Szegedi Tudományegyetem
Informatikai Intézet

Komplex hálózatok, gráf alapú adatbányászat és információ kinyerés valós rendszerekből

PhD értekezés tézisei

London András

Témavezető:

Dr. Pluhár András

Szeged
2017

1. Bevezető

A *„kisvilág" gráfok* [33] felfedezése jelentősen kibővítette a gráfelméleti kutatások irányát. A gráfok *komplex rendszerek* matematikai modellezésének legfontosabb eszközévé váltak stabil elméleti hátteret biztosítva ezen kutatásoknak. Ennek következtében az utóbbi években a gráf alapú adatbányászati és valós hálózatok vizsgálatára irányuló kutatások száma nagymértékben nőtt és jelenleg is ezt tartják a legbiztatóbb irányvonalnak relációs adatok [19] és komplex rendszerek [8] vizsgálatában. Komplex rendszerek gyakran gráffal (más néven hálózattal) modellezhetők, ahol a gráf pontjai (más néven csúcsai) a rendszer entitásait (szereplőit, ágenseit) reprezentálják, míg az élek (más néven linkek) a szereplő párok közti internakciókat (kapcsolatokat, hasonlóság mértékét) szemléltetik. Néhány kiváló összefoglaló értekezés a témában például Newman [31] és Boccaletti és társai [11] munkái. A „hálózatos megközelítés" egyrészt a hatalmas adatmennyiség és a bonyolult rendszer egyszerűsítésére és vizualizációjára szolgál, másrészt rendkívül hatékony a legfontosabb elemek, csoportok és a köztük lévő kulcsfontosságú interakciók feltárásában. Kissé leegyszerűsítve mondhatjuk, hogy míg az adatbányászat célja információ (ismeretanyag) kinyerése a rendelkezésre álló adatokból mintázatok és hasonlóságok feltárása által, addig a gráf alapú adatbányászat (vagy egyszerűen gráfbányászat) célja információ kinyerés gráffal reprezentált adatok, azaz hálózatok esetén.

A hálózatos modellezés és elemzés, valamint az adatbányászat azonos célja a vizsgált komplex rendszerről rendelkezése álló adatokból való információ kinyerése, összegzése és vizualizációja. Ez a folyamat egy matematikai modell (komplex hálózatos reprezentáció, vagy adatbányászati modell) megalkotásával kezdődik, majd egymást követő adatelemzési (illetve gráf elemzési) lépésekkel folytatódik. Ezen disszertáció célja bemutatni a szerző munkáját a területen, fókuszálva hálózatos, illetve gráf alapú adatbányászati modellek és módszerek megalkotására és alkalmazására valós problémák megoldásától motiváltan.

1.1. Valós hálózatok tulajdonságai

A szerző munkájában megalkotott, illetve alkalmazott algoritmusok és módszerek gráfokon értelmezettek, melyeket minden esetben valós komplex rendszerek modellezésére használunk. Formálisan egy irányítatlan (irányított) *gráf* (vagy *hálózat*) egy $G = (V, E)$ pár, mely két halmazt V-t és E-t, tartalmaz, ahol $V \neq \emptyset$, E pedig V nem rendezett (rendezett) párjainak egy részhalmaza. $V = \{1, 2, \ldots, n\}$ elemeit a gráf *pontjainak* (vagy csúcsainak), E elemeit pedig a gráf *éleinek* (vagy linkjeinek) nevezzük. A dolgozatban legtöbbször *súlyozott* gráfokkal dolgozunk, ami azt jelenti, hogy adott egy $w : E \to \mathbb{R}$ függvény, mely minden (i, j) élhez egy w_{ij} valós számot rendel.

Amióta Leonhard Euler 1736-ban megoldotta a königsbergi hidak problémáját megalkotva ezzel a gráfelméletet, a gráfokat számos különböző perspektívából vizsgálták és teljesen különböző területeken és problémákra alkalmazták. Az utóbbi évtizedekben tapasztalati úton egyértelművé vált, hogy a valós rendszereket (illetve relációs adatokat) modellező gráfok, azaz a valós hálózatok szerkezete nagyon különböző a korábban legtöbbet vizsgált gráfok, mint például az Erdős-Rényi véletlen gráf (1959) [18] szerkezetétől.

Watts és Strogatz [33], illetve Albert és Barabási [9] alapmunkáival (1998 és 1999) új lendületet kapott a komplex hálózatok tanulmányozása, mely hálózatok matematikai tárgyalásban lényegében valós rendszerek gráf modelljei. Komplex hálózatok a legkülönfélébb területeken jelennek meg, többek között a társadalomban, gazdaságban, biológiában és műszaki területeken egyaránt. Ezen hálózatokra jellemző a kiegyensúlyozatlan struktúra, időben folyamatosan változnak és komplexek abban az értelemben, hogy a globális strukturális (és funkcionalitási) tulajdonságaik általában nem következnek az egyes részeinek tulajdonságaiból és működéséből.

