Szegedi Tudományegyetem
Informatika Doktori Iskola

Hálózatalapú adatorientált módszerek alkalmazásvezérelt problémákra

PhD értekezés tézisei

Hajdu László

Témavezető:

Dr. Krész Miklós

Szeged
2021

1. Bevezetés

A hálózatok csodálatosak. Néhány közülük megtalálható az élet majdnem minden területén, a szociológiától kezdve a pénzügyi és biológiai folyamatokon keresztül egészen az emberi testig. A hálózatok jelenléte még az olyan entitások esetében is megfigyelhető, amelyek között természetes értelemben véve nem feltétlenül van kapcsolat, azonban egy a tulajdonságaik alapján definiált kapcsolati struktúrájukat leíró hálózat definiálható, lehetővé téve ezzel a tudás teljesen új módon történő kifejezését. A hálózat mint struktúra nemcsak érdekes és komplex matematikai kérdéseket vet fel, de segítségével rejtett, addig nem ismert tudás is kinyerhető a valós adatból. Az adatból, amely napjainkban az egyik legértékesebb nyersanyaggá lépett elő. A társadalommal kapcsolatos különböző tevékenységek és ezek mögöttes folyamatai hatalmas mennyiségű adatot hoznak létre, amely a napjainkban jelenlevő technológiai tudás segítségével a rendelkezésünkre áll. Mindazonáltal elmondható, hogy az adat a benne lévő tudás ismerete nélkül nem képvisel értéket, ezért az elmúlt évtizedekben a fő hangsúly azon volt, hogy hogyan tudunk kinyerni tudást és információt a meglévő adatokból. Következésképpen az adatelemzés és az adattudomány, valamint az adatközpontú módszertanok vezető kutatási területekké váltak mind a tudományos, mind az ipari területeken.

A dolgozatban a szerző hatékony algoritmusokat mutat be hálózattudományon alapuló modellekre épülő alkalmazás-orientált optimalizálási és adatelemzési feladatok megoldására. A dolgozat fő gondolata az, hogy ezeket a feladatokat a hálózatalapú megközelítés mentén kapcsoljuk össze. A feladatok egészen a meglévő rendszeren történő epidemiológiai modellezéstől, fertőzésterjedési mechanizmusok megértésén és fertőzés-/befolyásterjedés maximalizálásán vagy minimalizálásán át, a pénzügyi alkalmazásokig terjednek.

2. Alapvető definíciók és jelölések

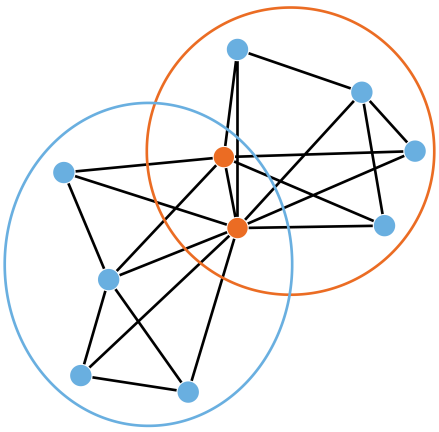
A dolgozat algoritmusainak és módszereinek többsége irányítatlan vagy irányított hálózatokon alapul. Ahhoz hogy definiáljunk egy irányítatlan (irányított) hálózatot, legyen $G(V, E)$ egy rendezett pár, ahol V a csúcsok halmaza, és E az élek vagyis rendezetlen (rendezett) csúcspárok halmaza. A csúcsok egy gráfban -tehát a V halmaz elemei- kapcsolódhatnak egymáshoz élekkel. Ha $u, v \in V$ és e egy olyan él, amely összeköti az u és v csúcsokat, akkor az e élt jelölhetjük $e = (u, v)$, $e_{u,v}$ vagy $e(u, v)$-ként, továbbá u és v *szomszédosak*. Legyen d_v vagy $deg(v)$ a v csúcs foka, amely a rá illeszkedő élek számát jelenti. Irányított hálózatok esetén az u és v csúcsok sorrendje az $e(u, v)$ él esetén az adott él irányát jelenti. Ebben az esetben az $e(u, v)$ él *kimenő* éle az u csúcsnak és *bemenő* éle a v csúcsnak. Továbbá amennyiben a G irányított, úgy minden v csúcsra szintén meghatározható a csúcs kifoka $d_{out}(v)$ valamint befoka $d_{in}(v)$. Egy adott csúcs szomszédai az irányított hálózatokban két csoportba oszthatók a rájuk illeszkedő adott él irányától függően. A *ki-szomszédok* v-hez kötődő élei a v csúcsból kifelé, míg a *be-szomszédok* kapcsolódó élei a v irányába mutatnak.

