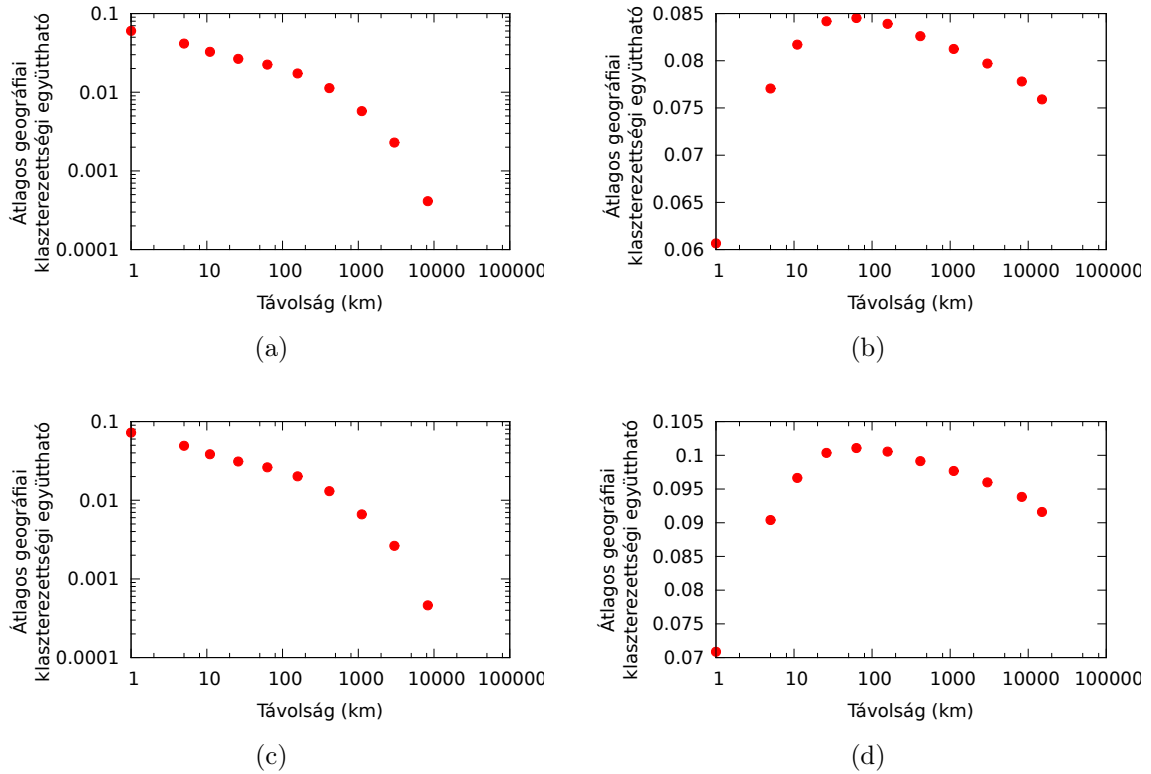
Az ábrákból látszik, hogy ha a küszöbindexnél kisebb súllyal rendelkező éleket szedjük ki, akkor a gráf átlagos klaszterezettségi együtthatója logaritmikus jelleggel csökken. Ez érthető, hiszen az emberek a földrajzilag hozzájuk közel eső emberekkel tartják a kapcsolatot, és a földrajzilag közeli barátaikat tudják bemutatni egymásnak. Azt találtuk, hogy interkontinentális méretben már elhanyagolhatóan kicsi az átlagos klaszterezettségi együttható értéke, vagyis nagyon kevés az esély arra, hogy egy ember két ismerőse, akik két másik kontinensen laknak ismerjék egymást.

Amikor a küszöbindexnél nagyobb súllyal rendelkező éleket vesszük ki, akkor a pontoknak egészen más jellege lesz. Itt nyilván, ha kicsinek választjuk a küszöbindexet (pár km) akkor a közepes hosszúságú élek sem lesznek benne a hálózatban, amik az ábrák szerint fontos szerepet játszanak a hálózat összetettségében. A görbének van egy maximuma, ez a teljes gráf esetében ~ 0.084 - 0.085 közé esik, míg a legnagyobb algráfnál ez az érték

magasabb ~ 0.1 - 0.12 közötti.

A pontsorozatokon megfigyelhetünk egy maximumhelyet, amire mindkét esetben ~ 65 km-es értéket határoztuk meg. Ez egy nagyváros vagy egy régió jellemző kiterjedése. Az adatainkból tehát azt kaptuk, hogy egy felhasználó két ismerőse akkor ismeri legnagyobb valószínűséggel egymást, ha a távolságuk a közös ismerőstől nem nagyobb mint 65 km. Vagyis egy ilyen maximális élhosszabból álló hálózat a legösszetettebb.

Ezt a mérőssorozatot elvégeztük úgy is, hogy az átlagos geográfiai együttthatót mértük a gráfjainkra.