A komplex hálózatok globális tulajdonságai (úgy mint például az *élsűrűség*, *átmérő*, *átlagos úthossz*, *fokszám eloszlás*, *közösség-* és *mag/periféria szerkezet*) mellett a hálózat pontjainak (és éleinek) értékelése és rangsorolása is intenzíven kutatott témává vált. A disszertáció szempontjából a legfontosabb hozzájárulás Sergey Brin és Larry Page, a Google alapítói nevéhez fűződik, akik egy ötletes véletlen bolyongást definiáltak a weben való szörfözési szokások modellezésére [14]. Az általuk PageRank-nek elnevezett algoritmus az egyik legelterjedtebb úgynevezett csúcs *centralitási mérték* lett a hálózatkutatásban. Megfelelően használva a módszert a PageRank pontos információt ad a komplex hálózat pontjainak fontosságáról a hálózatban betöltött szerepük alapján. Brin és Page-től függetlenül, Jon Kleinberg egy alternatív eljárást javasolt weblapok fontosságának meghatározására, amit HITS-nek (Hyperlink Induced Topic Search) nevezett el [25]. A PageRank és HITS stabil matematikai elméletre épülnek, és szinte mindig jól alkalmazhatóak komplex hálózatok esetén adatbányászati céllal [22].

2. Néhány valós probléma hálózatos megközelítése

Különböző valós életből jövő problémát vizsgáltunk hálózatos modellek segítségével. Általánosan elmondhatjuk, hogy a hálózatok elemzése és gráfalgoritmusok alkalmazása rendkívül hasznos a modellezett rendszerek szerkezetének és működésének megértése szempontjából. Módszereink segítségével képesek voltunk jelentéssel bíró információt kinyerni a rendelkezésre álló adatokból és témaspecifikus kérdésekre válaszokat adni. A következő pontokban összegezzük a fejezet fő eredményeit.

2.1. Tudománymetria és citációs hálózatok

Bár számos alkalmazás esetén a PageRank (PR) értékeket ki kell számolni a hálózat összes csúcsára, vannak olyan szituációk, amikor nem tudjuk kiszámolni, vagy nincs is szükségünk az összes értékre, hanem a teljes hálózat csak egy kisebb részén számolunk PageRank-et. Chen és társai [15] adtak először PageRank közelítő algoritmust, ami a cél csúcsok értékét számítja egy alkalmas részgráf segítségével, nagyfokú precizitással. A módszerük a cél csúcsokból indulva felépít egy részgráfot, majd PageRank algoritmust futtat különböző heurisztikákat alkalmazva a részgráf peremén lévő pontok PageRank értékeinek becslésére. Egy módosított algoritmust mutattunk be és alkalmaztunk Egerváry Jenő híres cikke [17] „körüli" citációs hálózaton. Ezzel egy alternatív (hálózat alapú) mértéket adtunk meg cikkek tudományos hatásának meghatározásra [5]. Az algoritmusunk a következő lépésekben foglalható össze:

1. *Részgráf építés*: A cél csúcsokból – melyek PageRank értékére kíváncsiak vagyunk – indulva egy részgráfot építünk fel az irányított élek mentén visszafelé haladva. Ez egy iteratívan mélyülő szélességi kereséssel valósítható meg, mely egy fix szintet (célpontoktól való távolságot) elérve megáll. A bejárt pontok által feszített gráfot tekintjük az PageRank közelítő algoritmus bemenetének.

2. *PR értékek becslése a peremen*: Egy heurisztikát alkalmazunk erre a célra. Minden peremen lévő pont PageRank értéke egy fix, értéket kap kezdetben, ami be-fokának és a felépített részgráf élszámának hányadosa.

3. *PR kiszámolása a célpontokra*: A PageRank algoritmust futtatjuk a részgráfon. Minden lépésben a peremen lévő pontok esetén a 2-es pontban becsült értéket használjuk minden iteráció során.

1. táblázat. PR érték, elérési valószínűségek és idézések száma az Egerváry citációs hálózatban.

Publikáció	PR-érték	PR-rangsor	RP-érték	RP-rangsor	Citációk száma	Citációs rangsor
Egervári [17]	0.891	4	0.009	2	39	65
Kuhn [26]	1.189	1	0.042	1	726	1
Ford, Fulkerson [20]	0.525	8	0.004	9	39	65
Bellman [10]	0.399	11	0.003	10	18	158

Ugyanezen célra definiáltuk az „elérési valószínűség" (RP) értéket is. Elmondhatjuk, hogy a hálózat alapú megközelítés jóval reálisabb képet ad a cikk hatásáról (fontosságáról), mint más tudománymetriai indexek. Néhány eredmény röviden összegezve az 1. táblázatban látható.

2.2. Közösségi közlekedési hálózatok modellezése

Öt magyar város (Debrecen, Győr, Miskolc, Pécs, Szeged) tömegközlekedési rendszerét modelleztük és elemeztük hálózatos módszerek segítségével. A választás az alábbi szempontok szerint történt: (i) különösen a 100,000 és 250,000 közötti populációval és jelentős tömegközlekedési hálózattal rendelkező város legyen; (ii) a tömegközlekedés jellege (földhasználat, gazdasági szerep) és megszervezése hasonló a vizsgált városok esetén, ugyanakkor (iii) a földrajzi adottságok (táj, vízrajz, város méret) különbözőek legyenek. A területek 162 és 462 km^2 közöttiek, azaz közép méretűnek tekinthetőek, a városok morfológiája pedig különböző.