A tézisfüzet tartalmának megértéséhez továbbá szükséges definiálni a teljes gráfok fogalmát. Egy adott G egyszerű gráfhoz (hurokél és párhuzamos él nélküli gráf), ahol $|V| = n$, az élek minimális száma 0 az *üres gráfok* esetén és $n(n-1)/2$ a teljes gráfok esetén, melyek jelölése K_n. A klikk egy olyan részgráf a G gráfban, melynek csúcsai teljes gráfot alkotnak. Következésképpen, a *k-klikk* egy olyan klikk, amelynek pontosan k darab csúcsa van. Egy G gráfban a *maximum klikk* a legnagyobb klikk, ami a legtöbb csúcsot tartalmazza, míg a *maximális klikk* olyan teljes részgráf, amelynek mérete nem bővíthető tovább. Irányított hálózatokban legyen d_{v_c} a v csúcs korlátozott kifoka a c klikkben, ami a csúcs adott klikken belüli kifokát jelenti.

Definíció szerint az irányított klikk minden olyan élt tartalmaz v_1-ből v_2-be, ahol $d_{v1_c} > d_{v2_c}$, továbbá nem tartalmaz irányított kört, és minden csúcs a klikkben különböző kifokkal rendelkezik. A definíciók tekintetében fontos még, hogy beszéljünk a zárt séta fogalmáról. A hálózatokon értelmezett séta éleknek egy olyan véges vagy végtelen alternáló sorozata, mely csúcsok sorozatát köti össze, így $e_{v_i,v_1}, e_{v_1,v_2}...e_{v_{j-1},v_j}$ egy olyan séta, amely v_i-ből v_j-be megy. Amennyiben az élek nem ismétlődnek, vonalról beszélünk. A séta *zárt*, ha a v_i és v_j megegyezik.

2.1. Közösségstruktúra hálózatokban

A komplex valamint valós hálózatok gyakran rendelkeznek olyan strukturális sajátossággal, amely azt vonja maga után, hogy a hálózat csúcsai olyan halmazokba csoportosíthatóak, ahol a kapcsolatok sűrűsége viszonylag magas. A hálózatok ezen tulajdonságát közösségstruktúrának nevezzük. A szakirodalomban a közösségek számos különböző definíciója ismert, azonban a legtöbb esetben olyan sűrű részgráfokként vannak értelmezve, ahol a csúcsok közötti élek száma relatíve magas a hálózat többi részéhez képest. A közösségek két típusba sorolhatók attól függően, hogy egy csúcs mennyi közösségnek lehet a tagja. Abban az esetben ha az adott csúcs definíció szerint csak egy közösségben szerepelhet, a struktúrát *nem átfedő közösségnek* vagy *klaszternek* nevezzük. Ezzel szemben ha az adott csúcs több közösséghez is tartozhat, a struktúrát *átfedő* közösségnek nevezzük.



1. ábra. Példa átfedő közösségre.

Az előzőleg definiált sűrű részgráfok hálózatból történő kinyerését *közösségkeresésnek* nevezzük és számítási szempontból nehéz feladatnak minősül. A klaszterezés és közösségkeresés története egészen a 70-es évekre vezethető vissza, így elmondható, hogy a szakirodalomban számos megoldás található, melyek a teljesség igénye nélkül klaszterezésre a következők [27, 25, 24, 26, 19, 11] valamint átfedő közösségkeresésre a következők [12, 22, 30, 28, 3, 21]. Elmondható, hogy a valós hálózatok esetén az átfedő közösségek gyakrabban jelennek meg annak köszönhetően, hogy valós entitások gyakran több csoportnak vagy közösségnek a részei. Ebből kifolyólag a disszertáció során leginkább átfedő közösségekkel foglalkoztunk. A közösségkereső módszerek részletes áttekintése megtalálható a [10] publikációban.

2.2. Diffúziós modellek

A hálózatok és struktúrájuk hatékonyan képesek támogatni valamint kifejezni a különböző valós folyamatok modellezését. A diffúziós modellek eredetüket tekinve az epidemiológiai felhasználási

területekről származnak, ahol a folyamatok különböző vírusok és járványok terjedését fejezik ki. Mindazonáltal elmondható, hogy a folyamatok az orvosi alkalmazások mellett számos további területről is származhatnak. Modellezhetjük az információ, csőd, elvándorlás terjedését is, vagy ahogy az a disszertációban is bemutatásra került, akár viselkedési mintákat illetve pszichológiai hatásokat is.