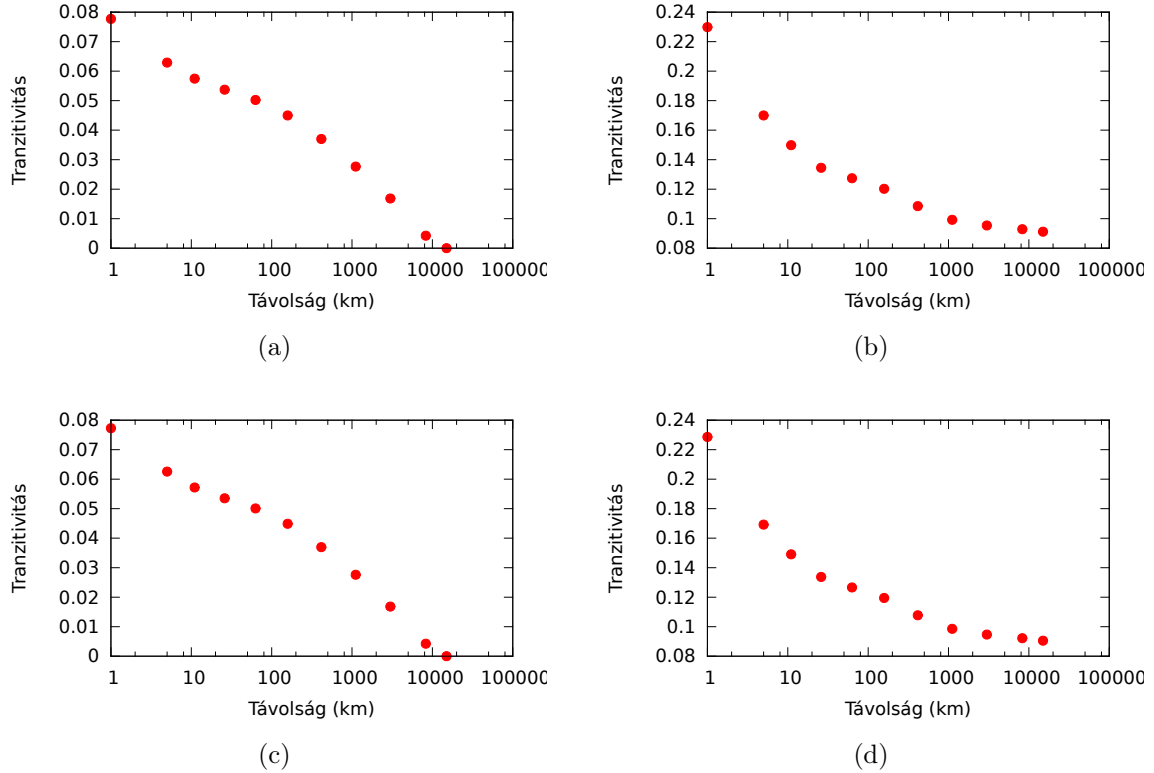


16. ábra. A különböző küszöbindexekre mért értékek az (a) esetben a kisebb éleket vesszük ki, a (b) eset amikor a küszöbindexnél nagyobb éleket vesszük ki. Ezek az ábrák a teljes gráfra vonatkoznak, míg a (c) és a (d) ugyanezek, csak a legnagyobb részgráfra.

Az ábrák hasonló jellegűek mint az átlagos klaszterezettségi együtttható mérések, de itt amikor a küszöbindexnél kisebb éleket vesszük ki, akkor nagyon gyorsan kezd esni a geográfiai klaszterezettségi együtttható értéke. Ez az átlagos geográfiai klaszterezettségi együtttható definíciójával magyarázható, ez ugyanis kisebb súlyt rendel a hosszú élt tartalmazó háromszögekhez, és egy idő után már csak ilyenek lesznek a vizsgált gráfokban. A másik esetben, amikor a küszöbindexnél nagyobb éleket vesszük ki, akkor hasonló görbét kaptunk, mint az átlagos klaszterezettségi együtttható mérések. A maximumhelyet itt

is ~ 65 km-nél lehet megállapítani. Úgy tűnik, hogy ezt a határt a legjobb klaszterezettségre ennek az együttthatónak a mérésével is meg lehet kapni, vagyis nem függ a szociális hálózat topológiájától.

A tranzitivitás küszöbindexektől való függését is megvizsgáltuk.



17. ábra. A különböző küszöbindexekre mért értékek az (a) esetben a kisebb éleket vesszük ki, a (b) eset amikor a küszöbindexnél nagyobb éleket vesszük ki a gráfból. Ezek az ábrák a teljes gráfra vonatkoznak, míg a (c) és a (d) ugyanezek, csak az óriáskomponensre.

