Szakdolgozat

Szociális hálózatok statisztikai elemzése

Rudolf Ádám

Fizikus MSc.,

Számítógépes fizika szakirány

Témavezetők:

Dr. Csabai István

egyetemi tanár

Szüle János

doktorandusz hallgató

Eötvös Loránd Tudományegyetem

Komplex Rendszerek Fizikája Tanszék

2014

Kivonat

A szakdolgozatban a `Twitter` API publikus Sample adatfolyamának archiválásával az ELTE Komplex Rendszerek Fizikája Tanszékén épített relációs adatbázisba gyűjtött adatok statisztikai elemzésével foglalkoztam. Az adatok gyűjtése során prioritást élvezett a *geo-tagged*, azaz földrajzi koordinátával rendelkező tweetek begyűjtése, így a hálózat földrajzi tulajdonságai vizsgálhatóvá váltak. A kölcsönös említési kapcsolatokból kialakított gráf távolságeloszlását számoltam, amiről azt találtam, hogy törésektől eltekintve hatványfüggvényt követ, és lokalizáltabb a kölcsönös követési gráfnál. Megvizsgáltam az eloszlás alakjának különböző tényezőktől való függését. Kimértem a nésűrűség fraktáldimenzióját a világra, valamint az USA-ra és Nagy-Britanniára, és szeparáltam a távolságeloszlás népsűrűségből, és szociális hatásokból származó alakját. A két említett gráf összehasonlításával azt találtam, hogy azokat nagy mértékben ugyanazok a felhasználók alakítják ki, de a kapcsolatrendszer lényegesen eltér. Főleg a fokszámok összehasonlításából azt találtam, hogy a kölcsönös említési gráf sokkal valószínűbben tartalmaz valódi embereket reprezentáló felhasználókat, és a 100-200-nál több kölcsönös követési kapcsolattal rendelkező felhasználók valószínűleg nem valódi személyek. Ezután hashtagek népszerűségének, és térbeli lokalizáltságának kapcsolatát vizsgáltam, és elemeztem a szórások eloszlását. Megmutattam, hogy a népszerűbb, és a kevésbé népszerű témák véletlenszerű pontokhoz, és véletlenszerűen választott tweetcsoportokhoz képest mennyivel lokalizáltabbak. Megmutattam, hogy rengeteg 0-hoz közeli szórású hashtag van, és hogy nem tudunk egyértelműen szétválasztani lokalizált, de népszerű témákat a többitől.

Köszönetnyilvánítás

Köszönettel tartozom témavezetőimnek, Csabai Istvánnak és Szüle Jánosnak, hogy lehetőséget kaptam bekapcsolódni a tanszéken folyó kutatómunkába, és hogy ötleteikkel, meglátásaikkal mindig lelkesen segítették munkámat.

Köszönet illeti Dobos Lászlót és Kondor Dánielt az adatbázis és a fejlesztőgépekhez való jogaim kezeléséért, és az adatbázis kezelésben nyújtott segítségükért.

Jelen kutatást a FuturICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

Budapest, 2014. május 30.

Tartalomjegyzék

1. Bevezetés **1**

2. A `Twitter` működése, fogalmai **3**

3. Az adatok **4**

3.1. Adatgyűjtés 4

3.2. Az elmentett adatok felépítése 5

4. Hálózatelméleti bevezető **6**

4.1. Gráfelméleti jellemzők 7

4.1.1. Betweenness centrality 7

4.1.2. Fokszám eloszlás 8

4.1.3. Klaszterezettség 9

4.2. A Barabási-Albert modell 11

4.2.1. Skálafüggetlen eloszlás 12

5. Statisztikai vizsgálat **14**

5.1. Sokféle kapcsolatrendszer 14

5.2. Távolságeloszlás 15

5.2.1. Távolságeloszlás az USA-ra és Európára 19

5.2.2. Távolságeloszlás alakjának s és r függése 21

5.3. Népsűrűség fraktáldimenziójának kimérése 29

5.4. A kölcsönös követési és kölcsönös említési gráf összehasonlítása 35

5.4.1. Egyszerű kvantitatív összehasonítási mértékek, kis kvalitatív elemzés 35

5.4.2. Metszet részletesebb elemzése 38

6. Témák (hashtagek) elemzése **43**

6.1. Hashtagek népszerűségének és lokalizáltságának vizsgálata 43

7. Összegzés **50**

Hivatkozások **53**

Nyilatkozat **57**

Ábrák jegyzéke

4.1 Példa a fokszámeloszlásra: egymást kölcsönösen említő felhasználókból alkotott gráf fokszámeloszlása logaritmikus skálán ábrázolva. 9

5.1 A kölcsönös említési gráf ($s = 1$ esetén) távolságeloszlása a kölcsönös említési gráféval összehasonlítva. 18

5.2 A kölcsönös említési gráf ($s = 1$ esetén) távolságeloszlása Európa és az Egyesült Államok területére szűkítve. 19

5.3 A világ távolságeloszlásának alakja különböző küszöbértékeket megszabva az oda és vissza irányú említések számának minimum értékére. (s) 21

5.4 A valószínűség sűrűség módosulása az s küszöbérték változtatásával. . 22

5.5 Oda-vissza irányú említések arányának eloszlása, és azonos eloszlással generált véletlenszerű számpárok arányának eloszlása. 23

5.6 Oda-vissza irányú említések arányán azonos eloszlással generált véletlenszerű számpárok arányához viszonyatva. 25

5.7 Említési gráf távolságeloszlása különböző oda-vissza említési arányokra. 26

5.8 A valószínűség sűrűség módosulása az r arányra szűrésekkel. 26

5.9 r arányok relatív gyakorisága $s = 10$-es küszöbérték mellett. 28

5.10 Oda-vissza irányú említések aránya azonos eloszlással generált véletlenszerű számpárok arányához viszonyatva $s = 10$ esetén. 28

5.11 `Twitter` felhasználók népsűrűsége. Koordinátáik a *friends-of-friends* algoritmussal lettek meghatározva[30] tweetjeik koordinátáinak klaszterezéséből a legnagyobb számosságú klaszer átlagaként. 29

5.12 Véletlenszerűen összekötött felhasználók távolságeloszlása, és gömbi egyenletes eloszlású pontok elméletileg számolt távolságeloszlása. A kettő közötti különbség abból fakad, hogy a `Twitter` felhasználók a Föld felszínén nem egyenletesen vannak elosztva. Az egyik legfontosabb különbség, hogy az egyenletes eloszlás négyzetesen, a valódi felhasználók eloszlása viszont majdnem lineárisan indul. 31

5.13 Véletlenszerűen összekötött felhasználók távolságeloszlásának eleje logaritmikus skálán ábrázolva. A kezdeti szakaszra illesztett egyenes meredeksége adja a pontok (felhasználók) korrelációs dimenzióját. . . 33

5.14 Véletlenszerűen összekötött felhasználók távolságeloszlásának eleje logaritmikus skálán ábrázolva, a távolságot közelítő formulával számolva. Ezt az ábrát jóval több pontra számoltam, mint a 5.13. ábrát. A kezdeti szakaszra illesztett egyenes meredeksége adja a pontok (felhasználók) korrelációs dimenzióját. 34
5.15 A kölcsönös említési, és kölcsönös követési gráfok különálló, és átfedő pontjainak (felhasználóinak), és éleinek aránya. A felhasználók között nagy az átfedés, de az élek igen nagy részben különböznek, vagyis többnyire ugyanazok a felhasználók alakítanak ki különböző jellegű követési, és említési kapcsolatokat. 37
5.16 A kölcsönös említési gráf és a kölcsönös követési gráf metszetének felhasználói a két gráf fokszámai által kifeszített térben. A nagy fokszámokra jellemző felhasználóknál inverzió látszik a korreláció helyett. 38
5.17 A kölcsönös említési gráf és a kölcsönös követési gráf metszetének felhasználói a két gráf fokszámai által kifeszített térben. A színek az adott ponthoz tartozó felhasználók számának logaritmusát jelzik. Egy nagyon erősen lecsengő csúcs van az origó közelében, továbbá a szintvonalak mintha Poisson görbéket követnének. Az ábra piros vonallal tartalmazza a változó méretű binekre említési, és zöld vonallal követési barátok számát is a másik barátok számának függvényében. . 39
6.1 Hasgtagek térbeli szórása a gyakoriság függvényében. Ezen a szakaszon mind CNT_{gt}, mint σ függvényében erősen lecsengő az eloszlás. . 45
6.2 A 6.1. ábrához hasonló módon ábrázolva a gyakoriságok 1 000 fölötti gyakoriságú hashtagekre. Ilyen gyakori hashtagekre már konstans az átlag, és a szórás eloszlásának maximuma az átlag körül van, és akörül nagyjából szimmetrikus. 46
6.3 Hashtagek átlagos szórása a népszerűség függvényében a hosszá illesztett véletlen pontokból, valamint véletlen tweetcsoportokból számolt hasonló görbékkel. 47
6.4 $CNT_{gt} > 1000$ hashtagek összesített σ_R szerinti eloszlása 50 km-es binmérettel. 49

1. Bevezetés

Számos tudományág történetében azt láthatjuk, hogy az egyre több alapvető törvényszerűség, és egyszerű rendszer leírása után a leírandó rendszerek egyre bonyolultabbá, egyre komplexebbé válnak. Ugyanígy van a fizikában is. A klasszikus newtoni mechanikától ma már nem sok alapvető újdonságot várhatunk. A modern fizika, mint kvantummechanika, és relativitáselmélet is ugyan még most is sok megválaszolatlan kérdést tartalmaz, ennek ellenére jól kiforrottnak számít. Az elmúlt évtizedekben azonban előtérbe került a sok elemből álló kölcsönható rendszerek, mint egységes egész vizsgálata a klasszikus módszer, a részekre bontás helyett. Az ilyen rendszerek napjainkban egyik legnépszerűbb megközelítési módja, ha azokat hálózatként modellezzük, vagyis matematikailag gráfként írjuk le.

Az emberi társadalom is tekinthető egy igen komplex rendszernek. Habár a társadalom esetében természetesen nem várhatunk a klasszikus fizikai értelemben vett egyszerű összefüggéseket, de a természettudományos megközelítés mégis hasznos lehet. Ennek értelmében méréseket kell végeznünk, és a mérések eredményeit modellek keretei között kell magyaráznunk.

Az elmúlt évtizedekben mind a hálózatkutatás[1], mind a számítógépes adatgyűjtés[2] szerepe igen nagyra nőtt a természettudományokban. A szociális kapcsolatok régóta tudományos vizsgálódás tárgya.[3] A számítógépes szociális hálózatok (pl. `Facebook`, `Twitter`) elterjedésével pedig eddig soha nem látott mértékben belelátunk az emberek közötti kapcsolatrendszer alakulásába. Az ilyen internetes szolgáltatásokat tekinthetjük egyfajta "mérőeszköznek", amik adatokat szolgáltatnak a valódi szociális hálózatokról, vagy azokhoz igen hasonló felépítésű rendszerekről. Ezen szolgáltatások használata közben a társadalmi érintkezéseknek, kapcsolatoknak, személyeknek mind nyoma van, így a rendszer szinte minden változása nyomon követhető, így vizsgálható, ellentétben a valódi beszélgetésekkel és ismerősi viszonyokkal. Ennek köszönhetően szociológusok, informatikusok, fizikusok előszeretettel foglalkoznak ezen infokommunikációs hálózatok kutatásával.

Jelen szakdolgozat témája a `Twitter` internetes mikroblog-szolgáltatás. A `Twitter` vizsgálatával széleskörű szakirodalom foglalkozik. Kulcsszavak alapján szinte azonnal detektálhatók fontosabb események, például földrengés[4], bizonyos módszerekkel pedig monitorozható felhasználók általános kedve területre lebontva.[5] Mindez hasznos lehet marketing célokra[6, 7, 8], előrejelzésekre, például egy politikai

választás kimenetele[9], a tőzsde viselkedése[10], vagy akár betegségek kitörése[11].

Mindezek, továbbá az, hogy naponta 200 millió felhasználó több, mint 400 millió tweetet küld[12], mutatják, hogy a `Twitter` egy bőséges információforrás, amit sokszínűen fel lehet használni, és az információterjedés egy fontos eszköze, így kapcsolati rendszerének feltárása, struktúrájának megismerése érthető igény.

A dolgozat felépítése a következő. A 2. és 3. fejezetekben röviden leírom a `Twitter` dolgozat szempontjából fontos terminológiáját, és bemutatom az általam használt adathalmaz felépítését és jellemzőit. E rövid bevezető után a 4. fejezetben az idevágó gráf- és hálózatelméleti alapokat ismertetem, és részletesebben bemutatom a szakdolgozat motivációját. A 5. fejezetben ezután a kölcsönös említési gráf tulajdonságait vizsgálom, főleg a távolságeloszlás alakján keresztül. Bemutatom, hogy milyen mértékben alakítják ki ezt az alakot a népsűrűség, és milyen mértékben szociális effektusok. Megmutatom, hogyan lehet kimérni a népsűrűség fraktáldimenzióját ezen adatok segítségével, végül a kölcsönös említési, és kölcsönös követési gráfokon végzek összehasonlítást arra keresve a választ, hogy milyen jellegű kapcsolatok reprezentálhatnak valódi ismeretségi kapcsolatot, és milyenek, amik valószínűleg csak a `Twitteren` léteznek. Végül a 6. fejezetben a *hashtagek*, azaz a felhasználók által bejelölt témákat vizsgálom népszerűségük, és térbeli lokalizáltságuk szerint.

2. A `Twitter` működése, fogalmai

A `Twitter` lényege a következő. A regisztrált felhasználók maximum 140 karakter hosszúságú üzeneteket (*tweeteket*) tehetnek közzé, amik az alapbeállítás szerint publikusak, és a felhasználó profiloldalán jelennek meg. A felhasználóknak lehetőségük van más felhasználókat *követni* (*follow*). Ha követünk egy felhasználót, az azt jelenti, hogy automatikus értesítést kapunk a tweetjeiről. Lehetőség van mások tweetjeit nekünk újra tweetelnünk, ezt *retweetelésnek* hívják. Ez hasonló, mint egy e-mail továbbítása, vagy egy *megosztás* a `Facebook`-on.

A tweet tartalmazhat linket képre, videóra, honlapra.

Lehetséges a tweetben egy másik felhasználó *említése* (*mention*) a "@" karakter segítségével, amit a megemlített felhasználó felhasználóneve követ. Ekkor az említett felhasználó értesítést kap az említésről. Ezt az üzenetek közvetlen címzésére szokták használni, vagyis az adott tweet - habár publikus - egy konkrét felhasználónak szól. Ilyen dedikált tweetekből gyakran teljes beszélgetések alakulnak ki felhasználók között.

A *hashtag* egy szó, ami előtt a "#" karakter áll. Általában témákat jelölnek meg vele, a tweetek címkézéseként szolgál, ami lehetővé teszi a tematikus keresést.

Fontos, hogy egyes tweetekkel együtt - készüléktől és beállítástól függően - a közzététel GPS koordinátája is beérkezik, ezzel vizsgálható akár felhasználók elhelyezkedése, mozgása is.[13] Ezeket *geo-tagged* tweeteknek nevezzük.

3. Az adatok

3.1. Adatgyűjtés

A `Twitter` publikus API-ja (*application programming interface*) az adott felhasználók követése mellett lehetővé teszi a teljes twitteres adatfolyam véletlenszerű mintavételezését is. Ez az összes üzenet kb. 0,5%-át jelenti, de a kicsi arány ellenére ez egy igen hasznos adatforrásnak bizonyult, véletlenszerűsége miatt pedig reprezentatív mintának tekinthető. Sok kutatás valós időben elemzi ezt az adatfolyamot[4, 14, 15, 16], ám az ELTE Komplex Rendszerek Fizikája Tanszékén relációs adatbázist terveztek és építettek arra a célra, hogy a folyamot archiválják hónapos, vagy éves nagyságrendű időn keresztül.[17] Az így nyert igen nagy mennyiségű adat alkalmasabb a hálózatok statisztikai és földrajzi jellemzőinek elemzésére. Lehetőség van továbbá az adatfolyam szűrésére, például csak a *geo-tagged* tweetek elmentésére.

Ha egy tweet beérkezik, akkor az API elküldi a küldő felhasználó adatait is. Retweet esetén pedig mind a kérdéses, mint az eredeti tweet adatait megkapjuk. Ez azt eredményezi, hogy az adatbetöltés előtti, meglepően régi tweetek is előkerülnek és betöltődnek az adatbázisba. Ez az adatfolyam kb. 50-60 tweetet eredményez másodpercenként, ez napi 4-4,5 millió tweetet jelent, ami kb. 10 GB naponta. Az adatbázisba való betöltés nem valós időben, hanem kb. havonta történik meg az időközben felgyülemlett adatokból.

Az adatbázis a Microsoft SQL Server 2012 motort használja, az adatok elérésére így az SQL lekérdezőnyelv segítségével történik.

3.2. Az elmentett adatok felépítése

Az adatok két fő táblába vannak rendezve, ezek a `"tweet"`, ami a tweetek, és a `"user"`, ami a felhasználók adatait tartalmazza. Minden tweet és felhasználó egyedi azonosítóval rendelkezik (`tweet_id` és `user_id`). Az adatbetöltések futásokba, azaz *runokba* vannak rendezve, amik szintén egyedi azonosítót kaptak (`run_id`). Ez azt a célt szolgálja, hogy meg lehessen különböztetni a különböző időpontokban, és/vagy különböző szűrésekkel végrehajtott adatbetöltéseket.

A `"tweet"` tábla tartalmazza többek között a tweetek azonosítóját, a tweetet közzétevő felhasználó azonosítóját, mint idegen kulcsot, a tweet szövegét unicode karakterekben (ami lehetővé teszi a nemzetközi használatot), a tweet közzétételének idejét és helyét (ha meg van adva) hosszúsági és szélességi koordinátákban, valamint egységgömbre vetített x, y, z koordinátákban. Ha a tweet válasz egy másikra, akkor az eredeti tweet azonosítója is beérkezik a tweettel.

A `"user"` tábla tartalmazza egyebek mellett a felhasználó azonosítóját, felhasználónevét (amit maga választ), egy leírást saját magáról, hogy engedélyezte-e a geo-tagged tweeteket, a követői, követettjei, tweetjei számát, valamint információkat a profiljáról, például a háttér színét. Ez a tábla akkor bővül, ha olyan tweet érkezik, amihez tartozó felhasználó még nincs benne a táblában.

Ezek mellett léteznek más táblák is: a `"tweet_hashtag"` a hashtaggel megjelölt kulcsszavakat, a `"tweet_retweet"` a retweetek, és az eredeti tweet azonosítóját tartalmazzák, a `"tweet_user_mention"` pedig azt, hogy melyik tweetben ki kit említett meg. Egy `"user_update"` tábla naplózza a felhasználókra vonatkozó adatok változásait. A tábláknak a leírtakon kívül más mezői is vannak, amik azonban lényegtelenek a dolgozat szempontjából.

Megjegyzendő, hogy a követési kapcsolatok felderítése bonyolultabb probléma, ám ennek részletezése nem témája ezen dolgozatnak. A lényeg az, hogy az adatbázisban erről is rendelkezésre állnak információk.

A fentebb leírt táblákból természetesen származtatható sokféle másik is, de ezek alkotják az alapját azoknak az adatoknak, amikkel a `Twitter`-rel kapcsolatos kutatások folynak a Komplex Rendszerek Fizikája Tanszéken

4. Hálózatelméleti bevezető

Szociális hálózatoknak nevezzük azokat a rendszereket, amiket adott szereplők, és a közöttük kialakult kétszereplős kapcsolatok alkotnak. A szereplők lehetnek személyek, vállalatok, csoportok, a közöttük lévő kapcsolatnak pedig tekinthetjük a barátságot, ismeretséget, közös publikációkat, stb.. Ezek természetes módon reprezentálhatók gráfokkal.

A gráfelmélet a matematika mára igen régi ága (megjelenését ugyanis Euler nevéhez, és az 1736-os évhez kötjük[18]). Ennek köszönhetően kiforrott, és széles körű eszközrendszere van, amiből e fejezetben bemutatok néhány alapvető fogalmat.

A gráf definíciója szerint pontok és élek halmaza, vagyis adott egy V halmaz (az angol *vertex*, azaz csúcs szóból), és egy E halmaz (az angol *edge*, azaz él szóból), ami a V halmaz önmagával vett Descartes szorzatának részhalmaza, azaz elemei a csúcsokból alakított rendezett párok. Mivel a párok rendezettek, ez megengedi az élek irányítását, de ennek alesetének tekinthető az irányítatlan gráf, ahol egy él két végén lévő csúcsok szerepei szimmetrikusak. A gráf tartalmazhat hurokéleket, amikor egy csúcs önmagával is össze van kötve, valamint definiálhatunk egy $w : E \to \mathbb{R}$ súlyfüggvényt, ami az élekhez valós számokat rendel, amit súlynak hívunk. Ezen súlyokkal reprezentálhatjuk az egyes csúcsok közötti kapcsolat erősségét.

