TDK DOLGOZAT

Életkor becslése az egocentrikus hálózat alapján

Tamás Gábor
Fizika BSc II. évfolyam
Természettudományi Kar

Témavezető:
Dr. Török János
egyetemi docens
Elméleti Fizika Tanszék
Természettudományi Kar

Budapesti Műszaki és Gazdaságtudományi Egyetem
2016

Tartalomjegyzék

1. fejezet

Motiváció

A komplex rendszerek vizsgálata mindig nagy kihívások elé állította a fizikusokat [1]. A legizgalmasabb ilyen rendszer, a társadalom, már régóta foglalkoztatja nem csak a szociológusokat, hanem a természettudósokat is [2]. Míg korábban elképzelhetetlen volt, hogy az emberek viselkedését nagy mintákon is vizsgálhassuk, ma ez már egyáltalán nem lehetetlen. Bár az emberi tevékenység meglehetősen nehezen mérhető, azonban a digitális világban járva mindannyian információmorzsákat hagyunk magunk után, melyeket összegyűjtve és rendszerezve új ismeretekhez juthatunk. A hatalmas mennyiségű összegyűjtött adat feldolgozása az információ- és kommunikációtechnológia (ICT) fejlődése révén vált lehetővé: létrejött a Big Data néven ismertté vált tudományterület, melynek célja az óriási adathalmazok szűrése és feldolgozása.

A hálózatközpontú szemléletmód számtalan interdiszciplináris területen jól alkalmazható: ilyenek például a biológia, az informatika, a közlekedés, a pénzügyek, valamint a TDK munkám tárgyát képező szociofizika is. A szociális hálózatok a való életben létező társadalmi kapcsolatrendszereket reprezentálják, így csúcsaik az egyéneknek, míg éleik és az élekhez rendelhető paraméterek valamilyen típusú és erősségű szociális kapcsolatnak felelnek meg.

A szociofizika a statisztikus fizikában alkalmazott módszereket felhasználva vizsgálja az emberi kapcsolatok hatásait, sikerrel tanulmányozza a konfliktusok fejlődését [3] és az információ terjedését [4] is. Mivel nagy embertömegeket érintő (és így akár veszélyeztető) kísérleteket nem lehet végezni, ezért elengedhetetlen a rendelkezésre álló adatbázisok megbízhatóságának vizsgálata, illetve az esetleges hiányosságok pótlása.

TDK munkám során is egy ilyen problémával foglalkoztunk, azaz egy sok esetben hiányzó adatot, az életkort kívántuk megbecsülni. Két okból kifolyólag esett rá a választásunk: egyrészt egyetlen egész szám segítségével leírható, így jól kezelhető, másrészt láttunk esélyt arra, hogy optimalizációs technikák és gépi tanulási eljárások használata nélkül, egy konkrét lépésekből álló módszerrel is sikerrel járhatunk. Az általunk megalkotott modell segítséget nyújthat olyan döntések meghozatalakor is, melyek során a rendelkezésre álló adatokra (például nézettségi, telefonálási, látogatottsági adatok) kell hagyatkozni, és az emberek korstruktúrájának kiemelkedő szerepe van. Azonban fontos kiemelni, hogy az életkor becslése mellett módszerünk a későbbiekben - természetesen megfelelő módosítások elvégzését követően - összetettebb jellemzők (lakhely, munkahely, látogatott oktatási intézmények) vizsgálata során is felhasználható lehet.

2. fejezet

Szociális hálózatok

A szociális hálózatokra olyan komplex rendszerekként tekintünk, melyekben az atomok helyén emberek állnak, a közöttük lévő élek pedig valamilyen szociális kapcsolatot reprezentálnak. Szocális kapcsolat alatt két ember közötti olyan kapcsolatot értünk, amelyre teljesül, hogy az egyik fél ellenszolgáltatás nélkül is hajlandó lenne egy kis szívességet megtenni a másik érdekében, valamint fennáll, hogy a két ember képes lenne 5 percen keresztül a magánéletéről beszélgetni [5].

2.1. Adatbázisok

A kutatás során az általunk megalkotott eljárás tesztelésére két adatbázis is rendelkezésünkre állt. Ezek egyike volt az iWiW közösségi oldalon regisztrált felhasználók adatai - köztük születési dátumuk - és kapcsolatrendszerük. Az iWiW fénykorában Magyarország legnagyobb közösségi oldalaként több, mint 4,5 millió felhasználóval rendelkezett, így az akkoriban internetelérással rendelkező lakosság [6] közel kétharmadáról van információnk, ami nagyon jó aránynak tekinthető. Az iWiW hálózatában az átlagos fokszám 220 körüli, ezért a hálózat sűrűnek mondható.

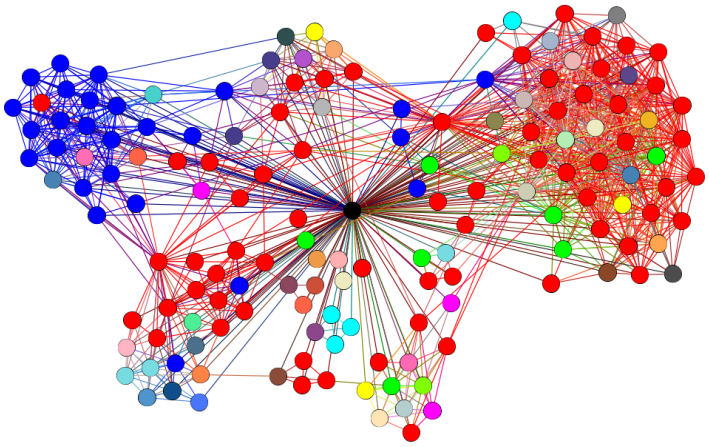
Emellett lehetőségünk nyílt megvizsgálni módszerünk hatékonyságát egy európai ország mobilszolgáltatójának 2008-ból származó adatain is. Ezen adathalmaz az ügyfelek életkorát és az általuk lebonyolított hívásokat (azok tényét rögzítve) tartalmazta. Az adott ország felnőtt lakosságának közel egynegyedéről állt rendelkezésünkre információ, azonban a hálózat az iWiW hálózatával összevetve ritkább, az átlagos fokszám

csupán 7 körüli.

Mindkét adatbázisról elmondható, hogy az adatok hiányosak voltak. Ennek oka az iWiW esetében az, hogy sokan nem - vagy hibásan - adták meg születési dátumukat a regisztráció során. Az összes felhasználó 34%-ának születési éve nem volt ismert. A telefonos adatok esetében elmondható, hogy azok egy megadott szolgáltatótól származnak, így ha az ügyfél az adott szolgáltató hálózatán kívülre telefonált, a hívott félről már nem rendelkezünk információval. Az ügyfelek körülbelül 56%-ának életkoradata volt ismeretlen.

2.2. Egocentrikus hálózatok

A szociális hálózatok viszonylag sűrűek és nagyok (több millió csúcsot is tartalmazhatnak), ezért az egész hálózat vizsgálata technikailag és elméletieg is nehézkes. Azonban erre nincs is szükség, hiszen az egyén jellemzőit az ismerősein keresztül tudjuk jól becsülni. A szociális hálózatnak azt a jól meghatározott részét, amely csak egy kiválasztott egyént és ismerőseinek kapcsolatrendszerét tartalmazza, egocentrikus hálózatnak nevezzük. Az egocentrikus hálózatban mindenki kapcsolódik az egohoz, a közösségeket pedig az ismerősök közötti kapcsolatok definiálják. A 2.1 ábrán egy egocentrikus hálózat látható.



2.1. ábra. Példa egy egocentrikus hálózatra, a fekete csúcs az ego, a színek különböző lakóhelyeket jelölnek

2.3. Közösségek a hálózatokban

Arra, hogy mikor is nevezhetjük egy adott hálózathoz tartozó csúcsok egy részét egyazon közösséghez tartozónak, több meghatározás is született az idők során [7]. Ez is jól mutatja, hogy nehéz definiálni a közösség fogalmát. Az általános hálózatos definíció az, hogy egy közösség tagjai több éllel kapcsolódnak egymáshoz, mint a hálózat más csúcsaihoz, azaz a közösség egy lokálisan sűrűn összekötött csoport. A 2.1 ábrán látható egocentrikus hálózatot a Gephi szoftverrel [8], a ForceAtlas 2 eljárás [9] segítségével ábrázoltuk. Az ábrán azonos színnel az azonos lakhelyű egyéneket jelöltük. A ForceAtlas 2 eljárás egy vonzó kölcsönhatást implementál, amely két csúcs között akkor lép fel, ha egy él összeköti őket. Egy ilyen egyszerű algoritmus is jól láthatóan definiál néhány közösséget, de önmagában ez a módszer nem alkalmas a közösségek teljes körű feltérképezésére.