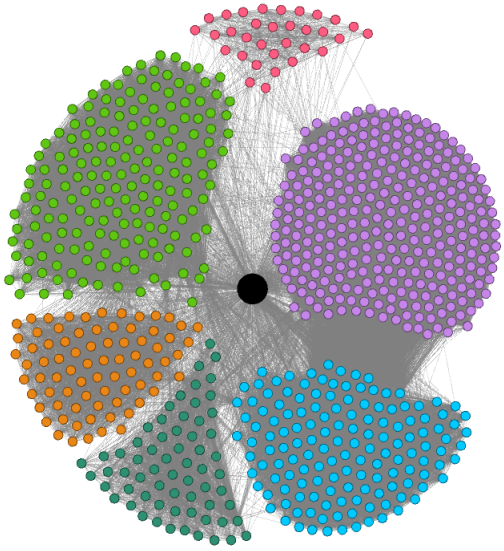
A diffúziós modellekkel kapcsolatban az első tárgyalandó témakör a compartmental modellek témaköre. Ezen módszerek diszkretizálják a különböző fertőző betegségek matematikai modelljét oly módon, hogy a populáció tagjait különböző csoportokra bontják az epidemiológiai állapotuk alapján. A témához kapcsolódóan fontos megemlíteni, hogy az első diffúziós modellek differenciál egyenleteken alapszanak és nem hálózatokon kerültek bevezetésre. A compartmental modellekkel kapcsolatban az első módszer amit meg kell említenünk az SIR amely Kermack illetve McKendrick által került bevezetésre 1927-ben [9]. Az SIR során a populáció tagjai három különböző állapotban lehetnek. Az S állapotban azon fogékony egyének vannak, akik ha egy fertőzött egyénnel találkoznak, átkerülnek a fertőzött állapotba. Az I csoport tagjai a fertőző egyének, akik kontaktba kerülve fogékony egyénekkel, képesek megfertőzni azokat. Az R állapotban azok vannak, akik előzőleg fertőzöttek voltak és meggyógyultak, immunisok vagy pedig elhunytak. Az előző koncepció egy egyszerűbb fajtája az SI modell, ahol az egyének csak S és I állapotban lehetnek, megtalálható a disszertáció második fejezetében. A hálózatalapú epidemiológiai modellek tekintetében jó összefoglalót adnak a következő szakirodalmak [8, 18].

Egy másik, a disszertáció témaköréhez szorosan kapcsolódó módszer a Független Kaszkád (IC), melynek matematikai definíciója Kempe, Kleinberg és Tardos nevéhez köthető [6]. A módszer során a csúcsok aktív illetve inaktív állapotban lehetnek attól függően, hogy fertőzöttek-e vagy sem. A fertőzés adott élvalószínűségek alapján történik, és a hálózat csúcsainak az állapotváltozása csak inaktívból aktívba történhet. Ahhoz, hogy definiáljuk a Független Kaszkád modellt, legyen $G(V, E)$ egy hálózat, ahol $\forall(v, u) \in E$ élen van egy $p(v, u)$ valószínűség, amely a kapcsolati erősséget reprezentálja és $0 < p(v, u) \leq 1$. A modellben a fertőzés iteratív módon történik és a folyamat egy kezdeti A_0 halmazból indul. A folyamat megértéséhez definiáljuk az A_i halmazt, mint az i-edik iterációban az A_{i-1} csúcsok által megfertőződött halmazt. A módszer terminál, amennyiben az újonnan megfertőződött csúcsok halmaza üres. A fertőzött csúcsok számának a várható értéke az A_0 kezdeti fertőzött halmaz esetén legyen $\sigma(A_0)$. A Független Kaszkáddal kapcsolatban egy optimalizálási probléma definiálható, amelynek a neve *fertőzésmaximalizálási probléma* [7] és ahol a cél egy olyan A_0 kezdeti fertőzött halmaz keresése, amely maximalizálja a fertőzött csúcsok várható értékét. A fertőzés értékének tehát a $\sigma(A)$-nak a kiszámítása #P- hard probléma [5], azonban szimulációval bármilyen pontosság elérhető.

3. A disszertáció főbb témakörei

A disszertáció fő tézispontjai 5 különböző témakör mentén csoportosíthatóak, melyek a következők:

- Közösségkeresés és fertőzések modellezése tömegközlekedési hálózatokon.
- Fertőzésterjedés és közösségek.
- Uplift Hálózati modell célzott beavatkozások optimalizálására.
- Időalapú hálózatok vizsgálata csalásdetektálásra a bankszektorban.
- Hálózatalapú műszakkiosztás.



2. ábra. Utasok közötti contact hálózat.

A továbbiakban a tézisfüzetben ezen témakörökhöz kapcsolódó kutatás eredményeit ismertetjük.

3.1. Közösségkeresés és fertőzések modellezése tömegközlekedési hálózatokon

Egy olyan új módszertan került bemutatásra, melynek a célja a tömegközlekedési hálózatok közösségstruktúrájának vizsgálata, valamint meglévő rendszerek epidemiológiai szempontok szerint történő kiértékelése. A módszer két új hálózatot hoz létre egy olyan contact hálózatból (2. ábra) amely az utasok közötti együttutazási kapcsolatokat írja le. A kapcsolódó hálózatok definíciói az 1. táblázatban láthatóak.

