A természet és a társadalom komplex hálózataiban található átfedő csoportosulások feltárása

Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society

PALLA Gergely [1,2], DERÉNYI Imre [2], FARKAS Illés [1,2],
POLLNER Péter[1] és VICSEK Tamás [1,2]

[1] MTA-ELTE Statisztikus Fizika és Biológiai Fizika kutatócsoport,
H1117 Budapest, Pázmány Péter Sétány 1/A, Magyarország, (tel.: +36-1-3722795, fax: +36-1-3722757,
e-mail: bioadmin@angel.elte.hu, honlap: http://angel.elte.hu/kutcsop)
[2] ELTE Biológiai Fizika tanszék, H1117 Budapest, Pázmány Péter sétány 1/A, Magyarország,
(tel.: +36-1-3722795, fax: +36-1-3722757, e-mail: bioadmin@angel.elte.hu, honlap: http://angel.elte.hu)

ABSTRACT

A fundamental question of great current interest is how to interpret the global organization of real-world networks as the coexistence of their structural sub-units (communities) associated with more highly interconnected parts. The existing methods used for large networks find disjunct communities, while most of the actual networks are made of highly overlapping and nested cohesive groups of nodes. Here we introduce the Clique Percolation Method enabling the extraction of overlapping communities on a large scale. We find that overlaps are indeed very significant, and the distributions we introduce to characterize the statistical features of community overlaps reveal novel universal features of networks. Finally, we find that the development of the modular structure of networks is driven by preferential attachment, in complete analogy with the growth of the underlying network of nodes.

ÖSSZEFOGLALÓ

A hálózatkutatás egyik alapvető fontosságú új területe a természetben található hálózatok csoportosulásainak feltárása. A hálózati csoportosulások általában olyan sűrűn kapcsolt szerkezeti alegységnek felelnek meg, melyben a csúcsok erősebben kötődnek egymáshoz, mint a hálózat többi részéhez. A nagyméretű hálózatokra jelenleg alkalmazott csoportosuláskereső módszerek diszjunkt csoportosulásokat találnak, ezzel szemben a valódi hálózatok többségében a csoportosulások egymást átfedik, és esetenként egymásba is ágyazódhatnak. Az általunk kifejlesztett Klikk Perkolációs Módszer egy hatékony megoldást nyújt nagyméretű hálózatok átfedő csoportosulásainak feltárására. Vizsgálataink szerint a természetben található hálózatok esetén a csoportosulások átfedése valóban szignifikáns, és az átfedések statisztikai tulajdonságainak jellemzésére bevezetett eloszlások a hálózatok új univerzális tulajdonságait tárják fel. A csoportosulások időfejlődésével kapcsolatos eredményeink szerint a hálózatok moduláris szerkezetének kialakulását preferenciális csatolódási mechanizmusok vezérlik, teljesen analóg módon az alapul szolgáló hálózatok növekedésével.

Kulcsszavak: Komplex hálózatok, átfedő hálózati modulok, csoportosulások, klikk perkoláció, preferenciális csatolódás

1. BEVEZETÉS

1.1. Hálózati csoportosulások

A hálózatok a természet és a társadalom leírásának igen általános és gyakran használt eszközei [1]. A rajtuk végbemenő folyamatok szempontjából meghatározó szerepe van a csoportosulásoknak (más néven moduloknak vagy klasztereknek). A csoportosulásoknak nincs egy általánosan elfogadott, egyértelmű definíciójuk, de általában olyan részgráfokat szokás csoportosulásnak elfogadni, amelyeken belül a csúcsok egymáshoz erősebben (sűrűbben) kapcsolódnak, mint a hálózat többi részéhez [5]. Egy egyszerű példa hálózati csoportosulásokra az emberi kapcsolatok hálózatában található családok, baráti körök, munkahelyi közösségek által

definiált csoportok. Ismert, hogy például a hírek egy-egy ilyen csoportosuláson belül (az egymással személyesen vagy akár telefonon gyakran beszélő emberek között) gyorsan terjednek, a csoportosulások között viszont jóval lassabban, ezért az emberi kapcsolatok hálózatán történő információáramlás szempontjából alapvető szerepük van a csoportosulásoknak. A csoportosulásoknak egy másik példája a világháló oldalainak hálózatában található olyan weboldalak, amelyek egymás közt sok mutatóval rendelkeznek. Ezek az oldalak gyakran földrajzilag egymáshoz közel találhatóak vagy hasonló témájúak, és a keresőprogramok számára hasznos lehet a feltérképezésük. A sejtjeink molekulái is sűrű csoportokba rendeződnek, ezek a csoportok számos érdekes felismerést és gyógyászati lehetőséget rejtenek. Ha a molekulák kölcsönhatási hálózatában (a csúcsok molekulák, az élek kölcsönhatások) találunk egy sűrű csoportot, akkor gyakran előfordul, hogy ennek a csoportnak egy korábban nem ismert, jól leírható önálló funkciója van az élő sejtben.