A szociális hálózatok szereplői így a gráf csúcsaival, a közöttük lévő kapcsolatok pedig éleivel reprezentálhatók, így matematikailag kezelhetővé válnak. A szociális hálózatok ilyen jellegű kutatásával a hálózatelmélet foglalkozik, ami a gráfelmélet eredményeit, és eszközeit használja valódi kapcsolatrendszerek elemzésére.

Sok, az interneten megjelenő szolgáltatás épp ilyen szociális hálózatok kiépítésére kínál lehetőséget. Jelen dolgozat témája, a `Twitter` is ilyen. Mivel nincs tematikája, ami leszűkítené a felhasználók csoportját az emberek egy bizonyos részhalmazára, továbbá elég népszerű, joggal várhatjuk tőle, hogy kapcsolatrendszere bizonyos szempontból hasonlít a valódi ismeretségi kapcsolatokra. Érdekes kérdés, hogy valamilyen adott információ hogyan terjed felhasználóról felhasználóra a `Twitter` hálózatán keresztül. Mely felhasználók játszanak központi szerepet a hírek terjedésében, és milyen jellegű kapcsolatok mentén terjed tovább az információ. Ez fontos lehet például abból a szempontból, hogy reklámcéllal mely felhasználókat érdemes megcélozni, vagy hogy bizonyos hírek terjedésének elősegítésében, vagy lassításában mely kapcsolatokat, vagy felhasználókat érdemes befolyásolni, figyelni.

4.1. Gráfelméleti jellemzők

Ebben az alfejezetben ismertetem azokat az alapvető gráfelméleti jellemzőket, amik szükségesek a dolgozat megértéséhez.

Két pont **szomszédos**, ha él köti össze őket. Szociális hálózatban általában valamilyen kapcsolatban lévő szereplőpárok szomszédosak, pl. barátok, családtagok, ismerősök. **Út**nak nevezzük pontok és élek olyan sorozatát, ahol minden él ott indul, ahol az előző végződött, és egy pont sem szerepelhet benne kétszer. Szociális hálózatokban ez egy ismerősöm ismerősének ismerőse, stb. jellegű kapcsolatrendszer. Jelentősége, hogy közvetett kapcsolatot teremt nem szomszédos pontok között. Mivel a világon az emberek az összes többinek csak elég kis részét ismerik, az információterjedés szempontjából igen fontos szerepet játszanak az utak, ugyanis az információ ezek mentén jut el távoli pontokba.

Út **hossza** az őt alkotó élek száma (súlyozott gráfban az élek súlyainak összege). Két pont **távolsága** az őket összekötő utak közül a legrövidebb hossza. Szociális hálózatokban ez a mérőszám ott játszik szerepet, hogy adott terjedő információnak egyik pontból eljutva a másikba hány szereplőn kell átugrania. A valóságban persze nem feltétlenül a legrövidebb úton terjed az információ, de egyszerűsítésként tekinthetjük így.

Egy gráf **sűrűségének** nevezzük a gráfban ténylegesen létező élek számának, és az összes lehetséges él számának hányadosát. Mivel az összes lehetséges él száma irányítatlan gráfban a pontpárok számával egyenlő, ami $\frac{|V|(|V|-1)}{2}$, a sűrűség a $\frac{2|E|}{|V|(|V|-1)}$ formulával számolható. (Irányított gráfban az élek lehetséges száma ennek kétszerese, tehát a sűrűség formulájából eltűnik a kettes faktor.)

4.1.1. Betweenness centrality

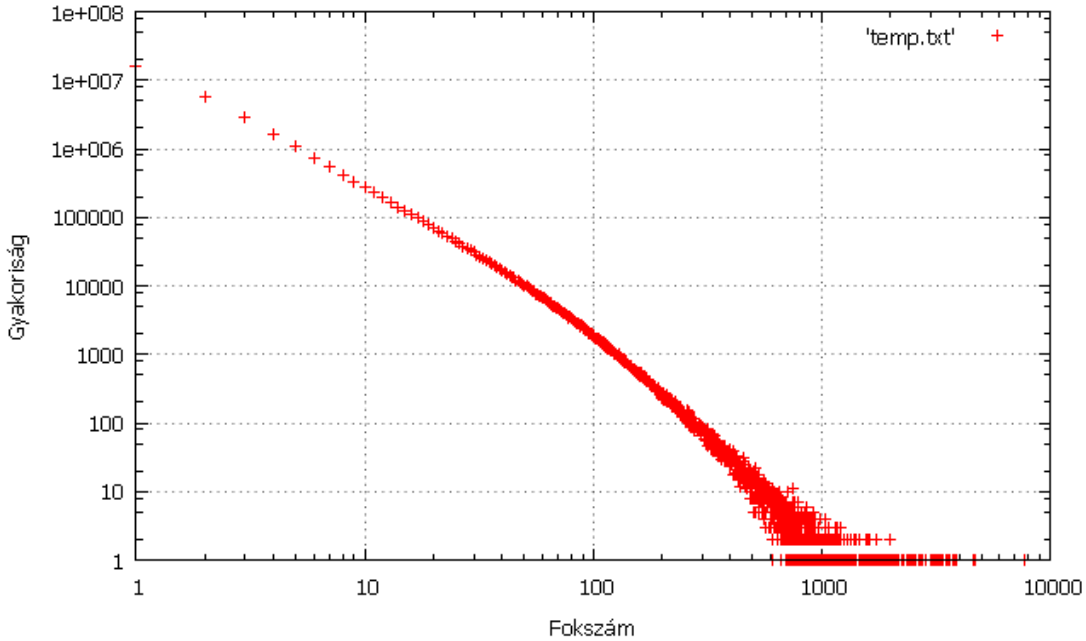
A hálózatelméletben különböző mérőszámok születtek pontok fontosságának számszerűsítésére, ezeket centralitásnak, vagy központiságnak nevezzük. Mivel a gráfok definíciójában nincsen semmilyen euklideszi tér, a középpont definiálása egészen más jellegű feladat, mint a geometriában, és nem triviális. Többféle centralitási jellemző is van, számunkra érdekes lehet az a mérőszám, aminek elnevezése az angol szakirodalomban **betweenness cetrality** (közöttiség centralitásnak lehetne fordítani, de magyar szakirodalom hiányában az angol kifejezést fogom használni, ugyanis az az elterjedt).[19] Ezt úgy számoljuk, hogy a gráfban létező összes pontpár-

ra meghatározzuk a közöttük vezető legrövidebb utat, majd minden ponthoz hozzárendeljük, hogy ezen utakból hány ment rajta keresztül. Mivel az előbbiekben kijelentettük, hogy az információ utak mentén terjed, a *betweenness centrality* méri azt, hogy egy-egy szociális szereplő hány távoli szereplő közötti információs útvonalban játszik szerepet, azaz mennyire fontos csomópont az információterjedési hálózaton. A *betweenness centrality* meghatározására léteznek algoritmusok, ám mivel a pontpárok száma $\mathcal{O}(|V|^2)$ nagyságrendű, ahol $|V|$ a pontok száma, a legrövidebb út megtalálási ideje két pont között pedig a legnaivabb algoritmus szerint szintén $\mathcal{O}(|V|^2)$-tel skálázik[20], a *betweenness centrality* meghatározása numerikusan igen bonyolult probléma. Habár született algoritmus $\mathcal{O}(|V||E|)$ futási idővel is[21], ahol $|E|$ az élek száma, nagy hálózatokra ez is nagyon nagy futási időt jelent. (Az általam vizsgált hálózatokban a pontok száma milliós, az élek száma tízmilliós nagyságrendű.) Később látni fogjuk, hogy skálafüggetlen hálózatokban jóval egyszerűbb dolgunk lesz a nagy *betweenness cetrality* értékkel rendelkező pontok megtalálásában.

4.1.2. Fokszám eloszlás

A gráfelméletben egy csúcs **fokszámának** nevezik a hozzá kapcsolódó élek számát, vagy másképp mondva a szomszédainak számát. Egy szociális hálóban ez pl. egy szereplő ismerősei számát jelenti. Irányított gráfban egy ponthoz külön tartozik kimenő, és bejövő fokszám, például a `Twitter` rendszerében az lehet ilyen, hogy hány embert követek, és hány ember követ engem. A fokszám is jelezheti a pont valamilyen értelemben vett fontosságát, ugyanis aki sok más szereplővel van kapcsolatban, sok kapcsolat mentén vehet részt az információterjedésben, és eltűnése nagyobb mértékben rendezi át a hálózat struktúráját, mint kevés kapcsolattal rendelkező ismerőseié.

A gráfelméletben **fokszámeloszlás**nak nevezzük azt a függvényt, ami minden előforduló fokszámhoz hozzárendeli a gyakoriságát - azaz hogy hány ilyen fokszámú csúcs van a gráfban. Esetünkben gyakorlatilag azt adja meg, hogy hány olyan szociális szereplő van a hálózatban, aki adott számú kapcsolattal rendelkezik. A fokszámeloszlás alakja jó és egyszerű statisztikai jellemzést ad a hálózat struktúrájáról, ugyanis a különböző hálózatmodellekben különböző, jól definiált alakot vesz fel[22].



4.1. ábra: Példa a fokszámeloszlásra: egymást kölcsönösen említő felhasználókból alkotott gráf fokszámeloszlása logaritmikus skálán ábrázolva.

4.1.3. Klaszterezettség

Szociális hálózatok gráfjaiban a szereplők hajlamosak a teljes hálózat sűrűségénél jelentősen sűrűbb részgráfokat alkotni[23, 24] (részgráf a V halmaz egy részhalmazából, és a közöttük húzódó élekből alkotott gráf). Az ilyen, összefüggő csoportokat nevezzük **klaszter**eknek. Az klaszterek megjelenése érthető a való életbeli példákból. Egy adott városon belül nagyobb a valószínűsége, hogy két ember ismeri egymást, mint különböző városok között. Klasztereket alkothatnak az egy iskolába járó, egy munkahelyen dolgozó emberek, vagy bármilyen, hasonló tulajdonsággal rendelkező szociális szereplők. A klaszterek jelentősége abban áll, hogy a benne résztvevő szereplők nagyobb hatással vannak egymásra, a klaszteren belül nagy sűrűsége miatt gyorsabban terjed az információ.

A klaszterezettség jellemzésére bevezetett egy kvantitatív tulajdonság a **klaszterezettségi együttható**. Ennek definiálásához először be kell vezetnünk a **kör**, és **háromszög** fogalmát. A kör egy olyan út, aminek kezdő- és végpontja megegyezik, más néven zárt út. A háromszög pedig egy 3 hosszúságú kör, vagyis három olyan pont, amik páronként szomszédosak. (Ezzel ez egy 3-ad rendű **klikk** is, ahol a k-ad rendű klikk a gráfnak egy k pontból álló, teljesen összefüggő részgráfja, azaz min-

den pont mindennel össze van kötve. Ezzel a definíció teljesíti a köznapi életben vett klikktől várható tulajdonságot, azaz hogy egy hálózaton belül egy összefüggő részcsoportot jelentsen.)

Az i-edik pont klaszterezettségi együtthatóját a következőképpen definiáljuk:

$$c_i = \frac{2T_i}{k_i(k_i - 1)} \tag{4.1}$$

Ahol T_i azon háromszögek száma, aminek az i-edik pont része, k_i pedig az i-edik pont fokszáma, azaz a szomszédainak száma. A nevezőben lévő kifejezés a szomszédokból alkotható párok számának kétszerese, a klaszterezettségi együttható tehát azt adja meg, hogy a pont szomszédaiból alkotott párok mekkora hányada van kapcsolatban egymással is. (Egy ember ismerőseiből mennyien ismerik egymást is.) Ha a pont és szomszédai klikket alkotnak, ez a szám 1, ha viszont csak a vizsgált pont teremt kapcsolatot a szomszédai között, akkor 0.

Az egész hálózat jellemzésére használhatjuk az **átlagos klaszterezettségi együtthatót**, ami c_i-k átlaga az összes pontra számolva:

$$C = \frac{1}{|V|} \sum_{i=1}^{|V|} c_i \tag{4.2}$$

Ez a szám szemléletesen azt adja meg, hogy egy véletlenszerűen kiválasztott pont szomszédai milyen valószínűséggel kötődnek egymáshoz, szemléletesen pl. egy véletlenszerűen kiválasztott személy ismerősei milyen valószínűséggel ismerik egymást is. Az átlagos klaszterezettségi együttható a valóságban megtalálható szociális hálózatokra lényegesen nagyobb, mint akkor várnánk, ha véletlenszerűen kötnénk össze a szereplőket. Ezt a való életbeli tapasztalatunk is mutatja, miszerint az ismerőseink igen nagy része egymást is ismeri, főleg bizonyos csoportokon belül (mint pl. a már említett iskola, munkahely, város).

4.2. A Barabási-Albert modell

A valóságos hálózatok leírására az idők során több modellt is alkottak. Ha nem tudunk semmit egy kapcsolat kialakulásának valószínűségéről, akkor a hálózatot a pontok között valamilyen p valószínűséggel behúzott élekkel modellezhetjük. Ezt a modellt Erdős és Rényi terjesztette fel az 1960-as években. Ez a modell azonban a valóságban megfigyelt szociális hálózatok klaszterezettségénél lényegesen kisebb értéket ad, aminek orvoslására Watts és Strogatz 1994-ben új modellt vezetett be egy szabályos, klaszterezett hálózat éleinek egy részének áthuzalozásával.[22]

1999-ben Barabási Albert-László és Albert Réka a megfigyelésekhez még jobban illő modellt alkotott. Modelljük gyökeresen különbözik az előzőektől abban a tulajdonságában, hogy nem statikus hálózatokat ír le, hanem figyelembe veszi azok időbeli fejlődését új pontok időben történő hálózathoz adásával.[1] Felismerésük lényege, hogy a hálózathoz hozzáadott új pontok a kapcsolataikat nem véletlenszerűen alakítják ki, hanem nagyobb valószínűséggel kapcsolódnak olyan pontokhoz, amik már eleve sok kapcsolattal rendelkeznek. Ezt nevezik a *gazdag egyre gazdagabb lesz* elvnek. Ezzel sikerült megfogniuk azt a valóságos effektust, hogy a hálózat időben dinamikusan változik, és hogy az új kapcsolatok kialakítása jelentősen befolyásolja a hálózat struktúráját. Mivel ez az egyszerű elv sok esetben teljesül, a Barabási-Albert modell alkalmas sok való életbeli hálózat leírására. Egyebek mellett tipikusan ilyenek a szociális hálózatok[3], és az internetes struktúrák[22].

A modell egy olyan hálózatot eredményez, ahol nagyon sok kis fokszámú pont lesz, és kevés nagyon nagy fokszámú. Ez a néhány pont felelős a hálózat egyben tartásáért, a pontpárok közötti utak jelentős része rajtuk megy keresztül. Ez azt is eredményezi, hogy bármely kis fokszámú pontból rövid úton eljuthatunk egy nagy fokszámúba, ahonnan megint rövid úton bármely kis fokszámúba, így a pontok átlagos távolsága a pontok számához képest meglepően kicsi lesz ($\mathcal{O}(\ln(|V|)$-vel skálázik). Ezt nevezzük *kisvilág-tulajdonságnak*. A kisvilág-tulajdonságot mutató gráfok átlagos klaszterezettségi együtthatója szignifikánsan nagyobb az azonos pontokból létrehozott gráfénál, aminek átlagos pont pár távolsága nagyjából megegyezik a kérdéses gráféval.

4.2.1. Skálafüggetlen eloszlás

A kisvilág-tulajdonságot mutató hálózatok fokszámeloszlása igen jellegzetes formát követ. Mivel kis fokszámú pontokból sok van, aztán az egyre nagyobb fokszámokra a gyakoriság drasztikusan csökken, a függvényalak egy erősen levágó függvény. Barabási és Albert megmutatták, hogy a fokszámeloszlás egy csökkenő hatványfüggvény szerint változik:

$$P(k) \sim k^{-\gamma} \tag{4.3}$$

Ahol γ egy (általános esetben nem egész) pozitív szám, k pedig a fokszám. Megmutatható, hogy $\gamma < 2$ esetben az eloszlásnak nincs elméleti várható értéke, $\gamma < 3$ értékekre pedig végtelennek tekinthető a szórása. Az ilyen fokszámeloszlással rendelkező hálózatokat skálafüggetlen hálózatnak hívjuk.[32]

A gyakorlatban a hatványfüggvény felismerése úgy történik, hogy az adatokat logaritmikus x és y tengelyen ábrázoljuk. Amennyiben az eloszlás hatvány alakú, a pontok egy egyenesen helyezkednek el, aminek meredeksége $-\gamma$. Ez illesztéssel meghatározható.

Számunkra fontos tulajdonsága az ilyen gráfoknak, hogy - ahogyan már utaltam rá -, a skálafüggetlen hálózatokban kialakítható utak jelentős része a nagy fokszámú pontokon megy át, ami azt is jelenti, hogy a nagy fokszámú pontoktól azt várjuk, hogy nagy lesz a *betweenness centrality* értékük. Goh, Kahng és Kim (2008)[25] megmutatták, hogy amennyiben $\gamma > 1$ és $\delta > 1$, ahol δ a *betweenness centrality* eloszlásának γ-hoz hasonló hatványkitevője, az adott fokszámú pontok átlagos *betweenness centrality* értéke a fokszám emelkedő hatványa szerint növekszik. Ez azt jelenti, hogy ha a konkrét értékre nem vagyunk kíváncsiak, hanem csak a *betweenness centrality* szempontjából fontos pontokat keressük, akkor - miután beláttuk, hogy skálafüggetlen hálózattal van dolgunk - a bonyolult és számításigényes algoritmusok helyett elég a nagy fokszámú pontokat megkeresni, ami egyszerű és triviális feladat. Szemléletesen ez azt jelenti, hogy azok a szociális szereplők vesznek részt sok információ áramlási útvonalban, akiknek sok kapcsolatuk van.

A nagy fokszámú pontok meghatározása egyszerű ugyan, de amint látni fogjuk, a `Twitter` esetében nem triviális, hogy milyen gráf szerint határozzuk meg a fokszámokat. Tehát ezután a fontos feladat, hogy olyan kapcsolatokat definiáljunk a `Twitter` rendszerén belül, amik valódi ismeretségi kapcsolatokat mutatnak, amik mentén egymást ténylegesen ismerő felhasználók valódi beszélgetéseket folytatnak. Ezután az eddigieknek megfelelően meg tudjuk határozni a csomópontként viselkedő felhasználókat.

5. Statisztikai vizsgálat

5.1. Sokféle kapcsolatrendszer

Mint az eddigiekből már kiviláglott, a `Twitter` rendszerében rengetegféle kapcsolatrendszert definiálhatunk. Ezeket gráfokkal reprezentáljuk, ami csúcsok valamilyen V és élek E halmaza. A csúcsok tipikusan felhasználók, ám a közéjük húzható élekre sok különböző definíciót alkothatunk. A legtermészetesebb kapcsolat, ha egy felhasználó követ egy másikat.[26] Ez irányított gráf, ám egyrészt ennek a gráfnak a feltárása nem triviális[17], másrészt nem reprezentál valós ismeretségi viszonyt[27], mivel a felhasználók gyakran hírességeket, cégeket, humoroldalakat követnek, vagy pedig pont ezek kezdenek követni felhasználókat reklámcéllal.

Definiálható kapcsolat kölcsönös követéssel is, ez már erősebb feltétel, és egy nem irányított gráfot eredményez. Retweet alapján is lehet éleket behúzni, ám hírességek tweetjeit gyakran retweetelik, és egyes tweetek témájuk miatt is népszerűvé válhatnak, a retweetelés tehát megint csak nem garantál személyes kapcsolatot. Az említések is egyirányú éleket jelentenek, ez már közelebbi kapcsolatot jelenthet, hiszen az üzenetet az egyik felhasználó a másiknak címezte, ám szintén problémát okozhat a hírességek, weblapok, cégek felhasználóinak gyakori említése.

Erre megoldás, ha a kölcsönös említéseket vizsgáljuk. Ez már egy oda-vissza irányú üzenetváltás, szemben például az oda-vissza irányú követéssel, ami lehet egy személytelen kattintás is, és ahogy Huberman, Romero és Wu megmutatták (2008), az említés, akár egy minimum említésszámmal szigorítva szorosabb kapcsolatot jelent a követésnél.[27] A dolgozatban a kölcsönös említési gráf statisztikai jellemzőit fogom vizsgálni. Nevezzük ezt a kapcsolatot *említési barátságnak*, a kölcsönös követést pedig *követési barátságnak*.

További előnye, hogy az említések számával lehet súlyozni az éleket, így árnyalva a kapcsolatrendszert. Ha ilyen módon élt húzunk a és b felhasználó közé, azaz a is említette b-t, és b is a-t, akkor jelölhetjük az $a \to b$ irányú említések számát m_{ab}-vel, a fordított irányúakat pedig m_{ba}-val. Bevezethetünk egy s küszöbértéket úgy, hogy csak azokat a kapcsolatokat tekintjük élnek, ahol m_{ab} és m_{ba} is nagyobb s-nél. Minél nagyobb s, ez egy annál szigorúbb feltétel. s szemléletes jelentése, hogy két felhasználót akkor tekintünk összekötöttnek, ha közöttük egy legalább s hosszúságú oda-vissza beszélgetés zajlott. Minél nagyobb s, annál valószínűbb, hogy

a két felhasználó között tényleges (akár személyes, akár csak online) ismeretség van. Érdekes mérőszám lehet még az $r = \min\{m_{ab}, m_{ba}\}/\max\{m_{ab}, m_{ba}\}$ arány. Valós beszélgetések esetén ez közel van 1-hez (vagy egzaktul annyi), ám ha az egyik felhasználó a másikat sokkal többször említette, mint az őt, akkor ez a szám kicsi. Ez jelezhet spamelést, vagy egyéb reklámtevékenységet, utalhat botok jelenlétére. Érdemes megvizsgálni, hogy a hálózat statisztikai viselkedése hogyan függ a kettő mérőszámtól.