1. táblázat. A fejezet során használt hálózatok definíciói

	Contact hálózat	Transfer hálózat	Community hálózat
Csúcsok	Utasok	Elemi utascsoportok	Utasok amelyek legalább két darab járaton utaztak
Élek	Az utasok között van él, ha ugyanazon járaton voltak jelen ugyanabban az időben (irányítatlan)	Az elemi utascsoportok össze vannak kötve, ha van közös utasuk (irányított)	Az utasok össze vannak kötve, ha legalább két járaton együtt utaztak (irányítatlan)
Attribútumok	Az együttutazás hossza és kezdete	Az elemi utascsoportok között "átszálló" utasok száma	Kapcsolati erősség, amely az utasok utazási mintáinak a hasonlóságát méri

A módszerünk három lépésben hozza létre a transfer hálózatot. Az első lépésben a $G(V, E)$ contact hálózatot bontjuk szét a járatok mentén, majd a második illetve a harmadik lépésben maximális klikkeket keresünk a Bron-Kerbosch algoritmus segítségével [1], és felépítjük az F transfer hálózatot (1. táblázat). A kapott hálózat hatékonyan ábrázolja a rendszernek a használatát, így a folyamat ezen részén detektáltuk az utasok által használt gyakori járatkombinációkat, illetve az utascsoportok mozgását. Ezen információk hasznosak lehetnek a tömegközlekedési társaságok számára az utasok szokásainak és viselkedésének a monitorozása szempontjából.

A kapcsolódó fejezet második részében bemutattunk egy új élsúlyozott hálózati struktúrát, amelyet H-val jelöltünk, és amely egy új link alapú kapcsolati erősséget definiál az utasok között. A kapcsolati erősség a hálózatban található élsúlyokat definiálja és figyelembe veszi az utaspárok együttmozgásait a járatok között, valamint bünteti a szokásos utazási mintákat, amelyek nem jelentenek erős rejtett kapcsolatot az utasok között. A módszer validálásra került valós bemeneten is, melynek során Twin Cities Minneapolis városának a tömegközlekedési hálózatán teszteltünk.

3.1.1. Járványterjedési alkalmazás

A tömegközlekedési hálózaton történő közösségek felderítésének az egyik alkalmazása az infrastruktúra biztonság témaköre. Az utasok közösségeinek megértése lehetővé teszi a tömegközlekedési rendszeren megjelenő fertőző betegségek hatékonyabb és pontosabb követését, mivel ezen alkalmazások egyik legfőbb kihívásai közé tartozik az egyazon járművön utazó utasok valós kapcsolatainak a felderítése. A hálózaton történő járvány szimulációhoz az SI compartmental modellt használtuk. Hasonlóan a [2]-ben található módszerhez, 100 utast random módon kezdetben fertőzöttnek tekintettünk. A modell probabilisztikus természete miatt az SI modellt $k = 10000$-szer futtattuk és a fertőzési valószínűségeket az ötödik időpillanatban tekintettük. A módszerünk kimenete a járatok vírus-előfordulási valószínűség szerinti rangsorolása volt.

3.2. Fertőzésterjedés és közösségek

A fejezetben a közösségkeresés és a diffúziós modellek kapcsolatát vizsgáltuk tovább. A fő célunk egy olyan új közösségalapú fertőzésmaximalizáló módszer bemutatása volt, amely benchmarking rendszerként is használható különböző közösségkereső módszerek rangsorolására. A fejezet első részében, miután bemutattuk a Független Kaszkád modelljét és a mohó algoritmussal történő fertőzésmaximalizálás módszerét, új közösségalapú redukciós technikák kerültek bemutatásra. Legyen $G(V*) \subset G(V)$ egy redukált csúcshalmaz, ahol a mohó algoritmus minden iterációban a $G(V*) \setminus A_0$ halmazból választ. Legyen $f(v) : v \to Z$ egy függvény, amely minden csúcshoz hozzárendel egy egész számot. A csúcsok az $f(v)$ értékük alapján rendezettek, és csak a magas $f(v)$ értékkel rendelkező csúcsok kerülhetnek bele a $G(V*)$ halmazba.

Ahogyan azt korábban bemutattuk, a közösségek sűrű, erősen kapcsolódó részgráfok, ahol a csúcsok kapcsolata szorosabb, mint a hálózat többi részében. Amennyiben egy ilyen részgráf elég sűrű, a fertőzés vagy befolyás könnyebben terjedhet a csúcsai között. Több sűrű részgráfot összekötő csúcsok speciális pozícióval rendelkeznek a hálózatokban, mely pozíció felhasználható a fertőzésmaximalizálás hatékonyságának növelésére. Legyen $f_c(v) : v \to Z$ egy függvény a közösségi értékre, amely minden csúcshoz hozzárendeli a csúcshoz kapcsolható közösségek számát. Az ötlet azon alapszik, hogy díjazzuk a sűrű részgráfok/közösségek közötti csúcsokat. A dolgozat során két különböző redukciós módszer került bemutatásra. Az elsőnek a lépései a következők:

1. Átfedő közösségek keresése a hálózatban.

2. A módszer meghatározza az$f_c(v) : v \to Z$ értékeket minden egyes csúcsra.

3. $G(V*)$ redukált keresési tér létrehozása az $f_c(v)$ értékek segítségével.