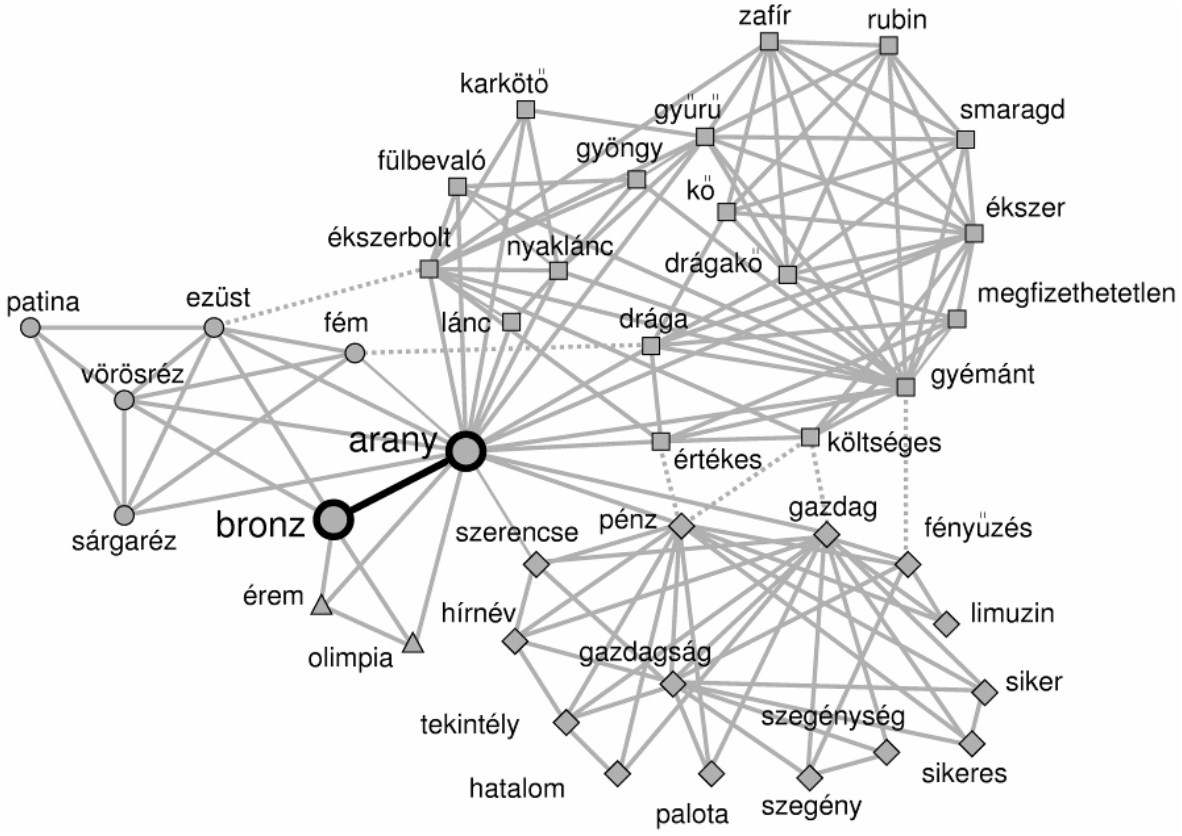
1.2. Csoportosuláskeresés élek eltávolításával

A csoportosulások (klaszterek) központi szerepe miatt a keresésükre kidolgozott módszerek számos tudományterületen és alkalmazásban használatosak. A jelenleg elterjedt klaszterezési eljárások döntő része egymástól elszigetelt, átfedéseket nem tartalmazó modulokat keres. Egyszerű (irányítatlan, súlyozatlan) gráfok esetén a klaszterezés leggyakoribb módja a gráf szétdarabolása izolált csoportosulásokra. Ilyenkor egy rögzített szabály szerint az éleket elkezdjük egyesével eltávolítani, majd egy ponton megállunk, és a megmaradt élek által összetartott gráf-komponenseket tekintjük az eredeti hálózat csoportosulásainak. Eltávolítandó élnek érdemes mindig a leggyengébb láncszemet választani, vagyis azt a kapcsolatot, amely a legnagyobb terhelésnek van kitéve, ha például az éleket drótoknak képzeljük, és véletlenszerűen választott pont párok között elektromos áramot folyatunk át a rendszeren. A terhelés ugyanis várhatóan a sűrű tartományokat összekötő éleken lesz a legnagyobb, így ezek eltávolítása során a sűrű tartományok többnyire érintetlenül maradnak. Kérdés azonban, hogy az élek eltávolításával mikor érdemes megállni. Szeretnénk azt az állapotot megtalálni, amikor az eredeti gráfban meglévő „sűrűsödések”" (csoportosulások) közé eső élek már eltűntek, de maguk a csoportosulások még épek. Ennek a problémának a megoldására vezette be Girvan és Newman a modularitás fogalmát [6]. Ez a mennyiség jellemzi, hogy a gráf pillanatnyi felosztása mellett hogyan viszonyul egymáshoz a csoportosulásokon belül ill. között futó élek száma az eredeti gráfban. A megállás pillanatát ezek után a modularitás maximumának elérésével határozhatjuk meg.

1.3. Csoportosuláskeresés átfedésekkel

Ha egy gráfot izolált csoportosulásokra bontottunk fel (például élek eltávolításával), akkor szükségszerűen kapott csoportosulások közül a hálózat minden csúcsa a legfeljebb egyhez tartozhat, tehát a csoportok között nincsen átfedés. Ezzel szemben a valódi hálózatokban gyakoriak a csoportosulások közötti átfedések: egy elem több csoportnak is tagja lehet [13]. Az egyik legismertebb példa erre az ismeretségi kapcsolatok hálózata. Ebben a hálózatban mindannyian több, egymástól eltérő szerepű csoportosulásnak is tagjai vagyunk. Példaként említhető családunk, iskolatársaink köre, baráti körünk, vagy munkatársaink. Két csoportnak természetesen több közös tagja is lehet, például a baráti körünk és az iskolatársaink csoportja számos közös taggal rendelkezhet. Szintén érdekes átfedő csoportosulásokat tartalmaz egy nyelv szóasszociációs hálózata (ld. 1. ábra). Ebben a hálózatban minden csúcs az adott nyelv egy-egy szavát jelöli, és két csúcs akkor van összekötve, ha az általuk jelölt két szót a vizsgálatokban megkérdezett személyek társították egymáshoz.

Érdekes megvizsgálni, hogy mi történik az ismeretségi hálózattal, ha felosztással próbálunk benne csoportosulásokat keresni, tehát átfedéseket nem engedünk meg. Tudjuk, hogy a legtöbb ember jó néhány csoportosuláshoz tartozik egyszerre, ezért akármilyen módon jelölünk ki a hálózatban számára egyetlen, a többivel nem átfedő klasztert, akkor abban az adott résztvevő több csoportosulásának töredékei együtt lesznek jelen. Például, ha az ismeretségi hálózatban a jelen cikk valamelyik szerzője számára egyetlen klasztert jelölnénk ki, akkor ebbe nagy valószínűséggel családtagok, iskola- és munkatársak egyaránt belekerülnének. Ez a megoldás két fontos hibalehetőséget rejt: a kijelölt egyetlen csoportosulásba belekerülnének olyanok is, akik nem ismerik egymást, például a vizsgált ember valamelyik családtagja és munkatársa (téves pozitív); számos családtag viszont ettől különböző klaszterbe kerülne (téves negatív).