5.2. Távolságeloszlás

Egy adott definícióval definiált kapcsolatot jellemezhetünk azzal, hogy az összekötött felhasználópárok tagjai milyen valószínűséggel milyen távol vannak egymástól. Ez egy valószínűségsűrűség függvény lesz a távolság függvényében. Vizsgálhatjuk az eloszlás karakterisztikáját, összehasonlíthatjuk formáját más gráfokra jellemző eloszlással, vizsgálhatjuk az eloszlás alakjának hely, s, vagy r függését is. Számszerű összefüggést tudunk adni a távolság és egy kapcsolat kialakulásának valószínűsége között. Mindezekhez azonban két nem triviális problémát is meg kell oldanunk.

Először is meg kell határoznunk a felhasználók koordinátáit. A felhasználók megadhatnak magukról egy *location* nevű adatot, ez azonban szöveges adat, és a felhasználók maguk döntik el, mit írnak ide. Gyakran üresen marad, vagy nem szabványos, humoros, értelmezhetetlen szöveget írnak be, így ez használhatatlan arra, hogy helyet rendeljünk hozzájuk. Használható helykoordinátánk csak egyes tweetekhez van. Mivel az ELTE-n több párhuzamos kutatás is folyik ezen adatokkal[28, 29], hozzáférhetünk más felhasználók által létrehozott táblákhoz is. Így rendelkezésünkre áll a `"user_location"` tábla, ami a következőképpen készült. Veszszük azokat a felhasználókat, akik tweetjeinek ismerjük a koordinátáit. Hozzájuk a friends-of-friends algoritmussal[30] 1 km-es küszöbindexszel klaszterezzük a tweetjeit, vagyis földrajzilag összefüggőnek tekinthető csoportokat alkotunk, méghozzá legalább 1-et, legfeljebb hármat. Ezek tipikusan lehetnek az otthon, illetve a munkahely. A klaszterezés C# nyelven íródott programmal történt. A program napszakot is rendelt a klaszterekhez, de az egyszerűség kedvéért itt a legnagyobb számosságú klaszter koordinátáit (a klaszterbe tartozó koordináták átlagát) vettük a felhasználó koordinátájának.

A felhasználók közötti távolságot úgy definiáltuk, hogy ha a Földet gömbnek tekintjük, akkor a mindkettejükön keresztülmenő főkör kisebbik ívhossza (ti. a legrövidebb geodetikus). A távolság számolására két módszer is adódik aszerint, hogy hogyan reprezentáljuk a Föld felszínén lévő pontot. Leírhatjuk egy Föld középpontjából a pont felé mutató vektorral, akkor a két vektor skalárszorzata alapján számolható az általuk bezárt szög, ami arányos a pontok felszínen mért távolságával. A pontot reprezentálhatjuk a hosszúsági és szélességi koordinátáikkal is. Ezek egyszerű koordinátatranszformációval egymásba transzformálhatók. A szélességi és hosszúsági koordinátákból a haversine formulával[31] számolhatunk távolságot. A haversine-formula:

$$d = 2 \cdot R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{5.1}$$

ahol φ_1 és φ_2 az egyik, illetve másik pont földrajzi szélességét, λ_1 és λ_2 pedig hosszúsági koordinátáit jelentik. Mindkét módszert C# rutinnal valósítottam meg, de az adatbázisban SQL függvény is rendelkezésemre állt.

Létrehoztam egy `"mutual_mention_all"` táblát a következőképpen. A `"user_mention"` tábla alapján leszámoltam, hogy az adott felhasználópárok adott irányba hányszor említik egymást (vagyis meghatároztam az m_{ab} és m_{ba} értékeket). Kiválogattam azokat, akik kölcsönösen említették egymást, vagyis mindkét irányú említés szerepelt az eredeti táblában. Fontos megjegyezni, hogy az önemlítéseket itt kiszűrtem, vagyis a kapott gráfban nem lesznek hurokélek. Ezzel megkaptam a kölcsönös említési gráf éllistáját.

Érdemes megjegyezni, hogy míg az irányított említési gráfban 97 174 876 felhasználó között húzódott 959 532 601 él, addig a kölcsönösséget megkövetelve már csak 29 003 802 felhasználó marad, 82 626 211 éllel, vagyis a felhasználók harmada, az éleknek pedig kevesebb, mint tizede (pontosabban mivel az oda-vissza élek 2-nek számítanak, kb. 17%-a) vesz részt a kölcsönös gráfban is. Ez mutatja, hogy a kölcsönösség megkövetelése egy erős megszorítás, vagyis tényleg szorosabb kapcsolatot jelent az egyirányú említésnél. Említeni engedély nélkül lehet bármelyik felhasználót, ám a kölcsönös említés már egyfajta visszaigazolást is jelent.

Ezután vettem a kapott gráfból azokat, akikhez a `"user_location"` táblában tartozott koordináta a fent leírt módon, így tovább szűkítve a gráfot $N_V = 2\,645\,369$ felhasználóra és $N_E = 3\,552\,931$ élre. Ez egy nagyon ritka gráf, a lehetséges éleinek $\frac{2 \cdot N_E}{N_V(N_V-1)} = 1,02 \cdot 10^{-6}$ része van csak behúzva. Az így kapott gráf éllistáját (az él két végén lévő felhasználó `user_id`-jei) tartalmazza a `mention_mutual_all` tábla, az oda és vissza említések számával (m_{ab}, m_{ba}), és mindkét felhasználó koordinátáival.

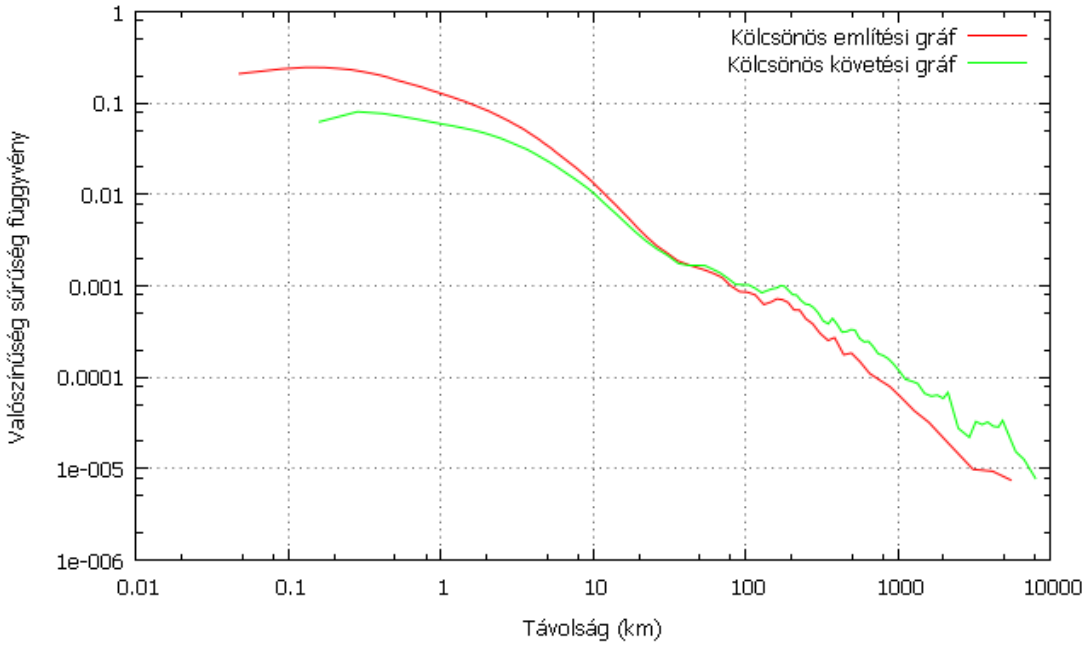
Innentől egyszerű SQL lekérdezésekkel elérhető az egyes pontok fokszáma (vagyis az egyes felhasználók említési barátainak száma), valamint szűréssel változtathatjuk s küszöbértéket. Kiszámoltam a fokszámeloszlást, amire azt kaptam, hogy nem követ pontosan hatványfüggvényt: szakaszonként hatvánnyal közelítve a kitevő folytonosan változik kb. -1,62-ről -3,37-ig. Mivel 2 alatti kitevőre a hatvány farkú eloszlásoknak nincs várható értéke, 3 alattira pedig szórása[32], durván mondhatjuk, hogy a kis fokszámú (kb. 5 alatti) felhasználók fokszámeloszlásának nincs elméleti várható értéke, és a kb. 10 alattiaknak pedig szórása.

Az adatbázis elérésére C# programot írtam, ami a sorba rendezett távolságadatokat lekérdezve kumulatív eloszlásfüggvényt (*Cumulative Distribution Function - CDF*) készít változó binmérettel. A binek méretét úgy állítottam be, hogy mindegyikbe az adatok $1/n$ része kerüljön. Ha nem jelzek mást, $n = 100$. A CDF-et ezután numerikusan deriválva megkapjuk a valószínűség sűrűség függvényt (*Probability Distribution Function - PDF*). Így a kapott függvény alakja sima ott is, ahol csak kevés adat áll rendelkezésünkre, és egy hisztogram nagyon zajos lenne. A kölcsönös említési gráf ($s = 1$ esetén) távolságeloszlása a kölcsönös követési gráféval összehasonlítva az 5.1. ábrán látható.

A kölcsönös követési gráfot a fentiekhez hasonló módon állítottam elő, egy párhuzamos kutatás részeredményeire támaszkodva.

Látszik, hogy néhányszor 10 km távolságnál a két ábra átmetszi egymást. Ez azt jelenti, hogy ennél a nagyságrendnél kisebb távolságokra az említési barátság, ennél nagyobbakra a követési barátság valószínűsége dominál. A néhányszor 10 km-es nagyságrend egy tipikus város nagyságrendje, úgyhogy az említési barátság lokalizáltabb, és inkább városon belül jellemző a követési barátsághoz viszonyítva.

A két eloszlás alakja viszonylag bonyolult, ám egymásra bizonyos szempontból hasonlítanak. Mindkettő közelíthető egyenesekkel bizonyos intervallumokon. Ez a logaritmikus skála miatt azt jelenti, hogy a felhasználók távolságeloszlása ezen a szakaszon jól közelíthető hatványfüggvénnyel. Ezeket az egyenes szakaszokat törések



5.1. ábra: A kölcsönös említési gráf ($s = 1$ esetén) távolságeloszlása a kölcsönös említési gráféval összehasonlítva.

választják el egymástól, pl. ahol a két görbe metszi egymást. Ahogy már említettem, ez a városok tipikus nagyságrendje (mivel az eloszlás alakjának kialakításában szerepet játszik minden város, ahonnan adatunk érkezett, nem mondhatunk egy konkrét méretet, a nagyságrend az, ami fontos). Ez azt jelenti, hogy egy városon belül más a kapcsolatok kialakulásának valószínűsége, mint városoknál nagyobb skálán. Továbbá mindkét görbén látszanak hasonló egyenetlenségek. Ezek a szárazföld, és a népsűrűség földrajzi adottságai miatt vannak: a legszembetűnőbb ilyen egyenetlenség egy gödör néhány ezer km-nél, ami főleg a követési gráf görbéjén látszik. Ezt az okozza, hogy a `Twitter` felhasználók döntő többsége Európában és az USA-ban van, így egy Európa, vagy USA nagyságrendű skálán sok kapcsolatot fogunk találni (kontinensen belüli, kb. 1 000-3 000 km), és az Európa-USA távolságon is (kontinens közi kapcsolatok, nagyon durván 8 000 km), azonban a kettő között kevesebbet (kb. 4 000-6 000 km), mert a legtöbb európai és egyesült államokbeli felhasználótól ilyen távolságra az óceán van.

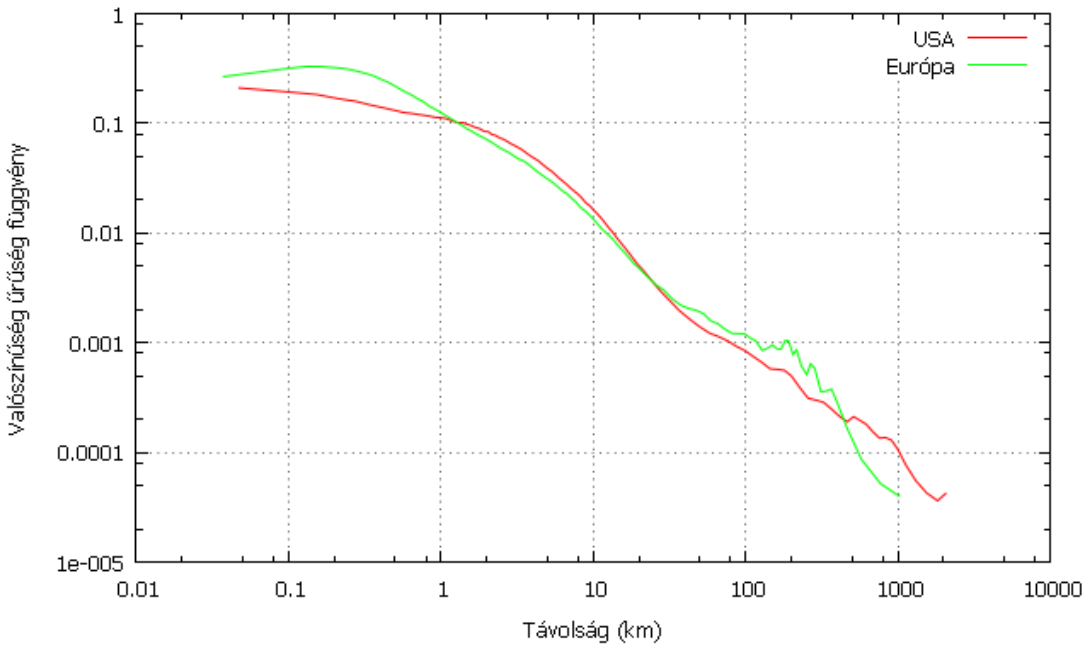
Az eloszlás logaritmikus adataira szakaszonként egyenest illesztettem. Három szakaszra tettem meg, de a legkisebb skála ($\sim$ 500 m alatti) olyan kicsi, hogy sok benne lévő pontra a koordináták, és a távolság számoló formula relatív hibája miatt

nem biztos, hogy bízhatunk benne. Továbbá, mivel a felhasználók helyzetét egy pontklaszter átlagaként határoztuk meg, a koordináták pontos értéke nem biztos, hogy egzaktul egy fizikai jelentőséggel bíró pontot jelent. A felhasználó lakóhelye, vagy bármilyen rá jellemző helyszín ettől kis mértékben eltérhet, ami pár szár méteres skálán már nagy relatív hibát jelenthet. Mindenesetre erre a távolságra a kitevő ellentettje $\gamma_1 = 0,534 \pm 0,0081$ lett.

A város nagyságrendű szakasz (pár km-től kb. 25 km-ig) kitevője $\gamma_2 = 1,44 \pm 0,024$. Itt egy kihagyás után a görbe megint egy egyenest látszik követni, ezek a nagy skálájú kapcsolatok, kb. 200 km fölött. Erre a kitevő az előzőhöz eléggé hasonló $\gamma_3 = 1,45 \pm 0,026$.

5.2.1. Távolságeloszlás az USA-ra és Európára

A 5.1. ábrán lévő görbét sok, különböző népsűrűségű és karakterisztikájú terület alakítja ki. Mivel, ahogy már említettem, `Twitter` felhasználók főleg Európában és az USA-ban vannak, érdemes ezekre a területekre leszűkítve is külön-külön megvizsgálni a kölcsönös említési gráf távolságeloszlását. A számításokat elvégeztem, a kapott görbéket a 5.2. ábra tartalmazza.



5.2. ábra: A kölcsönös említési gráf ($s = 1$ esetén) távolságeloszlása Európa és az Egyesült Államok területére szűkítve.

A konklúzió a képekből: mind Európában, mind az USA-ban, sőt az egész világon is a tárgyalt töréspontokon kívül az eloszlás hatványfüggvénnyel jól leírható. (Más területeken nem olyan elterjedt a Twitter, ezért nem vizsgáltam az adatokat máshova leszűkítve.) Mindegyik ábrán lehet látni egy kb. karakterisztikus városméret skálájú lecsengést, aztán onnantól egy másik kitevővel egy valószínűleg kontinens méretűt. Az eloszlás csak ezután kezd szórni és elromlani. Az USA-ban valamiért erősebb a lecsengés, mint Európában, és az egész világon, és az eloszlás nagyobb skálán marad hatványszerű.

Egyenest illesztettem mindkét görbe 3 szakaszára, a kapott meredekségeket, azaz az eloszlás adott szakaszának kitevőit a 5.1. táblázat tartalmazza. A táblázat második sora a városi, a harmadik a kontinens nagyságrendű léptékeket tartalmazza. Az első sor egy km-es nagyságrendű skála enyhe lecsengéssel, ami nem feltétlenül megbízható. Megjegyzem, hogy az európai eloszlás elején van egy emelkedő szakasz is, vagyis ez alapján bizonyos véges távolságú kapcsolatból van a legtöbb, a nála kisebbekből pedig kevesebb. Ez azonban már 100 m-es nagyságrend, ami a numerikus hibák, és egyéb effektusok miatt nem számottevő.

Tart. (km) (USA)	γ_{USA}	**Tart. (km) (Eu.)**	$\gamma_{\text{Eu.}}$
0,6 - 3,2	$0,822 \pm 0,0031$	0,03 - 1,8	$0,245 \pm 0,0084$
10 - 32	$1,43 \pm 0,024$	10 - 40	$1,67 \pm 0,015$
32 - 316	$0,8 \pm 0,43$	56 - 1000	$0,86 \pm 0,018$

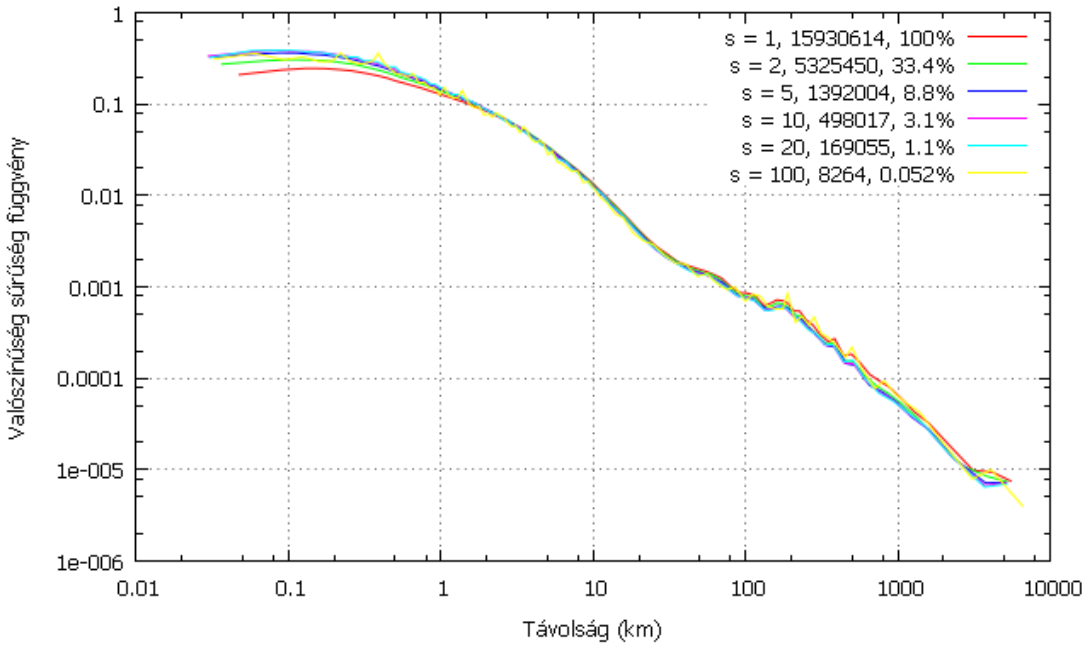
5.1. táblázat: Hatványkitevők a kölcsönös említési gráf távolságeloszlására Európára és az Egyesült Államokra szűkítve.

Összefoglalva: az USA-ban a városok átlagos mérete valószínűleg nagyobb, és a leghosszabb kapcsolatok is messzebb nyúlnak, mint Európában, azonban az eloszlás mindkét fontos skálán gyorsabban vág le, vagyis Európában az említési kapcsolatok messzebbre nyúlnak az USA-hoz képest, míg az Egyesült Államokban inkább közelebbi felhasználók létesítenek ilyen kapcsolatot Európához viszonyítva.