4. Mohó alapú fertőzésmaximalizálás a $G(V*)$ keresési tér felhasználásával.

A második redukciós algoritmus, amely bemutatásra került, egy egyszerűsített heurisztika, ahol az előzőekben definiált módszer negyedik lépését kihagytuk.

3.2.1. A módszer kiértékelése irányított hálózatokon

Először a módszerünket 1080 db irányítatlan hálózaton teszteltük, melyek generálása Andrea Lancichinetti és Santo Fortunato módszerének [20] segítségével történt. A közösségi érték kiszámítására a meglévő módszerek benchmarkolása céljából nyolc különböző, a szakirodalomban található irányítatlan közösségkereső módszert használtunk. A 2. táblázat a bemutatott módszereink hatékonyságát mutatja az eredeti mohó heurisztikához képest.

2. táblázat. Azon esetek száma, ahol az adott módszer a legjobb eredményt adta. A táblázatban az algoritmus három variánsa látható: Eredeti heurisztika 20%os csökkentett keresési térrel, 10%-os keresési térrel és az egyszerűsített heurisztika.

	20%-os keresési tér	10%-os keresési tér	Egyszerűsített heurisztika
Greedy(Kempe)	207	243	335
CPM	71	61	2
COPRA	10	1	0
GCE	0	0	0
Infomap	732	769	737
MOSES	2	0	0
OSLOM	2	2	6
SBM inference	27	0	0
SLPA	29	4	0

3.2.2. A koncepció további vizsgálata: Irányított Hub Perkoláció

Az alapötlet irányítatlan közösségkereső módszerekkel való tesztelése után, a célunk a módszertan irányított hálózatokon történő tesztelése volt két irányított módszer segítségével, melyek során további, a struktúrájukkal kapcsolatos információ nyerhető ki a csúcsokról. A fejezet során bemutattuk az Irányított Hub Perkolációs algoritmust, amely egy létező irányítatlan módszer [3] kiterjesztése. A módszer struktúrája az eredeti irányítatlanéhoz hasonló, azonban egy új paraméter került bevezetésre a folyamat végén, valamint a módszer irányítatlan részeit irányítottra cseréltük. Az irányított hub perkoláció segítségével definiáltuk a hub értéket a csúcsokra. Hasonlóan a közösségi értékhez, legyen $f(v)$ $f : v \to Z$ egy függvény, amely minden csúcshoz egy egész számot rendel. Legyen $f_{hv}(v)$ egy függvény, amely minden csúcshoz hozzárendeli a h_v hub értéket, amely azt fejezi ki hogy a csúcs mennyi irányított klikkben szerepel. Az irányított hub perkolációs módszerünket összehasonlítottuk az irányított klikk perkoláció [29] és az eredeti mohó algoritmus által adott eredményekkel mesterségesen generált és valós bemenetek esetén.

3.3. Uplift hálózati modell célzott beavatkozások optimalizálására

Egy új módszer, az Uplift hálózati modell került kifejlesztésre, amely képes célzott beavatkozások optimalizálására egy negatív globális hatás csökkentése érdekében. A módszer interdiszciplináris kutatásként került bemutatásra, ahol a cél a szervezetek és cégek dolgozóinak mentális állapotának növelése volt célzott pszichológiai beavatkozások segítségével. A módszer egy kiterjesztése az Általánosított független kaszkádnak [4]. Az Uplift network model formális definíciójához legyen $G = (V, E)$ egy irányított vagy irányítatlan hálózat, ahol $\forall(v, u) \in E$ él rendelkezik egy $p(v, u)$ valószínűséggel, ahol $0 < p(v, u) \leq 1$ és $\forall v \in V$ csúcsra definiálható egy a priori $v^{apriori}$ és egy uplift v^{uplift} valószínűség, ahol mindkettő egy 0 és 1 közötti valószínűség, tehát $0 \leq v^{apriori} \leq 1$ és $0 \leq v^{uplift} \leq 1$. A $v^{apriori}$ valószínűség azt jelenti, hogy az adott csúcson nem történt beavatkozás, míg a v^{uplift} az adott csúcs beavatkozás utáni valószínűségét jelenti. A módszer célja egy olyan csúcshalmaz keresése, amely elemein a $v^{apriori}$ valószínűségek v^{uplift} valószínűségre történő csökkentése maximalizálja a fertőzés értékének a különbségét egy olyan referenciaszimulációhoz képest, ahol nem történt beavatkozás, így minimalizálva a globális fertőzöttséget.

3.3.1. Pszichológiai esettanulmány

A módszer egy olyan szociális hálózaton került tesztelésre, amelyet 14 norvég idősek otthonából származó dolgozók adatai alapján építettünk fel. A dolgozók mentális állapotának felmérése és az eredmények kiértékelése a WHO-5 kérdőív [31] segítségével történt. A célzott beavatkozások eredményeit összehasonlítottuk a random módon történő intervenciók eredményeivel. A pszichológiai esettanulmány eredményei a 3. táblázatban láthatóak.

3. táblázat. A WHO-5 százalékos pontszám átlagos egy főre eső növekedésének az összehasonlítása.