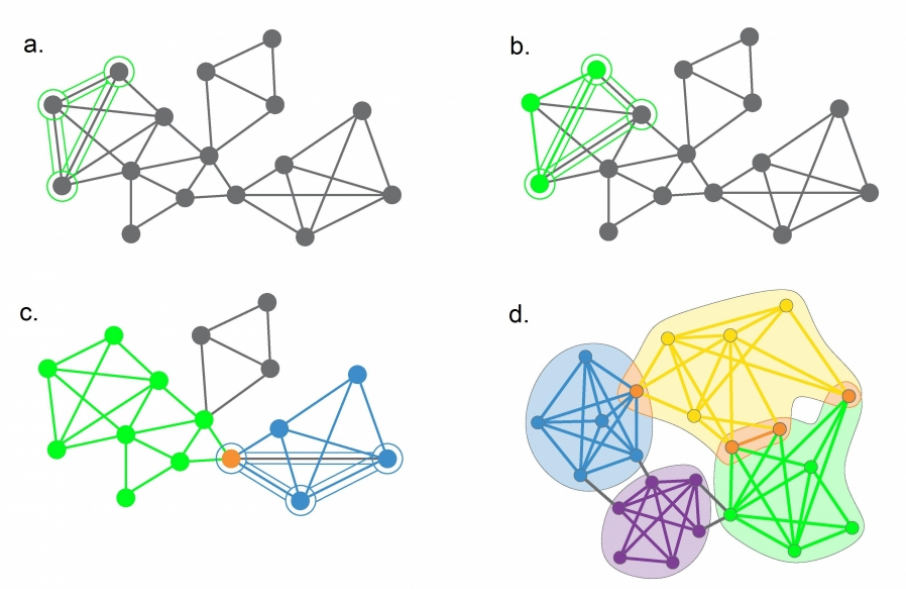
Alapvető kérdés azonban, hogy léteznek-e egyáltalán közösségek egy adott hálózaton belül? Tekintve, hogy még egyértelmű definíciót sem tudunk adni a közösség fogalmára, elméletileg nem bizonyítható, hogy valóban léteznek. Azonban - és ez a fizika területén nem ismeretlen jelenség - mérések segítségével (ami esetünkben a közösségfelismerő algoritmusok futtatását jelenti) mutathatunk példát a létezésükre, azaz az elméletek működőképességét adatok segítségével a gyakorlatban is vizsgálhatjuk. Számtalan esetben nyert már bizonyítást, hogy valóban van létjogosultsága a közösségek felismerésének, mivel ezek a közösségek nagy valószínűséggel hasonló viselkedésmintákat fognak mutatni [7].

Bár a hálózatban rejlő közösségek kódolva vannak az egyének kapcsolatrendszerében [7], de könnyen belátható, hogy felismerésük nem egyszerű feladat, hiszen a legtöbb esetben egy ember egyszerre több közösségnek is tagja, tehát a közösségek átfednek egymással. A közösségek felderítésére a korábbiakban számtalan módszer született, melyeket többféle szempont alapján is kategorizálhatunk: az eljárás lehet globális vagy lokális, a megtalált csoportok lehetnek diszjunktak vagy átfedőek. Mivel az egocentrikus hálózatok kis méretűek, ezért a globalitás nem lényeges szempont, azonban fontos, hogy a közösségek átfedőek legyenek, mint azt a 2.1 ábra is szemlélteti; például családtag és iskolatárs, munkatárs és barát is lehet valaki egy személyben. Ezért a 2.3.1

és a 2.3.2 fejezetekben két olyan közösségfelismerő algoritmust mutatunk be, amelyek a közösségek közötti átfedéseket is figyelembe tudják venni.

2.3.1. Klikk perkoláció

A klikk perkolációs algoritmus (melynek implementációja a CFinder számítógépes program) [10] egy hierarchikus közösségfelismerő eljárás. Klikknek nevezzük a gráf egy olyan részgráfját, amely egy k csúcsból álló teljes gráf. A klasszikus perkoláció során összefüggő fürtöket keresünk, így a k-klikk perkoláció során k-klikk fürtök megtalálása a cél. Ezek olyan tartományok, amelyeken belül egy k-klikkből kiindulva bármely csúcsot el tudunk érni úgy, hogy minden előtte lévő klikkhez $k-1$ csúcsunk csatlakozik ($k=2$ esetén visszakapjuk a hagyományos perkolációt, ahol egy klikk egy élnek felel meg, azaz a 2-klikk perkoláció a hálózat összefüggőségét vizsgálja). Ahogy k értékét növeljük, úgy egyre sűrűbben összekötött tartományokat tudunk felderíteni, amelyek hierarchikusan egymásba vannak ágyazva. A módszer működését a 2.2 ábra mutatja be. A 2.2a ábra egy 3-klikk perkoláció szerinti klikket mutat, míg a 2.2b ábra azt szemlélteti, hogy ebből az algoritmus hogyan halad tovább, összekötve azt egy szomszédos klikkel. A 2.2c és a 2.2d ábrák rendre az adott hálózatban a $k=3$, illetve a $k=4$ esetekben felismert közösségeket mutatják.



2.2. ábra. A klikk perkolációs algoritmus működésének szemléltetése [7]

A klikk perkoláció több olyan tulajdonsággal rendelkezik, amelyek előnyössé teszik a társadalmat reprezentáló hálózat vizsgálata során. Hierarchikus módszerként figyelembe tudja venni a társadalom hierarchikus voltát, segítségével az átfedő közösségeket is meg tudjuk találni, valamint nem heurisztikus, azaz jól definiált lépéseket hajt végre. Nagy minták vizsgálata során azonban nehezen használható, mivel futásideje a csúcsok számával exponenciálisan növekszik.

2.3.2. Ahn-módszer

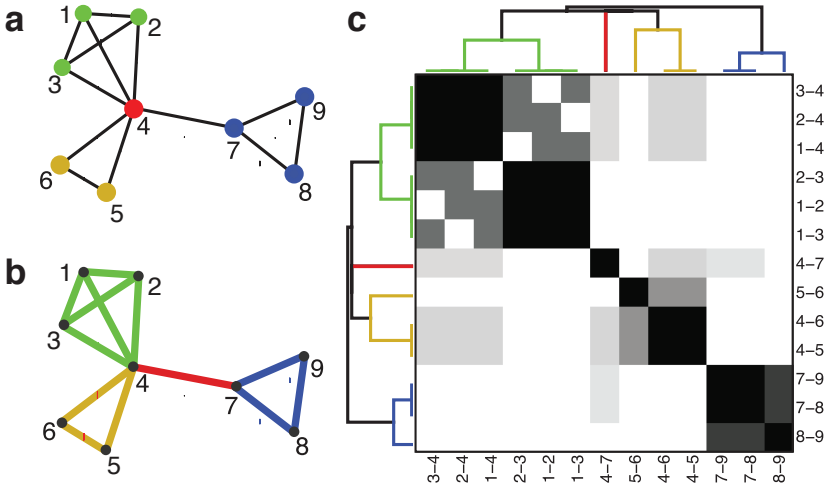
Az Ahn-módszer [11] legfontosabb újítása a szélesebb körben elterjedt közösségfelismerő módszerekhez képest, hogy a csúcsok helyett az éleket sorolja csoportokba. Bár első látásra ez kevésbé tűnik természetesnek, azonban így figyelembe tudjuk venni az átfedéseket is a különböző közösségek között, míg ez a csúcsok csoportokba sorolásakor nem lehetséges. A szociális hálózatban maga a közösség definiálja a kapcsolat típusát, ezért is lehet hasznos a linkalapú szemlélet.

Az Ahn-módszer ún. hierarchikus klaszterezést hajt végre [7, 12]. Ennek első lépése a hálózatra vonatkozó hasonlósági mátrix felépítése, azaz annak megadása, hogy két él mennyire hasonlít egymáshoz. Két él hasonlóságának számszerű jellemzésére a (2.1) egyenlet szerinti S mennyiség szolgál, ahol e_{ik} és e_{jk} rendre az i és k, valamint a j és k csúcsokat összekötő éleket jelöli, míg az n_+ kifejezés az adott csúcshoz tartozó szomszédos csúcsok listáját takarja, amelyhez még hozzávesszük az adott csúcsot is.