Elsőként végeztünk átfogó összehasonlító hálózatelemzést (a modern hálózatkutatás eszközeivel) magyar városok tömegközlekedési hálózatain. Ehhez első lépésként a rendszer gráf modelljét kellett megadni, amit többféleképpen tettünk meg. Általánosan, az állomásokat és megállókat reprezentáltuk a gráf csúcsaival, majd irányított éleket határozunk meg a megállók között közlekedő járatoknak megfelelően. Súlyokat rendeltünk mind az élekhez, mind a pontokhoz a járatkapacitásokat és a csúcsidei menetrendeket figyelembe véve. A munka egyik fő jelentőségének az adatgyűjtést és a hálózatok automatikus generálását tekintjük. A hálózatok globális és lokális tulajdonságait hasonlítottuk össze, és láttuk, hogy a kisvilág tulajdonság megjelenik mind az úthossz, mind pedig egyes csúcs centralitási értékek eloszlása esetén. A hálózatok topológiája és egyes lokális tulajdonságai közti különbségek főleg történeti, földrajzi és gazdasági okokból fakadnak. Rá tudtunk világítani néhány inkonzisztenciára és szervezésbeli problematikára, továbbá közlekedés mérnökök által is megerősített érzékenynek tűnő csomópontokat és útvonalakat határoztunk meg [23].

2.3. Oktatási adatok bányászata

Az *oktatási adatbányászat* területe az oktatásban képződő adatok elemzésével és elemzési módszerek fejlesztésével foglalkozik [32]. A cél elsősorban mintázatok és általános jellemzők megtalálása nagyméretű és változatos oktatási adatokon. Az oktatási adatbázisok magas fokú komplexitása, a rendelkezésre álló adat mennyiségének nagysága és az eredmények interpretálásának nehézsége miatt ilyen típusú elemzések elvégzése általában nehéz feladat.

Különböző hálózatos reprezentációkat javasoltuk oktatási adatok vizsgálatára és példákon keresztül megnéztük, hogy melyek a potenciális adatbányászati eszközök, melyekkel vizsgálhatóak ezen rendszerek. Rámutattunk, hogy milyen információ nyerhető ki az adatokból az adatbányászati és adatvizualizációs módszerek segítségével [6]. A gráfreprezentációtól függően négy különböző hálózatos modellt mutattunk be, majd az egyiket részletesebben vizsgáltuk egy esettanulmány keretében. A tanulók fejlődésének vizsgálatához és rangsorolásukhoz egy módosított

PageRank algoritmust adtunk meg. Azt kaptuk, hogy a tanulók páronkénti összehasonlításával nyert hálózaton a a tanulókhoz rendelt PageRank értékek alapján egészen jó sorrendet tudunk felállítani a tanulók között a tanulmányi teljesítményükre vonatkozóan és folyamatosan követni tudjuk a tanulmányi fejlődésüket [4]. Végül a gráfalapú adatbányászati technikáknak további előnyeire mutattunk rá oktatási adatok vizsgálatában.

3. Gazdasági rendszerek hálózatos modelljei

Gazdasási rendszerek hálózatos modellezése a gazdasági jelenségek mélyebb megértésére irányul a hálózattudomány fogalmainak és elemzési eszköztárának segítségével [24]. A fejezetben gazdasági modellezésre alkalmas hálózatos megközelítéseket mutattunk be. Esettanulmányaink között egy nemzetközi kereskedelmi hálózat időbeli dinamikájának vizsgálata, továbbá portfólió optimalizálásra is használható, részvény-korreláció alapú hálózatok elemzése szerepel. A következő pontokban összegezzük a fejezet fő eredményeit.

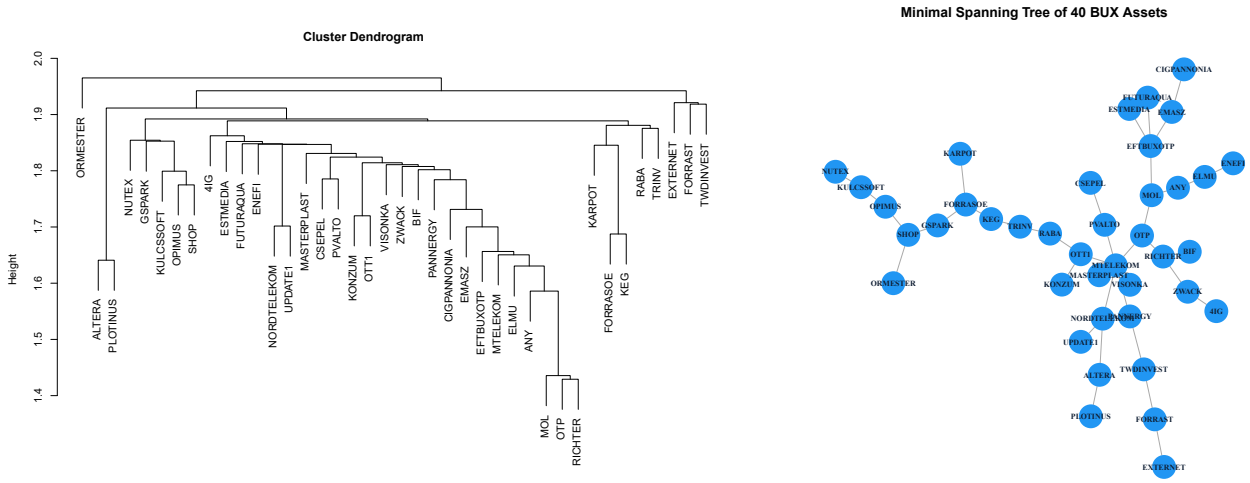
3.1. Kereskedelmi hálózatok

A kereskedelmi hálózatokat egy leegyszerűsített, de a rendszer komplexitását megőrző modellben vizsgálhatjuk, ahol az adott kereskedelmi rendszer országait tekintjük a hálózat pontjainak, míg két ország egymással való kereskedelmi viszonyát az adott országoknak megfelelő pontok közötti élekkel reprezentáljuk az importból, illetve exportból származó pénzáramokon keresztül. Az így kapott hálózati modellek matematikai értelemben irányított és súlyozott gráfok, ahol az él iránya a pénzmozgás irányát, súlya annak nagyságát mutatja.