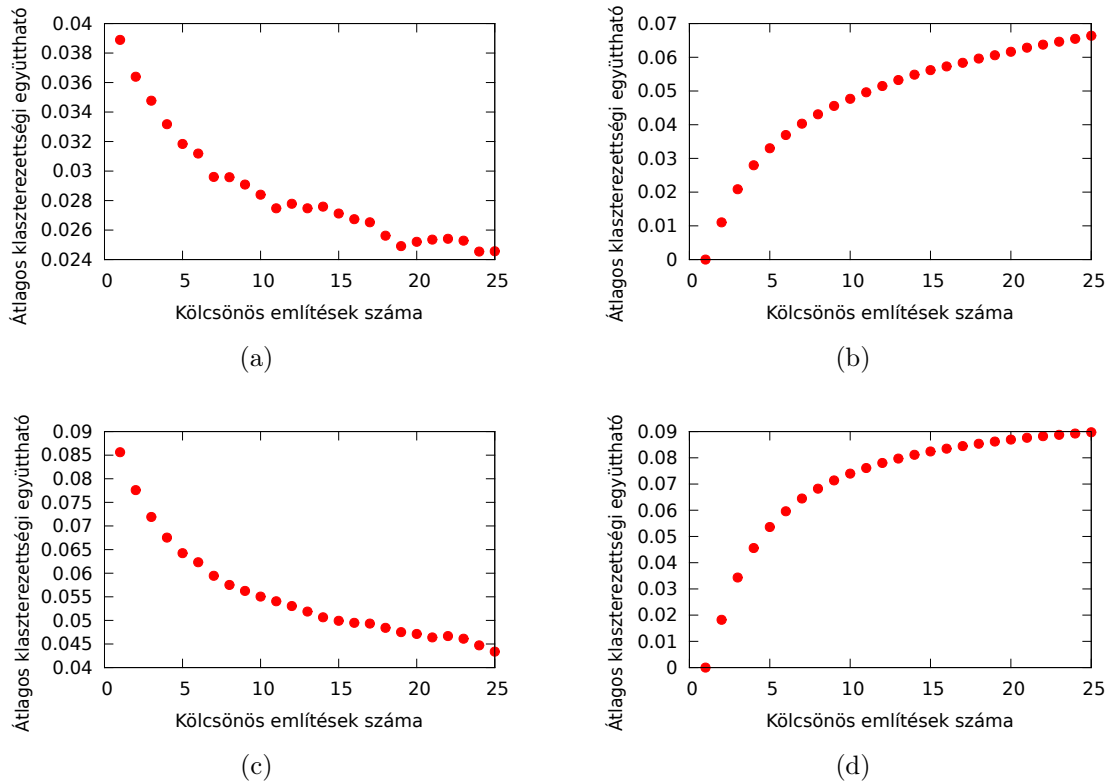
Itt az (a) és (c) ábrákon az előzőekhez hasonló nagyjából logaritmikus csökkenést tapasztalunk, akkor amikor a küszöbindexnél kisebb éleket hagyunk el. Ez annyira nem meglepő, hiszen például az átlagos klaszterezettségi együtttható vizsgálatok az eredményt kaptuk, hogy egy felhasználó két barátja közötti kapcsolat valószínűsége a távolsággal logaritmikusan csökken. A háromszögek relatív számának tehát ezt a tendenciát kellene követnie, és mi pont ezt tapasztaltuk itt is. A görbe jelleg tehát ugyanaz a két esetben. Ezek az ábrák ugyanazt az eredményt adják, mint amit az átlagos klaszterezettségi együttthatóra kaptunk. A (b) és (d) ábrák ellenben különböznek azokról az ábráktól, amiket az átlagos klaszterezettségi együtttható vagy az átlagos geográfiai együtttható vizsgálatok mutattunk be. Azok a görbék növekedtek, majd elértek egy maximum értéket, ami után csökkenő tendenciával megkaptuk a teljes gráfra mért értéket. A tranzitivi-

tás mérésekor azonban a mért értékek egy csökkenő sort adnak ki. Itt azt tapasztaltuk, hogy ha elhanyagolunk éleket a tranzitivitás nő, még akkor is, ha csak a kis élhosszúságú kapcsolatokat hagyjuk benne a hálózatban. Ekkor a tranzitivitás értéke ~ 2.5 -szöröse az eredeti értéknek. Ahogy növeljük a küszöbindexet úgy csökken a mért érték, vagyis a közeli felhasználók fontosabbak a hálózat összetettségében.

A dolgozatban már említettük, hogy az átlagos klaszterezettségi együttható és a tranzitivitás nem független egymástól. A két mennyiség mérésekor azonban eltérő eredményeket tapasztaltunk. A mérések közötti eltérések a két mennyiség közötti különbséggel magyarázhatóak. Amíg az átlagos klaszterezettségi együttható minden ponthoz rendel egy értéket és azokat átlagolja ki, addig a tranzitivitás számolásakor két, a hálózatra jellemző globális tulajdonságot osztunk egymással. Egy pont vagy részt vesz háromszögekben, vagy nem. Ha nem vesz részt, és kiszedjük a hálózatból, akkor a tranzitivitás számolásánál csak a nevező értékét változtatja meg, vagyis ha sok ilyen felhasználót veszünk ki, akkor a tranzitivitás nő. Mi azt találtuk, hogy ha csak a kis élsúlyú felhasználókat tartom a hálózatban akkor is nő ez az érték, akkor viszont nem, ha csak ezeket hagyom ki. A mért érték mindig akkor csökkent, amikor a nagy élsúlyú kapcsolatokat tartottuk meg, vagy csak ilyenekkel egészítettük ki a gráfokat. Vagyis az egymástól távoli felhasználók rontják a hálózat összetettségét. Az egymáshoz közeli felhasználók klaszterezettségben játszott szerepe fontosabb. Ezt a jelenséget a tranzitivitás vizsgálatával sokkal jobban ki lehet mutatni, mint az átlagos klaszterezettségi együttható mérésével.