Beavatkozások száma százaléka	Átlagos javulás random intervenció esetén	Átlagos javulás célzott intervenció esetén	Célzott és random közötti különbség
10 (3.6%)	0.74	0.83	0.09
20 (7.2%)	1.43	1.61	0.18
50 (18.0%)	3.55	3.94	0.38
100 (36.0%)	6.85	7.64	0.79
200 (71.9%)	13.58	14.64	1.06

Annak ellenére, hogy a pszichológiai állapot és kedv személyek közötti terjedése viszonylag alacsony, a módszerünk képes volt az összpontszám növelésére csupán az intervenciós stratégia optimalizálásával.

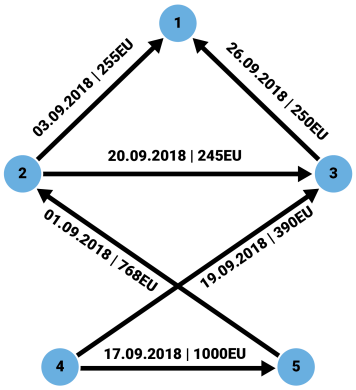
3.4. Időalapú hálózatok vizsgálata csalásdetektálásra a bankszektorban

Egy olyan módszer került bemutatásra, amely képes speciális minták temporális hálózatokból történő kinyerésére. Mivel a probléma háttere egy banki alkalmazásból származik és a definiált minták is erősen kötődnek ehhez a területhez, az algoritmust egy pénzügyi esettanulmány

segítségével mutattuk be, ahol a célunk utalási körökhöz köthető gyanús minták detektálása volt. Az utalási kör definíció szerint legyen egy zárt séta, ahol:

- A kör mentén a csúcsok ismétlődése megengedett, míg az éleké nem.
- Legyen $t_e^0, t_e^1 ... t_e^{n-1}, t_e^n$ egy érvényes kör menti időbélyegek sorozata, ahol $t_e^i \leq t_e^{i+1}, i = 0...n-1$ tehát az időbélyegek időben egymás után következnek.
- Legyen a_e^0 az első utaláshoz köthető összeg és a_e^i az ezt követő tranzakciók összegei, ahol $i = 1...n$. Legyen $a_e^0 \cdot (1-\alpha) \leq a_e^i \leq a_e^0 \cdot (1+\alpha)$, amely azt jelenti, hogy a különbség az első tranzakció összege és a kör menti többi tranzakció összege között nem lehet nagyobb egy előre definiált $0 \leq \alpha \leq 1$ által meghatározott értéknél. Az α a módszer paramétere, amely a felhasználó által lehet definiálva.

A 3. ábrán látható gráf öt különböző ügyfél tranzakcióit tartalmazza. A kördefiníció alapján az 1-2-3-1 zárt séta mentén az időbélyegek növekvő sorrendben vannak, és egy körként definiálható, amennyiben $\alpha \geq 0.1$, mindazonáltal például az 5-2-3-4-5 zárt séta nem kör az időbélyegek és a kör mentén található összegek miatt.



3. ábra. Példa tranzakciós gráfra, amely utalási kört tartalmaz.

Az algoritmus először rendezi a hálózatban található éleket a rajtuk lévő időbélyegek alapján, így utalások idő szerint rendezett listáját kapjuk. A második lépésben a módszer egy módosított mélységi bejárást alkalmazva köröket gyűjt ki a hálózatból. Minden kör a kezdőutalásával kerül azonosításra, tehát miután egy adott kezdeti utalásból minden kört megtaláltunk, az algoritmus beállítja az utalást végleg látogatottra annak érdekében, hogy további körök ne tartalmazzák az adott utalást.

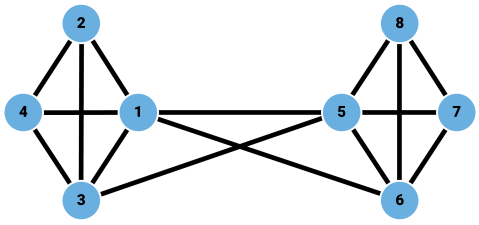
3.4.1. Valós esettanulmány banki tranzakciós hálózaton

Az módszerünket 2016-ból származó valós anonimizált adatokon teszteltük. A módszer egy csalásdetektáló rendszer része, amely 2018 óta működik egy ismert magyar banknál. A pénzintézet visszajelzései alapján, a bemutatott módszer 2-4 gyanús csaláshoz köthető aktivitást detektál minden hónapban és az elmúlt években sikeresen felderített több csaláshoz köthető ügyfélkört is.