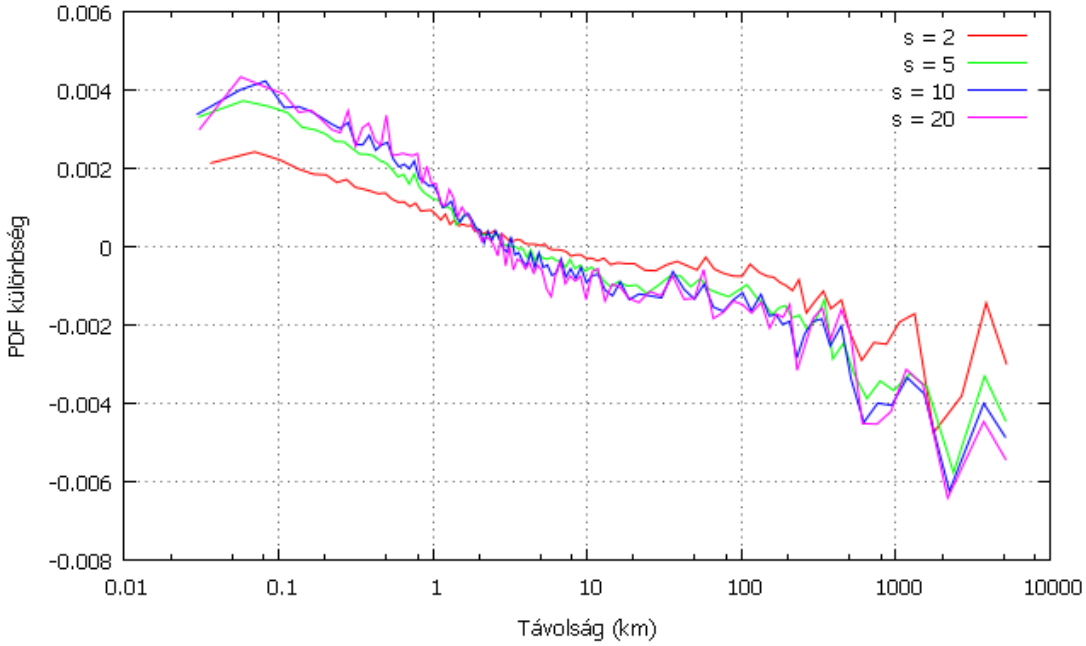
5.2.2. Távolságeloszlás alakjának s és r függése

A fejezet korábbi részében bevezettem a kölcsönös említési gráf éleinek jellemzésére két mennyiséget: az s-t, azaz az oda-vissza említések minimum számát, valamint r-t, az oda-vissza említések arányát (kisebbik/nagyobbik). Felmerülhet, hogy ezektől vajon számottevően függ-e a távolságeloszlás alakja. Elképzelhető például, hogy a hosszú beszélgetések (nagy s) főleg valódi, térben közeli ismerősök között zajlik, míg a rövidek távolabbi felhasználó párok között is. Ugyanilyen eltérés elképzelhető a kiegyensúlyozott ($r \approx 1$), és aránytalan (kis r) beszélgetések között is.

Először azt vizsgáltam, hogy s különböző értékeire hogyan viselkedik az eloszlás. Hosszabb beszélgetések, vagy rövidebbek lokalizáltabbak-e, netán nem jelent különbséget a változás? Egyéb szűrés nélkül különböző s küszöbértékekre kirajzoltam a távolságeloszlás alakját. Az jelmagyarázatban jelöltem az eloszlást kialakító kapcsolatok számát, ami erősen csökken s függvényében, valamint ezek százalékos arányát az összes kapcsolathoz képest. (A különböző s értékek eloszlásának konkrét alakját nem használom sehol, így az árbáját nem közlöm, csak megjegyzem, hogy egy hatványfüggvénynél gyorsabban, de egy exponenciális függvénynél lassabban csökken.)



5.3. ábra: A világ távolságeloszlásának alakja különböző küszöbértékeket megszabva az oda és vissza irányú említések számának minimum értékére. (s)



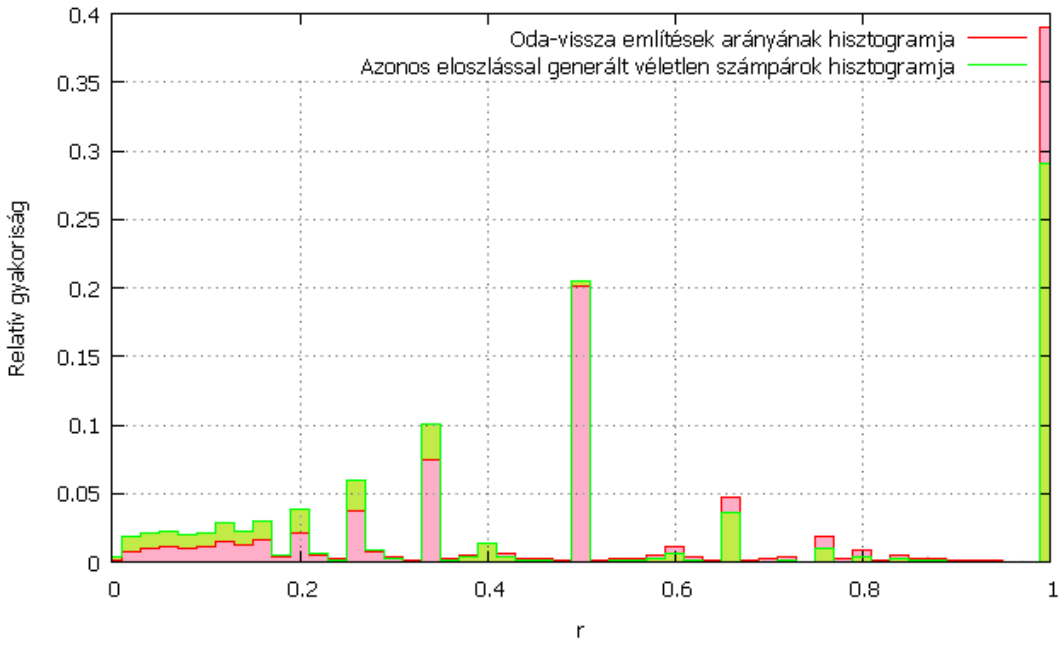
5.4. ábra: A valószínűség sűrűség módosulása az s küszöbérték változtatásával.

A 5.3. ábrán 6 különböző küszöbérték mellett ábrázoltam a távolságeloszlás alakját. A normálás miatt ezek is összehasonlíthatóak. Először csak kevés élt szűrünk ki, aztán a jelentős részüket. Azt látjuk, hogy a küszöbértéket növelve a kis távolságok valószínűsége növekszik, míg a nagyoké csökken, tehát a hosszabb beszélgetések a rövidekhez képest valószínűbben jönnek létre közelebbi felhasználók közt, de az eltérés nem drasztikus. Az $s = 5$-nél nagyobb küszöbértékekre már olyan kis mértékben változik az eloszlás, hogy nem is látható a grafikonon. A pontok számának csökkenésével a binértékek relatív szórása megnő, ezért a kb. $s = 100$, és afölötti értékekre már széttöredezik a görbe, a kevés távolságérték statisztikailag nem fog stabil halmazt alkotni. Ezek alapján azt mondhatjuk, hogy kb. $s = 5$-ös küszöbbel érdemes szűrni, ha ki akarjuk szűrni a véletlenül kialakuló pár oda-vissza említésből létrejött kapcsolatokat.

Néhány küszöbértékre ábrázoltam ugyanezen görbék, és az $s = 1$ görbe különbségét, hogy látszódjon, a küszöb változtatásával mennyivel módosul a távolságeloszlás. Ezek a görbék a 5.4. ábrán láthatóak (logaritmukus x, de lineáris y skálán). Az eredeti görbéket változó binmérettel generáltam, ennek két következménye volt. Az első, hogy a különböző függvények binjei nem ugyanoda estek, ezért a kivonáshoz lineárisan interpoláltam. A másik, hogy a binek nagy része 10 km alatt van, így a

különbségek ott a legjelentősebbek, ennek kiküszöbölésére a kapott különbségértékeket beszoroztam a binmérettel. (Vagyis a valószínűség sűrűség függvények binekre vett integráljainak különbségét vettem. Ezeknek összegének elméletileg 0-t kell adnia. Ezt ellenőriztem, és numerikus hibáktól eltekintve valóban 0 lett.) Ezen az ábrán is az látszik, hogy $s = 5$ fölött a görbék már kevéssé különböznek, és hogy néhány km-es skálán van a kritikus pont, aminél kisebb távolságokra s egyre szigorúbb megszorításaira növekszik a kapcsolatok valószínűsége, fölötte pedig csökken.

Az eloszlás r függésének vizsgálatára először is hisztogramot készítettem az r értékek előfordulási gyakoriságáról 0,02-es binmérettel, és 1-re normáltam a bin értékeket. A hisztogramot a 5.5. ábra tartalmazza.



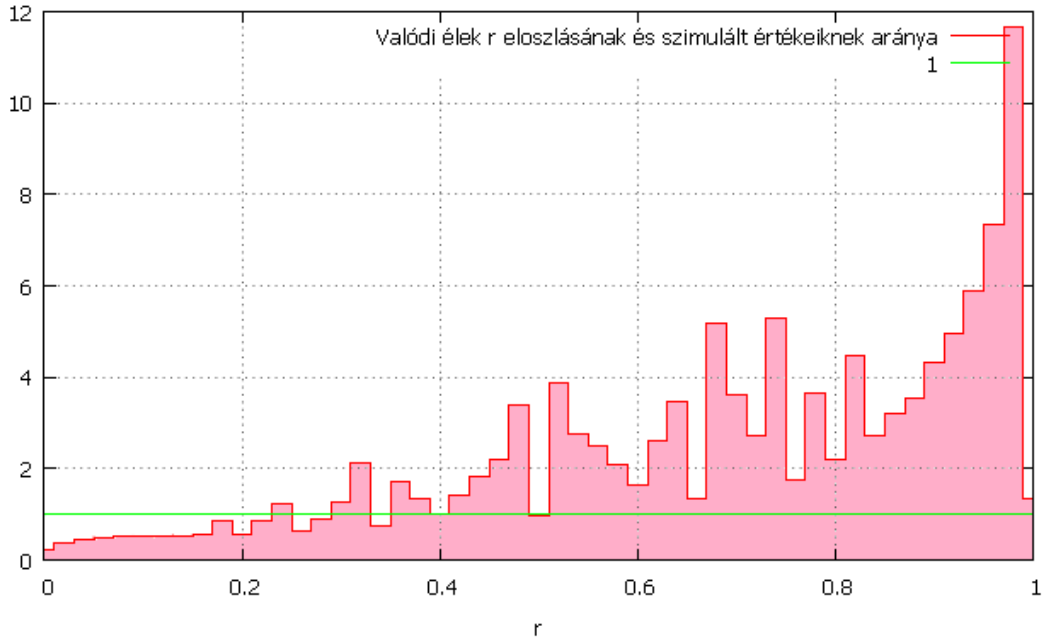
5.5. ábra: Oda-vissza irányú említések arányának eloszlása, és azonos eloszlással generált véletlenszerű számpárok arányának eloszlása.

Feltűnő, hogy van egy nagyon magas csúcs 1-nél, a második legmagasabb 0,5-nél, aztán kb. 1/3-nál, és 2/3-nál, és több csúcs is szimmetrikusan, szabályosan helyezkedik el. Vajon milyen effektus okozza, hogy rendkívül sok beszélgetés zajlik ilyen arányokkal?

Ha megvizsgáljuk az m értékek eloszlását (amibe m_{ab}, és m_{ba} is beletartozik), vagyis hogy az összes adott irányú említésszám hányszor fordul elő, azt látjuk, hogy sok más értékhez hasonlóan hatványszerűen levág. Összesen 1 238 féle különböző m_{ab}, vagy m_{ba} fordul elő. Habár 10 000-nél nagyobb szám is van a halmazban, az 1-szer történő említések az összesnek 49%-át teszik ki, és az összes m több, mint 90%-a 9 alatt van. Ez azt jelenti, hogy ha ezekből arányokat akarunk alkotni, akkor a legtöbb 1/1-es lesz, és az egyre nagyobb számok egyre kisebb valószínűségekkel fordulnak elő. Így az 1/1, 1/2, 1/3, 2/3, 3/4 arányok a leggyakoribbak között lesznek, ilyen sorrendben csökkenő valószínűséggel, és pl. a 25/36 aránypár igen ritkán fordul elő.

Feladatunk a 5.5. ábra hisztogramjából kiszűrni ezt az effektust. Ennek érdekében a fent említett m eloszlást létrehoztam, és a gyakoriságokkal súlyozott valószínűségekkel véletlenszerű számokat generáltam, amikből véletlenszerű párokat alkottam (szám szerint 60 000 000-t), és megalkottam a hisztogramjukat a fentiekkel megegyező bineléssel. Az összehasonlítás kedvéért ezt is lenormáltam. A kapott hisztogramot szintén tartalmazza a 5.5. ábra.

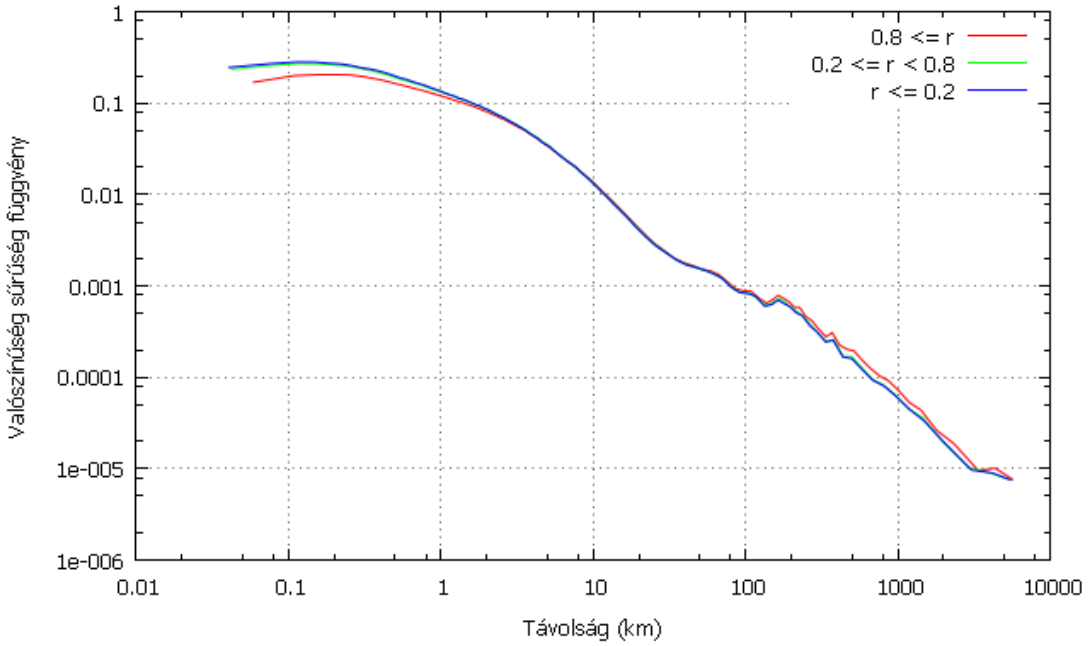
A valódi kapcsolatrendszerből fakadó effektust az jelzi, hogy az említési gráfból számolt értékek hogyan térnek el a véletlenszerű értékekkel szimulált adatoktól, így a kettőt elosztottam egymással minden binre, és az így kapott értékeket a 5.6. ábrán ábrázoltam. A konstans 1 értéket is behúztam, hogy látszódjon, mi lenne az elméletileg várható érték. Az ábrán az látszik, hogy 1-hez közeledve a beszélgetések részaránya egyre nő, vagyis minél aszimmetrikusabb a beszélgetés, annál kevésbé valószínű. (Habár az 1-hez tartozó bin értéke megint visszacsökken 1 körülire.) Az ábra tanúsága szerint egyáltalán nincs kiemelkedés az aszimmetrikus beszélgetéseknél.



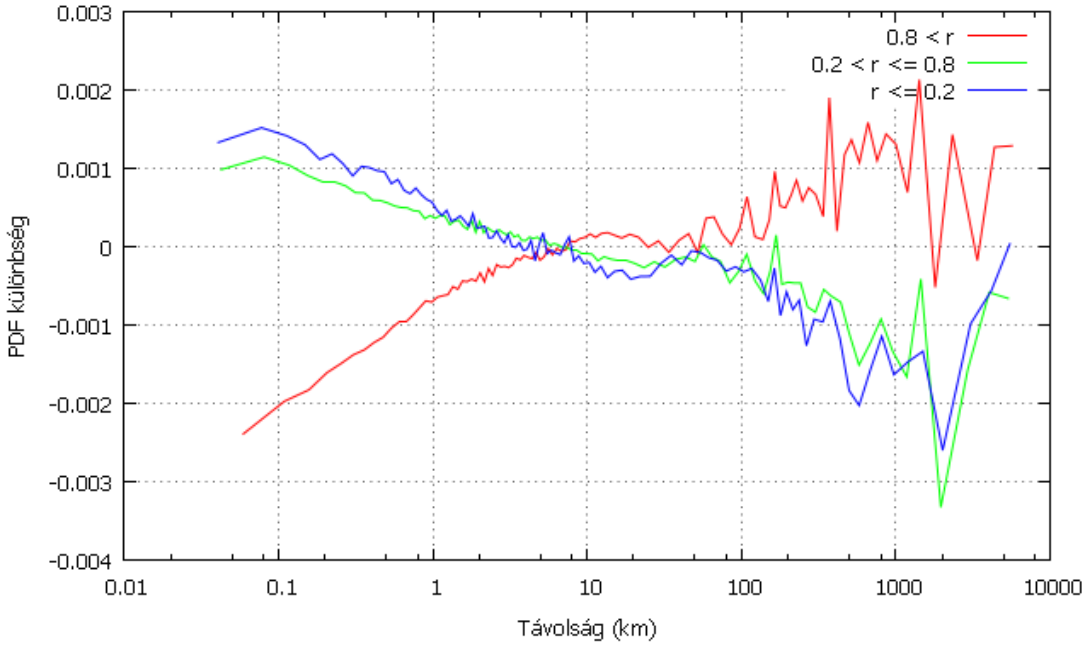
5.6. ábra: Oda-vissza irányú említések arányán azonos eloszlással generált véletlenszerű számpárok arányához viszonyatva.

Nem tudjuk r szerint egyértelműen csoportokra bontani a gráf éleit, de azért megvizsgálhatjuk, hogy a legszimmetrikusabb, és a legaszimmetrikusabb kapcsolatokra hogyan tér el a távolságeloszlás. 0,2-nél és 0,8-nál 3 tartományra osztottam a linkeket. Mindhárom tartományban sok élet találunk, a véletlenszerűen várhatóhoz képest azonban ezek a tartományok jól szétosztják a kapcsolatokat.

A 5.7. ábrán ábrázoltam a három tartományra szűkített távolságeloszlást. A két alsó tartomány alig különbözik, azonban a 0,8 fölötti arány, vagyis a legkiegyensúlyozottabb beszélgetések láthatóak kevésbé lokalizáltak, mint az aszimmetrikusabbak. Ez meglepő, hiszen a valódi beszélgetésektől várhatnánk, hogy 1-hez közeli értékeket produkálnak, és hogy valószínűbben jönnek létre közelebbi felhasználók között. Az eltérés azonban nem jelentős.



5.7. ábra: Említési gráf távolságeloszlása különböző oda-vissza említési arányokra.



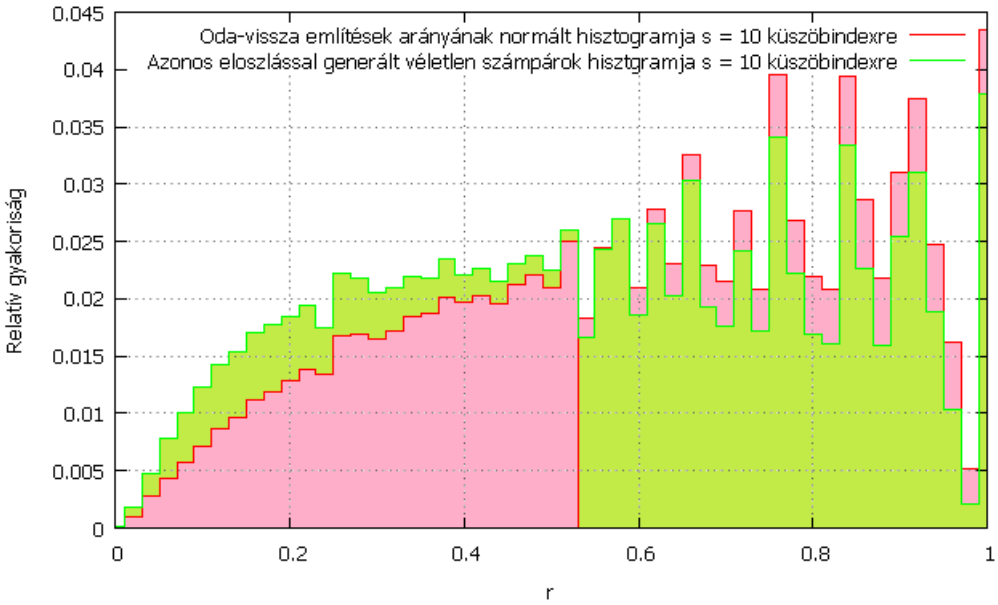
5.8. ábra: A valószínűség sűrűség módosulása az r arányra szűrésekkel.