1. ábra

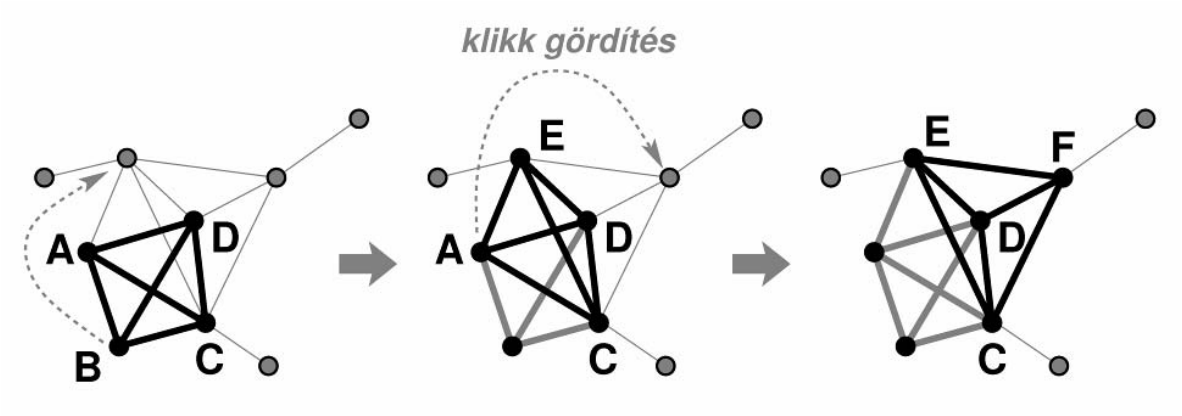
A University of South Florida Free Association Norms angol nyelvű szó asszociációs hálózatban a gold (arany) szóhoz talált négy csoportosulás magyar nyelvű megfelelői. A körökkel jelölt csoportosulás fémekkel kapcsolatos, a háromszögekkel jelölt csoportosulás olimpiai érmekkel, míg a másik két csoportosulás a jólét illetve az ékszerek köré rendeződik. A vastagított él és két csúcs több csoportosuláshoz tartozik, azaz csoportosulások átfedésében találhatóak. A pontozott vonalak csoportosulások közötti kapcsolatokat mutatnak.

A vázolt probléma kiküszöbölésére fejlesztettünk ki egy klikk perkoláción alapuló csoportosuláskereső módszert, mely természetes módon engedi meg a csoportosulások közti átfedéseket. Vizsgálataink szerint számos valódi hálózatban (például tudományos együttműködési, szó asszociációs ill. fehérje kölcsönhatási gráfokban) a csoportosulások között az átfedések valóban jelentősek és a bevezetett új statisztikus jellemzők segítségével nem triviális skálázási és korrelációs tulajdonságok találhatóak.

2. A KLIKK PERKOLÁCIÓS MÓDSZER

2.1. A módszer ismertetése

A hálózatokban található átfedő csoportosulások keresésére egy lehetséges eljárás a 2. ábrán illusztrált klikk perkolációs módszer (Clique Percolation Method, rövidítve CPM) [11]. Ez a módszer egyenként k darab csúcsból álló, teljesen összekötött részgráfokat (k-klikkeket) használ a csoportosulások feltérképezéséhez. Két k-klikket akkor mondunk szomszédosnak, ha csak egyetlen csúcsban különböznek egymástól, azaz k-1 csúcsuk közös. Egy, a CPM segítségével kapott csoportosulás azokból a k-klikkekből épül fel, melyek közül bármelyikből eljuthatunk bármely másikba szomszédos k-klikkeken keresztül.



2. ábra

A klikk-perkolációs módszer (CPM) [2] bemutatása egy kis hálózaton k=4 méretű klikk esetén. Az ábrán sötét színnel jelölt k-klikk sablon a gördülés során bejárja a hálózat A-B-C-D-E-F csoportosulását. Minden gördítési lépésben a sablon egyetlen csúcsa mozdul el és a többi k-1=3 csúcs helyben marad: ez a k-1 csúcs a gördítés előtti és utáni klikk közös része.

Ez a megközelítés nagyon közel áll a csoportosulások eredeti megfogalmazásához: a csoportosulásokon belül sok kapcsolatot szeretnénk, a csoportosulások között viszont keveset, hiszen k darab csúcs akkor van a lehető legsűrűbben összekötve, ha egy k-klikket alkotnak. A bevezetett k-klikk szomszédság segítségével definiálhatjuk a k-klikkek hálózatát is, ahol az egyes csúcsok az eredeti hálózat k-klikkjeinek felelnek meg, és két csúcs között akkor van él, ha a megfelelő k-klikkek szomszédosak. A csoportosulások ebben a képben a k-klikk hálózat összefüggő komponenseinek felelnek meg. Ezeket a statisztikus fizikában perkolációs klasztereknek szokás nevezni, innen származik a módszer elnevezése. Mivel az eredeti hálózatban egy csúcspont egyszerre több k-klikk perkolációs klaszternek (csoportosulásnak) is tagja lehet, ezért a CPM természetes módon engedi meg a csoportosulások közti átfedéseket.

A vázolt csoportosulás definíció jól szemléltethető k-klikk sablon gördítésén keresztül (2. ábra). A k-klikk sablon izomorf egy k-klikkel, és ráilleszthető a gráf bármelyik k-klikkjére, majd onnan egy lépésben tovább gördíthető egy szomszédos k-klikkre. A csoportosulások így olyan részgráfoknak felelnek meg, melyek bejárhatók k-klikk sablon gördítéssel.

2.2. Optimális paraméter-beállítás

Módszerünk direktben súlyozatlan hálózatokra alkalmazható, hiszen a fent bemutatott csoportosulás-definíció sehol sem használta az élek súlyát. Amennyiben súlyozott hálózatot kívánunk a CPM segítségével analizálni, az él-súlyokat oly módon vehetjük figyelembe, hogy bevezetvén egy w^* súlyküszöböt a w^*-nál gyengébb éleket nem vesszük figyelembe. A súlyküszöb növelésével a csoportosulások mérete csökken és csak a legerősebben összekapcsolt részek maradnak meg. Hasonló effektust okoz k növelése is, a nagyobb k-hoz tartozó csoportosulások kisebbek, de ugyanakkor kohézívebbek is. A w^* és k paraméterek változtatása hasonlít egy mikroszkóp felbontásának beállításához.