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \tag{2.1}$$

Így az algoritmus a 2.3a ábrán látható mintahálózatból a 2.3c ábrán látható hasonlósági mátrixot hozza létre. Ezt követi az agglomeratív hierarchikus osztályozás, amely során folyamatosan újabb éleket adunk hozzá az aktuálisan vizsgált közösséghez. A 2.3a ábra alapján megállapítható, hogy például az 1-3 és 1-2 élek környezete teljesen megegyezik, így ezek egy szintre kerülnek a hierarchiában. Ezzel szemben a 4-6 és 5-6 élek környezete már jobban eltér, ezért azok különböző szinteken állnak. Az osztályozás eredményeként egy dendrogram (hierarchikus fa) áll elő, melyet egy jól meghatározott horizontális vonallal elvágva megkapjuk a hálózatban megtalálható közösségeket.

A módszer javaslatot ad a vágás helyére is: egy statisztikailag releváns elvágási sűrűséget (partition density) definiál, és annak maximumánál vágja el a dendrogramot. A mintahálózat alapján felépített, és a felismert közösségek szerint beszínezett hierarchikus fa a 2.3c ábrán, a mátrix mellett látható.

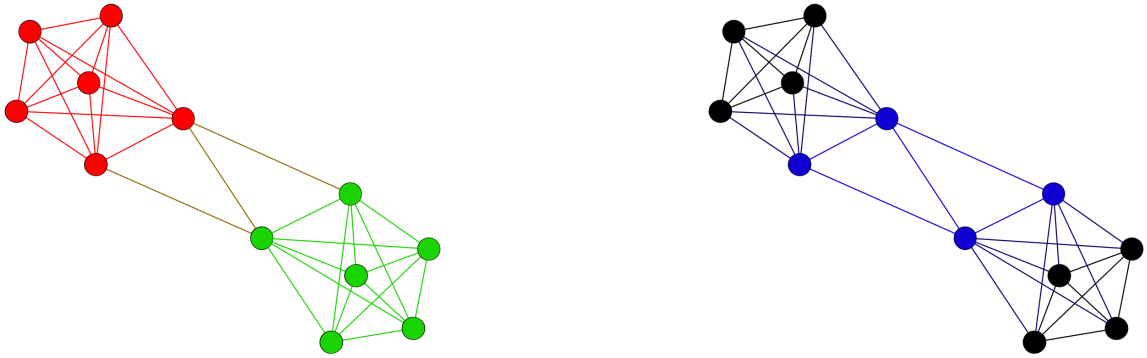


2.3. ábra. Az Ahn-módszer bemutatása egy egyszerű példán keresztül [11]

Az Ahn-módszer futásideje a csúcsok számával négyzetesen nő, azaz futása gyorsabb, mint a klikk perkolációs algoritmusé. Ezzel együtt rendelkezik mindazon tulajdonságokkal, amelyek a klikk perkolációt is alkalmassá teszik a társadalmat reprezentáló hálózatokban a közösségek felderítésére, sőt, gyakoriak az olyan esetek, amelyekben a két eljárás közül csak az Ahn-módszer talál meg egy adott közösséget. A 2.4 ábrán egy ilyen helyzetet mutatunk be.

A 2.4 ábrán látható hálózatot a 3-klikk perkolációs algoritmus egy összefüggő közösségként ismeri fel, míg a 4-/5-/6-klikk perkoláció a piros és zöld színnel jelölt közösségeket találja meg. A kék csúcsok által alkotott közösséget azonban egyik esetben sem ismeri fel a klikk perkoláció. Ezzel szemben az Ahn-módszer külön-külön felismeri a két nagy közösséget (piros és zöld), és külön csoportként tünteti fel a 2.4b ábrán látható, kék színnel jelölt csúcsokat is. Úgy gondoljuk, hogy a két nagy közösség közötti kapcsolatok valóban egy független csoport jelenlétére utalnak.

A jelentősen kisebb futásidő és a hálózat közösségeiről adott kép nagyobb részletessége miatt munkánk során az Ahn-módszer segítségével derítettük fel a vizsgált hálózatokban rejlő közösségeket.



(a) Két nagy közösség

(b) Összekötő közösség

2.4. ábra. Az Ahn-módszer előnyei

2.4. Társadalmi predikciók

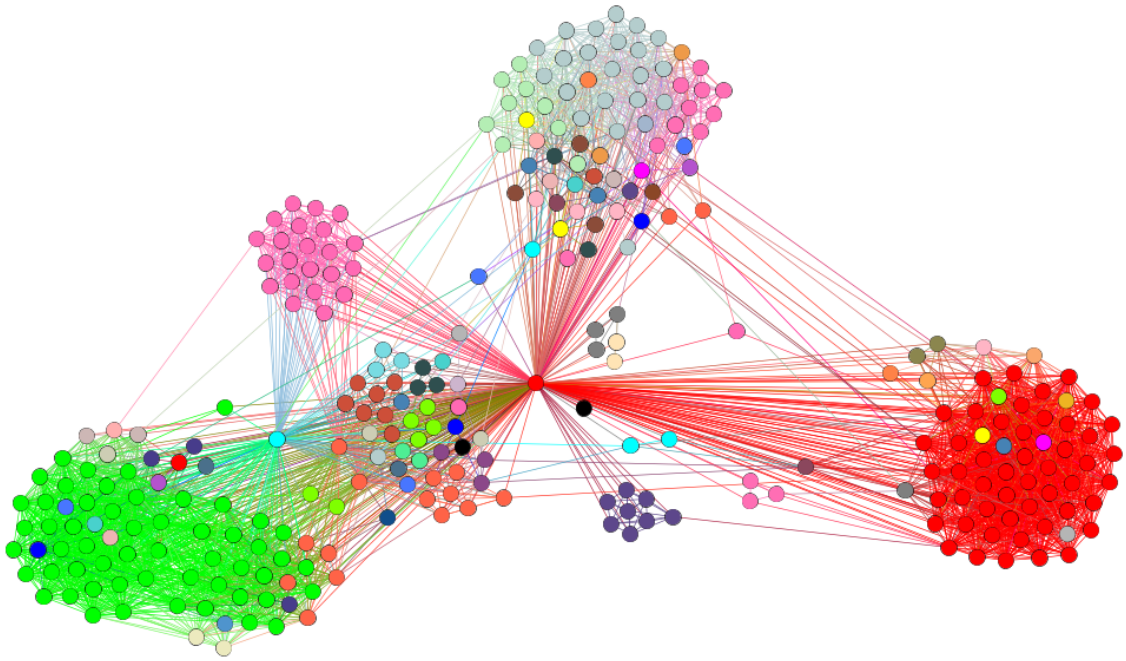
Társadalmi predikciónak nevezzük az olyan eljárásokat, melyek során egy kidolgozott elmélet segítségével valamit előre kívánunk jelezni. A vírusterjedés esetében például létezik jól működő predikciós eljárás [13]. Azonban az emberek tulajdonságait, viselkedését megbecsülni sokkal nehezebb. A predikciós problémák bevett megoldási módja az, hogy első lépésként feltárjuk a probléma befolyásoló tényezőit, azokat egy-egy értékkel jellemezzük, majd a paramétereket lineáris regressziós eljárással optimalizáljuk [14]. Mivel a befolyásoló tényezőket gyakran nagyon sok paraméter írja le, ezért előfordulhat, hogy végül elvész az a néhány hatás, amely döntően befolyásolja a vizsgált rendszert. Bár a lineáris regresszión alapuló optimalizációs eljárást széles körben használják, számunkra ez nem járható út, mivel a legtöbb szociális kapcsolatokat leíró adatbázisban nem áll rendelkezésre megfelelően nagy számú paraméter.

3. fejezet

Az életkor meghatározásához használt eljárás

Módszerünk lényege, hogy az egyén ismerőseinek kapcsolatrendszeréből (egocentrikus hálózat) felderített közösségek átlagéletkora szoros kapcsolatban van az egyén életkorával. Az ismerősök általában hasonló korúak, mint az ego, vagy körülbelül 25 évvel idősebbek/fiatalabbak (gyerek-szülő generáció) [15], ami hisztogram alapú technikával jól szétválasztható. Az életkor meghatározására általunk kidolgozott és használt eljárás a már ismertetett Ahn-módszer [11] segítségével felismert közösségeken alapul. A hálózatban található közösségek felismerésére a témavezetőm kutatócsoportja által létrehozott közösségfelismerő programot használtuk [16].