A gráfjaink másik súlyozása a felhasználók közötti kölcsönös említések száma volt. Az élek elhagyásának a módszerével az így súlyozott gráfokra is elvégeztük a mérést. Ebben az esetben csak az átlagos klaszterezettségi együtthatót vizsgáltuk.

Itt azt lehet megfigyelni az ábrákon, hogy ha a küszöbindexnél kisebb éleket vesszük ki, akkor ismét csökkenést tapasztalunk az átlagos klaszterezettségi együtthatóban. A másik esetekben pedig ha a küszöbindexnél nagyobb éleket vesszük ki, akkor a görbe növekedni fog. Az ábrákból jól látszik a fontossága a gyenge és a közepesen erős kapcsolatoknak. 7-es kölcsönös említéseknel küszöböt lehet meghatározni, ugyanis ez után a küszöb után már nem változik az átlagos klaszterezettségi együttható értéke gyorsan. Tekintsük innentől a erős kapcsolatoknak azokat az éleket, amelyeknek a súlya nagyobb mint hét. A másik szignifikáns eltérés rögtön a görbe elején van, míg az óriáskomponensre ~ 0.085 az átlagos klaszterezettségi együttható értéke, addig a teljes gráfra ez az eredeti érték fele. Ezzel megkaptunk még egy fontos küszöböt, ami után a görbe lassabban változik. Legyenek azok a gyenge kapcsolatok, ahol 1 a kölcsönös említések száma.



18. ábra. Az átlagos klaszterezettség együtthatók értéke a küszöbindexek függvényében. Az (a) ábrán az eredeti gráfból a megadott küszöbindexnél nagyobb éleket, míg a (b) ábrán a küszöbindexnél kisebb éleket tartottuk meg. A (c) és a(d) ábra is ezt mutatja, csak a legnagyobb összetett komponensből alkottunk itt kisebb hálózatokat.

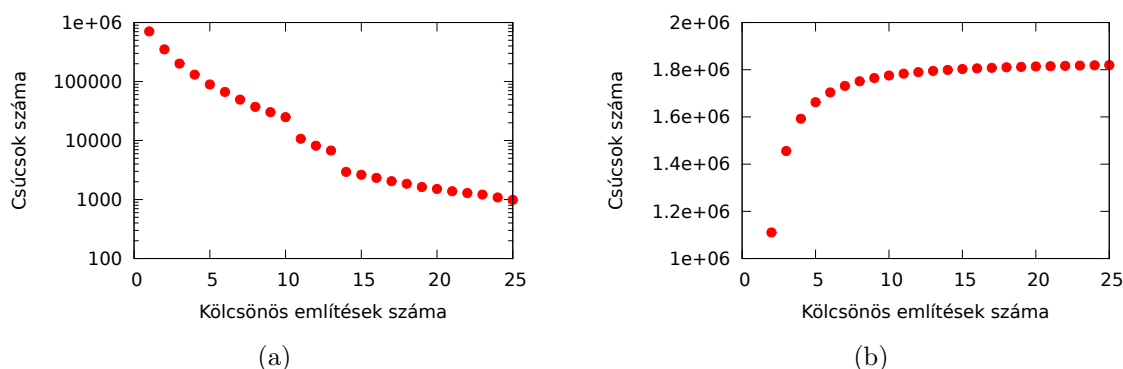
Ezzel a módszerrel definiálhattuk a hálózathoz tartozó kapcsolati erősségeket:

- Gyenge kapcsolat van két felhasználó között, ha pontosan egyszer említették meg egymást a tweetjeikben.
- Közepesen erős kapcsolat van közöttük, ha a kölcsönös említések száma 2-7 között van.
- Erős kapcsolat pedig az, ha két felhasználó 7-nél többször említette meg egymást.

Célunk volt megvizsgálni, hogy a gyenge kapcsolatok fontos szerepet játszanak-e a társadalmunk összetettségében. A fenti ábrák alapján ezt még nem mondhatjuk ki. Egy elfogadott elmélet szerint [15] a gyenge kapcsolatok tartják össze a szociális hálózatot. Ez úgy történhet, hogy az egymás felé erős kapcsolatokkal kötődő egyének által alkotott klaszterek között, a klaszterek között jelen levő gyenge kapcsolatok teremtenek kapcsolatot. Vagyis feltételezhető, hogy a gyenge kapcsolatok többsége nem vesz részt háromszögekben, de ettől még fontos szerepük van a hálózat összetartásában.