Ha egy konkrét csúcshoz tartozó csoportosulások érdekelnek minket, akkor azokat érdemes egy szélesebb w^* és k tartományban megvizsgálni. Ilyenkor csúcsról csúcsra más és más paraméter-értékeknél fogjuk a legérdekesebb képet látni. Ugyanakkor a globális csoportosulás-szerkezet vizsgálatához valamilyen kritérium szerint fixálni kell a súlyküszöböt és k-t. Az általunk használt kritériumot a perkoláció ihlette, és azon alapszik, hogy lehetőleg a legtöbb információt hordozó csoportosulás-szerkezetet nyerjük ki. Amennyiben túl alacsony w^* és k paramétereket választunk, a rendszer „perkolál”, azaz megjelenik egy óriás-csoportosulás, mely magába foglalja a hálózat túlnyomó részét, elfedvén a csoportosulás-szerkezet lokális részleteit. Ezzel szemben túl magas paraméter-értékeknél csak elszórtan találunk néhány kisméretű csoportosulást, hiszen csak a legerősebben összekapcsolt, legkohézívebb részek maradnak meg. Az ideális paraméterválasztás valahol a két véglet között található: adott k értékhez w^*-t úgy kell beállítani, hogy még éppen ne jelenjen meg egy óriás-csoportosulás.

3. A CPM ALKALMAZÁSA NAGYMÉRETŰ HÁLÓZATOKON

3.1. A vizsgált rendszerek

A CPM segítségével megvizsgáltuk három nagyméretű, természetben illetve a társadalomban található hálózat átfedő csoportosulás szerkezetét. A tanulmányozott rendszerek a következők voltak:

- A Los Alamos Condensed Matter archívum preprint gyűjteményéből nyert társszerzőségi hálózat [15], melyben minden n szerzős cikk 1/(n-1)-el növeli a szerzők közti kapcsolatok erősségét. (Összesen 30739 csúcs, 136065 él).
- A South Florida Free Association norms list-ből kapott szóasszociációs hálózat [16], melyben két szó közt lévő kapcsolat súlya arányos azzal a gyakorisággal, mellyel a megkérdezettek az egyik szóról a másikra asszociáltak a tesztek során. (Összesen 10617 csúcs, 63788 él).
- A *Saccharomyces cerevisiae* (sarjadzó élesztő) egysejtű modellszervezet fehérje kölcsönhatási hálózata [17], melyben két fehérje között akkor van kapcsolat, ha a kísérletek során kölcsönhatottak egymással. (Összesen 2609 csúcs, 6355 él).

Mindhárom hálózat esetében a CPM segítségével kapott csoportosulásokhoz természetes módon lehetett jelentést, funkciót társítani. A társszerzőségi hálózat esetén egy szerző különböző csoportosulásai általában a különféle érdeklődési területeinek feleltek meg, hiszen gyakran előfordul, hogy más-más témában másokkal működünk együtt. A szóasszociációs hálózatban egy szó csoportosulásai a szó különféle jelentéséhez kötődtek, ezt illusztrálja az 1. ábra. Végül a fehérje kölcsönhatási hálózat esetén a csoportosulások a sejtműködés során ellátott különféle funkcióknak feleltek meg.

A csoportosulásszerkezet globális vizsgálatánál az optimális súlyküszöb- és k paramétereknek a következő értékek adódtak: a társszerzőségi hálózat esetén $w^*=0.1$, k=6, a szóasszociációs hálózat esetén $w^*=0.015$, k=4, végül a fehérje kölcsönhatási hálózat esetén k=4 (itt az élek súlyozatlanok voltak).

3.2. Csoportátfedési statisztikák

A csoportok közti átfedések jellemzéséhez három statisztikai eloszlást vezettünk be. Az első az s^{ov} átfedési méret eloszlása, ahol két csoport közti átfedés mérete a közös csúcsok számával egyenlő. Vizsgáltuk a csúcsok m tagsági index eloszlását is, ahol egy csúcs tagsági indexe alatt azon csoportosulások számát értjük, melyhez a csúcs hozzátartozik. Végül tanulmányoztuk a csoportosulások gráfjának d^{com} fokszámeloszlását is, ahol a csoportosulások gráfját az átfedések révén származtattuk. Ebben a gráfban egy-egy csúcs egy-egy csoportosulásnak felel meg, és két csúcs akkor van összekötve, ha a megfelelő csoportosulások átfedik egymással. Így egy csoportosulás fokszáma megegyezik azon egyéb csoportosulások számával, melyekkel átfed. E három, a csoportátfedéseken alapuló statisztika mellett vizsgáltuk a csoportosulások s^{com} méreteloszlását is.

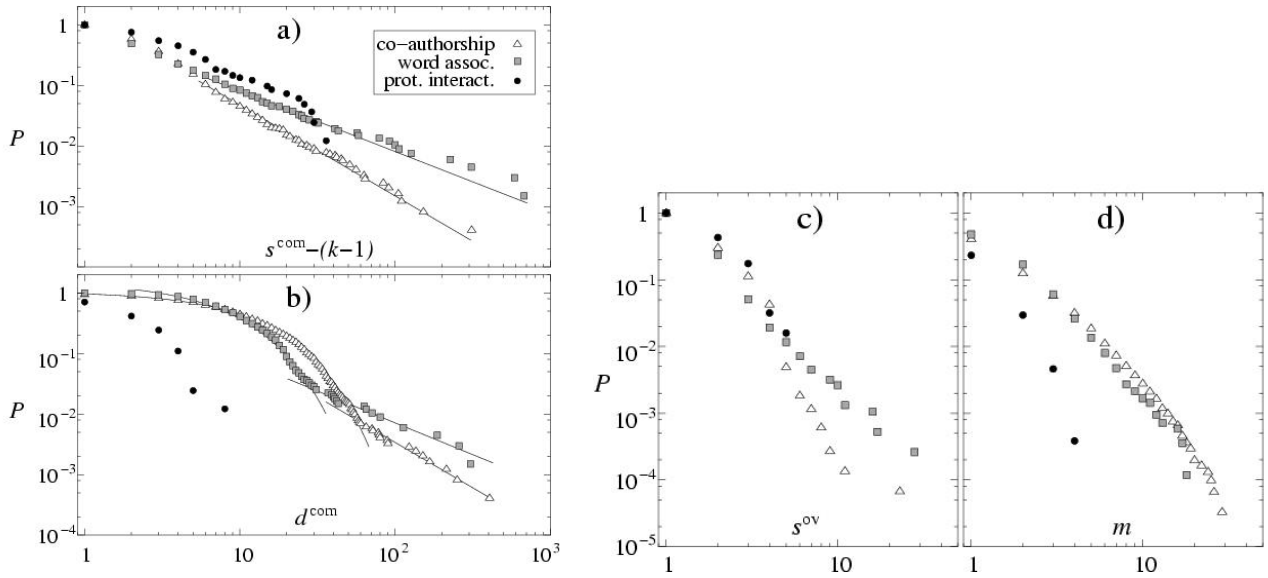
A kapott eredményeket a 3. ábra mutatja be. Ismert, hogy a korábbi, átfedéseket tiltó csoportosulás keresők segítségével kapott csoportosulások méreteloszlása hatványszerű. A 3a. ábrán látható, hogy a CPM által nyújtott teljesebb képben (ahol a csoportok közti átfedések is megengedettek) ez a tulajdonság megmarad, a csoportosulások méreteloszlása egy körülbelül 1.6-os hatványkitevővel csökken a nagy méretek felé.