A következőkben a módszer működésének leírását minden lépés során egy valós mintapéldán keresztül - egy 300 ismerőssel rendelkező iWiW felhasználó egocentrikus hálózatának segítségével - szemléltetjük. Ezen felhasználó egocentrikus hálózatát a 3.1 ábra mutatja, melyen különböző színekkel jelöltük az Ahn-módszer segítségével felfedezett közösségeket. Az ábrát a 2.2 fejezetben leírtak szerint készítettük el. A színezés során a legnagyobb létszámú közösségtől haladtunk a kisebbek felé, egy közösséget ugyanazzal a színnel jelöltünk, és ha egy csúcsot már beszíneztünk, később nem színeztük át.



3.1. ábra. A közösségek alapján rendezett egocentrikus mintahálózat

3.1. Adatok szűrése

Módszerünk vizsgálata során az adatbázisokból megadott szempontok alapján válogattuk ki a tanulmányozandó felhasználókat. Az első és legfontosabb kritérium az volt, hogy az egyének életkorának ismertnek kellett lennie, hiszen csak így tudtuk összevetni az algoritmusunk által becsült életkorértéket egy valódi adattal. A kiválogatás során 10 és 80 év közötti felhasználókat vettünk figyelembe. Emellett a rendelkezésünkre álló számítási kapacitás korlátozottsága miatt az összes egyén egocentrikus hálózatát nem tudtuk vizsgálni, ezért csak adott számú ismerőssel rendelkező felhasználókat választottunk ki. Mivel az iWiW hálózata sűrű, ezért abból meghatározott számú ismerőssel rendelkező felhasználókat tudtunk kiválogatni, és így csak egy-egy konkrét ismerősszám tartozott egy-egy kategóriához a módszer tesztelése során. Ezzel szemben a telefonos hálózatban megfelelően nagy számú kapcsolattal rendelkező ügyfelek csak korlátozottan álltak rendelkezésünkre, így az ismerősszámokra vonatkozóan tartományokat jelöltünk ki, és az ezen tartományokból létrehozott kategóriákban külön-külön vizsgáltuk módszerünk hatékonyságát. A 3.1 és a 3.2 táblázatokban a kiválasztott egok darabszámát tüntettük fel.

Ismerősök száma	Darabszám
50	5408
100	7060
200	5735
300	4247

3.1. táblázat. A kiválasztott egyének száma - iWiW

Ismerősök száma	Darabszám
60-69	11142
70-79	6115
80-89	3372
90-109	3448
110-149	2200
150-249	1037
250-499	245

3.2. táblázat. A kiválasztott egyének száma - telefonos adatok

3.2. Közösségek elemzése

Első lépésként minden kiválasztott egyén esetében a témavezetőm kutatócsoportja által létrehozott közösségfelismerő program [16] segítségével felderítettük az egocentrikus hálózatban megtalálható közösségeket. A program futásidejének csökkentése érdekében előzetesen minden egocentrikus hálózatból eltávolítottuk az egot és a hozzá tartozó éleket. Ez a végeredményt nem befolyásolta, hiszen az ego - definíció szerint - az egocentrikus hálózat minden csúcsához kapcsolódik, így nem jelent információveszteséget, ha nem tagja az egocentrikus hálózatnak. A program futásának eredményeként az összes kiválasztott egyén esetében rendelkezésünkre állt az azok egocentrikus hálózatában megtalálható közösségeket tartalmazó adatfájl. Következő lépésként ezen adatfájlokat dolgoztuk fel.

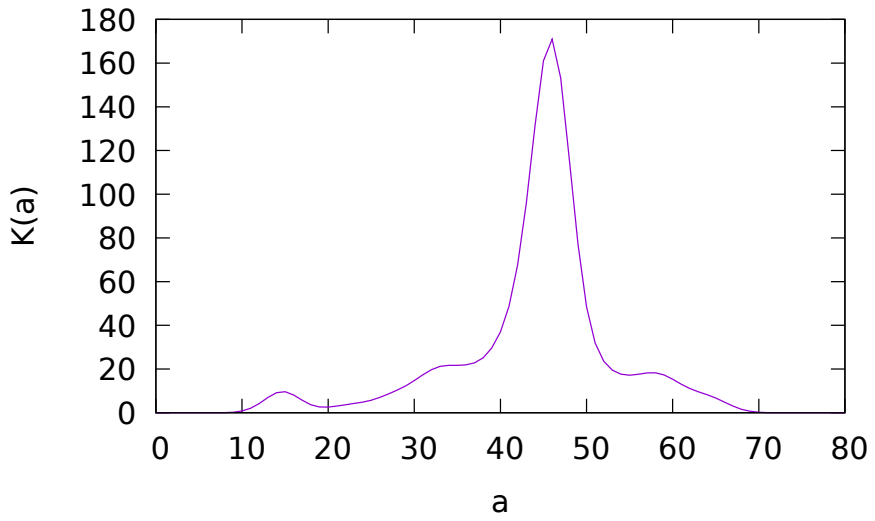
Egy adatfájlon végighaladva minden egyes közösségre meghatároztuk az azt alkotó egyének életkorának átlagát. Csak azokat a közösségeket tudtuk figyelembe venni,

amelyekben legalább egy egyén életkora ismert volt. Ezt követően kiszámítottuk a kapott értékek mozgóátlagát. A mozgóátlag-számítás során a (3.1) egyenlet szerinti Gauss-függvényt felhasználva, a (3.2) összefüggés szerinti diszkrét konvolúciós eljárást hajtottuk végre, ahol a_i az i-edik csoport átlagéletkorát, míg C a csoportok számát jelöli, σ_g pedig a Gauss-függvény szórása. 1 és 3 közötti σ_g értékekkel tesztelve az algoritmusunkat megállapítottuk, hogy a kapott eredmény szempontjából nincs nagy jelentősége (maximum két-három százalékos különbséget okoz) σ_g pontos értékének, ezért vizsgálatainkat σ_g=2 érték mellett végeztük.

$$g(a) = \frac{1}{\sigma_g \sqrt{2\pi}} \mathrm{e}^{-\frac{a^2}{2\sigma_g^2}} \tag{3.1}$$

$$K(a) = \sum_{i=1}^{C} g(a - a_i) \cdot 1 \tag{3.2}$$

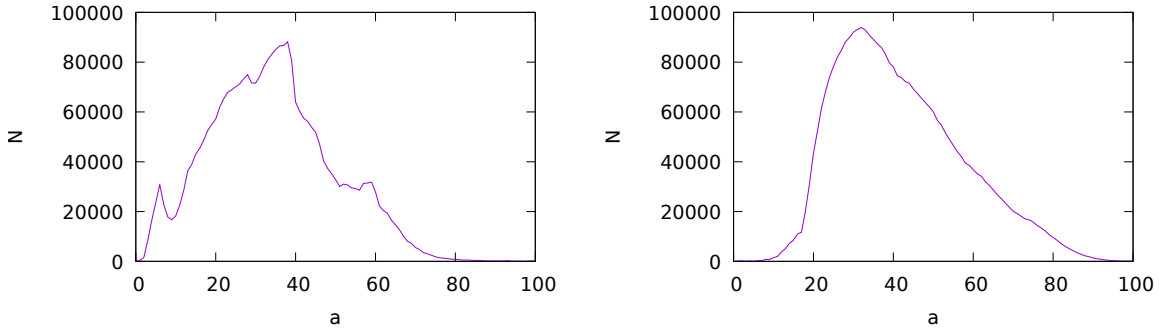
A mozgóátlag-számítás során az egyes csoportok átlagéletkorát ugyanolyan súllyal vettük figyelembe, mivel egy-egy érték relevanciája nem hozható egyértelmű kapcsolatba a csoport méretével. Minden egyes ego esetében hisztogramon ábrázoltuk a kapott mozgóátlagot, így téve szemléletessé a közösségek korstruktúráját. A 3.2 ábrán a mintahálózat (3.1 ábra) alapján elkészített hisztogram látható.