3.5. Hálózatalapú műszakkiosztás

Bemutattunk egy új hálózatalapú heurisztikát a műszakkiosztási feladat megoldására. Ahhoz, hogy definiáljuk a problémát, legyen C a dolgozók vagy munkavállalók, S pedig az elvégzendő műszakok halmaza. A cél a dolgozók műszakokhoz rendelése úgy, hogy a költség minimális legyen. Következésképpen, legyen $f = S \rightarrow C$ egy hozzárendelés, ahol minden műszakhoz pontosan egy ember van hozzárendelve. Az alap foglalkoztatási költsége egy dolgozónak a szerződésében van definiálva. Optimális esetben minden munkavállaló az elvárt munkaideje szerint dolgozik, azonban abban az esetben, ha a munkaidő meghaladja az elvárt munkaidőt, a munkaadó köteles túlóradíjat biztosítani a munkavállaló részére. Így a költség definiálható a $cost = \alpha * overtime + \beta * employment\ cost$ képlettel, ahol az overtime a túlfoglalkoztatottság, az employment cost a foglalkoztatási költség valamint az α és a β előre definiált súlyok. A cél olyan megoldást találni, amely minimalizálja az előzőleg definiált formulát, valamint megfelel a különböző szabályozásoknak, amelyek tipikusan a különböző nemzetek egyedi törvényei által definiáltak.

4. ábra. Példa két egymás utáni naphoz kötődő konfliktus gráfra.



A módszerünk az optimalizálási feladatot egy olyan konfliktus gráf segítségével ábrázolja, ahol a csúcsok a műszakok és akkor vannak összekötve, ha nem végezheti őket egyazon dolgozó. A 4. ábrán egy példagráf látható. A bemutatott módszer egy kétlépéses gráfszínezési heurisztika, amely kiszínezi az előzőleg definiált gráfot, ahol a színek a dolgozók egy halmazának felelnek meg. Az algoritmus lépései a következők:

1. Kezdeti műszakkiosztás.

a. Dolgozók számának becslése.

b. Szabadnapminták generálása.

c. Konfliktus gráf felépítése.

d. A hálózat kezdeti színezése.

2. A hálózat újraszínezése Tabu Search algoritmus segítségével.

3.5.1. A módszer kiértékelése és valós esettanulmány

A tesztelés során a módszer viselkedését mind mesterségesen generált, mind pedig valós adatokon is kiértékeltük. A valós adat a szegedi tömegközlekedési vállalattól származik. A módszer eredményeit összevetettük az egészértékű programozási modell által adott eredményekkel kis méretű bemenetek esetén. Következtetésképpen elmondható, hogy a módszerünk képes volt

felülmúlni az IP modellt a futásidő tekintetében, valamint sok esetben az elméleti alsó korlátot (lehető legjobb megoldás) is elérte, ezzel gyors és hatékony megoldást szolgáltatva az eredeti problémánkra.

4. A szerző munkájának tézispontszerű összefoglalása

A következő fejezet a disszertáció fő eredményeit összegzi. A 4. táblázatban a disszertáció tézisei, fő fejezetei és a kapcsolódó publikációk közötti kapcsolat látható.

4. táblázat. Kapcsolat a tézispontok, fejezetek és kapcsolódó publikációk között.

	[15]	[13]	[14]	[23]	[16]	[17]
I.	•					
II.		•	•			
III.				•		
IV.					•	
V.						•
Fejezet	2.	3.	3.	4.	5.	6.

I. Bemutattunk egy módszert, amely képes a tömegközlekedési rendszerek sebezhetőségének vizsgálatára az utasok közötti együttutazási hálózat segítségével. A fejezet első részében egy új közösségdefiníciót vezettünk be, amely az utasok tömegközlekedés használata közben létrejövő kapcsolatait fejezi ki, valamint a tömegközlekedési rendszer használatát elemeztük és gyakori járatkombinációkat detektáltunk, segítve ezzel a tömegközlekedési vállalatokat a szolgáltatásaik fejlesztésében. Az utasok közösségeinek detektálása és a rendszer elemzése után megvizsgáltuk egy járvány során előforduló lehetséges forgatókönyveket, valamint a járvány szempontjából veszélyes járatok is detektálásra kerültek a tömegközlekedésen történő vírusterjedés csökkentése érdekében. A módszerünket Minneapolis tömegközlekedési rendszerén teszteltük. A tézisponthoz kapcsolódó publikáció a következő helyen megtalálható [15].

II. Egy érdekes koncepciót mutattunk be, amelynek célja a fertőzési modellek és közösségkeresés összekapcsolása. A fejezet fő ötlete azon a megfigyelésen alapszik, hogy a fertőzési folyamatok könnyebben terjedhetnek egy közösségben, tehát egy sűrű részgráfban, mint a hálózat többi részében. Egy olyan új módszertant mutattunk be, ahol egyrészről a közösségkereső módszerek segítségével képesek voltunk hatékonyabb fertőzésmaximalizálási eljárások kifejlesztésére, másrészről a fertőzésmaximalizálás segítségével a közösségkereső módszerek fertőzési szempontból rangsorolhatóak. A koncepció bemutatása után a módszerek tesztelésre kerültek random és valós adatokon, valamint a tárgyalt közösségkereső módszereket is rangsoroltuk. A fejezet alapjául szolgáló kutatás megtalálható a következő publikációkban [13, 14].