Ennél összetettebb, nagyon érdekes viselkedést mutat a csoportosulások fokszámeloszlása (3b. ábra). Ezek az eloszlások két jól elkülöníthető részre tagolódnak: kis fokszámoknál exponenciálisan indulnak, majd egy ponton átváltanak hatványfüggvénybe, és a nagy fokszámok felé a csoportméret eloszlás hatványkitevőjével megegyező kitevővel csengenek le. Mindhárom hálózat esetén a csúcsok fokszámeloszlása hatványszerű; az imént ismertetett eredmény szerint egy magasabb szerveződési szinten, a csoportosulások szintjén, ehhez hasonló viselkedést tapasztalhatunk, (hiszen nagy fokszámok esetén a csoportosulások fokszáma is hatványfüggvény szerint cseng le). Emellett a csoportosulások szintjén eltérések is tapasztalhatók a csúcsok szintjéhez viszonyítva, ugyanis a kis fokszámoknál látott exponenciális rész nincs jelen a csúcsok esetén. Az, hogy a csoportosulás fokszámeloszlás farkának hatványkitevője megegyezik a csoportosulásméret eloszlás hatványkitevőjével, azzal magyarázható, hogy egy nagyméretű csoportosulás fokszáma jól becsülhető úgy, hogy feltesszük, hogy átlagosan minden csúcs egy δ járulékkal növeli a csoportosulás fokszámát, és így a csoportosulás fokszám egyszerűen a méret és δ szorzatával egyenlő.

A 3c. ábra szerint a csoportosulások átfedési méret eloszlása közel van egy nagy kitevőjű hatványfüggvényhez. Érdekes jelenség, hogy a k-1 méretet meghaladó átfedések is előfordulnak kis számban. Ez természetesen csak úgy lehetséges, hogy az átfedéseknek megfelelő részgráfok nem teljes részgráfok, (azaz bennük nincs mindenki mindenkivel összekötve). Végezetül a 3d. ábrán a csúcsok tagsági index eloszlását mutatjuk be, ezek az eloszlások szintén közel vannak egy gyorsan csökkenő hatványfüggvényhez.

Összehasonlításként megvizsgáltuk a csoportosulás-szerkezetet jellemző eloszlásokat a tanulmányozott hálózatoknak megfelelő véletlen hálózatokban is. Ezeket a véletlen hálózatokat az eredeti hálózatokból állítottuk elő az élek sorozatos véletlenszerű átkötésével, az egyes csúcsok fokszámának megtartása mellett. A kapott véletlen hálózatok rendkívül szegényes csoportosulás-szerkezetet mutattak, csak elvétve lehetett bennük

egy-két apró csoportosulást találni. Ez a jelenség alátámasztja azt, hogy az eredeti hálózatokban talált gazdag csoportosulás-szerkezet nem a módszerünk mesterséges melléktermékeként állt elő, hanem valóban a tanulmányozott hálózat belső korrelációit jeleníti meg egy igen szemléletes és áttekinthető módon.



3. ábra

A CPM segítségével a kapott csoportosulásokra jellemző kumulatív eloszlások a társszerzőségi hálózatban (háromszögek), a szóasszociációs hálózatban (négyzetek), valamint a fehérje kölcsönhatási hálózatban (körök). a) a csoportosulásméret-eloszlás, b) a csoportosulások fokszámeloszlása (a csoportosuláshálózat fokszámeloszlása), c) a csoportátfedések méret-eloszlása és d) a csúcsok tagsági indexének eloszlása. (Az ábra átvétel a [12] publikációból).

	N	$\langle d^{com} \rangle$	<C>	<r>
Társszerzőségi	2450	12.10	0.44	58%
Szóasszociációs	670	11.33	0.56	72%
Fehérje kölcsönhatási	82	1.54	0.17	26%

1. táblázat. A feltárt csoportosulások további jellemzői. N a csoportosulások számát jelöli, $\langle d^{com} \rangle$ a csoportosulások átlagos fokszámának felel meg, <C> a csoporthálózat átlagos klaszterezettségi mutatója, míg <r> egy csoportosulás azon tagjainak átlagos hányada, melyek még legalább egy másik csoportosulásnak is tagjai.

Az 1. táblázatban a csoportosulások további statisztikai jellemzőit tüntettük fel. Az első oszlop a csoportosulások összesített N számát mutatja, míg a második oszlop a csoportosuláshálózat $\langle d^{com} \rangle$ átlagos fokszámának felel meg. A harmadik oszlop a csoportosuláshálózat <C> átlagos klaszterezettségi mutatóját tünteti fel. Általánosan, egy hálózat adott csúcsának C klaszterezettségi mutatója a csúcs szomszédjai közt található élek száma osztva a csúcs szomszédjai közt lehetséges élek számával, ezért C mindig nulla és egy közé esik. Ezt a mennyiséget a csúcsokra átlagolva kapjuk az átlagos klaszterezettségi mutatót, mely a táblázat tanúsága szerint meglehetősen magas értékeket vesz fel a csoportosuláshálózatok esetén. Ez azt mutatja, hogy ha két csoportosulás átfed egy közös harmadikkal, akkor nagy valószínűséggel egymással is átfednek. Ez leggyakrabban olyan konfigurációban fordul elő, mikor az érintett három csoportosulás közösen osztozik az átfedési tartományon.