Tanulmányunkban az EU országok és a gazdasági nagyhatalmak egymás közti kereskedelmi adatait vizsgáltuk hálózatos megközelítésben [30]. Elemzéseink azt mutatják, hogy habár az GDP arányos export valamennyi európai ország esetén nőtt az EU-s csatlakozást követően, a korábbi KGST országok esetén ezt a növekedést elsősorban az Oroszország és Kína felé irányuló exportnövekedés okozta. Különböző rangsoroló algoritmusokat alkalmazva (totál országonkénti export, PageRank, HITS) a hálózaton láttuk, hogy a Pareto-elv itt is érvényesül, azaz a teljes exportmennyiség jelentős részét csupán néhány ország adja. A hálózati módszerek is azt mutatják, hogy azon országok, melyek esetén az export volumen alapvetően alacsony, de GDP arányosan magas, jelentős gazdasági függésben vannak a gazdasági nagyhatalmaktól melyek felé irányul exportjuk jelentős része. Rámutattunk, hogy a hálózatok erős mag/periféria szerkezetet mutatnak. Azt találtuk, hogy az EU esetén egy stabil mag (Németországgal, Franciaországgal és az Egyesült Királysággal, mint vezető hatalmak) és egy szétesőperiféria található (többek közt a volt KGST, a Balkán és dél-európai országokkal, melyek különböző okokból tartoznak oda). Ezután modularitás optimalizáló közösségkeresést alkalmazva határoztuk meg a hálózat közösségeit és vizsgáltuk ezek időbeli változását. Azt kaptuk, hogy a kereskedelmi hálózatban a periférián lévő országok az Oroszország, illetve Kína által fémjelzett klaszterekben helyezkednek el, szemben a magban lévőkkel, amely országok Németország, illetve USA középpontú klaszterekben helyezkednek el, rávilágítva ezzel a valós gazdasági viszonyokra.

3.2. Pénzpiacok korreláció alapú hálózatai

A pénzpiacokon egy cég teljesítményét a részvényeinek árfolyama hivatott tükrözni, a cég értéke pedig a kapcsolódó részvényár és a forgalomban lévő részvények számának szorzata. Bár a cégek



(a) (b)

1. ábra. Cimkézett hierarchikus fa melyet az ún. „single linkage” klaszterező eljárással kaptunk, és a társított minimális feszítő fa a Budapesti Értéktőzsde 40 részvénye esetén.

közti tényleges (pl. tulajdonosi, szerződési) kapcsolatok nem ismertek általában, természetesnek tűnik feltételezni, hogy ezen kapcsolatok a részvényárfolyamok közti korrelációkban tükröződnek.

A hálózatkutatás alapvető eszközeinek segítségével vizsgáltuk a *Markowitz-féle portfólió kiválasztás* problémáját [28], mely a következőképp fogalmazható meg. Adott n kockázatos részvény, melyekből egy portfóliót állítunk össze meghatározva az egyes i $(i = 1, \ldots, n)$ részvényekbe fektetett p_i $(i = 1, \ldots, n)$ tőkearányt (feltételezve, hogy $\sum_i^n p_i = 1$). A $\mathbf{p} = (p_1, \ldots, p_n)$ portfólió várható hozama $r_p = \sum_{i=1}^n p_i r_i = \mathbf{p}\mathbf{r}^T$, kockázatát pedig a variancia- és kovarianciákon keresztül számoljuk, azaz $\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n p_i p_j \sigma_{ij} = \mathbf{p}\Sigma\mathbf{p}^T$. A képletekben r_i az i-edik részvény várható hozama (melyet a korábbi hozamokból becslünk), σ_{ij} az i-edik és j-edik részvény napi záró ár idősorából számolt kovarianciája egy adott időintervallumon. A cél olyan $\mathbf{p}$ portfólió meghatározása, mely esetén a σ_p^2 kockázat minimális, feltéve egy adott elvárt r_p hozamot.

Általában a célfüggvényben szereplő kovariancia mátrix egyrészt túl nagy (n^2 különböző elemet tartalmaz), másrészt rendkívül „zajos”, ezért különböző technikákat alkalmaztunk, hogy leválasszuk a mátrix azon részét, mely robusztus a statisztikai bizonytalansággal (véletlenszerűséggel) szemben, illetve, hogy csökkentsük a benne lévő elemek számát. A használt módszerek a *véletlen mátrixok* elméletén [29], illetve *hierarchikus klaszterezési* eljárásokon alapulnak [7] (illusztráció az 1. ábrán). A szűrési eljárásokon túl a várható hozamok számításához is több módszert kipróbáltunk.

Két adatsort, nevezetesen a Budapesti Értéktőzsdén jegyzett és a Yahoo Finance-en IT szektor részvények idősorait vizsgáltuk és ezeken hasonlítottuk össze a portfólió optimalizálás eredményességét a szűrési módszerek és hozambecslések különböző variációi esetén. Sok bootstrap kísérletet végrehajtva láttuk, hogy összhangban korábbi tesztekkel, a klasszikus Markowitz megoldás javítható szűrési technikákat alkalmazva a kovariancia mátrixon. A realizált hozamok, és a portfólió megbízhatósága is javul, utóbbit a becsült és a tényleges kockázat közti különbség csökkenése mutatja [2, 21].