4.3. A hálózat gyenge kapcsolatokról való függése

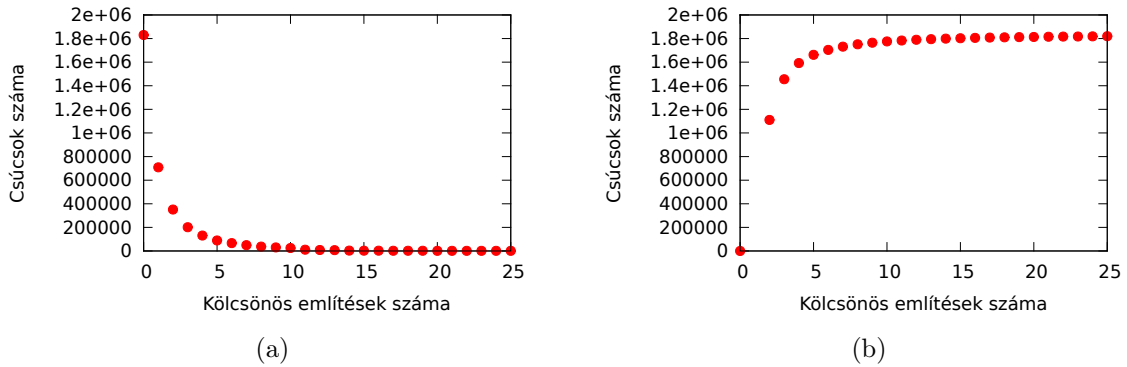
Gyenge kapcsolatnak definiáltuk azokat az éleket, amiknek a kölcsönös említések szerinti súlyozása 1. A dolgozatban korábban említettük, hogy a gyenge kapcsolatok fontos szerepet játszanak a társadalmi hálózatok összetartásában.[15] Ez azt jelenti, hogy ha a módszereinkkel ki tudjuk mutatni ezt a jelenséget, akkor azok megbízhatóak, illetve az adatsorunk is jól reprezentálja a tényleges szociális hálót. A hálózat gyenge kapcsolatokról való függését az élek elhagyásának a módszerével vizsgáltuk. Azt néztük meg, hogy hogyan esik szét a hálózat, ha a bizonyos kölcsönös említésszámnál nagyobb, illetve kisebb éleket elhanyagoljuk. Ennél a módszernél a gráfból kiesett izolált pontokat is beleszámoltuk a kapott algráfok számába. Ezenkívül azt vizsgáltuk még, hogy hogyan változik a megmaradt gráfok legnagyobb algráfjainak a csúcsainak a száma a küszöbindexek változtatásának a függvényében. Ezeket a vizsgálatokat csak az óriáskomponensre végeztük el. Vizsgáljuk először az adott küszöbindexhez tartozó legnagyobb algráf csúcsainak a változását!



19. ábra. Az x tengelyen a küszöbindexek, míg az y tengelyen az szerepel, hogy adott küszöbindexnél a legnagyobb megmaradt algráfnak mennyi csúcsa van. Az (a) ábrán az eredeti gráfból a megadott küszöbindexnél nagyobb éleket, míg a (b) ábrán a küszöbindexnél kisebb éleket tartottuk meg.

Ennél a módszernél ahogyan vesszük ki a hálózatból az egyre erősebb kapcsolatokat, úgy csökken a legnagyobb algráf csúcsainak a száma. Ha a gyenge kapcsolatokon kívül minden más súlyú kapcsolatot a hálózatban hagyunk, akkor a legnagyobb algráf ~ 900000 felhasználóból áll. Ezzel szemben, ha a küszöbindexnél nagyobb éleket távolítjuk el, akkor azt tapasztaljuk, hogy ha csak a gyenge kapcsolatokat hagyjuk a gráfban, már akkor ~ 1100000 felhasználóból áll a legnagyobb algráf, vagyis a gyenge kapcsolatokról álló legnagyobb összekötött komponens máris nagyobb, mint az ennél erősebb kapcsolatokról alkotott struktúra. Ezeken az ábrákon kihagytuk a 0-ás küszöbindexhez tartozó ponto-

kat, valamint az (a) ábrán az egyik tengely logaritmikus. Az adatpontokat másképpen ábrázolva sokkal szemléletesebb ábrákat kapunk.