3.3. A CPM legfontosabb tulajdonságainak összegzése

A fent ismertetett eredmények tükrében elmondhatjuk, hogy a CPM egyfelől egy flexibilis eszköz átfedő hálózati csoportosulások feltárására, hiszen lehetőségünk van ráfókuszálni egy adott csúcs környezetére, és megvizsgálni a kiválasztott csúcs csoportosulásait különféle paraméter-beállítások mellett. A lokális csoportosulás-analízissel párhuzamosan a teljes hálózat csoportosulás-szerkezetét is tanulmányozhatjuk, ezen a téren egy nagyon fontos új megközelítést kínál módszerünk a csoportosuláshálózat létrehozásán keresztül. Az álta-

lunk tapasztalt skálázás a csoportosulás fokszámeloszlás esetén egy új megvilágításba helyezi a vizsgált rendszerek hierarchiájának kérdését. Eredményeink szerint a csúcsok szerveződési szintjéről a csoportok szerveződési szintjére történő váltáskor továbbra is hatványszerű marad a fokszámeloszlás lecsengése, mindemellett megjelenik egy karakterisztikus fokszám, ami alatt a fokszámeloszlás exponenciális. A több szerveződési szinttel rendelkező komplex rendszerekről alkotott képünkben ez arra világít rá, hogy a különböző szerveződési szintek egyfelől hasonlítanak egymásra (a rendszer bizonyos mértékben önhasonló) [18], másfelől minden szintnek van egy külön sajátossága, ami megkülönbözteti a többi szinttől.

4. PREFERENCIÁLIS CSATOLÓDÁS CSOPORTOSULÁS SZINTEN

A csoportosuláshálózat fokszámeloszlásának hatványszerű lecsengése felvet egy érdekes kérdést. Ismert, hogy számos, a természetben és társadalomban található hálózat fokszámeloszlásának hatványszerű lecsengése a preferenciális csatolódási szabállyal van szoros összefüggésben [2]. Egy preferenciális csatolódási szabály szerint növekvő hálózatban egy új csúcs becsatlakozásakor a már meglévő csúcsok a fokszámukkal arányos valószínűséggel kapnak élt az új csúcshoz. Így a nagy fokszámú régi csúcsok nagyobb eséllyel tudják továbbnövelni kapcsolataik számát mint a kis fokszámúak. Meg lehet mutatni analitikusan, hogy egy ilyen mechanizmus szerint fejlődő hálózat fokszámeloszlása hatványszerű lesz, és történtek mérések is valós hálózatokon, melyek alátámasztották a preferenciális csatolódási szabály teljesülését a vizsgált hálózatok növekedése során [19]. Ezek alapján természetesen adódik az a kérdés, hogy vajon beszélhetünk-e preferenciális csatolódási szabályról a csoportosulások szintjén is, hiszen a csoportosulások fokszámeloszlása is hatványfarkú.

A társszerzőségi hálózat esetében a hálózat időfejlődése is nyomon követhető, ugyanis a megjelent cikkek havi bontásban tárolták (összesen 146 hónapon át). Ezen hálózat esetén a következő két kérdést vizsgáltuk empirikusan [22]: vajon tapasztalható-e preferenciális csatolódási mechanizmus az egyes csúcsok csoportosulásokhoz történő csatlakozása során, és tapasztalható-e preferenciális csatolódási mechanizmus a csoportosuláshálózat növekedése során?

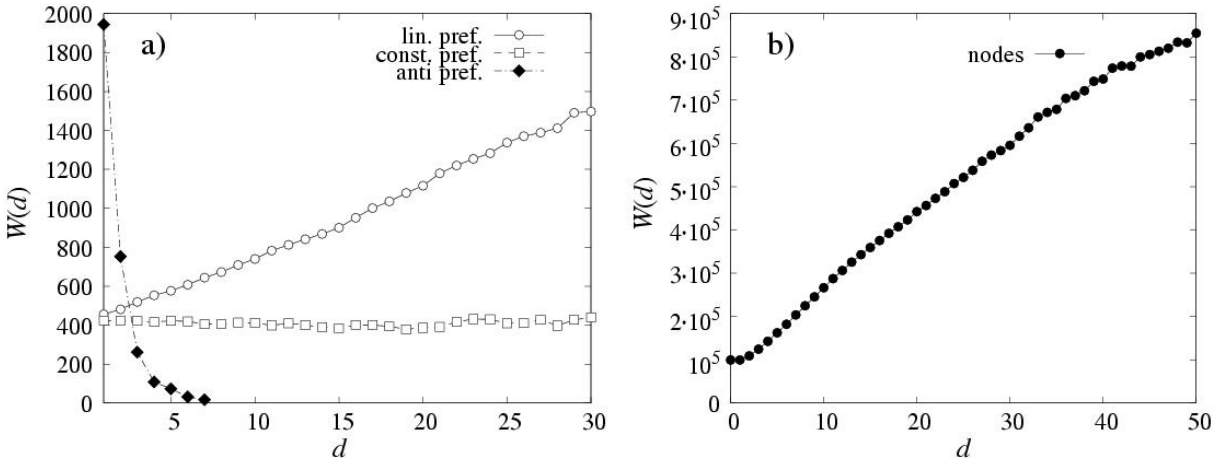
4.1. A preferenciális csatolódás kimutatásának módszere

A preferenciális csatolódási szabály kimutatására a következő általános módszert dolgoztuk ki [22], mely alkalmas az eloszlási trend megállapítására olyan, kis esetszámmal rendelkezésre álló adatok esetén is, ahol a hagyományos statisztikai eljárások [19] nem alkalmazhatók. Legyen ρ egy tulajdonság (pl. méret, vagy fokszám), és tegyük fel, hogy a vizsgált csatolódási mechanizmus szempontjából ρ értéke irreleváns. Ilyenkor a csatolódások során nagy átlagban a pontosan ρ tulajdonság eloszlásával fognak a csoportosulások kiválasztódni. Ellenben ha a csatolódás a nagy (vagy kis) ρ értékeket preferálja, akkor a nagy (vagy kis) ρ-val rendelkező csoportosulások nagyobb valószínűséggel fognak szerepelni, mint amit a ρ eloszlása alapján kapnánk. Egy ilyen eltérést úgy lehet kimutatni, hogy minden t időpontban meghatározzuk a ρ kumulatív eloszlását, $P_t(\rho)$-t, valamint a t és t+1 időpontok közt a csatolódások során kiválasztott csoportosulások normálatlan, kumulatív ρ eloszlását, $w_{t\to t+1}(\rho)$-t. Egy konkrét ρ^* esetén a $w_{t\to t+1}(\rho^*)$ értéke azon csoportosulások számával egyenlő, melyek kiválasztódtak a t és t+1 között a csatolódások során és ρ értékük t–ben nagyobb volt, mint ρ^*. A ρ szerint egyenletes csatolási preferenciától való eltérés kimutatásához egyszerűen fel kell összegezni az időlépések során $w_{t\to t+1}(\rho)$ és $P_t(\rho)$ hányadosát:

$$W(\rho) = \sum_{t=0}^{t_{\max}-1} \frac{w_{t\to t+1}(\rho)}{P_t(\rho)} \qquad (1)$$