A 5.4. ábrához hasonlóan a 5.8. ábrán ábrázoltam, hogy r arány bizonyos tartományokra szűkítésével a távolságeloszlás mennyit változik az eredetihez képest. A kritikus távolság itt is néhány km-es nagyságrendnél van, azonban meglepő módon a kiegyensúlyozott, szimmetrikus beszélgetésekre szűrve valószínűbben maradnak bent a távolabbi beszélgetések, vagyis a 0,8-as aránynál kisebb r-rel rendelkezők a lokalizáltabbak.

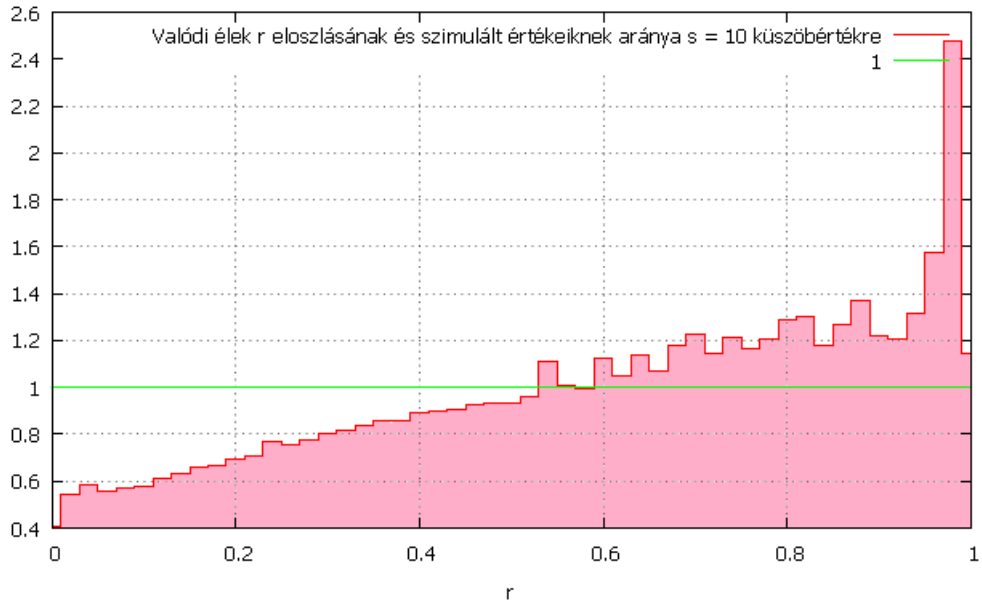
Megjegyzendő, hogy - ahogyan a korábbiakból kiviláglik - az 1-nél, 1/2-nél, stb. lévő nagy csúcsokat főleg a kis említésszámú élek hozzák létre. Mivel s növelésével pont a kis említésszámmal rendelkező éleket szűrjük ki, ettől várhatnánk a nagy csúcsok eltűnését. Azonban, ha m eloszlását hatványfüggvénynek tekintjük, s növelése csak az eloszlás elejének levágását jelenti, ami után a megmaradt legkisebb érték számossága ugyanúgy jelentősen nagyobb a többinél, vagyis az 1-nél lévő csúcs megmarad. Azonban az 1/2, 2/3, stb. helyek kicsúcsosodása helyett pl. $s = 10$-re a 10/11, 11/12, stb. csúcsok jelennek meg, tehát az 1 körüliek. $s = 10$-re is ábrázoltam az r-ek eloszlását a 5.5. ábrához hasonló módon a 5.9. ábrán, szintén a szimulált értékekkel együtt.

A 5.5. ábrához hasonló módon a mért és szimulált értékek hányadosát is kiszámoltam, amit a 5.10. ábra tartalmaz. Azon kívül, hogy kisebb az adatok szóródása, nagyjából ugyanaz mondható el róla, mint $s = 1$ esetén, azaz a szimmetrikus beszélgetések valószínűbbek, és nincs csúcs kis r-eknél, azaz nem húzhatunk egyértelmű határt a szimmetrikus és az aszimmetrikus beszélgetésekre akkor sem, ha csak hosszabb beszélgetéseket tekintünk (amivel kiszűrtük a beszélgetések igen jelentős hányadát).

További megjegyzendő tény, hogy az az s érték, ami fölött már nem változik jelentősen a távolságeloszlás alakja függ a mintavételezési időtől, és aránytól. Az azonban látszik, hogy s konkrét értékétől független az, hogy hol van az a kritikus pont, ami elválasztja az s szigorításával növekvő, és csökkenő valószínűségű távolságokat. Az esetleges felhasználások során s értékét a problémához illesztetten kell meghatározni, ám a jelen adathalmazból jövő eredmények azt mutatják, hogy a kérdéses érték meglepően kicsi.

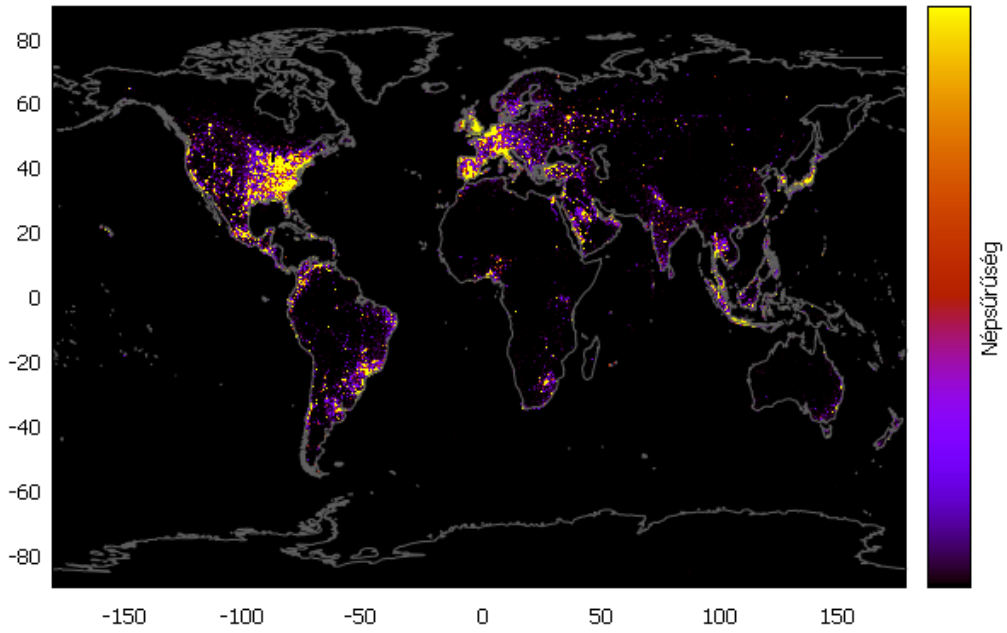


5.9. ábra: r arányok relatív gyakorisága $s = 10$-es küszöbérték mellett.



5.10. ábra: Oda-vissza irányú említések aránya azonos eloszlással generált véletlenszerű számpárok arányához viszonyatva $s = 10$ esetén.

5.3. Népsűrűség fraktáldimenziójának kimérése



5.11. ábra: `Twitter` felhasználók népsűrűsége. Koordinátáik a *friends-of-friends* algoritmussal lettek meghatározva[30] tweetjeik koordinátáinak klaszterezéséből a legnagyobb számosságú klaszer átlagaként.

Az egyes gráfok távolságeloszlásának alakját több effektus alakítja ki.

Egyrészt egy gömbfelületen gömbi egyenletes eloszlással szétszórt pontok távolságeloszlásának van egy meghatározott alakja. A gömbi egyenletes eloszlás azt jelenti, hogy egy adott felületen a pontok számának várható értéke arányos a felület nagyságával. Az így elosztott pontok egymástól való távolságának valószínűség sűrűség függvénye, vagyis távolságeloszlása könnyen számolható analitikusan, és Monte-Carlo-szimulációval is.

Ezt az alakot megváltoztatja az a tény, hogy a pontok (felhasználók koordinátái) a Föld felszínén nem egyenletesen oszlanak el. (Pl. kontinens nagyságrendű és kontinens közi nagyságrendű távolságokon sok kapcsolat van, de a kettő között (kontinensről óceánig vezető távolság) jóval kevesebb.) Nyilván többnyire csak a szárazföldön fogunk felhasználókat találni, és a szárazföldön is valamilyen változó sűrűségben helyezkednek el - a városokban többen, közöttük kevesebben. Egyes országokon belül, ahol a `Twitter` használati hajlandóság nem változik jelentősen területileg, feltételezhetően ez a sűrűség (a `Twitter` felhasználók népsűrűsége) arányos a nép-

sűrűséggel. Ezeknek a pontoknak a távolságeloszlása "ráül" a gömbi egyenletes eloszlású pontok távolságeloszlására. Ha ismerjük a felhasználók helyeit, véletlenszerűen kiválasztott párok összekötésével alkotott gráf távolságeloszlásával kis számításigénnyel becsülhetjük ezen pontok távolságeloszlását.

A harmadik effektus pedig abból fakad, hogy az egyes gráfokat vizsgálva előző bekezdésben említett sűrűséggel eloszló pontok nem véletlenszerűen vannak összekötve, hanem valamilyen bonyolult szociális effektusok által kialakítva.

Az eddig számolt távolságeloszlásokban mindhárom effektus szerepet játszott, azonban mivel külön-külön tudjuk számolni ezeket az eloszlásokat, az effektusokat szeparálhatjuk.

Ha egy gömbön egyenletesen vannak elosztva pontok, akkor egy adott ponttól bizonyos s távolságra eső pontok számának arányát jelöljük $f(s)ds$-sel! Az általánosság megszorítása nélkül vehetjük a vizsgált pontot az északi sarknak. E sarkba, és egy másik vizsgált P pontba mutató egyenesek által bezárt szöget (polárszöget) jelöljük ϑ-val! Belátható, hogy $f(s)ds$ arányos a P-n átmenő szélességi kör szélű, ds szélességű sáv területével. Ez a terület $R \cdot \sin(\vartheta) \cdot ds$, vagyis $f(s) \sim R \cdot \sin(\vartheta) = R \cdot \sin(s/R)$. $f(s)ds$-nek a $[0, \pi R]$ intervallumra integrálva 1-et kell adnia:

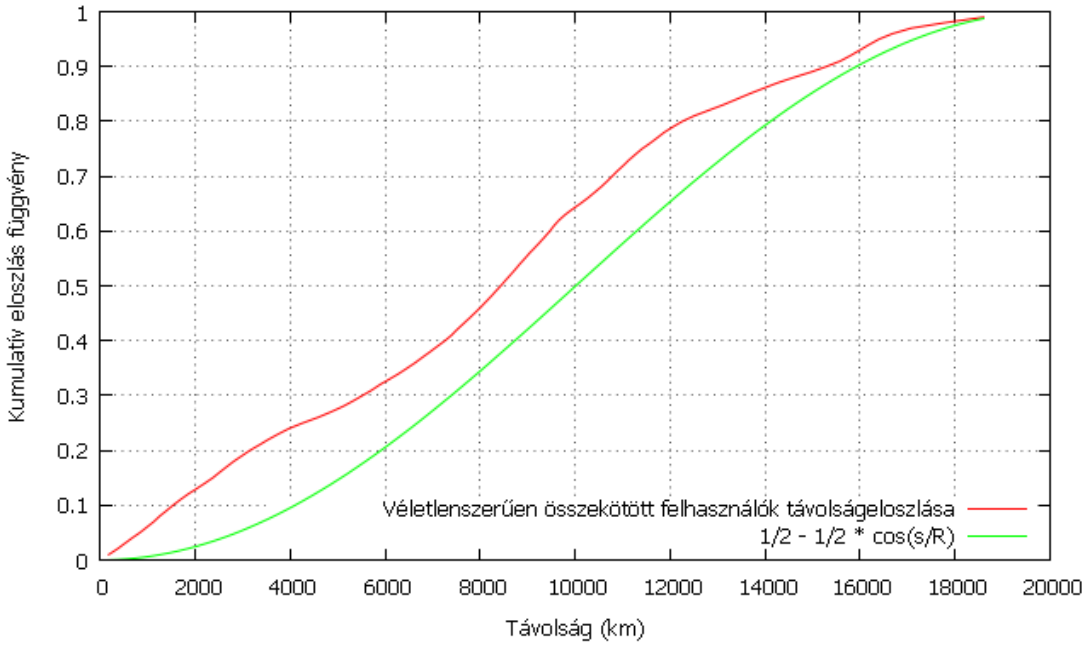
$$1 = \int_0^{\pi R} f(s)ds = CR \int_0^{\pi R} \sin(s/R)ds = -CR[\cos(s/R)]_0^{\pi R} = 2CR^2 \Rightarrow C = 1/2R^2$$

$$f(s) = \frac{\sin(s/R)}{2R} \tag{5.2}$$

Eloszlása pedig:

$$F(s) = \int_0^s f(s')ds' = \left[-\frac{1}{2}\cos(s'/R)\right]_0^s = 1/2 - 1/2 \cdot \cos(s/R) \tag{5.3}$$

A felhasználók valódi eloszlásának számolásához a felhasználók geo adatait tartalmazó táblából beolvastam az összes felhasználót, amik közül teljesen véletlenszerűen párokat választottam ki, szám szerint 10 milliót. Ezekhez az eddigiekkel azonos módszerrel kiszámoltam a távolságeloszlást. Ennek eredménye az egyenletes eloszlás távolságeloszlásával a 5.12. ábrán látható.



5.12. ábra: Véletlenszerűen összekötött felhasználók távolságeloszlása, és gömbi egyenletes eloszlású pontok elméletileg számolt távolságeloszlása. A kettő közötti különbség abból fakad, hogy a `Twitter` felhasználók a Föld felszínén nem egyenletesen vannak elosztva. Az egyik legfontosabb különbség, hogy az egyenletes eloszlás négyzetesen, a valódi felhasználók eloszlása viszont majdnem lineárisan indul.

A gömbi eloszlástól való eltérések érthetőek a földrajzból: 5 000 km-nél a sűrűségfüggvényben van egy nagy minimum, (ami a 5.12. ábrán egy laposabb szakaszként jelenik meg) ami valószínűleg a kontinens-óceán távolságnak felel meg. Az, hogy a mért görbe minden helyen nagyobb az egyenletes gömbi eloszláshoz tartozónál, azt mutatja, hogy a felhasználók jóval lokalizáltabban helyezkednek el, mintha egyenletesen lennének szétosztva a Föld felszínén.

A káoszelméletben gyakran használt dimenzióméréték a korrelációs dimenzió, amit v-vel szokás jelölni. Ez egy általánosított dimenziófogalom, ami egy térben elosztott pontrendszer dimenzióját méri, és visszaadja az egyszerű, egész dimenziós esetekben várható dimenziószámot, de felvehet tört értékeket is fraktálok esetén.

Vezessük be a

$$C(\varepsilon) = \lim_{N\to\infty} \frac{g}{N^2} \tag{5.4}$$

korrelációs integrált, ahol g azon pontpárok száma, amik távolsága ε-nál kisebb.

Formálisan g-t a

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{N} \Theta(\varepsilon - dist(gp_i, gp_j))$$

alakban írhatjuk föl, ahol $dist(gp_i, gp_j)$ a gp_i és gp_j pontok távolsága a Föld felszínén, $\Theta(x)$ pedig a Heaviside-féle egységugrás függvény. N a pontok száma, vagyis $C(\varepsilon)$ az ε-nál kisebb távolságú és a lehetséges összes pár számának hányadosa, tehát egy normált kumulatív eloszlás. (Ebben az $N \to \infty$ határérték miatt N^2 mellett N-t elhanyagoltuk a pontpárok számában.) A határértéket véve az összegből integrál lesz, innen a korrelációs integrál elnevezés. A fent számolt $F(s)$ ennek a becslése a linkek csak egy adott részének vizsgálatával, de mivel a mintavételezés véletlenszerű volt, $F(s)$-t $C(\varepsilon)$ torzítatlan becslésének tekinthetjük, ezek után pedig a kettőt azonosítani fogom egymással.

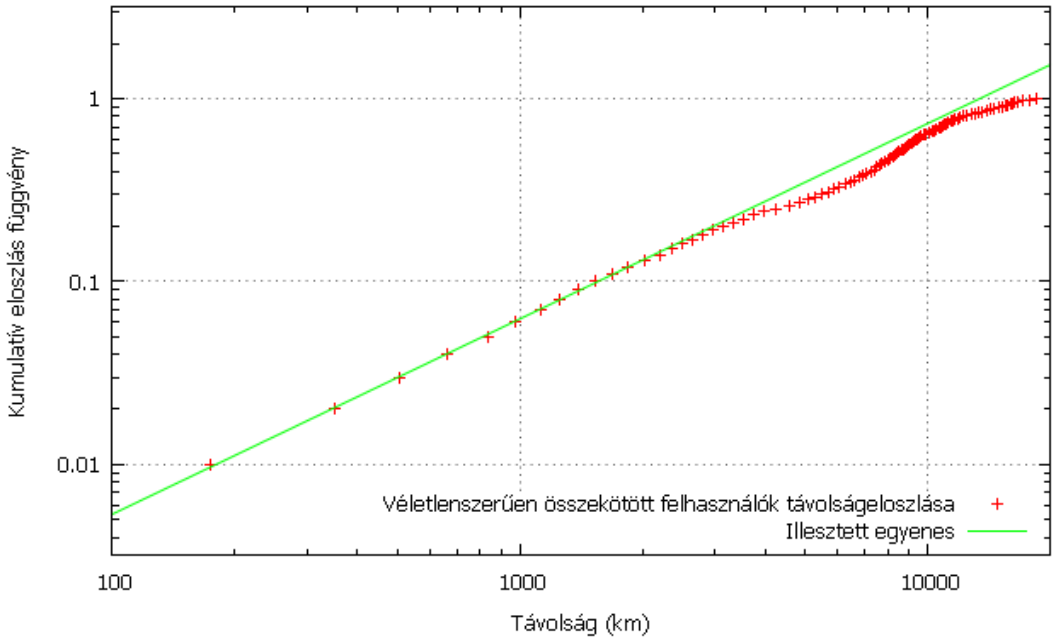
Ahogy a pontok száma növekszik, és a köztük lévő távolság csökken, igaz lesz a következő összefüggés kis ε-okra:[35]

$$C(\varepsilon) \sim \varepsilon^v$$

Vagyis, ha $F(s)$-t kis s-ekre hatványfüggvénnyel közelítjük, a kezdeti szakasz log-log diagramjára illesztett egyenes meredeksége lesz a népsűrűség korrelációs dimenziója.

Ezt megtettem a fent számolt $F(s)$-sel. Ábrázolva a 5.12. ábra elejét log-log skálán, egyenest illeszthetünk a kezdeti szakaszra. Ezt kb. 3,26-ig tettem meg, ami kb. 1 800 km-nek felel meg. Ekkora skálán igaz a hatványszerű viselkedés, azaz a kihozott dimenzió a népsűrűségre. Az egyenes meredeksége: $v = 1,067 \pm 0,0095$. Az ábrát az illesztett egyenessel a 5.13. ábra tartalmazza.