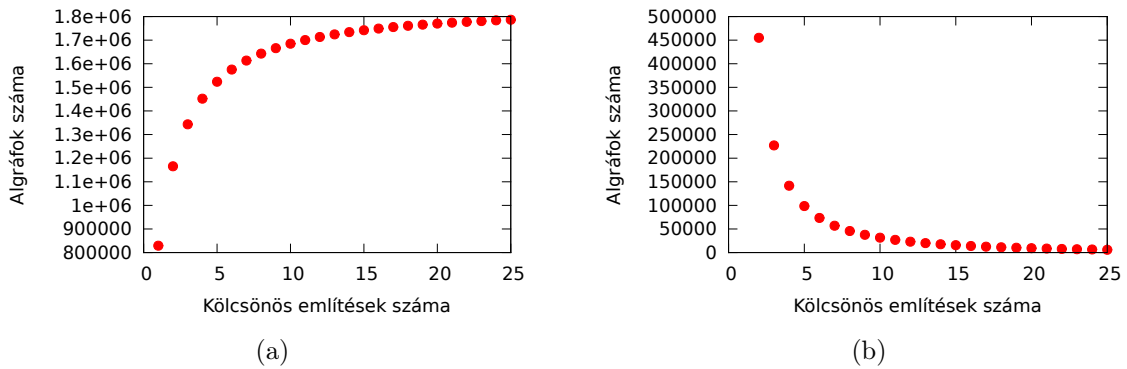


20. ábra. Az x tengelyen a küszöbindexek, míg az y tengelyen az szerepel, hogy adott küszöbindexnél a legnagyobb megmaradt algráfnak mennyi csúcsa van. Az (a) ábrán az eredeti gráfból a megadott küszöbindexnél nagyobb éleket, míg a (b) ábrán a küszöbindexnél kisebb éleket tartottuk meg.

Ezek az ábrák már jól látható, hogy a megmaradt legnagyobb összetett komponensek csúcsszámának a legnagyobb változása az első két pont között történik. Ez pont azt jelenti, hogy a legnagyobb szerepe az 1-es említéssel súlyozott éleknek van, vagyis a gyenge kapcsolatoknak. Még a közepesen erős kapcsolatok tartományában is erősen változik a pontsorozat, de az erős kapcsolatok már alig befolyásolják a legnagyobb összetett komponens csúcsszámát.

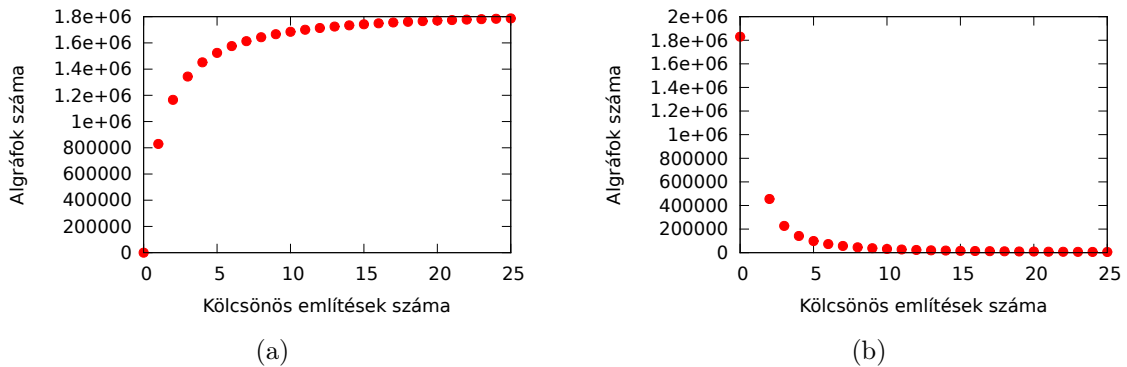
A másik vizsgálati módszerünkön azt néztük meg, hogy hogyan esik szét a hálózat, ha a bizonyos kölcsönös említésszámnál nagyobb, illetve kisebb éleket elhanyagoljuk. Itt nem úgy végeztük el a csonka gráfok generálását, hogy az élek kiszedése után ne maradjanak izolált pontok, mint a többi esetben, hanem úgy, hogy a gráfban benne hagytuk az izolált pontokat is. Így mindkét eset felső határa a legnagyobb algráf csúcsainak a száma.

Látszik, hogy ha csak a gyenge kapcsolatokat hanyagoljuk el a hálózatból, akkor az ~ 800000 db algráfra esik szét. Ez után a küszöbindexek növelésével hiába vesszük ki a többi kapcsolatot is, annyira nem változik az algráfok száma. Amikor csak a gyenge kapcsolatokat tartjuk meg, akkor csak ~ 450000 db algráf jelenik meg, ami majdnem feleakkora, mint az előbbi esetben kapott részhálózatok száma. A küszöbindex növelésével, ahogy adódnak hozzá az egyre erősebb kapcsolatok a hálózathoz, az algráfok száma csökken.



21. ábra. Az x tengelyen a küszöbindexek, míg az y tengelyen az szerepel, hogy adott küszöbindexnél mennyi algráfra és izolált pontra esik szét a hálózat. Az (a) ábrán az eredeti gráfból a megadott küszöbindexnél nagyobb éleket, míg a (b) ábrán a küszöbindexnél kisebb éleket tartottuk meg.

Az ábrázolásoknál itt is kihagytuk a 0-ás küszöbindexhez tartozó értékeket, ha ezeket is ábrázoljuk, akkor ismét szemléletesebb ábrákat kapunk.



22. ábra. Az x tengelyen a küszöbindexek, míg az y tengelyen az szerepel, hogy adott küszöbindexnél mennyi algráfra és izolált pontra esik szét a hálózat. Az (a) ábrán az eredeti gráfból a megadott küszöbindexnél nagyobb éleket, míg a (b) ábrán a küszöbindexnél kisebb éleket tartottuk meg.