cs

Ha a csatolódás szempontjából ρ értéke irreleváns (ρ szerint egyenletes csatolási preferencia), akkor a W(ρ) egy konstans függvény lesz. Ellenben ha a nagy (vagy kis) ρ értékek preferáltak, akkor W(ρ) növekvővé (vagy csökkenővé) válik. A vázolt módszert növekvő modell-hálózatokon teszteltük, melyeket a fokszámmal lineáris preferenciális csatolódással, fokszámtól független csatolódással, valamint fokszám szerint antipreferált csatolódással növesztettünk. Amint azt a 4a. ábra mutatja, a tesztek során rendre visszakaptuk, hogy a csúcsok fokszámának függvényében W(d) növekvő, konstans, illetve csökkenő tendenciát mutat az alkalmazott preferenciától függően. Emellett megvizsgáltuk magának a társszerzőségi hálózatnak a fejlődését is, és a 4b. ábra tanúsága szerint a csúcsok preferenciálisan csatlakoznak be a hálózatba a fokszám szerint.



4. ábra

A preferenciális csatolást kimutató empirikus módszer tesztelése. a) Eredmények fokszám szerinti lineáris preferenciával növesztett hálózat esetén (körök), fokszámtól független preferenciával növesztett hálózat esetén (négyzetek) és fokszám szerint anti-preferenciával növesztett hálózat esetén (rombuszok).
b) A társszerzőségi hálózat növekedésére kapott eredmény szerint a csúcsok a fokszámmal preferenciálisan kapcsolódnak be a hálózatba. (Az ábra átvétel a [22] publikációból).

4.2. A csoportosulások időfejlődésére kapott eredmények

A csoportosulások időfejlődése során a ρ helyébe az s^{com} csoportméretet, illetve a d^{com} csoportosulás fokszámot helyettesíthetjük aszerint, hogy a csoportméret, vagy a csoport fokszám szerinti preferenciális csatolódás létére vagyunk kíváncsiak. A csoportosuláshálózatba bekapcsolódó új csoportosulások csatolási mechanizmusának vizsgálata során így időlépésenként a kiválasztott régi csoportosulások $w_{t\to t+1}(s^{com})$ és $w_{t\to t+1}(d^{com})$ normálatlan kumulatív méret- illetve fokszámeloszlását kellett a $P_t(s^{com})$ és $P_t(d^{com})$ kumulatív méret- és fokszámeloszlással elosztani és a kapott hányadosokat az időlépések során felösszegezni:

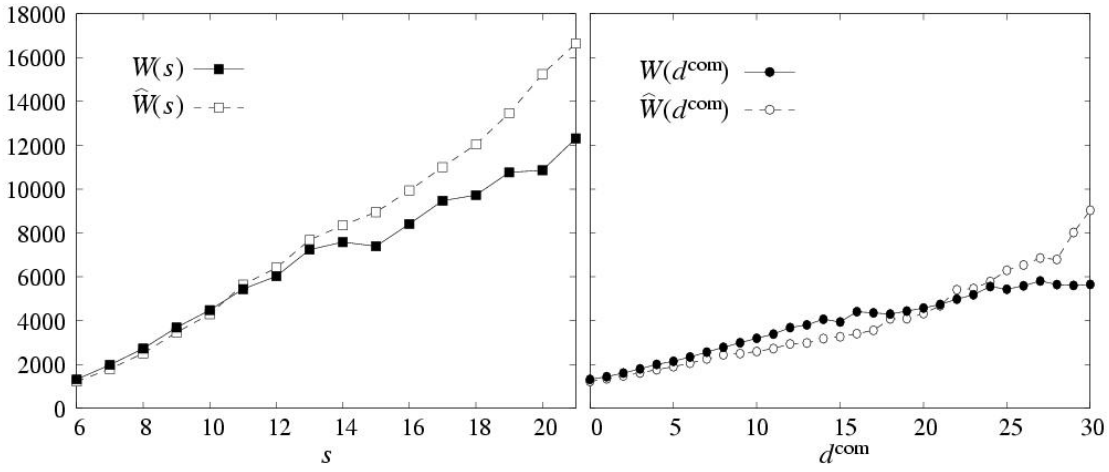
$$W(s^{com}) = \sum_{t=0}^{t_{max}-1} \frac{w_{t\to t+1}(s^{com})}{P_t(s^{com})} \tag{2}$$

$$W(d^{com}) = \sum_{t=0}^{t_{max}-1} \frac{w_{t\to t+1}(d^{com})}{P_t(d^{com})} \tag{3}$$

Hasonlóan, az új tagok megjelenésének vizsgálatakor időlépésenként azon csoportosulások $w^*_{t\to t+1}(s^{com})$ és $w^*_{t\to t+1}(d^{com})$ normálatlan kumulatív méret- illetve fokszámeloszlását határoztuk meg, melyek új tagokra tettek szert t és t+1 között. Ezt a két eloszlást a $P_t(s^{com})$ és $P_t(d^{com})$ kumulatív méret- és fokszámeloszlással osztottuk el és a kapott hányadosokat az időlépések során felösszegeztük:

$$\hat{W}(s^{com}) = \sum_{t=0}^{t_{max}-1} \frac{w^*_{t\to t+1}(s^{com})}{P_t(s^{com})} \tag{4}$$

$$\hat{W}(d^{com}) = \sum_{t=0}^{t_{max}-1} \frac{w^*_{t\to t+1}(d^{com})}{P_t(d^{com})} \tag{5}$$



5. ábra

A csoportosulások időfejlődésére kapott eredmények. Jól láthatóan mind a csoportosulásméret, mind a csoportosulás fokszám szerint preferenciálisan csatlakoznak az egyes csúcsok a csoportokhoz (fehér szimbólumok), valamint preferenciális csatolódási szabály szerint növekszik a csoportosulások hálózata is (fekete szimbólumok). (Az ábra átvétel a [22] publikációból).

A társszerzőségi hálózatra kapott eredmények az 5. ábrán láthatók. Mind a négy W(ρ) típusú görbe határozottan emelkedik, ami alapján az alábbi két következtetést vonhatjuk le:

- A csoportosuláshálózat növekedése során egy, még kapcsolatok nélküli csúcs (csoportosulás) a csoportosulásmérettel és csoportosulás fokszámmal preferenciálisan fog a csoportosuláshálózatba bekapcsolódni.
- A hálózatban egy csúcs, mely még egyetlen egy csoportosulásnak sem tagja, a csoportosulásmérettel és csoportosulás fokszámmal preferenciálisan fog egy csoportosuláshoz csatlakozni.