3.2. ábra. A közösségek számának eloszlása a mintahálózatban, az életkor függvényében

3.3. A megfelelő csúcs kiválasztása

A hisztogramon a legtöbb esetben jól lokalizálható csúcsokat fedezhetünk fel. Ezek száma lehet egy, kettő, három, vagy akár háromnál több is, az adott egocentrikus hálózat korstruktúrájától függően. A mintahálózat hisztogramján (3.2 ábra) egy nagy, illetve két kisebb csúcsot láthatunk. Megfigyeléseink szerint a csúcsokhoz tartozó életkorértékek valamelyike nagy eséllyel kapcsolatba hozható az egyén életkorával. Ezt igazolandó egy olyan algoritmust hoztunk létre, amely megkeresi a csúcsokat, és kiválasztja azok közül a legrelevánsabbat. Ahhoz, hogy a csúcsokat kvantitatív módon tudjuk jellemezni, egy megfelelő mennyiség definiálására volt szükség. Az, hogy milyen mennyiség lenne a legjobban alkalmas erre a célra, egyáltalán nem triviális. Elsőként - feltételezve, hogy a valódi életkorhoz közeli átlagéletkorú csoportok száma a legnagyobb - a legmagasabb csúcsot kerestük, azaz a csúcsokat a hozzájuk tartozó $K(a)$ értékekkel jellemeztük. Bár ez első ránézésre logikus gondolatnak tűnhet, azonban több hisztogram vizsgálata során világossá vált, hogy a középkorúakhoz tartozó csúcsok sok esetben akkor is a legmagasabbak voltak, ha az ego valójában 20-25 évvel fiatalabb vagy idősebb volt. Ez valószínűleg annak köszönhető, hogy mindkét hálózatban felülreprezentáltak voltak a középkorúak, ami a 3.3a és a 3.3b számú ábrákon jól látható. Az ábrákon az a életkor függvényében ábrázoltuk az ezen a életkorral rendelkező egyének N számát az adott hálózatban.



(a) iWiW

(b) Telefonos adatok

3.3. ábra. Egyének száma az életkor függvényében

További vizsgálataink során a legcélszerűbb mennyiségnek a (3.3) összefüggés szerinti, v-vel jelölt mennyiség adódott, ahol h a csúcshoz tartozó $K(a)$ érték, w pedig a csúcshoz tartozó félérték-szélesség.

$$v = \frac{h}{w} \tag{3.3}$$

Az így definiált mennyiség segítségével jól meg tudjuk ragadni azt, hogy egy csúcs mennyire „éles” vagy „lapos”. Emellett v-nek a (3.4) összefüggés szerinti arányossága áll fenn a szórásnégyzettel.

$$\frac{1}{v} \sim \sigma^2 \tag{3.4}$$

Az általunk megírt algoritmus lokális maximumhely-keresés segítségével azonosítja a csúcsokat, majd minden csúcsra vonatkozóan kiszámítja az adott csúcshoz tartozó v értéket. Végül a legmagasabb v értékkel rendelkező csúcsot kiválasztva kapjuk meg a legrelevánsabb csúcsot. Az ezen csúcshoz tartozó életkorérték az algoritmus által becsült életkora az egonak.

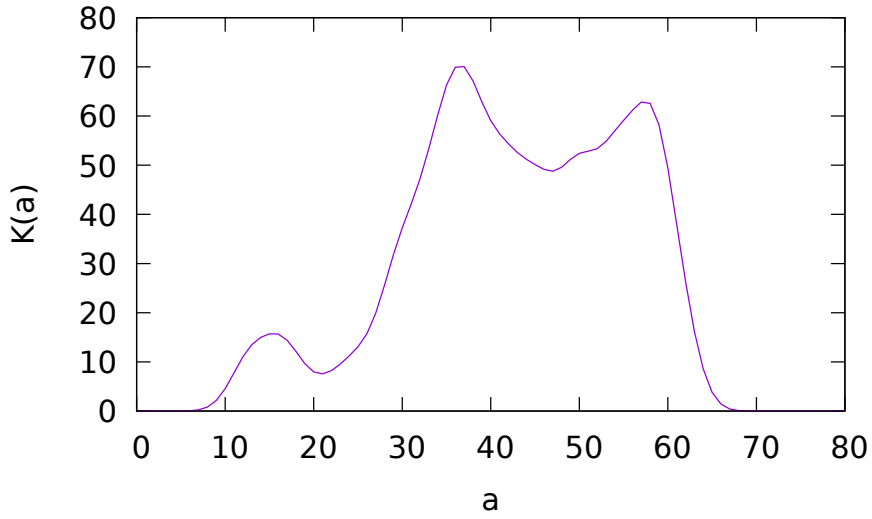
A mintahálózathoz tartozó adatokon (hisztogram: 3.2 ábra) lefuttatva az algoritmut, a 3.3 táblázat szerinti eredményt kaptuk. Ebben az esetben algoritmusunk becslése kiváló, a becsült életkor megegyezik a valódi életkorral (amely éppen 46 év).

Rang	Életkorérték (év)	v
1.	46	24,42
2.	15	1,61
3.	58	0,49

3.3. táblázat. A mintahálózathoz tartozó hisztogramon megtalált csúcsok

A hisztogram sajnos nem mindig mutat ilyen szép és egyértelmű képet. A 3.4 számú ábrán egy olyan ellenpéldát mutatunk be, amely jól szemlélteti módszerünk korlátait. Az ellenpélda hisztogramján (3.2 ábra) lefuttatva az algoritmut, a 3.4 táblázatban olvasható eredményeket kaptuk. A hisztogramon két nagy csúcs is látható, melyek meglehetősen hasonló alakúak. Ezt jól tükrözi az algoritmus által kiszámolt, az 1. és

2. ranggal rendelkező csúcsokhoz tartozó v értékek kis különbsége is, azonban a legrelevánsabbként kiválasztott csúcs sajnos nem a valódi életkorhoz (59 év) tartozó csúcs.



3.4. ábra. A közösségek számának eloszlása az életkor függvényében - ellenpélda

Rang	Életkorérték (év)	v
1.	37	2,12
2.	57	1,85
3.	15	1,57

3.4. táblázat. Az ellenpéldaként bemutatott hisztogramon megtalált csúcsok

3.4. A kapott életkor pontosságának becslése

Abban az esetben, ha az ego valódi életkora mégsem lenne ismert, a becsült életkor pontosságát nem tudnánk megállapítani. Ahhoz, hogy ez mégis lehetővé váljon, egy olyan mennyiség definiálására van szükség, amely kvantitatív módon képes jellemezni az eredmény pontosságát. Esetünkben csak az egyes csúcsokhoz tartozó v értékekre támaszkodhattunk, így kézenfekvőnek tűnt, hogy a sikeresség mércéjeként az első két csúcshoz tartozó életkorértékek v_1/v_2 hányadosát tekintsük. Amennyiben ez a hányados kellően nagy, a második legjobb csúcshoz tartozó v érték jelentősen elmarad az

első csúcshoz tartozóétól. Ekkor valóban az első csúcsot tekinhetjük releváns csúcsnak, tehát egy ilyen esetben kapott életkor várakozásaink szerint közel esik az egyén valódi életkorához.

4. fejezet

Eredmények

Az általunk megalkotott algoritmus több korábbi eljárástól [17, 18] eltérően nem használ gépi tanulási módszereket, így a program futása gyors. A tesztelést a 3.1 fejezetben ismertett módon megszűrt adatokon végeztük.

4.1. iWiW

4.1.1. A legrelevánsabb csúcs vizsgálata

Az iWiW adatainak vizsgálata során, az általunk kidolgozott módszer segítségével elért eredményeinket a 4.1 táblázatban foglaltuk össze.

Hiba (év)	Sikeresség (%)			
	50	100	200	300
± 1	38,68	67,46	76,22	68,17
± 2	47,21	74,94	85,09	81,49
± 3	52,40	78,05	87,25	85,73
± 4	55,62	79,76	88,65	87,85
± 5	58,41	81,36	89,52	89,15
± 6	60,89	82,54	90,31	90,30

4.1. táblázat. Az életkorbecslés sikeressége adott számú ismerős esetén - iWiW

A sikerességi adatok vizsgálata során megfigyelhető, hogy 200 ismerős mellett volt a legsikeresebb az algoritmus, attól távolodva csökkent a sikeres életkorbecslés aránya. Nem mindegy azonban, hogy milyen irányú távolodást tekintünk. A csökkenés mértéke az ismerősszám növelésével kisebb mértékű volt, mint az ismerősszám csökkentése során. Érdemes megemlíteni, hogy a 200-as ismerősszám esik a vizsgált kategóriák közül a legközelebbb az iWiW hálózatának átlagos fokszámához.