4. Hálózatos modellek értékelési és rangsorolási problémákra

Entitások páronkénti összehasonlításán alapuló értékelő és rangsoroló módszerek számos területen jelennek meg. Miután a gráf alapú (hálózat alapú) rangsorolás a Google kereső motor központi összetevőjévé vált, számos más területen is megjelent társadalmi hálózatok vizsgálatától egészen út- és elektromos hálózatok optimalizálási kérdéseinek vizsgálatáig [27]. A következő pontokban összegezzük a fejezet fő eredményeit.

4.1. Értékelés és rangsorolás sportok esetén

A negyedik fejezetben először egy *időfüggő PageRank* módszert (tdPR) adtunk meg és alkalmaztunk egy egyetemi asztalitenisz bajnokság résztvevőinek rangsorolására [3]. A tdPR módszer szerint egy játékos rangsorolása nem pusztán a győzelem-vereség mutatóján múlik, hanem számít, hogy az adott játékos mennyire erős ellenfelet győzött le (és rekurzívan, az ellenfél mennyire erős ellenfeleket győzött le, stb.), és az is, hogy mikor. Ebből következően ahhoz, hogy előrébb lépjen valaki a rangsorban egy aktuálisan erős(ebb) játékost kell legyőzni, azonban a gyengébb ellenfelek elleni győzelmek nem számítanak jelentősen a rangsorolásban. Az időfüggés további célja, hogy a játékosoknak rendszeresen kelljen játszani, ugyanis a régebbi eredmények folyamatosan elavulttá válnak, rontva ezzel a rangsorbeli pozíciót. A tdPR-t több széles körben elterjedt értékelési módszerrel hasonlítottuk össze. Megfigyeltük, hogy a módszerünk jó predikciós erővel bír és pontosabb képet ad a játékosok relatív erősorrendjéről, mint több széles körben elterjedt módszer. Feltételezünk egy (részben a rangsorolás által irányított) önszerveződési mechanizmust is, mely az „eredménygráf" fejlődését meghatározza. Egyértelmű, hogy a játékosok olyan mérkőzéseket szeretnének játszani melyek várhatóan izgalmasak (közeli tdPR értékű játékosokat keresnek ellenfélnek), továbbá a rangsorban is előrébb szeretnének lépni (aktuálisan magasabb tdPR értékűvel érdemes játszani). Bár a játékosok csak a rangsort (és az eddigi eredményeket) ismerik, a bajnokság későbbi evolúciója (lejátszott mérkőzések sorozata) mégis jól modellezhető egy gráf folyamattal. Ez a megfigyelés egy speciális gráfnövekedési folyamat leírására ad ötletet, ahol a pontok összekötése a PageRank értékeik függvényében történik. Ez a jelenség összefüggésben lehet egy elit kialakulásával sportok esetén, ugyanakkor megjegyezzük, hogy ezen hipotézis vizsgálata jövőbeni kutatások tárgyát képezi.

4.2. Valószínűségi előrejelzések sportok esetén

Egy új valószínűség alapú előrejelző modellt alkottunk meg és alkalmaztunk sportmérkőzések lehetséges kimeneteleinek becslésére. A módszer lineáris algebrai értékelő módszerek segítségével, korábbi végeredmény adatokat felhasználva rendel valószínűségeket az egyes kimenetelekhez. Szemben azon módszerekkel, melyek a csapatok aktuális, egymáshoz viszonyított erejét becsülik a múltbeli eredmények alapján, mint például a híres *Bradley-Terry model* [13], amely a maximum likelihood módszer segítségével ad előrejelzést, mi egy *előrenéző* hálózatalapú algoritmust javasoltunk. A modellünk alapfeltevése, hogy ha a csapatok (vagy játékosok) aktuális értékelése jól tükrözi a csapatok közti erősorrendet, akkor egy adott jövőbeni mérkőzést tekintve egy lehetséges kimenetel annál valószínűbb, minél kevésbé változik meg a csapatok közti erősorrend (értékelő vektor).

Tegyük fel, hogy az r-edik ($r = k, \ldots, R-1$) játéknap előtt (valamely k-tól kezdve, amikor már tudunk erősorrendet állítani, azaz értékelni) a csapatok (legyenek $V = (1, \ldots, n)$) értékelő vektora $\phi^{r-1}(V) = (\phi_1^{r-1}, \ldots, \phi_n^{r-1})$. A kulcsötlet, hogy az r-edik játéknapon ha i és j csapat játszik, akkor a meccs lehetséges eredményei (melyekre az $< x : y >$, $x, y = 0, 1, \ldots$ jelölést

használjuk) után kiszámolt $\phi^r(V)$ és $\phi^{r-1}(V)$ vektorok közötti különbség fordítottan arányos az eredmények bekövetkezési valószínűségével. Ezt a változást a következőképp számoljuk:

$$\delta^r_{xy} = \mathrm{dist}\big(\phi^{r-1}(V), \phi^r(V) \;\big|\; \text{végeredmény} =< x : y > \big)$$

ahol $\mathrm{dist} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ valamilyen távolságfüggvény. Kísérleteinkben egy időfüggő PageRank algoritmust használtunk, dist pedig az Euklideszi távolság volt. A hazai pálya előnye effektust szintén figyelembe vettük a következőképp. Minden i csapatra külön tekintettük a hazai-i és vendég-i csapatot, kapva így egy $2n \times 2n$-es eredménymátrixot. Ez a mátrix lényegében egy páros gráf szomszédsági mátrixát írja le, ahol az irányított súlyozott élek a csapatok közti mérkőzések végeredményeit jelentik. Európai labdarúgó bajnokságok eredmény adatain végeztünk elemzéseket és megfigyeltük, hogy a modell összességében jól teljesített, több esetben jobb prediktív pontosságot adva, mint a Bradley-Terry módszer legfejlettebb változatai. Megjegyezzük továbbá, hogy a modellben szereplő paraméterek finomhangolása és a módszer részletes vizsgálata jövőbeni kutatások tárgyát képezi.