Itt is látszik, hogy legjobban az elején, de a közepesen erős kapcsolatok tartományában is jelentősen változik a kapott pontsorozat. Az erős kapcsolatok tartományában létrejövő változás azonban elhanyagolható.

Mind a két esetben azt láthattuk, hogy ha manipuláljuk a gyenge kapcsolatokat, akkor a gráf összetettségében drasztikus változás jön létre. A gyenge kapcsolatok nem csak önmagukban alkotnak nagy kiterjedésű gráfokat, hanem összetartják a különböző erősebb kapcsolatokból alkotott algráfokat. Amikor csak a gyenge kapcsolatokat vettük ki, akkor nagyjából kétszer annyi algráf keletkezett, mint amikor csak a gyenge kapcsolatokat

tartottuk a gráfban. Ezeken a méréseken látszik, hogy tulajdonképpen mennyire játszanak fontos szerepet a gyenge kapcsolatok a szociális hálózatok klaszterezettségében. Ezen kívül beláttuk, hogy a vizsgálati módszerünk alapján levont konklúziók alátámasztják a szakirodalomban elfogadott elméletet. Vagyis jónak tekinthetjük mind az adatszettünket, mind a vizsgálati módszerünket.

5. Összegzés

A dolgozatunkban tehát egy geoszociális hálózatból alkotott gráfot, és annak a legnagyobb összefüggő komponensét vizsgáltuk. A vizsgálataink célja az volt, hogy behatóbb ismereteket szerezzünk a hálózatok klaszterezettségéről a felhasználók földrajzi elhelyezkedésének és kapcsolati erősségüknek a függvényében. Több módszert is bemutattunk, amivel a geoszociális hálózatok klaszterezettségét lehet vizsgálni. Ezek közül a geoszociális klaszterezettségi együttható megállapítására egy python alapú programot is írtunk, amivel a jövőben esetleg kiegészíthetjük a networkx gráfkezelő csomagot³. A munkánk során főként azt vizsgáltuk, hogy hogyan változik az átlagos klaszterezettségi-, geográfiai klaszterezettségi együttható illetve tranzitivitás értéke, ha a gráfokat különböző féleképpen manipuláljuk.

Az egyik módszerünk a véletlen hálózatok létrehozása volt. Ekkor azt találtuk, hogy a véletlen hálózatokra mért értékek a legtöbb esetben megközelítőleg normáeloszlást adnak, kivéve a legnagyobb algráfokból generált hálózatok, amikor geográfiai klaszterezettségi együtthatót mértünk rajtuk. Ezt az effektust nem tudtuk megmagyarázni teljesen, a mélyebb megértéséhez más vizsgálatokat is kellene végeznünk. Azt a feltételezésünket viszont alátámasztottuk minden esetben, hogy a valódi hálózat értékei magasabbak, mint a véletlen hálózatoké. Számszerűen $\sim 6-7$ -szer magasabbak a valódi hálózatra mért értékek. Ezek az eredmények a társadalomban jelenlevő önszerveződéssel magyarázhatóak, vagyis a felhasználók maguktól szerveződnek csoportokba.

Egy másik módszerünk az, hogy a hálózatból bizonyos küszöbindexnél kisebb vagy nagyobb súllyal rendelkező élek elhagyásával hoztunk létre új gráfokat. A súlyok a felhasználók földrajzi távolságai vagy a közöttük levő kölcsönös említések száma volt. Amikor a távolságokkal súlyozott gráfokat vizsgáltuk, azt találtuk, hogy a hálózatok klaszterezettsége függ a felhasználók közötti távolságoktól. Az egymástól távoliak rontják azt, míg egy bizonyos értéknél közelebbiek növelik. Mi ezt az értéket ~ 65 km-nek állapítottuk meg. Ekkor a legnagyobb a szociális hálózatok klaszterezettsége. Itt azonban nem a konkrét

³A networkx úgynevezett open-source programcsomag, vagyis a felhasználók is fejleszthetik

érték az igazán fontos (lehetne 30 vagy 100 km is, a görbéknek ezen a tartományon van maximuma, nem egy jól meghatározott értéknél), hanem az, hogy milyen nagyságrendbe esik. Ilyen nagyságrendben vannak a nagyvárosok, vagy a régiók átlagos kitejedése.

A kölcsönös említésekkel súlyozott gráfokból pedig a kapcsolatok erősségének a hatásait vizsgáltuk a klaszterezettségre. Ekkor három különböző erősségű kapcsolatot definiáltunk, amik közül is legfontosabbnak az úgynevezett gyenge kapcsolatokat ítéltük, ahol a felhasználók közötti kölcsönös említések száma 1. Az ilyen kapcsolatok kritikus fontosságúak a hálózat összetettségének a szempontjából, ezeknek a kapcsolatoknak a jelenlétére a hálózat nagyon érzékeny. Vizsgálatainkkal tehát alátámasztottuk azt az elméletet, hogy a gyenge kapcsolatoknak fontos szerepe van a hálózat összetartásában.