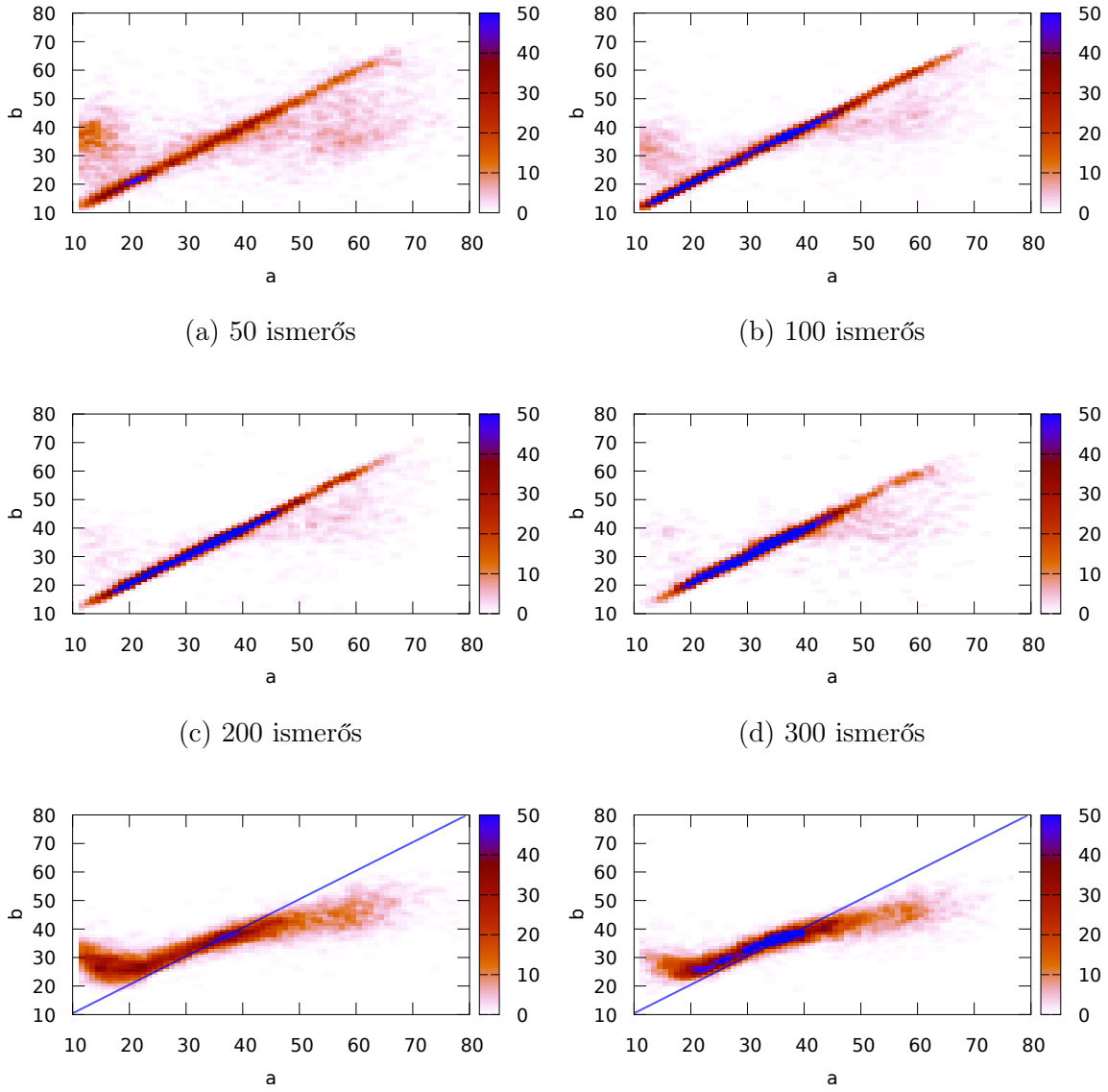
Ahhoz, hogy módszerünk hatékonyságát pontosabban ítélhessük meg, az ismerősök életkorának egyszerű átlagolásával is megbecsültük az egyének életkorát, az így kapott eredményeket a 4.2 táblázat tartalmazza. Jól látható, hogy az általunk kidolgozott, az egocentrikus hálózatban megtalálható közösségeken alapuló módszer sokkal jobban teljesít az egyszerű átlagoláshoz képest, főként akkor, ha pontos, ±2 éves becslésekre vagyunk kíváncsiak (34 % vs. 85 %, 200 ismerős esetén).

Hiba (év)	Sikeresség (%)			
	50	100	200	300
± 1	13,52	15,67	20,19	21,94
± 2	21,47	25,61	33,57	37,74
± 3	28,90	34,49	46,50	52,48
± 4	34,89	42,29	56,86	63,74
± 5	39,89	49,19	64,69	71,96
± 6	44,66	54,87	70,58	78,15

4.2. táblázat. Az életkorbecslés sikeressége egyszerű átlagolással - iWiW

A 4.1 ábrákon a becsült b életkort ábrázoltuk az a valódi életkor függvényében. Az általunk kidolgozott módszer segítségével becsült eredmények a 4.1a-4.1d, míg az átlagolással számítottak a 4.1e-4.1f grafikonokon láthatóak. Az, hogy egy vékony, ±2 év szélességű sávot látunk a 4.1a-4.1d ábrákon, módszerünk sikerességét tanúsítja. A hibás életkorbecsléseket tekintve megfigyelhetjük, hogy algoritmusunk a hálózatban megtalálható egyének életkorainak mediánja ($\sim$ 40 év) alatt felül-, míg a medián felett alulbecsülte a valódi életkort.

Továbbá azt is észrevehetjük, hogy ezekben az esetekben a hibásan becsült értékek vagy éppen a medián közelébe, vagy a valódi élekortól körülbelül ±25 évre esnek. Utóbbi megfelel a korábbi kutatások [15] eredményeinek.



(a) 50 ismerős

(b) 100 ismerős

(c) 200 ismerős

(d) 300 ismerős

(e) 100 ismerős, egyszerű átlagolás

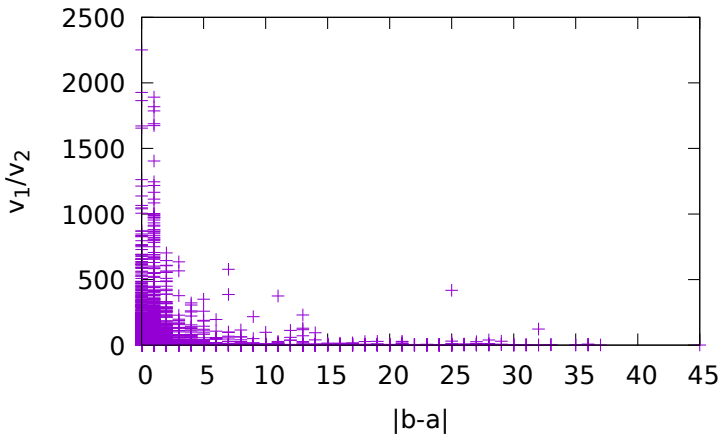
(f) 200 ismerős, egyszerű átlagolás

4.1. ábra. A becsült életkor a valódi életkor függvényében - iWiW

Ezzel szemben az átlagolás során kapott ábrák (4.1e-a 4.1f) gyökeresen eltérnek a módszerünkkel kapott eredményeket ábrázoló 4.1a-4.1d grafikonoktól. Teljesen más a becsült értékekhez tartozó pontok elhelyezkedése. Leolvasható, hogy az egyszerű átlagolás egy körülbelül 30 év szélességű sávban (közelítőleg $b \in [20; 50]$) ad eredményeket,

azaz a $[10; 80]$ intervallum felét sem fedik le az átlagolással kapott becslések. Az ábrán látható sáv jóval szélesebb, és jelentősen eltér a kék színnel berajzolt $b = a$ egyenestől is. Nagyjából az átlagos életkort kapjuk vissza nagyon kicsi korfüggéssel, amely a fiataloknál erősebb, hiszen rájuk jobban jellemző, hogy saját kortársaikkal vannak kapcsolatban.

A módszerünk segítségével kapott eredmények pontosságának becslését a 3.4 fejezetben tárgyaltak szerint kíséreltük meg. A 4.2 ábrán látható, hogy várakozásainknak megfelelően a nagy értékekhez meglehetősen pontos becslések tartoznak. Fordítva ez azonban sajnos nem igaz, azaz kis értékek esetén nem minden esetben pontatlan a becslés. Továbbá megállapíthatjuk, hogy a pontok egy jelentős része az origó közelébe esik, így sok olyan kis hibájú becslésünk van, melyekre a v_1/v_2 hányados kicsi.



4.2. ábra. Az eredmények pontosságának becslése 200 ismerős esetén - iWiW

Módszerünk átlagosan kifejezetten jól teljesít, azonban sajnos egy konkrét becslés hitelességéről semmilyen információt nem tudunk adni. A jövőben mindenképpen szeretnénk ezt a problémát is orvosolni.