A pontosság kedvéért ugyanezt a mérést elvégeztem kis területekre lebontva is. Ez azt jelenti, hogy a fent leírt számításokat újra elvégeztem, de úgy, hogy egy távolságértéket csak akkor vettem bele a listába, ha az legfeljebb 100 km volt. A normálás végett ilyenkor leszámoltam az összes vizsgált pontpárt, és azok arányát, amik 100 km alatt voltak. Ezzel $F(s)$ elejének pontosabb verzióját kaptam meg. Még egy közelítés volt a 0 távolságokhoz közeli értékek numerikus hibájának kiszűrésére: itt a felületet síkkal közelítettem, és a távolságnak a két vektor különbségének abszolút értékét vettem. Itt már csak 1 000 000 kapcsolatot vettem fel. Ennek az

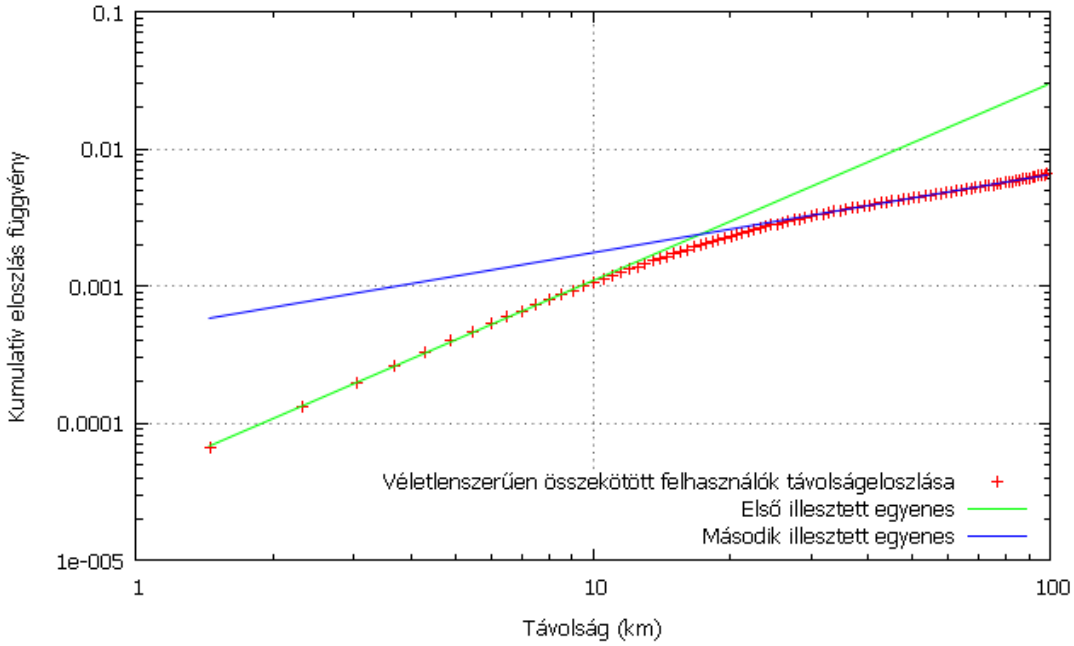


5.13. ábra: Véletlenszerűen összekötött felhasználók távolságeloszlásának eleje logaritmikus skálán ábrázolva. A kezdeti szakaszra illesztett egyenes meredeksége adja a pontok (felhasználók) korrelációs dimenzióját.

eloszlásnak log-log skáláját a 5.14. ábra tartalmazza. Erre az ábrára két egyenest is illesztettem: kb. 10 km-ig egy $v_1 = 1,437 \pm 0,0085$, aztán kb. 30 és 100 km között egy $v_2 = 0,570 \pm 0,0024$ dimenzió jellemző. Nem meglepő, hogy más távolságokon más értékeket kapunk, hiszen már nem új tény, hogy a lakosság multifraktál szerűen oszlik el. A kis skála valószínűleg a városon belüli eloszlás dimenziója, a nagyobb pedig maguknak a városok eloszlásának a fraktáldimenziója.

A korrelációs dimenzió a Rényi-féle általánosított dimenziók egy speciális esete. Ezek a D_q dimenziók a q paraméter értékének függvényében változnak, a v korrelációs dimenzió pedig a D_2-vel egyezik meg.

Egy 1996-os cikkben Appleby kimérte[36] külön az USA és Nagy-Britannia népsűrűségének általánosított dimenzióit másféle módszerrel, többféle q értékre, régebbi népszámlálási adatokra támaszkodva. Sorrendben $v_{\mathrm{USA}}^{\mathrm{cikk}} = 1,26 \pm 0,05$ és $v_{\mathrm{GB}}^{\mathrm{cikk}} = 1,29 \pm 0,01$ értékeket kapta. Én kiszámoltam az USA-ra és Nagy-Britanniára ugyanezt a fent említett módszerrel (500 km-ig 5 000 000 pontra, sűrűbb 1 000 binre osztva), de a részeredményeket és ábrákat már nem közlöm. Az eredményeim: USA: $v_{\mathrm{USA}} = 1,238 \pm 0,0022$, ami kb. 20 km-es skálán illeszkedett nagyon jól, és $v_{\mathrm{GB}} = 1,38 \pm 0,018$.



5.14. ábra: Véletlenszerűen összekötött felhasználók távolságeloszlásának eleje logaritmikus skálán ábrázolva, a távolságot közelítő formulával számolva. Ezt az ábrát jóval több pontra számoltam, mint a 5.13. ábrát. A kezdeti szakaszra illesztett egyenes meredeksége adja a pontok (felhasználók) korrelációs dimenzióját.

Az USA-ra az eltérés hibán belül van, Nagy-Britanniára viszont meglepően magas értéket kaptam, és meglepően kis (kisebb, mint 20 km) skáláig. Ez azért lehetett, mert egyszerűen egy téglalapra szűrtem, és Hollandia belelógott a sarokba, ahol igen közkedvelt a `Twitter`. Lehetőség van a felhasználók pontosabb régió szerinti szűrésére, azonban ez nagy számításigénye miatt nehezen kivitelezhető feladat. Fontos megjegyezni, hogy ezek az értékek csak néhány 10 km-es skálán érvényesek, vagyis városokon belül. Az eloszlások bizonyos törések után is hatványfüggvényt követnek, de lényegesen kisebb kitevővel. (Az USA-ban 0,8, Nagy-Britanniában 1 körülivel.) Ennek az lehet a magyarázata, hogy ennél nagyobb skálán a népsűrűség jóval ritkább, és kiterjedt területek helyett utak, folyók mentén rendeződik.

A kölcsönös említési gráf távolságeloszlásának csökkenését tehát részben magyarázza ez az effektus (szemléletesen, hogy távolabb már más azoknak a száma, akivel kapcsolatot lehet létesíteni). Ha a kumulatív eloszlásfüggvény hatványszerű, és kitevője v, akkor a deriváltja, azaz a valószínűség sűrűség függvény is hatványszerű, $v-1$ kitevővel. A 5.2. fejezetben a távolságeloszlásra illesztett függvényekből számolt γ értékek az eloszlás kitevőjének ellentettjét jelentették, így a távolságeloszlás ki-

tevőjének egy $-\delta = v - 1 + \gamma$ részét nem magyarázza a népsűrűség távolságeloszlása, azaz a távolságeloszlás $p(s) \sim s^{-\gamma}$ függvény egy s^{v-1} és egy $s^{-\delta}$ függvények szorzataként áll elő, mint a népsűrűségből, és a szociális jellemzőkből jövő effektusok összeadódása. Ez a kitevő városon belül durván $\delta_1 \approx 1,9$ körül van, és $\delta_2 \approx 1$ körül nagyobb skálán.

Mindezekből az látszik, hogy 1) a `Twitter` felhasználók eloszlása valóban arányos a népsűrűséggel azokban az országokban, ahol népszerű a szolgáltatás, 2) könnyen használható módszerrel, mindenki számára elérhető adatokból számolható a népsűrűség fraktáldimenziója, illetve maga a népsűrűség, és így egyéb tulajdonságai is, és 3) ez részletesebben megtehető, mint korábban a népszámlálási adatokra támaszkodva, ugyanis ezzel a módszerrel szétválik a városon belüli, valamint a város szintű struktúra, míg a népszámlálás adatokkal csak város szintű részletességgel vizsgálható a probléma.

5.4. A kölcsönös követési és kölcsönös említési gráf összehasonlítása

Rendelkezésünkre áll a kölcsönös követési, illetve a kölcsönös említési gráf, azok néhány jellemzőjével együtt. Megvizsgálhatjuk néhány szempontból, hogy ezek miben hasonlítanak, és miben térnek el egymástól.

5.4.1. Egyszerű kvantitatív összehasonítási mértékek, kis kvalitatív elemzés

A következőkben a kölcsönös említési, és kölcsönös követési gráfot fogom összehasonlítani. Mindkét gráf élei (E), valamint pontjai (V) halmazokat alkotnak. Ezeknek képezhetjük metszetét, unióját és szimmetrikus differenciáját (unió $\setminus$ metszet). Ezek segítenek majd az összehasonlításban. Metszet, unió, stb. alatt ezeket fogom érteni.

Leszámolhatjuk a felhasználók, valamint a linkek számát mindkét gráfban, és a metszetekben is.

Jelöljük a felhasználók számát $|V|$-vel, az élek számát $|E|$-vel. Az 1 indexelje az említési gráfot, a 2 a követési gráfot, az $\cap$ pedig a metszet gráfot! Ekkor a számosságok:

$$|E_1| = 15\ 930\ 614 \qquad |E_2| = 48\ 707\ 444 \qquad |E_\cap| = 2\ 665\ 971$$
$$|V_1| = 5\ 900\ 120 \qquad |V_2| = 3\ 333\ 010 \qquad |V_\cap| = 2\ 268\ 478$$

Az irodalomban használt néhány gráfhasonlósági mérőszám (*Vertex-Edge Overlap - VEO*[33], *Graph Edit Distance - GED*[34]) által inspirálva vizsgáljunk meg pár általunk bevezetett mérőszámot is! A metszetet indexeljük a $\cap$ jellel, az uniót az $\cup$ jellel, az eltérő elemeket pedig a $\triangle$ jellel! ($|\cap| + |\triangle| = |\cup|$.) Az 1 most is az említési, a 2 a követési gráfot jelenti. A jelölésrendszerben az E az élekre, a V a pontokra vonatkoztatott mérőszámokat jelenti, indexben az elvégzett művelet van.

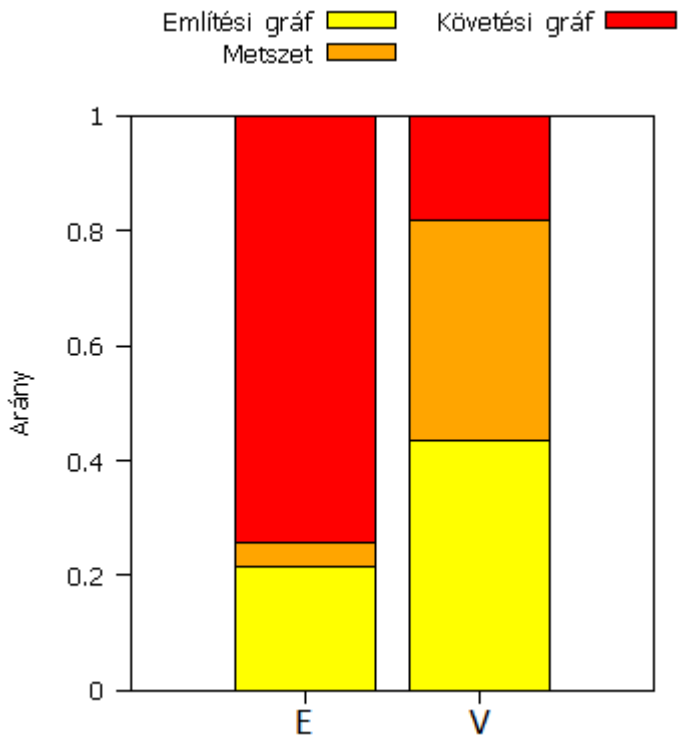
Az eltérő, valamint megegyező elemek aránya az összesre vonatkoztatva:

$$E_{|\triangle|/|\cup|} = 0,9570 \qquad E_{|\cap|/|\cup|} = 0,0430$$
$$V_{|\triangle|/|\cup|} = 0,6743 \qquad V_{|\cap|/|\cup|} = 0,3257$$

Ebből az látszik, hogy az összes él kb. 4%-a van jelen mindkét gráfban, viszont a pontokra ez már több, mint 32%.

$$E_{|2|/|1|} = 3,0575$$
$$E_{|\cap|/|1|} = 0,1673$$
$$E_{|\cap|/|2|} = 0,0547$$

$$V_{|2|/|1|} = 0,5649$$
$$V_{|\cap|/|1|} = 0,3845$$
$$V_{|\cap|/|2|} = 0,6806$$



5.15. ábra: A kölcsönös említési, és kölcsönös követési gráfok különálló, és átfedő pontjainak (felhasználóinak), és éleinek aránya. A felhasználók között nagy az átfedés, de az élek igen nagy részben különböznek, vagyis többnyire ugyanazok a felhasználók alakítanak ki különböző jellegű követési, és említési kapcsolatokat.

Összefoglalva: a követési gráf egy kb. fele olyan népes gráf, mint az említési, de sűrűbb. A sűrűség -, azaz a létező élek és a lehetséges összes él számának hányadosa - az említési gráfra $\frac{2\cdot|E_1|}{|V_1|(|V_1|-1)} = 9,1525 \cdot 10^{-7}$, a követési gráfra pedig $\frac{2\cdot|E_2|}{|V_2|(|V_2|-1)} = 8,7690 \cdot 10^{-6}$, tehát egy lehetséges oda-vissza követés pár kb. tízszer akkora valószínűséggel jön létre, mint egy ugyanilyen kölcsönös említési kapcsolat.

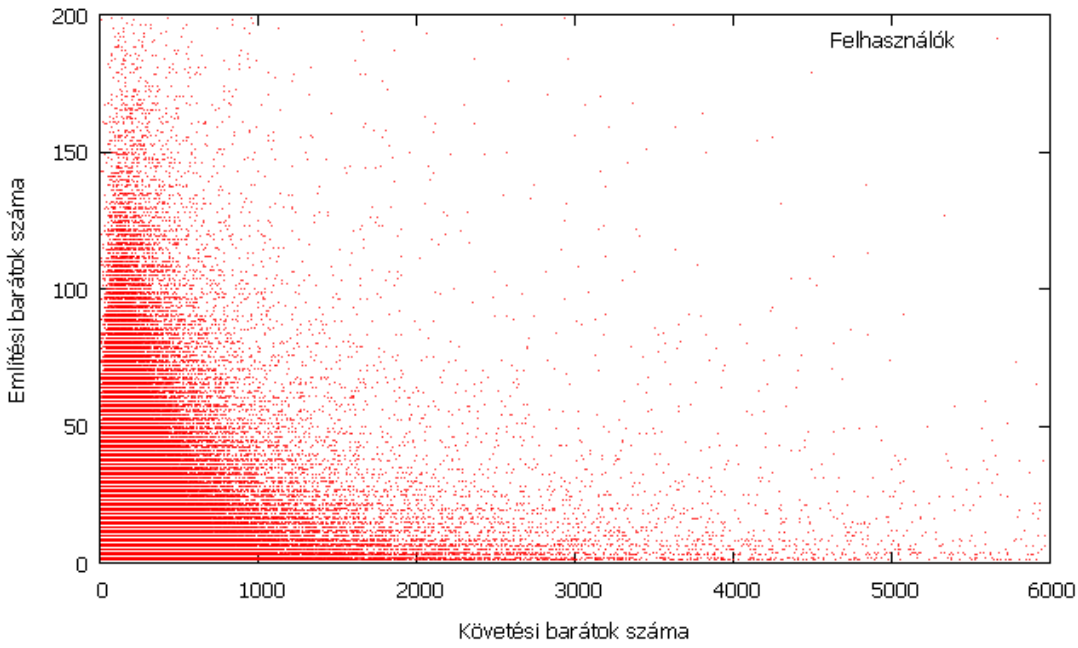
A sűrűbb követési gráf pontjainak több, mint 2/3-a benne van a metszet halmazban, vagyis pongyolán mondhatjuk, hogy ezek a pontok nagyjából a részhalmazát alkotják az említési gráf pontjainak. Ez azt jelenti, hogy a két gráfot (legalábbis az általunk "látott" részeiket) igen nagy arányban ugyanazok a felhasználók alkotják, azonban ugyanezen emberek a két kapcsolatból egy merőben eltérő rendszert építenek ki.

Megjegyzendő, hogy az adatgyűjtésben nagy prioritású cél volt a követési gráfban résztvevő felhasználók felderítése, így a mért arány kicsit eltolódhat a valóstól, de ez a gráfok sűrűségéről, és a kapcsolatrendszerek különbözőségéről levont konklúziót ez nem befolyásolja.

5.4.2. Metszet részletesebb elemzése

Szűkítsük le mindkét gráfot a metszetpontokra.

Egy lehetséges mód az összehasonlításra a fokszámok korrelációjának vizsgálata. Ehhez ábrázoltam a metszetből minden felhasználót egy pontként úgy, hogy a két fokszám a két koordinátája, így kapva egy szórási diagramot, amit a 5.16. ábrán láthatunk.

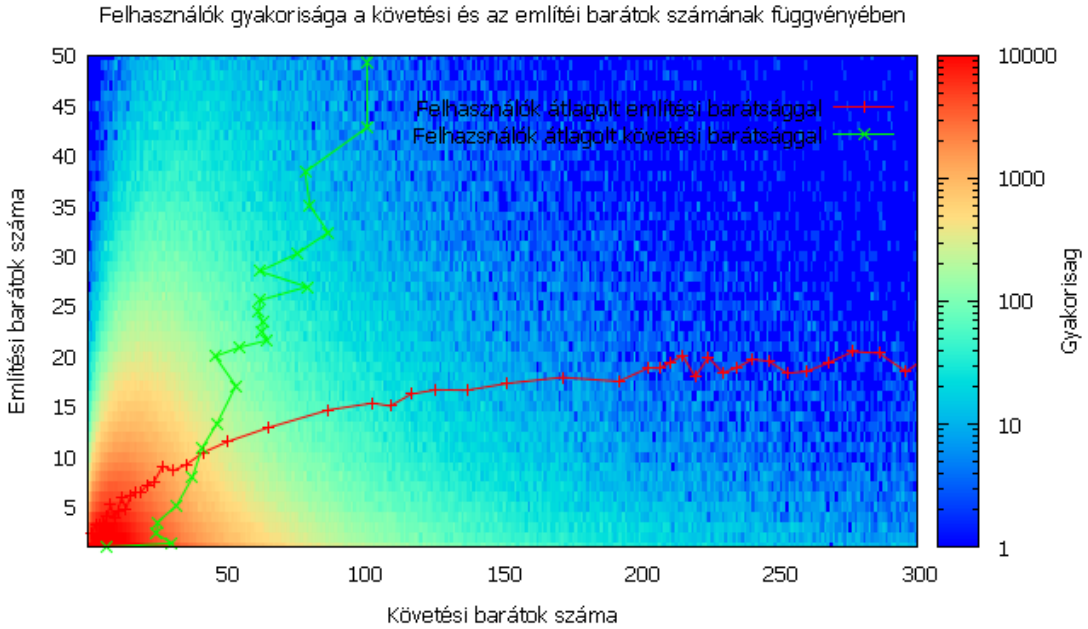


5.16. ábra: A kölcsönös említési gráf és a kölcsönös követési gráf metszetének felhasználói a két gráf fokszámai által kifeszített térben. A nagy fokszámokra jellemző felhasználóknál inverzió látszik a korreláció helyett.

A kis számoknál még nem látszik az adott koordinátájú pontok multiplicitása, pedig az információt rejthet. Ami viszont elsőre szembeszökő, az az, hogy azokhoz a felhasználókhoz, akikhez sok követési barát tartozik, többnyire kevés az említési barát, és fordítva, vagyis az intuitíve várható kb. lineáris kapcsolat helyett nagy számokra egy inverzió figyelhető meg.

Mindkét gráf fokszámeloszlásából várható, hogy a legtöbb felhasználó az origó környékén tömörül, csak ez a 5.16. ábrán nem látszik. A 5.17. ábrán színekkel ábrázoltam az egyes pontokhoz tartozó felhasználók számát, az origó körüli tartományt kiemelve. A színskála a gyors lecsengés miatt logaritmikus.

A szemléletesség kedvéért bizonyos binekbe osztva átlagoltam a binekbe kerülő értékeket a követési, és említési barátok szerint is. A binek méretét úgy állítottam be, hogy egy intervallumon egy binbe adott számú felhasználó kerüljön, de az ábra különböző intervallumain ez a szám különböző. A binméreteket úgy választottam, hogy elég nagy legyen ahhoz, hogy a függvények menete sima legyen, de elég kicsi ahhoz, hogy kellően sűrűn tartalmazzon információt a függvény menetéről. Mindkét görbét -, azaz gyakorlatilag az említési barátok számának átlagát a követési barátok számának függvényében, valamint a követési barátok számának átlagát az említési barátok számának függvényében - tartalmazza a 5.17. ábra.



5.17. ábra: A kölcsönös említési gráf és a kölcsönös követési gráf metszetének felhasználói a két gráf fokszámai által kifeszített térben. A színek az adott ponthoz tartozó felhasználók számának logaritmusát jelzik. Egy nagyon erősen lecsengő csúcs van az origó közelében, továbbá a szintvonalak mintha Poisson görbéket követnének. Az ábra piros vonallal tartalmazza a változó méretű binekre említési, és zöld vonallal követési barátok számát is a másik barátok számának függvényében.

A 5.17. ábrán is látszik, hogy a két mennyiség között nem egyértelműen lineáris a kapcsolat, de az egyszerű inverziónál is többet mondhatunk. Első ránézésre úgy látszik, az adott számú felhasználóval rendelkező pontok (vagyis az adott valószínűségű pontok) Poisson görbék mentén helyezkednének el.

Látszik, hogy mindkét mennyiség átlaga a másik függvényében nagyjából lineárisan indul, aztán a követési barátok függvényében az említési barátok száma (piros görbe a 5.17. ábrán) szaturációba megy át valahol 100 és 200 között. Ez ugyanaz a nagyságrend, ahol Huberman, Romero és Wu szerint a felhasználók aktivitása telítődik a követőik számának függvényében.[27] Százas nagyságrendű követési barátig tehát a kétféle kapcsolat száma arányos egymással, ami nagyjából elvárható egy valódi embertől. Ebbe a nagyságrendbe esik az úgynevezett Dunbar-szám, ami azt mutatja, hogy egy ember nagyjából hány emberrel tud fenntartani stabil szociális kapcsolatot, vagyis hány valódi ismerőse lehet. Robin Dunbar erre a számra 150-et javasolt.[37] Mindezek alapján kijelenthetjük, hogy a magánembereket reprezentáló, valódi ismeretségi kapcsolatokkal rendelkező felhasználók többnyire azok, akik a 100-as nagyságrendnél kevesebb követési baráttal rendelkeznek. Az ennél több kapcsolattal rendelkező felhasználók relatíve kevés említési baráttal rendelkeznek, viszont irreálisan sok követési baráttal, tehát valószínűleg nem valódi emberek, vagy kapcsolataik nem reprezentálnak valódi ismeretséget. A határ nem éles, inkább a 100-200 körüli nagyságrend a fontos.