5. Valós rendszerek páros gráfos modelljei

Komplex rendszerek egy speciális, mégis rendkívül fontos osztályát adják azok a rendszerek, melyek páros (más néven kétrészes) gráffal modellezhetők. Az ilyen gráfok csúcsai két osztályba (legyenek A és B) sorolhatóak, élek pedig csak A és B között lehetnek, osztályon belül nem. Páros gráfok természetes módon jelennek meg társadalmi hálózatoktól biológiai rendszerekig. A következő pontokban összegezzük a fejezet fő eredményeit.

5.1. Közösségkeresés páros gráfokban statisztikus validációval

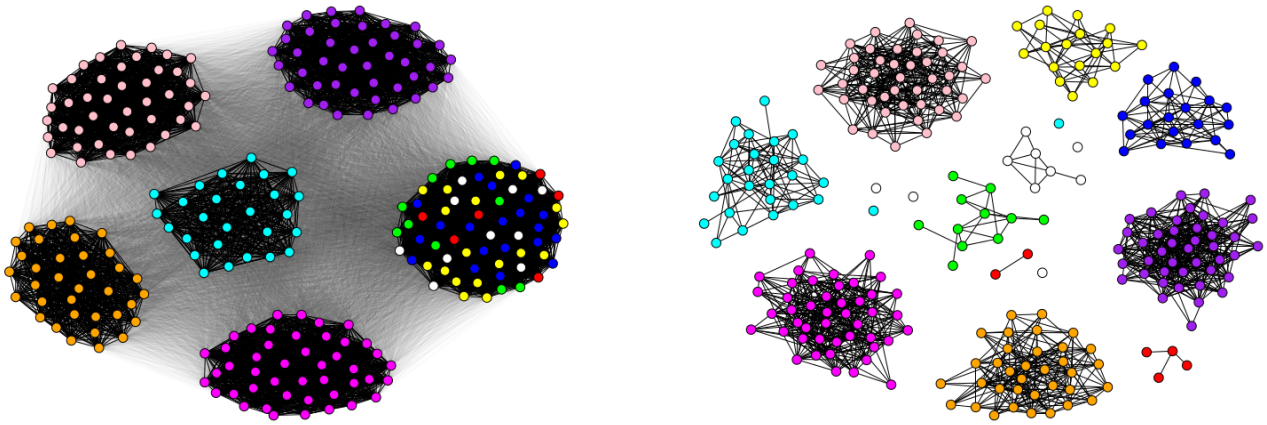
Egy módszertant mutattunk be mely páros gráfok esetén értelmezett közösségek magjainak (mag pontjainak) megtalálásra alkalmas [12]. Ehhez statisztikailag érvényesített féloldali projekciót használtunk. Ahhoz, hogy a projekcióval kapott gráfban egy élt statisztikailag érvényesítsünk, minden a_i és a_j A-beli pont (és hasonlóan minden b_k és b_ℓ B-beli pontra) között, szemben a legtöbb projekciós eljárással, csak akkor húzunk be élt, ha a közös szomszédaik száma (B-ben) nem konzisztens a null hipotézissel mely adott eloszlás szerinti véletlen számú szomszédot feltételez. A hipotézisteszteléshez az egyoldalú hipergeometrikus tesztet használtunk. A null hipotézis az, hogy annak a valószínűsége, hogy a_i és a_j közös szomszédainak száma pontosan x hipergeometrikus eloszlást követ, azaz

$$H(x|m, d_i, d_j) = \frac{\binom{d_i}{x}\binom{m-d_i}{d_j-x}}{\binom{m}{d_j}},$$

ahol m a pontok száma B-ben, d_i (d_j) pedig a_i (a_j) pont foka. Ezután a szokásos módon egy p-értéket definiálunk minden (a_i, a_j) párhoz:

$$p_{ij} = 1 - \sum_{x=0}^{n_{ij}-1} H(x|m, d_i, d_j),$$

aghol n_{ij} a megfigyelt szomszédok száma. Ha a p-érték kisebb vagy egyenlő, mint egy adott α szignifikancia szint, abból következik, hogy a megfigyelés inkonzisztens a null hipotézissel, azaz elvetjük azt, így a_i és a_j közi élt érvényesnek tekintjük a projekcióban. Ugyanakkor olyan



2. ábra. Közösségek az adjacencia, illetve a Bonferroni projekció esetén

tesztek, melyek helytelenül elvetik a null hipotézist (azaz elsőfajú hibát vétenek) nagyobb eséllyel jellenek meg ha párhuzamosan több tesztet végzünk. Ahhoz, hogy ezt kiküszöböljük, egy közös szignifikancia szintet használunk minden teszt (pontpár) esetén. A legerősebb megszorítást a *Bonferroni korrekció* adja, ami minimalizálja a hamis pozitív teszteket (azaz az elsőfajú hibák számát), viszont nem mindig garantál megfelelő statisztikai pontosságot (vagyis sok hibás negatív tesztet, azaz másodfajú hibát eredményez).