4.1.2. Több csúcs együttes vizsgálata

Módszerünk javítására több lehetőségünk is van, például a csoportokat lehetne szűrni, vagy esetleg súlyozni az életkoreloszlás alapján, illetve a legjobb csúcs kiválasztásánál lehetnénk hatékonyabbak. Ez utóbbit könnyen meg tudjuk vizsgálni azzal, ha nem csak egy csúcsot, hanem a legjobb három csúcsot vesszük figyelembe, ezzel felderítve az

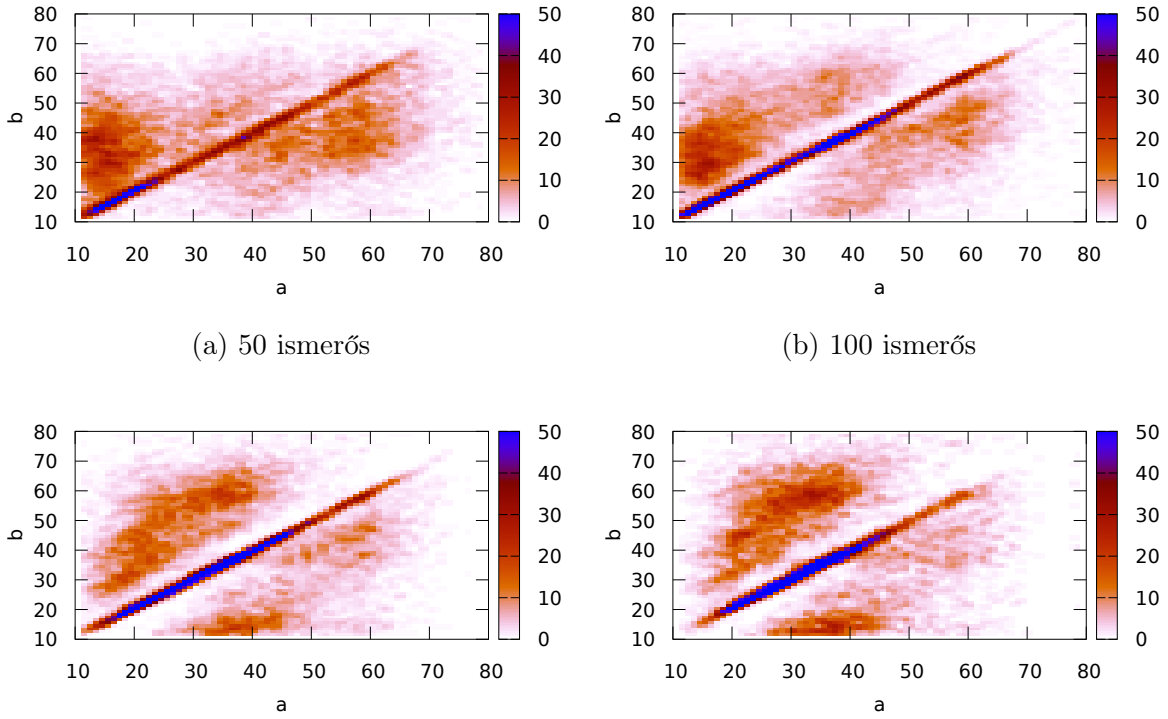
ebben rejlő tartalékot. Ezt úgy valósítottuk meg, hogy ha a kapott három csúcs közül az egyik ±2 távolságban volt a valódi életkortól, akkor azt sikeres becslésnek tekintettük. Az így kapott kibővített sikerességi arányokat - ±2 év maximális hiba mellett - a 4.3 táblázatban tüntettük fel.

Ismerősszám	Kibővített sikeresség (%)
50	62,63
100	85,24
200	90,44
300	85,73

4.3. táblázat. Az életkorbecslés kibővített sikeressége - iWiW

Látható, hogy maximum 5-15 % javulást lehet így elérni, ami különösen látványos a kis fokszámú egyének esetében. A jövőben meg kívánjuk vizsgálni, hogy a kis egocentrikus hálózatoknál hogyan lehet jobban kiválasztani a releváns csúcsot.

A 4.3 ábrákon egy közös grafikonon ábrázoltuk a három legmagasabb v értékkel rendelkező, azaz a legrelevánsabbnak tekinthető csúcsokhoz tartozó életkorértékeket a valódi életkor függvényében. Látható, hogy körülbelül ±25 év távolságban a becsült értékek „felhőszerűen” követik a valódi értékeket [15].



(a) 50 ismerős

(b) 100 ismerős

(c) 200 ismerős

(d) 300 ismerős

4.3. ábra. A legrelevánsabb csúcsokhoz tartozó életkorértékek - iWiW

4.2. Telefonos adatok

A telefonszolgáltatótól kapott adatok vizsgálata során elért eredmények elmaradtak az iWiW adatain történő tesztelések eredményétől. Ennek több oka is lehet:

- nem áll rendelkezésre az összes kapcsolati adat, mivel a telefonos adatok egy adott szolgáltatótól származnak - így ha valaki más szolgáltatónál lévő telefonszámot hívott, a hívott félről már nem lesz elérhető információ
- az iWiW-vel ellentétben az emberek nem csak ismerőseikkel állnak kapcsolatban, hanem gyakran ismeretlenekkel is, akik így irrelevánsnak mondhatók az életkor meghatározásának szempontjából
- a mobilszolgáltatóval megkötött szerződés szerinti tulajdonos sokszor nem egyezik meg a telefon használójával

- főképp a 20-30 éves korosztály használta 2008 körül aktívan a mobiltelefonját, ezért az idősebb emberekről egyszerűen kevesebb adat áll rendelkezésre
- a nagy fokszámú egyének nagy valószínűséggel hivatali ügyintézésre használják telefonjukat, ezért romlik erősen a becslés nagy fokszámok mellett

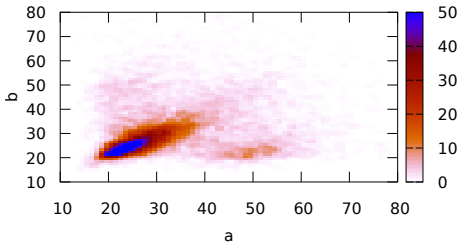
Az adatokat elemezve látható, hogy a kapcsolatok számának növelésével a sikeresség jelentősen romlott. Az iWiW hálózatában az átlagos fokszám körülbelül 220 volt, míg a telefonos adatok esetében körülbelül 7, azaz a telefonos adatok között már egy 60 ismerőssel rendelkező egyén egocentrikus hálózata is nagyon nagynak számít, valószínűleg nem csak ismerősökből áll. Ennek tükrében nem meglepő, ha a telefonos adatok esetében több száz kapcsolat mellett alig tudunk mondani valamit az egyén életkoráról. Azonban 100 alatti ismerősszámra eredményeink elfogadhatónak tekinthetők.

A telefonos adatok vizsgálata során elért eredményeinket a 4.4 táblázatban foglaltuk össze.

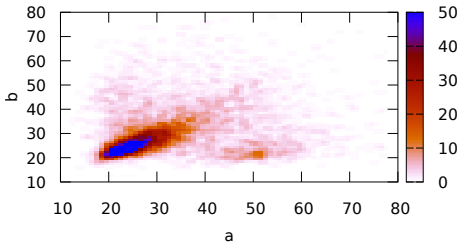
Hiba (év)	Sikeresség (%)						
	60-69	70-79	80-89	90-109	110-149	150-249	250-499
± 1	22,31	23,01	22,74	21,65	20,54	15,87	17,62
± 2	34,87	35,10	34,95	33,33	32,01	25,80	22,13
± 3	44,17	44,96	44,18	42,63	40,71	31,65	27,46
± 4	51,33	52,06	51,48	50,03	47,77	38,95	34,02
± 5	56,84	57,27	56,71	55,01	53,51	44,11	39,34
± 6	61,21	61,13	61,46	59,18	58,01	50,34	45,08

4.4. táblázat. Az életkorbecslés sikeressége adott számú ismerős esetén - telefon

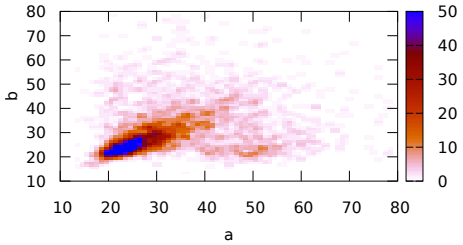
A 4.1 ábrákon kategóriánként ábrázoltuk a becsült b életkort az a valódi életkor függvényében.