Ezzel ellentétben az említési barátok számának függvényében mérve a követési barátok átlagát, a kapott függvény a végéig lineárisan emelkedik, kb. 60-ig, ami kb. a 10 000 legtöbb említési baráttal rendelkező felhasználó átlaga. Ez azt jelenti, hogy ennél több említési barátja a felhasználók csak nagyon kis részének van. (A metszetbeli felhasználók kevesebb, mint 0,4%-ának van 70-nél több említési barátja, míg a követési barátok száma nagyjából 2 000-ig skálázik. Az említési barátok száma tehát gyakorlatilag minden felhasználóra a Dunbar-szám alatt marad.) Az adatokra egyenest illesztve azt kapjuk, hogy a követési barátok számának átlagát ($\bar{N}_f$) az említési barátok számának (N_m) függvényében jó közelítéssel kapjuk a $\bar{N}_f = 1,68 \cdot N_m + 23$ formulával. (Az átlagolások miatt, ahogy az a 5.17. ábrán is látszik, ez a függvény nem a $\bar{N}_m(N_f)$ függvény inverze.)

A lineáris emelkedés, és az inverzió látszólagos ellentmondására két feloldás is lehet:

1.) bármelyik fokszámnak (pl. N_x) a másik (N_y) egy rögzített értékére egy hosszú farkú eloszlása van, ami minden N_x értékre ugyanaz, csak mivel a kis N_x értékekre több pont van, ezért a farokba is több kerül.

2.) lehetséges, hogy a felhasználókat két csoportra oszthatjuk: a nagy többséget alkotó átlagfelhasználókra, akikre a lineáris viselkedés jellemző, valamint a cégek, hírességek, robotok felhasználóira, amik az igen magas fokszámokért felelnek, és jellegüktől függően vagy a követési, vagy az említési barátok száma nagy, míg a másik kicsi.

Az 1)-nek ellentmond az, hogy - mint láttuk - bármilyen is az eloszlások alakja, a különböző értékekre különböző mintaátlagokat kapunk, vagyis nem ugyanaz az eloszlás húzódik végig egyik tengely mentén sem. A 2)-t viszont alátámasztja az, hogy a felhasználók a követési barátok száma szerint kettéválnak.

Hogy valami kvalitatív összefüggést adhassunk a két fokszám kapcsolatára, egyenest illesztettem az egyes felhasználókból álló adatsorra, ahova az 50-nél kevesebb követési, és 15-nél kevesebb említési baráttal rendelkező felhasználókat vettem bele (1 759 539 pont, ami az összesnek 77,6%-a, vagyis a pontok legnagyobb része részt vett az illesztésben, de csak a lineáris tartományban lévők). Az illesztés azt adta, hogy az említési barátok száma kb. $0,0672 \pm 0,00022$-szerese, vagyis kevesebb, mint a tizede a követési barátok számának.

További érdekes tény, hogy a 5.17. ábra szerint az említési barátok számának maximuma 0 körül van minden N_f-re, de a követési barátok számának maximuma emelkedik, ahogyan N_m is emelkedik (ez az y tengely melletti völgy az ábrán). Ez azt jelenti, hogy aki sok követési kapcsolattal rendelkezik, az legvalószínűbben nem rendelkezik említési kapcsolatokkal, vagy csak nagyon kevéssel, azonban aki sok említési kapcsolattal rendelkezik, az sokkal valószínűbben rendelkezik valamennyi véges követési kapcsolattal, mint 0-val. Tehát akit sokan követnek, és aki sokakat követ, az inkább csak a követési kapcsolatokra "hajt", de akik sok kölcsönös említési kapcsolattal rendelkeznek, azoknak "automatikusan jár" követési kapcsolat is. Ha valódi kapcsolatokkal rendelkező valódi felhasználókról azt feltételezzük, hogy a két fajta kapcsolataik száma nagyjából arányos egymással, akkor ez alapján azt mondhatjuk, hogy kölcsönös említési kapcsolatot inkább valódi felhasználók hoznak létre, míg kölcsönös követési kapcsolatot ettől eltérő viselkedésű felhasználók is.

Ez, tovább az, hogy 1) a $\bar{N_f}(N_m)$ függvény 1-nél nagyobb, 2) a $\bar{N_m}(N_f)$ illesztés alapján 50 alatt kb. $0,20 \pm 0,011$, 10-ig kb. $0,150 \pm 0,0085$, azaz bőven 1 alatti meredekséggel indul, és 3) a felhasználó párokra illesztett egyenes meredeksége $0,0672 \pm 0,00022$, 4) hogy N_m kb. 60-ig, N_f kb. 2000-ig megy, mutatja, hogy említési baráti kapcsolatot létrehozni jóval "nagyobb költség", mint követési barátot. Egy tipikus felhasználó jóval több követési, mint említési baráttal rendelkezik, vagyis utóbbi jóval szorosabb kapcsolatnak számít, és az oda-vissza említések száma miatt jóval árnyaltabb rendszert alkot. Újra megjegyzem, hogy az adatgyűjtés során prioritást élvezett a követési kapcsolatok feltárása, így az egyenesek meredekségének konkrét értéke nem bír jelentőséggel. Ez azonban a véletlen mintavételezés miatt csak a tengelyek átskálázódását jelenti: az átlagok menete, és az y tengely melletti völgy megmarad, továbbá igaz az, hogy a két gráfok ugyanazon felhasználók között különböző kapcsolatokat alkotnak. Az illesztett egyenesek meredeksége eltérhet ugyan, de ez csak kvantitatív különbséget jelent a két barátság számának összefüggésében.

Összefoglalva az eddigieket, kijelenthetjük, hogy a kölcsönös említési kapcsolat jóval inkább valódi felhasználókra jellemző, hogy az ilyen kapcsolati rendszer nagyon ritka, és hogy a 100-200 körüli követési baráttal rendelkező felhasználók nagy valószínűséggel nem valódi személyekhez tartoznak. Így javaslom a kölcsönös említési kapcsolatot, mint valódi, személyes ismeretségi kapcsolat bevezetését. Ez a kapcsolatrendszer felfedheti a felhasználók információ terjedés szempontjából fontos klikkjeit, csomópontjait és csatornáit. Ez hasznos lehet azon felhasználók megtalálásában, akiket meg kell célozni, ha a világon, vagy egy adott célcsoportban el szeretnénk terjeszteni valamilyen információt. Ehhez tehát a `Twitter` API véletlenszerűen mintavételezett adathalmazában a "@" jellel kezdődő említéseket kell vizsgálni, és az így kölcsönösen kialakult "említési barátok" közül azokat a felhasználókat kell keresni, akik sok ilyen kapcsolattal rendelkeznek.

6. Témák (hashtagek) elemzése

Az egy-egy tweet témájának megjelölésére szolgáló *hashtag* kiválóan alkalmas nem csak a témák szerinti keresésre, de egy adott téma népszerűségének mérésére, térbeli elterjedtségének vizsgálatára, időbeli felfutásának elemzésére is. Az adott tweetekben közösen megjelölt témák alapján vizsgálhatjuk azt, hogy milyen témák járnak együtt, és hogy ezek milyen hatással vannak egymásra.

6.1. Hashtagek népszerűségének és lokalizáltságának vizsgálata

Az adatbázisban a `tweet_hashtag` tábla tartalmazza a hashtaggel megjelölt szavakat. Minden sor egy ilyen szó egy előfordulása. Megtudhatjuk, hogy melyik felhasználó melyik tweetjében, mikor jelölte meg a szót. A tábla 649 644 487 sort tartalmaz, amikben 38 944 829 különböző szó fordul elő (kisbetű és nagybetű ugyanannak számít), vagyis egy szót átlagosan 16,68-szor jelöltek meg. Ezt leszűkíthetjük a koordinátával rendelkező (*geo-tagged*) tweetekre, itt már csak 20 954 552 téma fordul elő összesen 209 728 706 alkalommal, a helymegjelöléssel ellátott tweetekben egy hashtaget átlagosan 10,0087, majdnem pontosan 10 alkalommal tettek közzé. Kiszámoltam ezek hisztogramját, amiből kiderült, hogy hatványfüggvényt követ, egyenest illesztve a logaritmikus adatokra azt kaptam, hogy ez az $N(CNT_{gt}) = (15,96 \pm 0,024) \cdot CNT_{gt}^{-1,878 \pm 0,0059}$ formulával közelíthető.

Megszámoltam minden egyes hashtagre az előfordulási valószínűséget az összes tweet között (CNT), valamint a helymegjelöléssel rendelkező tweetek körében (CNT_{gt}). Nyilván $CNT_{gt} \leq CNT$. A továbbiakban ezekkel fogok dolgozni. Minden témára kiszámoljuk a kettő gyakoriság arányát, és ezeket témák szerint átlagoljuk. Ez 0,8962, (súlyozatlan átlag) vagyis egy adott hashtagre valamivel kevesebb, mint 90% a valószínűsége, hogy annak egy előfordulása koordinátázott lesz. Az összes előfordulásra átlagolva a CNT_{gt}/CNT hányadost (népszerűséggel súlyozott átlag) ez a szám 0,7594, vagyis véletlenszerűen kiválasztva egy adott hashtag előfordulást, valamivel több, mint 75% a valószínűsége, hogy koordinátás tweetben lett közzétéve. Ez azt jelenti, hogy a kevésbé népszerű hashtagek hajlamosabbak koordinátázott tweetekben előfordulni.

Korábban már bevezettem a Föld felszínén mért távolság számolására egy módszert. Erre támaszkodva bevezethetünk egy Föld felszínén mért szórást. A statisztikából ismert standard deviancia mintájára tekintsük ezt az átlagtól való négyzetes eltérések átlagának négyzetgyökének, és jelöljük σ-val! A pontok átlagát könnyen számolhatjuk, ha a pontokat a Föld középpontjából a pont felé mutató egységvektorral reprezentáljuk. (Ezen vektor koordinátái, és a szélességi-hosszúsági koordináták egyszerű koordinátatranszformációval átszámolhatók egymásba.) Ekkor ugyanis a pontok átlagának tekinthetjük a pontokat reprezentáló vektorok összegének normált változatát, ami szintén egy egységvektor, iránya megegyezik a vektorok átlagának irányával, vagyis a pontrendszer intuitíve várható közepén lévő Föld felszíni pontba mutat.) Ekkor egy Föld felszíni ponthalmaz szórása:

$$\sigma_R = \sqrt{\frac{\sum_{i=1}^{N} dist^2(gp_i, gp_{avg})}{N}} \tag{6.1}$$

ahol $dist(gp_i, gp_j)$ a gp_i és gp_j pontok távolsága a Föld felszínén, a 5.1. egyenletben leírt haversine-formulával számolva, N pedig a beolvasott pontok száma. gp_i az i-edik pont a felszínen, gp_{avg} pedig a fent leírt módon meghatározott átlag pont. A szórás mértékegységét abban kapjuk, amiben a $dist$ függvényünk felveszi az értékét. A radiánban jelölt szórást fogom σ-val jelölni, a km-ben mértet pedig σ_R-rel, jelezve, hogy a kettő közötti összefüggés: $\sigma_R = R \cdot \sigma$, ahol $R = 6\,372,797$ km a Föld átlagos sugara. A km-ben mért szórás szemléletesebb, ezért ezt fogom használni.

Azt szeretnénk megtudni, hogy a hashtagek térbeli eloszlása milyen összefüggésben van a népszerűséggel. Mivel a szórást csak a koordinátákkal megjelölt tweetek határozzák meg, σ_R és CNT_{gt} összefüggését fogom vizsgálni.

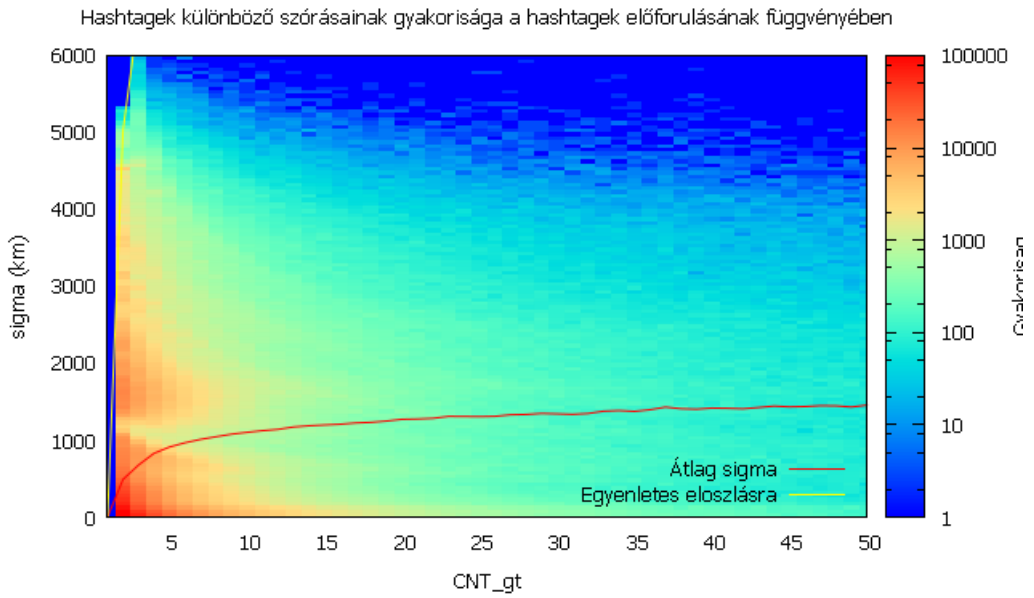
A 6.1. ábrán ábrázoltam a σ_R-eket a CNT_{gt}-k függvényében minden pontra. Minden CNT_{gt}-re 50 km-es binmérettel binekbe osztottam a σ_R értékeket és a 5.4.2. fejezet 5.17. ábrájához hasonló módon a binekbe eső adatpárok számát színekkel jelöltem logaritmikus színskálán.

Itt is számoltam CNT_{gt}-hez átlag σ_R-t, 200-nál kisebb CNT_{gt}-re egyesével, fölötte 20-as binekbe osztva, mert a kevés adatpont miatt különben nagyon nagy lett volna a szórás. Az átlag menetét szintén tartalmazza a 6.1. ábra. A hatványszerű eloszlás miatt kis gyakoriságú tagekből lesz a legtöbb, ám a szórás eloszlása változik CNT_{gt} függvényében: nem sokkal 1 000 km fölött látszik egy völgy, ami valószínűleg a kontinens közi hézag miatt van. A völgy, és a csúcsok CNT_{gt} növekedésével enyhén

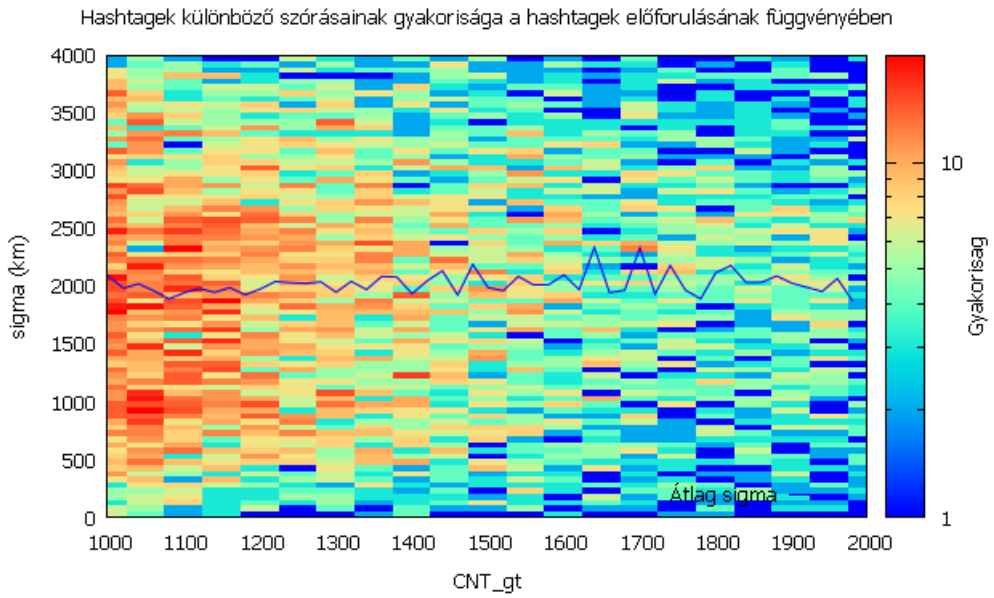
csökkennek, ami azért lehet, mert pár szétszórt pontnak nagy a szórása, de ha valamivel többet veszünk, akkor az újonnan hozzáadott pontok nagy valószínűséggel már közelebb lesznek egy már meglévőhöz, ami kicsit összehúzza a szórást. Rögzített CNT_{gt}-re a földrajzi mintázaton kívül egy lecsengő eloszlása van a σ értékeknek.

A 6.2. ábrán ugyanezt ábrázoltam 1 000 fölötti gyakoriságú hashtagekre. Az 50 és 1 000 közötti gyakoriságú hashtagek így kimaradtak, de a kis és nagy multiplicitása hashtagek mintázatának sajátosságai ezeken az intervallumokon látszanak legjobban. A binméret, és a színskála a jóval kevesebb pont miatt meg lett változtatva. Ilyen gyakori hashtagekre már rögzített CNT_{gt}-re a σ eloszlás nem lecsengő, hanem az átlag körüli nagyjából szimmetrikus eloszlású, és CNT_{gt} függvényében nem változik, csak a pontok száma csökken.

Érdekes lehet a szempontunkból, hogy elkülöníthető-e nagy számban közzétett, de kis területre lokalizált hashtagek csoportja. Ez azért lehet fontos, mert népszerű, ám helyhez kötött eseményekhez kötődnek, pl. olimpia, elnökválasztás, fesztivál, vagy esetleg valami katasztrófa, vagy más nem várt esemény. Az ilyen hashtagek detektálása használható lehet ilyen helyhez kötött események detektálásában.



6.1. ábra: Hasgtagek térbeli szórása a gyakoriság függvényében. Ezen a szakaszon mind CNT_{gt}, mint σ függvényében erősen lecsengő az eloszlás.



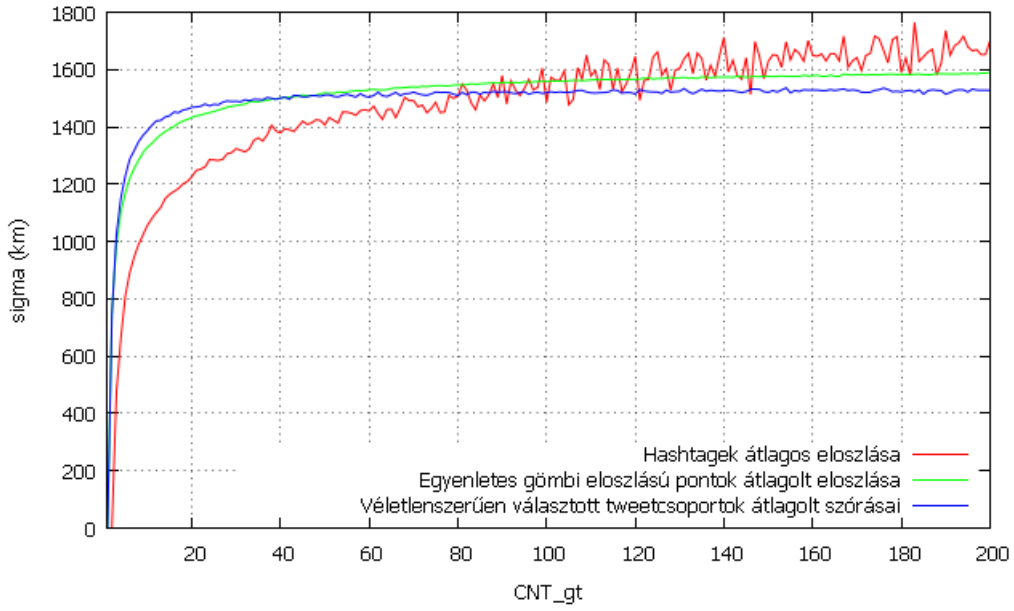
6.2. ábra: A 6.1. ábrához hasonló módon ábrázolva a gyakoriságok 1 000 fölötti gyakoriságú hashtagekre. Ilyen gyakori hashtagekre már konstans az átlag, és a szórás eloszlásának maximuma az átlag körül van, és akörül nagyjából szimmetrikus.

Az átlag (mint az várható) 0-ból indul. Az 1 multiplicitású pontok szórása csak 0 lehet. Az átlag ezután emelkedik, és néhány 10 után elég kicsi lesz a meredeksége, és 1 000-es nagyságrendnél nagyjából konstansnak tekinthető. Vagyis, ha a közzétételek helyei egy véges területre korlátozódnak, akkor kb. 1 000 témamegjelölés alatt a közzétételek már nagyjából lefedik ezt a területet, és az új üzenetek már nem új helyen fognak megjelenni, így nem növelik a szórást.