Elmondható, hogy a módszertanunk által megtalált közösségek magjai jelentős információt adnak, és robusztusak a páros gráf hibás, illetve hiányzó éleivel szemben. A statisztikai robusztusságot először mesterséges benchmark hálózatokon mutattuk meg. Rámutattunk, hogy ez a fajta információ kiszűrési eljárás szükségképpen növeli a közösségkeresés statisztikai precizitását, megtalálja a közösség stabil elemeit. Az eljárás használatát nagy, és nem teljesen megbízható adatokon annak ellenére is javasoljuk, hogy a közösségkeresés statisztikai pontossága csökken bizonyos esetekben. Két esettanulmányt is bemutattunk, az egyiket az IMDB adatbázisból nyert film-színész, a másikat pedig egy kutató-publikáció páros hálózaton.

5.2. Páros gráfok pontjainak értékelése és rangsorolása

A PageRank és HITS algoritmusok egy páros gráfokra vett általánosítását mutattuk be és alkalmaztuk bórkostolási adatsorokra a kóstolok szakértelmének értékelése és ebből adódó rangsorolásuk céljából [1]. Célunk a kóstolók szakértelmének mérése volt. Az algoritmusunk által kapott eredményeket egyszerűbb statisztikai módszerekkel mért eredményekkel vetettük össze. A valós adatokon végzett kísérleteink eredményei azt mutatják, hogy a co-HITS algoritmus az a priori tudásunkkal konzisztens eredményt ad a legtöbb esetben. Ezen felül finomabb különbségeket ad, szemben más módszerek által adott túl nagy differenciák helyett, a kóstolók között. Fontos előnye a hálózat alapú módszernek, hogy akkor is jól működik, amikor a kóstolók nem minden bort kóstolnak meg, így egy folyamatosan változó online kóstolási adatbázis esetén is képes lehet objektív rangsorolást adni a kóstolókról és a borokról is.

A szerző munkájának tézispontszerű összefoglalása

A következő pontokban összegzem a disszertáció fő eredményeit. A 2. táblázat mutatja az egyes publikációim megjelenését az egyes tézispontokban és fejezetekben.

	[16]	[5]	[23]	[6]	[4]	[30]	[21, 2]	[3]	[12]	[1]
I.	•	•	•	•	•	•		•		•
II.							•		•	
III.									•	
IV.									•	•
Chapter	1, 2, 3	1	1	1	1	1	3	4	5	5

2. táblázat. Kapcsolat a tézispontok, publikációk és fejezetek között

I. Rámutattam, hogy számos valós komplex rendszer modellezhető hálózatokkal és a gráfos adatbányászatot javasolom első lépésként ilyen rendszerek vizsgálatában. Minden esettanulmányban megmutattam, hogy a megfelelő adatok összegyűjtése (és szükséges tisztítása) után hogyan lehet a hálózatos megközelítést alkalmazni jelentős információ kinyerésére a modellezett rendszerből. Új adatbányászati módszereket mutattam be, melyek általában létező, egyre szélesebb körben elterjedt módszerek módosításai. Speciálisan egy lokális PageRank közelítő módszert és egy általánosított, páros gráfokon működő HITS algoritmust adtam meg hálózat pontjainak érékelésére és rangsorolására. Elmondható, hogy a felhasznált és újonnan fejlesztett módszerek jól alkalmazhatóak számos valós, hálózattal modellezhető probléma esetén a tudománymetriától, közlekedési és oktatási adatok vizsgálatán keresztül kereskedelmi rangsorok felállításáig.

II. Különböző perspektívákból tárgyaltam, hogy valós adatokkal dolgozva hogyan lehet mérni a modellezett rendszerben jelen levő a statisztikai bizonytalanságot (más néven zajt). Különböző módszereket adtam meg, illetve használtam olyan információ kiszűrésére, mely robusztus a statisztikai bizonytalansággal (véletlenszerűséggel) szemben, azaz robusztus a hibás adatokra vagy más zajforrásra nézve. Nevezetesen, hálózatalapú, vételen mátrixok elméletét felhasználó és statisztikai (hipotézistesztelés) alapú módszereket adtam meg és használtam részvényárfolyam korrelációk által definiált hálózatokra portfólió optimalizálási céllal, illetve páros gráfokon való közösségkeresési céllal. Eredményeim mutatják, hogy ezen technikák a portfólió probléma esetén javítanak a klasszikus megoldáson, közösségkeresés esetén pedig jó statisztikai pontossággal tudják meghatározni a közösségek magjait.

III. Páros gráfokkal foglalkozva bemutatom, hogy a páros gráfok szerkezete által kódolt információ jól használható a páros gráf egyes színosztályaiban lévő pontok klaszterezésére. Monte-Carlo szimulációk eredményei mutatják, hogy a közösségek magjai általában stabilak az adathibák vagy más zajforrásokra (pl. hiányzó élek) vonatkozóan és jó statisztikai precizitással megtalálhatóak, bár a metodológia statisztikai értelemben nem mindig pontos. A közösségek mag pontjainak megkereséséhez statisztikailag validált féloldali projekciókat használtam. A módszer robusztus a közösségkereső algoritmusra és az adatbázis hibás és hiányzó elemire vonatkozóan. Valós adatokon való esettanulmányok szintén bemutatásra kerültek. Páros gráfoknál maradva, a HITS értékelő-rangsoroló algoritmus alkalmazásának egy általános keretét írtam le, majd alkalmaztam valós (borkóstolási) adatokon. Az eredmények a módszer jó használhatóságát támasztják alá.