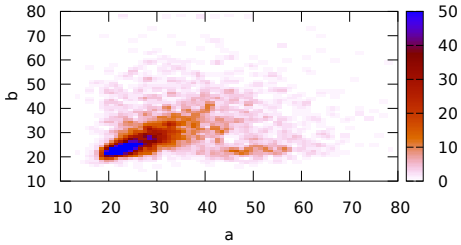
(a) 60-69 ismerős



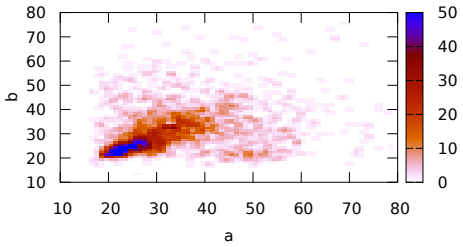
(b) 70-79 ismerős



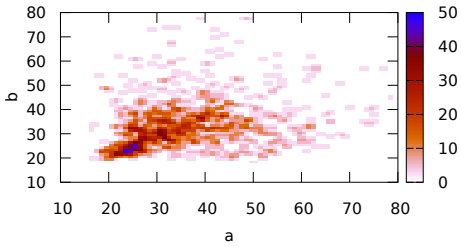
(c) 80-89 ismerős



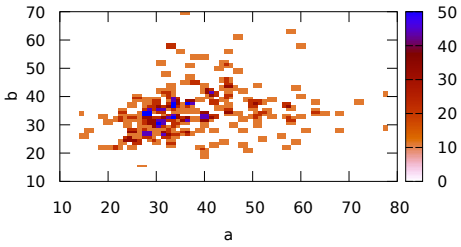
(d) 90-109 ismerős



(e) 110-149 ismerős



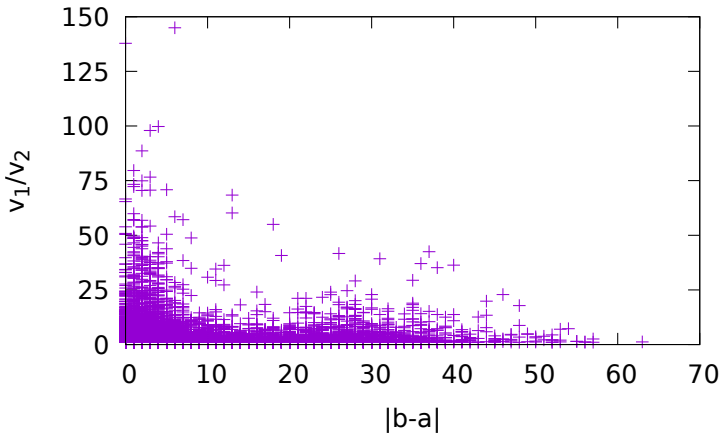
(f) 150-249 ismerős



(g) 250-499 ismerős

4.4. ábra. A becsült életkor a valódi életkor függvényében - telefon

A kapott értékek pontosságának becslését a 3.4 fejezetben tárgyaltak szerint kíséreltük meg. Az iWiW-hez hasonlóan ez a telefonos adatok esetében sem vezetett eredményre, mint az a 4.5 ábrán is látható.



4.5. ábra. Az eredmények pontosságának becslése 60-69 ismerős esetén - telefon

Ahhoz, hogy módszerünk hatékonyságát pontosabban ítélhessük meg, az ismerősök életkorának egyszerű átlagolásával is megbecsültük az egyének életkorát, a kapott eredményeket a 4.5 táblázat tartalmazza. Jól látható, hogy az általunk kidolgozott, az egocentrikus hálózatban megtalálható közösségeken alapuló módszer sokkal jobban teljesít az egyszerű átlagolásnál.

Hiba (év)	Sikeresség (%)						
	60-69	70-79	80-89	90-109	110-149	150-249	250-499
± 1	6,40	6,40	5,79	6,18	4,55	5,45	4,51
± 2	10,55	10,28	9,47	9,12	7,06	8,57	6,15
± 3	14,24	14,29	12,65	12,35	9,79	12,46	7,79
± 4	17,64	17,11	15,86	14,83	12,34	15,09	10,25
± 5	20,29	20,04	17,90	17,45	14,21	17,04	13,93
± 6	22,85	22,62	19,83	19,76	16,26	18,89	15,16

4.5. táblázat. Az életkorbecslés sikeressége átlagolással - telefon

5. fejezet

Összefoglalás, kitekintés

Célunk az volt, hogy az egyének életkorát egy egyszerű módszer segítségével minél pontosabban megbecsüljük. Ehhez egy olyan algoritmust alkottunk meg, amely nem használ gépi tanulási módszereket, így a program futása gyors, nem sztochasztikus.

Módszerünk lényege, hogy az egyén ismerőseinek kapcsolatrendszeréből (egocentrikus hálózat) felderített közösségek életkora szoros kapcsolatban van az egyén életkorával. Az ismerősök általában hasonló korúak, mint az ego, vagy körülbelül 25 évvel idősebbek/fiatalabbak (gyerek-szülő generáció) [15], ami hisztogram alapú technikával jól szétválasztható, a releváns csúcsot pedig annak szórása alapján választjuk ki. Az algoritmus hatékonyságát az iWiW-től és egy telefonszolgáltatótól származó adatokon teszteltük. Az iWiW-en az algoritmus meglehetősen hatékonynak bizonyult, az esetek $\sim$ 80%-ában maximum 2 év hibával visszaadta a helyes életkort. A legnagyobb hatékonyságot akkor értük el, amikor az ego iWiW-es kapcsolatrendszere legjobban hasonlított a valódi szociális hálózatára, azaz a kapcsolatok száma 200 körül volt.

A módszer a telefonos adatok vizsgálata során is alkalmazható volt, bár kisebb hatékonysággal. Ennek egyik legvalószínűbb oka az lehet, hogy a telefonhívások között sok munkaügyben történt hívás is megtalálható, amik nem tekinthetőek valódi szociális kapcsolatnak. A jövőben a közösségek megfelelő szűrésével, illetve a hisztogram elemzésének finomításával pontosabb eredmények elérése is lehetségessé válhat. Emellett hasonló elv alapján - természetesen az eljárás megfelelő módosításával - összetettebb tulajdonságok (így például a lakóhely, a látogatott oktatási intézmények, a munkhely) vizsgálatára is lehetőség nyílhat.

Irodalomjegyzék

[1] S. Wolfram, *A new kind of science*. General science, Wolfram Media, 2002.

[2] I. Asimov, *Az Alapítvány trilógia*. Gabo Kiadó, 2010.

[3] J. Török, G. Iñiguez, T. Yasseri, M. San Miguel, K. Kaski, and J. Kertész, „Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment," *Physical Review Letters*, vol. 110, no. 8, p. 088701, 2013.

[4] Z. Ruan, G. Iniguez, M. Karsai, and J. Kertész, „Kinetics of social contagion," *Physical Review Letters*, vol. 115, no. 21, p. 218702, 2015.

[5] R. Dunbar, „The social brain hypothesis," *Brain*, vol. 9, no. 10, pp. 178–190, 1998.

[6] Eurostat, „Households with broadband access." URL: http://ec.europa.eu/eurostat/tgm/table.do?tab=table&plugin=1&language=en&pcode=tin00073. Online; accessed 20-October-2016.

[7] A.-L. Barabási, *Network Science*. Cambridge University Press, 2016.

[8] M. Bastian, S. Heymann, and M. Jacomy, „Gephi: An open source software for exploring and manipulating networks," 2009.

[9] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, „Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PLoS ONE*, vol. 9, no. 6, p. e98679, 2014.

[10] I. Derényi, G. Palla, and T. Vicsek, „Clique percolation in random networks," *Physical Review Letters*, vol. 94, no. 16, p. 160202, 2005.

[11] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, „Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[12] Wikipedia, „Hierarchical clustering." URL: https://en.wikipedia.org/wiki/Hierarchical_clustering. Online; accessed 20-October-2016.

[13] R. Pastor-Satorras and A. Vespignani, „Epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, no. 14, p. 3200, 2001.

[14] M. Mestyán, T. Yasseri, and J. Kertész, „Early prediction of movie box office success based on wikipedia activity big data," *PLoS ONE*, vol. 8, no. 8, p. e71226, 2013.

[15] V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. Dunbar, „Sex differences in intimate relationships," *Scientific Reports*, vol. 2, 2012.

[16] Y. Murase, J. Török, H.-H. Jo, K. Kaski, and J. Kertész, „Multilayer weighted social network model," *Physical Review E*, vol. 90, no. 5, p. 052810, 2014.

[17] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, „Predicting age and gender in online social networks," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 37–44, ACM, 2011.

[18] D. Nguyen, N. A. Smith, and C. P. Rosé, „Author age prediction from text using linear regression," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 115–123, Association for Computational Linguistics, 2011.