Ezen a ponton megjegyezzük, hogy amint azt már az előző fejezetben is említettük, ebben a hálózatban a csoportosulások mérete és a csoportosulások fokszáma a nagy méretek és fokszámok felé erősen korrelált egymással. Ezért ha egy csatolódási mechanizmus preferenciális akár a csoportméret, akár a csoport-fokszám szerint, akkor preferenciálisnak kell lennie mindkettő szerint.

Eredményeink szerint a társzerzőségi hálózat időfejlődését hasonló mechanizmusok vezérlik mind a csúcsok, mind a csoportosulások szintjén. A csoportosulások hálózatának növekedése a preferenciális csatolódási szabály szerint történik, teljesen analóg módon az alapul szolgáló hálózat növekedésével. Ez a jelenség egy további megerősítése a rendszer különböző szerveződési szintjei közt tapasztalható hasonlóságnak.

5. A CPM EGYÉB ALKALMAZÁSAI

A http://www.cfinder.org címről ingyenesen letölthető a szerzők által kifejlesztett CFinder (Clique and Community Finder) program [23], amely a Klikk Perkolációs Módszer használatával csoportosulásokat keres és – több más elemzéssel együtt – a talált csoportosulások hálózatát bemutatja. A program tudományterülettől függetlenül alkalmazható minden olyan adatrendszer elemzésére, amely hálózatként ábrázolható. A felhasznált bemenő adatfájl a hálózat éleit sorolja fel, minden sorban a hálózat két, egymással összekötött csúcsának a nevét kell megadni. A program Windows, Linux és Macintosh számítógépeken egyaránt használható.

Néhány érdekes alkalmazás, amely a [12] publikáció megjelenése óta eltelt másfél év alatt a CFinder felhasználásával született: a sarjadzó élesztő (egysejtű modellszervezet) fehérje-fehérje kölcsönhatási hálózatában korábban nem ismert csoportok és új fehérje funkciók azonosítása [23], majmok agyának látókérgében az ott található idegsejt kapcsolatok hálózata alapján az egyes területek szerepének elemzése [24], könnyűzenei előadók csoportosulásainak vizsgálata és szociológiai értelmezése [25], valamint daganatos elváltozásokban a sejt megváltozott működéséért felelős fehérjék keresése [26].

KÖSZÖNETNYILVÁNÍTÁS

A szerzők köszönetüket fejezik ki Barabási Albert-Lászlónak a hasznos beszélgetések, valamint Adamcsek Balázsnak a CFinder program grafikai felületének kidolgozásában való közreműködéséért. A szerzők kutatásait az OTKA D048422, F047203, T049674 és K60456 jelű pályázatai támogatják.

HIVATKOZÁSOK

[1] Watts, D. J. & Strogatz, S. H.:Collective dynamics of 'small-world' networks. *Nature* **393** (1998) 440--442 .
[2] Barabási, A.-L. & Albert, R.: Emergence of scaling in random networks. *Science* 286 (1999) 509--512.
[3] Albert, R. & Barabási, A.-L.: Statistical mechanics of complex networks. *Rev. Mod. Phys*. **74** (2002) 47—97.
[4] Mendes, J. F. F. & Dorogovtsev, S. N.: *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[5] Blatt, M., Wiseman, S., & Domany, E.: Super-paramagnetic clustering of data. *Phys. Rev. Lett*. **76** (1996) 3251--3254.
[6] Girvan, M. & Newman, M. E. J.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** (2002) 7821-7826.
[7] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D.: Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101** (2004) 2658--2663.
[8] Spirin, V. & Mirny, L. A.:Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* **100** (2003) 12123--12128.
[9] Scott, J.: *Social Network Analysis: A Handbook, 2nd ed*. (Sage Publications, London, 2000).
[10] Watts, D. J., Dodds, P. S., & Newman, M. E. J.: *Identity and search in social networks. Science* **296** (2002) 1302--1305.
[11] Derényi I., Palla G., Vicsek T. Clique percolation in random networks. *Phys. Rev. Lett.* **94** (2005) 160202:1-4.
[12] Palla G., Derényi I., Farkas I., Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435** (2005) 814-818.
[13] Faust, K.: Using Correspondence Analysis for Joint Displays of Affiliation Networks. *Models and Methods in Social Network Analysis* (Eds Carrington, P., Scott, J., & Wasserman, S.) Ch. 7 (Cambridge University Press, New York, 2005).
[14] Gavin, A. C. *et al*.: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415** (2002) 141--147.
[15] Warner, S.: E-prints and the Open Archives Initiative. *Library Hi Tech* **21** (2003) 151--158.
[16] Nelson, D. L., McEvoy, C. L., & Schreiber, T. A.: *The University of South Florida word association, rhyme, and word fragment norms.* http://www.usf.edu/FreeAssociation/
[17] Xenarios, I. *et.al*., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., & Eisenberg, D.: DIP: the Database of Interacting Proteins. *Nucl. Ac. Res*. **28** (2000) 289--291.
[18] Song, C., Havlin, S., & Makse, H. A.: Self-similarity of complex networks. *Nature* **433** (2005) 392--395.
[19] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert & T. Vicsek: *Physica A* **311** (2002) 590--614.
[20] H. Jeong, Z. Néda & A.-L. Barabási: Measuring preferential attachment for evolving networks. *Europhysics Letters* **61** (2003) 567--572.
[21] M. E. J. Newman: Clustering and preferential attachment in growing networks. *Phys. Rev. E* **64** (2001) 025102.
[22] Pollner, P., Palla, G., & Vicsek, T.: Preferential attachment of communities: The same principle, but a higher level. *Europhys. Lett*. **73** (2006) 478--484.
[23] Adamcsek B., Palla G., Farkas I. J., Derényi I., Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22** (2006) 1021-1023.
[24] Négyessy L., Nepusz T., Kocsis L., Bazsó F. Prediction of the main cortical areas and connections involved in the tactile function of the visual cortex by network analysis. *Eur. J. Neurosci.* **23** (2006) 1919-1930.
[25] R. D. Smith. The network of collaboration among rappers and its community structure. *Journal of Statistical Mechanics – Theory and experiment* (2006) Art. No. P02006.
[26] P. F. Jonsson, T. Cavanna, D. Zicha, P. A. Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics* **7** (2006) Art. No. 2.