IV. Sportolók és sportcsapatok értékelésének és rangsorolási problémakörét tárgyaltam hálózatelemzési megközelítésben. Egy új, időfüggő PageRank algoritmus fejlesztettem ki, mely a mérkőzések eredményeiből nyert ún. eredménygráf segítségével rangsorolja a bajnokság szereplőit. Elmondható, hogy a módszer tisztább képet ad mint több, széles körben használt módszer és jobban teljesít predikciós erejét tekintve. Egy új, értékelés alapú előrejelzési

keretmódszert is bemutattam. A módszert (az egyes paramétereinek és az implementációk alapos optimalizálása nélkül) a híres Bradley-Terry módszer egy fejlett változatával, illetve fogadóirodák fogadási odds-ai által adott predikciókkal hasonlítottam össze. Egy időfüggő és PageRank módszert használva az eredmények azt mutatják, hogy a módszer alkalmas jó valószínűség alapú előrejelzéseket adni, sőt, több esetben jobban teljesít mint a Bradley-Terry modell predikciós pontosságát tekintve.

Válogatott referenciák

[1] **A. London** and T. Csendes. Hits based network algorithm for evaluating the professional skills of wine tasters. In *8th International Symposium on Applied Computational Intelligence and Informatics*, pages 197–200. IEEE, 2013.

[2] **A. London**, I. Gera, and B. Bánhelyi. Testing portfolio selection models using various estimators of expected returns and filtering techniques for correlation matrices (submitted). 2017.

[3] **A. London**, J. Németh, and T. Németh. Time-dependent network algorithm for ranking in sports. *Acta Cybernetica*, 21(3):495–506, 2014.

[4] **A. London** and T. Németh. Student evaluation by graph based data mining of administrational systems of education. In *Proceedings of the 15th International Conference on Computer Systems and Technologies*, pages 363–369. ACM, 2014.

[5] **A. London**, T. Németh, A. Pluhár, and T. Csendes. A local pagerank algorithm for evaluating the importance of scientific articles. *Annales Mathematicae et Informaticae*, 44:131–141, 2015.

[6] **A. London**, Á. Pelyhe, C. Holló, and T. Németh. Applying graph-based data mining concepts to the educational sphere. In *Proceedings of the 16th International Conference on Computer Systems and Technologies*, pages 358–365. ACM, 2015.

[7] M. R. Anderberg. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, volume 19. Academic press, 2014.

[8] A.-L. Barabási. The network takeover. *Nature Physics*, 8(1):14–16, 2012.

[9] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[10] R. Bellman. Mathematical aspects of scheduling theory. *Journal of the Society for Industrial & Applied Mathematics*, 4(3):168–205, 1956.

[11] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308, 2006.

[12] C. Bongiorno, **A. London**, S. Micciché, and R. N. Mantegna. Core of communities in bipartite networks (submitted to Physical Review E). *arXiv preprint arXiv:1704.01524*, 2017.

[13] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[14] S. Brin and L. Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56(18):3825–3833, 2012.

[15] Y.-Y. Chen, Q. Gan, and T. Suel. Local methods for estimating PageRank values. In *Proceedings of the 13th International conference on Information and Knowledge Management*, pages 381–389. ACM, 2004.

[16] P. Csermely, **A. London**, L.-Y. Wu, and B. Uzzi. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 1(2):93–123, 2013.

[17] J. Egerváry. Mátrixok kombinatorikus tulajdonságairól. *Matematikai és Fizikai Lapok*, 38:16–28, 1931.

[18] P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[19] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, 1996.

[20] L. R. Ford and D. Fulkerson. Solving the transportation problem. *Management Science*, 3(1):24–32, 1956.

[21] I. Gera, B. Bánhelyi, and **A. London**. Testing the markowitz portfolio optimization method with filtered correlation matrices. In *Proceedings of the Middle-European Conference on Applied Theoretical Computer Science*, pages 44–47, 2016.

[22] D. F. Gleich. Pagerank beyond the web. *SIAM Review*, 57(3):321–363, 2015.

[23] A. Háznagy, I. Fi, **A. London**, and T. Németh. Complex network analysis of public transportation networks: a comprehensive study. In *4th International Conference on Models and Technologies for Intelligent Transportation Systems*, pages 371–378. IEEE, 2015.

[24] M. O. Jackson. *Social and economic networks*. Princeton University Press, 2010.

[25] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[26] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[27] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[28] H. Markowitz. *Portfolio selection: Efficient diversification of investments. Cowles Foundation monograph no. 16*. New York: John Wiley & Sons, Inc, 1959.

[29] M. L. Mehta. *Random matrices*. Academic Press, 2004.

[30] Á. Merza, **A. London**, I. M. Kiss, A. Pelle, J. Dombi, and T. Németh. On the possible use of network science in the analysis of world trade (in Hungarian). *Közgazdasagi Szemle (Economic Review) LXIII. évfolyam*, pages 79–98, 2016.

[31] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[32] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker. *Handbook of educational data mining*. CRC Press, 2010.

[33] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.