Hivatkozások

- [1] A.Szalay and J. Gray. "Science in an exponential world". *Nature*, 440, pp. 413-414, 2006. DOI:10.1038/440413a
- [2] <http://home.web.cern.ch/about/computing>
- [3] <http://cas.sdss.org/dr7/en/>
- [4] L. Dobos, J. Szüle, T. T. Bodnár, T. Hanyecz, T. Sebők, D. Kondor, Zs. Kallus, J. Stéger, I. Csabai, G. Vattay (2013). A multi-terabyte relational database for geo-tagged social network data. arXiv:1311.0841
- [5] Cho, E., Myers, S., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. KDD '11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (Retrieved from <http://dl.acm.org/citation.cfm?id=2020579>)
- [6] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? Proceedings of the 19th International World Wide Web (WWW) Conference, April 26-30, 2010, Raleigh NC (USA).
- [7] D. Kondor, I. Csabai, L. Dobos, J. Szüle, N. Barankai, T. Hanyecz, T. Sebők, Zs. Kallus, G. Vattay (2013). Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages. arXiv:1311.1169

- [8] Zs. Kallus, N. Barankai, D. Kondor, L. Dobos, T. Hanyecz, J. Szule, J. Steger, T. Sebok, G. Vattay, I. Csabai (2013). Regional properties of global communication as reflected in aggregated Twitter data. arXiv:1311.1484
- [9] J. P. Huchra and M. J. Geller. Groups of galaxies. i - nearby groups. *Astrophysical Journal*, 257:423–437, June 1982.
- [10] http://en.wikipedia.org/wiki/Great-circle_distance
- [11] <http://networkx.github.io>
- [12] Scellato, S., & Mascolo, C. (2010). Distance matters: Geo-social metrics for online social networks. WOSN'10 Proceedings of the 3rd conference on Online social networks. (Retrieved from <http://dl.acm.org/citation.cfm?id=1863198>)
- [13] Erdős, Paul; A. Rényi (1960). "On the evolution of random graphs". *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 17–61.
- [14] R. Albert; A.-L. Barabási (2002). "Statistical mechanics of complex networks". *Reviews of Modern Physics* 74: 47–97. arXiv:cond-mat/0106096.
- [15] Granovetter, M. (1973). The strength of weak ties. *American journal of sociology*. Retrieved from <http://www.jstor.org/stable/10.2307/2776392>

1. Függelék

Tudományos Diákköri dolgozatomban témámat a szociális hálózatok összetettsége, és ennek a földrajzi távolságoktól való függése adja. Kutatásomhoz bekapcsolódtam az egyetemen Twitter adatokat elemző kutatócsoport munkájába. A csoport feladata egy Twitter adatbázis felállítása és az adatok elemzése. Témavezetőmmel ezen adatok alapján terveztük a társadalom klaszterezettségét a felhasználók egymástól való távolságától és az egymás közötti kapcsolati erősségektől való függését vizsgálni. A munka során az alábbi önállóan megoldandó részfeladatokat kaptam:

1. A vizsgált gráfot tartalmazó tábla felállítása az adatbázis többi táblája alapján
2. A gráf vizsgálatához használt programok megírása
3. A kapott eredmények kiértékelése

A gráfot tartalmazó tábla felállításához szükséges számolások már a rendelkezésemre álltak. Ezek a felhasználók földrajzi helyzeteinek a meghatározása, valamint a közöttük való távolság kiszámítására használt függvény. Ezeket ugyanis más kutatásokhoz is használták. Ennél a részfeladatnál, tehát az adatbányászat volt a dolgom.

A programok megírását önállóan végeztem, témavezetőm csak ellenőrizte a kapott eredményeket. Ide tartozott a véletlen hálózatok generáló program megírása, illetve az általános geográfiai klaszterezettségi együttható mérésének a leprogramozása. Ezt úgy tettem meg, hogy a networkx programcsomag forráskódját átírtam.

A geográfiai klaszterezettségi együttható vizsgálatát témavezetőm javasolta. A vizsgálati módszerek közül pedig a véletlen hálózatok generálását és az eredeti hálózat azokkal való összevetését. A másik fő vizsgálati módszerünket, a különböző súlyú élek elhagyását viszont önállóan találtam ki. Nincsen tudomásom arról, hogy ilyen vizsgálati módszert mások alkalmaznak.

Az adatok kiértékelését önállóan végeztem. Ennél a feladatnál témavezetőm átnézte a kapott konklúziókat és felhívta a figyelmemet az esetlegesen téves következtetésekre.

A gráf ábrázolások elkészítésekor a kutatócsoport által írt programokat használtam, ebben az esetben az én feladatom csak a kapott ábrák értelmezése. Emiatt nem fektettem nagy hangsúlyt a dolgozatban a vizualizációra.

Munkám lényeges újdonsága az élek elhagyásának vizsgálati módszerében, illetve az ebből kapott eredményekben mutatkozik meg. Fő eredményünk az a távolságküszöb, ahol a legösszetettebb a hálózat tudomásom szerint nem található meg a szakirodalomban, vagyis ezt mi mértük meg először.