Maga a Föld is egy ilyen véges terület, így a Föld felszínén gömbi egyenletes eloszlással véletlenszerűen elosztott N db pont szórása sem emelkedik egy érték fölé N növelésével. Monte-Carlo módszerrel meghatároztam ugyanezt a görbe menetet egy R sugarú gömbön véletlenszerűen elosztott pontokra. Marsaglia (1972)[38] által levezetett módszerrel ponthalmazokat hoztam létre különböző számossággal, majd kiszámoltam ezek standard devianciáját az eddigivel egyező módszerrel. Minden N értékhez kb. 10 000/N-szer ismételtem meg a számítást, és a kapott eredményeket átlagoltam. Ezt az eredményt tekintettem rendre az N-hez tartozó $\sigma_{\text{Véletlen}}$ értéknek. Ez a módszer a kis relatív hibát, és a rövid futásidőt is biztosította. A kapott görbéket egy konstans szorzóval Newton-Raphson módszerrel a hashtagekből mért átlagolt görbére illesztettem a $0 \leq CNT_{gt} \leq 200$ intervallumon. Szorzófaktornak 0,1517-et kaptam, vagyis a maximum 200 gyakoriságú hashtagek szórása átlagosan 15,2%-a

ugyanennyi véletlenszerűen szétszórt ponténak. Ezzel számszerűsítettük azt, hogy egy hashtag előfordulásai mennyivel lokalizáltabbak a véletlenszerűen szétszóródó pontoknál. A két görbét a 6.3. ábra tartalmazza.

Nyilvánvaló, hogy a hashtagek eloszlása igen távol áll a gömbi eloszlástól, ezért az előbb leírt számítást elvégeztem valódi tweetkoordináták véletlenszerűen kiválasztott N számosságú csoportjaival, és ugyanolyan módon átlagoltam őket, ahogyan fentebb az egyenletes eloszlású véletlen pontokat. Ezeket σ_{Tweet}-tel jelölöm. A tweetek véletlenszerűen kiválasztott csoportjainak szórásához a 0,2088 szorzófaktort kaptam. Ez az összehasonlítás azért fontos, mert ezen pontok szóródásában már benne van a földrajz, és a népsűrűség. A hashtagek eloszlása ugyanis várhatóan kisebb a véletlenszerűen elosztott pontoknál, hiszen a felhasználók csak véges méretű területekről tesznek közzé tweeteket. Azonban a 200-nál nem nagyobb népszerűségű hashtagek szórása a valódi tweetek szórásának is kb. ötöde. Ebből arra következtethetünk, hogy nem véletlenszerűen oszlanak el azon a területen sem, ahonnan ténylegesen posztolnak a felhasználók, hanem területileg is szerveződnek. A σ_{Tweet} görbéjét a 6.3. ábra szintén tartalmazza.

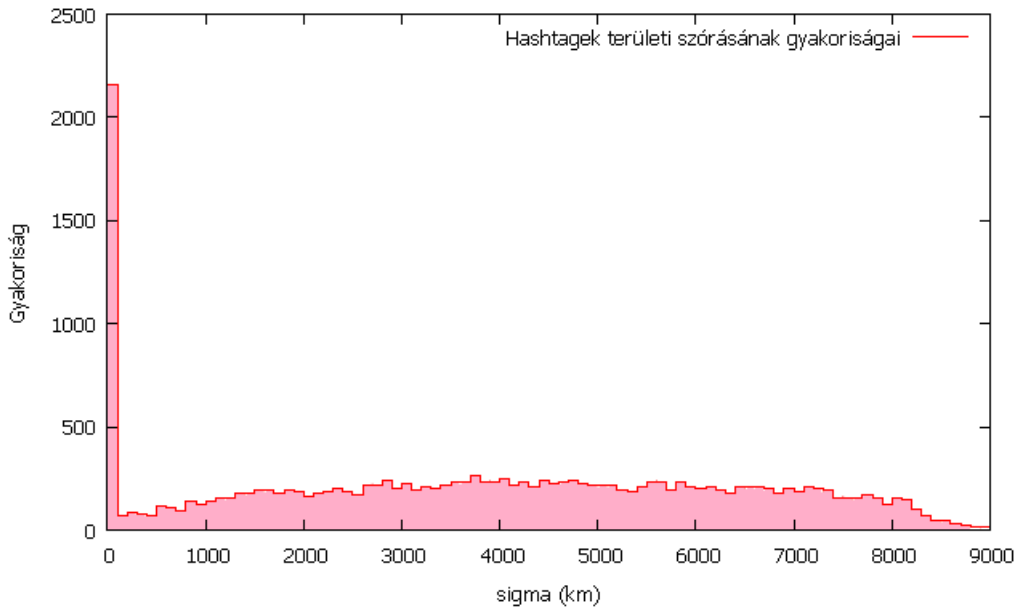


6.3. ábra: Hashtagek átlagos szórása a népszerűség függvényében a hosszá illesztett véletlen pontokból, valamint véletlen tweetcsoportokból számolt hasonló görbékkel.

Ezek az arányok, ahogy a 6.3. ábrán is látszik, valójában függenek CNT_{gt}-től. A görbék eltérő alakja amiatt van, mert az egyenletes eloszlásnál kisebb területen szétszórt pontok geometriája jobban hasonlít egy síkra, így az eloszlás alakja eltolódik. A gömbi alakkal történő illesztés tehát közelítés, de ad egy szemléletes képet a két szóródás számszerű eltéréséről.

Az átlagolt σ_R értékek görbéiből megtudhatjuk, hogy mekkora területre korlátozódnak a hashtagek, és a véletlenszerűen szétszórt pontok. Az átlag σ értékeket átlagolva az $CNT_{gt} > 1000$ pontokra kb. 2 100 km-t kapunk, vagyis egy ekkora szórású területet elérve az átlag nem növekszik tovább. 1 000-től 10 000-ig a véletlenszerű pontokat is legeneráltam, amik szórása N függvényében szinte tökéletesen konstans eredményt produkált: 10 830 km-t, tized százalék nagyságrendű ingadozással. Vagyis a nagy számú hastagek szórása már stabilan kb. 19,4%-a véletlen pontoknak. Ugyanez a szám a tweetekből számolt értékekre 7 950 km, így az arány 26,4%.

Elkészítettem szintén 50 km-es binmérettel az összes 1000 fölötti népszerűségű hashtag hisztogramját, ez a 6.4. ábrán látható. Ezen az látszik, hogy a szórások eloszlása tényleg egy szimmetrikus eloszlás, azonban a legkisebb binben ($\sigma_R < 50$ km) a többinél jelentősen több hashtag van. Az ábrák részletes közlése nélkül összefoglalom, hogy a legkisebb bint km-es felbontásban is megvizsgáltam, és a csúcsot nagyrészt a 0-hoz közeli szórású tagek alkotják. Ez jó eséllyel a (numerikus hibáktól eltekintve) pontosan ugyanonnan kiposztolt tagek miatt van, például ha egy étterem, vagy más hely minden tweetjében hashtagként jelöli meg saját magát, vagy az akcióit. Az, hogy nagyon 0-ra koncentrálódik a csúcs, továbbá az, hogy az ilyen kis szórású hashtagek CNT_{gt} szerinti eloszlása nem különbözik jelentősen más szórású hashtagek eloszlásának alakjától azt jelzi, hogy ezek nem a népszerű, de lokalizált események nyomát jelzik. Hasonlóan, ahogy a kölcsönös említési gráf élei sem váltak szét egyértelműen r szerint, itt sem tudunk egyértelmű határt húzni sem σ_R, sem CNT_{gt} szerint.



6.4. ábra: $CNT_{gt} > 1000$ hashtagek összesített σ_R szerinti eloszlása 50 km-es binmérettel.

A hashtagek lokalizáltságáról tehát számszerű összehasonlítást tudtunk adni a véletlenszerűen szétszórt pontokhoz, és a tweetekhez képest. Megmutattam a kevésbé népszerű, és a népszerűbb hashtagek szórásának viselkedését, és arra kerestem a választ, hogy ily módon detektálhatunk-e helyhez kötött eseményeket, amiről sokan beszélnek, de egyértelműen nem sikerült ilyet azonosítani. Egy esetleges részletesebb vizsgálat már nem képezi részét ennek a dolgozatnak.

Lehetőség van hashtagek egymással való kapcsolatát is vizsgálni oly módon, hogy azokat tekintjük összekapcsoltnak, amik együtt lettek közzétéve egy tweetben. Ezt lehet súlyozni az együtt előfordulások számával, így a hashtagekből is építhetünk egy súlyozott, irányítatlan gráfot. A fokszámeloszlás és klaszterezettség vizsgálata után lehetőség van a gráf földrajzi tulajdonságait, például a szórások korrelációját vizsgálni. Vizsgálhatjuk, hogy az összekapcsolódó témák népszerűsége milyen hatással van egymásra az idő függvényében. Az ilyen vizsgálatok azonban már nem férnek bele a jelen dolgozat kereteibe.

7. Összegzés

Dolgozatomban a hálózatkutatás és statisztika eszközeinek felhasználásával vizsgáltam az internetes `Twitter` mikroblog szolgáltatás által létrehozott szociális hálózatot. Az adatokat az ELTE Komplex Rendszerek Fizikája Tanszékén gyűjtötték relációs adatbázisba a `Twitter` API ingyenes Sample adatfolyamát használva. Én ezen az adathalmazon dolgoztam.

A dolgozat első felében az elméleti alapozás után a kölcsönös említési gráf statisztikus viselkedését vizsgáltam, főleg az összekötött felhasználópárok távolságeloszlásán keresztül. Megmutattam, hogy ezen adatokból hogyan mérhető ki a népsűrűség fraktáldimenziója, és hogy ez milyen összefüggésben van a távolságeloszlással, azaz hogy a távolságeloszlás milyen mértékben áll össze földrajzi, illetve társadalmi hatások szorzataként. A gráfot részletesen összehasonlítottam a kölcsönös követési kapcsolatokból kialakított gráffal. Megmutattam, hogy nagyrészt ugyanazok a felhasználók között a két gráf gyökeresen eltérő kapcsolatrendszert eredményez.

Megmutattam, hogy sok felhasználó rendelkezik ezres nagyságrendű kölcsönös követési kapcsolattal, ami sokkal több, mint ahány ismerőssel egy ember átlagosan rendszeres kapcsolatot tart (Dunbar szám, ≈ 150). Ezen a szakaszon a felhasználók említési kapcsolatainak számának módusza 0 körül van, átlaga pedig nem növekszik a követési kapcsolatok számával. Ezzel ellentéten az említési kapcsolatok száma a vizsgált adathalmazban nem növekedett a Dunbar szám fölé, ezek a felhasználók tipikusan valamilyen véges számú követési kapcsolattal rendelkeznek, aminek átlaga és maximuma is növekszik az említési kapcsolatok számának növelésével. Ez azt is jelenti, hogy aki sok kapcsolattal rendelkezik az egyik fajtából, nem feltétlenül rendelkezik tipikusan sokkal a másikból. Mindezekből arra következtettem, hogy a kölcsönös említési kapcsolat sokkal inkább olyan tulajdonságokat mutat, mint egy valódi ismeretségtől elvárható, így a hétköznapi, bizalmi információk terjedését e gráf mentén várhatjuk (ami sokkal ritkábba a követési gráfnál). Ez főleg igaz, ha megkövetelünk egy oda-vissza is igaz minimum említésszámot. A távolságeloszlás alakját vizsgálva úgy találtam, hogy ezt a küszöbértéket nem érdemes 5-nél többre állítani, mert efölött növelve a küszöböt, már azonos valószínűséggel vesztünk különböző hosszúságú éleket. Különösen nagyot választva pedig az adathalmaz olyan kicsi lesz, hogy a távolságok relatív szórása igencsak megnő.

Megmutattam, hogy a kölcsönös említési gráf jó közelítéssel skálafüggetlen, ami - ahogy a 4. fejezetben megmutattam, - azt jelenti, hogy a nagy fokszámú pontok fontos szerepet játszanak az információáramlásban.

Így javaslatot teszek egy, a felszín alatt húzódó kapcsolatrendszer definiálására az egymást oda-vissza legalább 5-ször említő felhasználókból kialakítva. Az ezen hálózatban résztvevő felhasználók közül pedig a nagy fokszámúak viselkednek információáramlási csomópontként.

Dolgozatom második részében a felhasználók által tweetjeikben megjelölt témák, azaz a hashtagek népszerűségének és térbeli szórásának kapcsolatát vizsgáltam. Megmutattam, hogy a szórásuk a kb. 1 000-es nagyságrendű gyakoriság fölött már konstansnak tekinthető, eloszlásuk pedig az átlag körül szimmetrikus. Kis gyakoriságú hashtagekgre a szórás átlaga növekszik, maximuma viszont csökken, az eloszlás alakján pedig a korábbi fejezetekhez hasonlóan visszatükröződnek a földrajzi tulajdonságok. Számszerű összefüggést adtam a hashtagek térbeli lokalizáltsága, és a Föld felszínén véletlenszerűen elosztott pontok, valamint véletlenszerűen kiválasztott tweetek lokalizáltsága között mind kis, mind nagy népszerűségű hashtagekre. Nem meglepő módon a kis számú hashtagek szórása kb. 15,2%-a a véletlenszerűen elosztott pontokénak, és 20%-a a véletlenszerűen választott tweetekének. Népszerű hastagekre ez rendre 19,4% és 26,4%. Megmutattam, hogy szignifikánsan sok hashtag van 0, vagy ahhoz közeli szórással. Azt is megmutattam, hogy nincsenek egyértelműen elkülönülő népszerű, de lokalizált témák.

Jelen szakdolgozat PDF formátumban online elérhető a `http://adamrudolf.web.elte.hu/MSc_szakdolgozat.pdf` címen. A problémák megoldására írt C# osztályokat és függvényeket tartalmazó névtér forráskódja szintén szabadon elérhető a `http://adamrudolf.web.elte.hu/letoltesek/Csharp/MSc/TwitterTools.rar` címen angol nyelvű XML dokumentációval együtt. Az ebből készült online html dokumentáció elérhető a `http://adamrudolf.web.elte.hu/dokumentacio/html/index.html` címen, PDF formátumban letölthető innen: `http://adamrudolf.web.elte.hu/dokumentacio/latex/refman.pdf`. A szakdolgozathoz készült SQL kódok a dolgozat fejezetei szerint csoportosítva szintén elérhetőek a `http://adamrudolf.web.elte.hu/letoltesek/SQL/MSc_szakdolgozat.txt` címen.

A jövőben lehetőség van másfajta kapcsolatrendszerek, például retweetelés alapján alkotott kapcsolatok geo adatinak ehhez hasonló elemzésére. A hashtagek terén meg lehet vizsgálni részletesebben, hogy a szórások eloszlását milyen mértékben okozzák földrajzi tulajdonságok. Lehetőség van hashtagek kapcsolatának vizsgáltára aszerint, hogy mely témák lettek azonos tweetekben megjelölve. Vizsgálható a klaszterezettség, illetve az ily módon összekötött hashtagek szórásainak korrelációja. Vizsgálható az időfüggés is, például hogy egy kevésbé ismert hashtag egy népszerűvel való összekapcsolódása után mennyivel nő meg a népszerűsége.

Hivatkozások

[1] Barabási, A.-L., Albert, R. (1999), *"Emergence of scaling in random networks"*, Science 286, 509-511.

[2] A.Szalay and J. Gray. *"Science in an exponential world"*. Nature, 440, pp. 413-414, 2006. DOI:10.1038/440413a

[3] J. Travers, S. Milgram. *"An Experimental Study of the Small World Problem"*. Sociometry, Vol. 32, No. 4, pp. 425-443., 1969.

[4] T. Sakaki, M. Okazaki, Y. Matsuo. *"Earthquake shakes Twitter users: real-time event detection by social sensors"*. Proceedings of the 19th international conference on World wide web. Pages 851-860. 2010.

[5] C. Biever. *"Twitter mood maps reveal emotional states of America"*. New Sciencist. Vol. 207, Issue 2771, page 14. 2010.

[6] H. Thomases. *"Twitter Marketing: An Hour a Day"*. John Wiley & Sons, 2009.

[7] K. Lacy. *"Twitter Marketing For Dummies"*. John Wiley & Sons, 2011.

[8] T. Weinberg, C. Pahrmann. *"Social Media Marketing: Strategien für Twitter, Facebook & Co."* O'Reilly Germany, 2012.

[9] A. Bermingham, A. F. Smeaton. *"On using Twitter to monitor political sentiment and predict election results"*. Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP), 2011.

[10] J. Bollen, H. Mao, X. Zeng. *"Twitter mood predicts the stock market"*. Journal of Computational Science, Vol. 2, Issue 1, pages 1-8., 2011.

[11] C. St Louis, G. Zorlu. *"Can Twitter predict disease outbreaks?"* BMJ 2012;344:e2353

[12] H. Moore. *"Twitter files for IPO in first stage of stock market launch"*. The Guardian. September 2013.

[13] 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)

[14] J. Benhardus, J. Kalita. *"Streaming trend detection in twitter".* IJWBC, pages 122-139, 2013.

[15] A. Bruns, Y. E. Liang. *"Tools and methods for capturing twitter data during natural disasters".* First Monday, 2012.

[16] M. Hu, S. Liu, F. Wei, Y. Wu, J. T. Stasko, K.-L. Ma. *"Breaking news on twitter".* In CHI, pages 2751-2754, 2012.

[17] L. Dobos, J. Szüle, T. T. Bodnár, T. Hanyecz, T. Sebok, 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

[18] Biggs, N.; Lloyd, E. and Wilson, R. (1986), *"Graph Theory"*, 1736-1936, Oxford University Press

[19] Freeman, L. (1977). *"A set of measures of centrality based on betweenness".* Sociometry 40: 35-41. doi:10.2307/3033543.

[20] Leyzorek, M. et al. (1957). *"Investigation of Model Techniques"* - First Annual Report - 6 June 1956 - 1 July 1957 - A Study of Model Techniques for Communication Systems. Cleveland, Ohio: Case Institute of Technology.

[21] Brandes, U. (2001). *"A faster algorithm for betweenness centrality".* Journal of Mathematical Sociology 25: 163-177. Retrieved 10.11.2011.

[22] Vespignani, A., Pastor-Satorras, R. (2004) *"Evolution and Structure of the Internet: A Statistical Physics Approach"* Cambridge University Press New York, NY, USA 2004 ISBN:0521826985

[23] P. W. Holland, S. Leinhardt (1971). *"Transitivity in structural models of small groups".* Comparative Group Studies 2: 107-124.

[24] D. J. Watts, S. Strogatz (June 1998). *"Collective dynamics of 'small-world' networks"*. Nature 393 (6684): 440-442. Bibcode:1998Natur.393..440W. doi:10.1038/30918. PMID 9623998.

[25] K. I. Goh, B. Kahng, D. Kim (12 Dec 2001). *"Universal Behavior of Load Distribution in Scale-Free Networks"*. Physical Review Letters. 87.278701 (27). doi:10.1103/physrevlett.87.278701.

[26] 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).

[27] B. A. Huberman, D. M. Romero, F. Wu *"Social networks that matter: Twitter under the microscope",* Cornell University Library, Ithaca, NY 14850, arXiv:0812.1045 2008.

[28] D. Kondor, I. Csabai, L. Dobos, J. Szüle, N. Barankai, T. Hanyecz, T. Sebok, Zs. Kallus, G. Vattay (2013). *"Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages."* arXiv:1311.1169

[29] 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

[30] J. P. Huchra, M. J. Geller. *"Groups of galaxies. i - nearby groups"*. Astrophysical Journal, 257:423-437, June 1982.

[31] http://en.wikipedia.org/wiki/Haversine_formula

[32] M. Mewman (2005), *"Power laws, Pareto distributions and Zip's law."* Contemporary Phys 2005, 46, 323

[33] P. Papadimitriou, A. Dasdan, H. Garcia-Molina, *"Web graph similarity for anomaly detection,"* JISA, vol. 1, no. 1, p. 1167, 2008.

[34] H. Bunke, P. J. Dickinson, M. Kraetzl, W. D. Wallis, *"A Graph-Theoretic Approach to Enterprise Network Dynamics (PCS)"*. Birkhauser, 2006.

[35] P. Grassberger, *"Generalized Dimensions of Strange Attractors"*. Physics Letters A 97 (6): 227?230, 1983.

[36] Appleby, S. (1996), *"Multifractal Characterization of the Distribution Pattern of the Human opulation."* Geographical Analysis, 28: 147-160. doi: 10.1111/j.1538-4632.1996.tb00926.x

[37] Purves, D. (2008). *"Principles of cognitive neuroscience. Sinauer Associates Inc."*

[38] Marsaglia, G. (1972) *"Choosing a Point from the Surface of a Sphere."* The Annals of Mathematical Statistics 43, no. 2, 645–646. doi:10.1214/aoms/1177692644.

NYILATKOZAT

Név: Rudolf Ádám András

ELTE Természettudományi Kar, szak: Fizikus MSc.

NEPTUN azonosító: HCB5XG

Szakdolgozat címe:
Szociális hálózatok statisztikai elemzése

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

Budapest, 2014. május 30.

a hallgató aláírása