III. Egy pszichológiai esettanulmány segítségével mutattunk be egy az Általánosított Független Kaszkádon alapuló fertőzésmaximalizálási modellt. A fejezet első részében az új modell kifejtése mellett egy olyan módszer kifejlesztésére koncentráltunk, amely képes minimalizálni egy, a hálózaton már jelenlévő fertőzés terjedését. A fejezet második részében egy

olyan esettanulmányt vizsgáltunk, ahol a cél a mentális állapot javítása volt célzott intervenciók segítségével. A modellt és a módszert egy olyan szociális hálózaton teszteltük, amelyet norvég idősek otthonaiban dolgozók adatai alapján építettünk fel. A kapcsolódó tudományos publikáció megtalálható a következő helyen [23].

IV. Egy új módszer került bemutatásra temporális hálózatok elemzésére banki környezetben történő csalásdetektálás céljából. A kutatás motivációja a pénzügyi szektorból származik, ahol speciális hálózati motívumok detektálása felderíthet különböző csalásokhoz köthető tevékenységeket. A kapcsolódó rész fő feladata az volt, hogy az általános pénzügyi csalások tárgyalása után bemutassunk egy újszerű körkereső módszert, amely képes a körutalások detektálására banki tranzakciós hálózatokban. A módszert 2016-ból származó valós adatokon teszteltük, valamint egy ismert magyar bank csalásdetektáló rendszerének a része 2018 óta. A tézisponthoz kapcsolódó publikáció megtalálható a következő helyen [16].

V. A műszakkiosztási problémát oldottuk meg egy olyan heurisztika segítségével, amely egy hálózatot használ az optimalizálási feladat keresési terének ábrázolására. A módszerünk célja a foglalkoztatottság költségének csökkentése, valamint egy olyan megoldás létrehozása volt, amely megfelel a törényileg előírt szabályozásoknak. A kétlépéses módszerünket random és egy tömegközlekedési vállalat által szolgáltatott valós bemeneteken teszteltük. A fejezet végén a módszer által adott eredmények összehasonlításra kerültek a matematikai modell által adott eredményekkel. A fejezethez kapcsolódó kutatás megtalálható a következő helyen [17].

Selected references

[1] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, Sept. 1973.

[2] A. Bóta, L. Gardner, and A. Khani. Identifying critical components of a public transit system for outbreak control. *Networks and Spatial Economics*, 08 2017.

[3] A. Bóta and M. Krész. A high resolution clique-based overlapping community detection algorithm for small-world networks. *Informatica*, 39:177–187, 01 2015.

[4] A. Bóta, M. Krész, and A. Pluhár. Approximations of the generalized cascade model. *Acta Cybernetica*, 21:37–51, 01 2013.

[5] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1029–1038, 09 2010.

[6] K. David, K. Jon, and T. Éva. Maximizing the spread of influence through a social network. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137-146, 07 2003.

[7] K. David, K. Jon, and T. Éva. Influential nodes in a diffusion model for social networks. *Lecture Notes in Computer Science*, 3580:1127–1138, 07 2005.

[8] O. Diekmann and J. Heesterbeek. Mathematical epidemiology of infectious diseases: Model building, analysis and interpretation. *Wiley Series in Mathematical and Computational Biology, Chichester, Wiley*, 01 2000.

[9] I. Foppa. W.o. kermack and a.g. mckendrick: A seminal contribution to the mathematical theory of epidemics (1927). *A Historical Introduction to Mathematical Modeling of Infectious Diseases*, pages 59–87, 01 2017.

[10] S. Fortunato. Community detection in graphs. *Physics Reports*, 486, 06 2009.

[11] M. Girvan and M. Newman. Community structure in social and biological networks. *proc natl acad sci*, 99:7821–7826, 11 2001.

[12] S. Gregory. Finding overlapping communities in networks by label propagation. *New journal of Physics*, 12(10):102018, 2010.

[13] L. Hajdu, A. Bóta, and M. Krész. Community based influence maximization in the independent cascade model. *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 237–243, 2018.

[14] L. Hajdu, A. Bóta, and M. Krész. Evaluating the role of community detection in improving influence maximization heuristics. *(Under Review)*, 2021.

[15] L. Hajdu, A. Bóta, M. Krész, A. Khani, and L. M. Gardner. Discovering the hidden community structure of public transportation networks. *Networks and Spatial Economics*, 20, 03 2020.

[16] L. Hajdu and M. Krész. Temporal network analytics for fraud detection in the banking sector. *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, pages 145–157, 2020.

[17] L. Hajdu, A. Tóth, and M. Krész. Graph coloring based heuristic for crew rostering. *Acta Cybernetica*, (4):643–661, 2020.

[18] M. Keeling and K. Eames. Networks and epidemic models. *Journal of the Royal Society, Interface / the Royal Society*, 2:295–307, 10 2005.

[19] B. Kernighan and S.-D. Lin. An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.*, 49:291–307, 1970.

[20] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 80:016118, 08 2009.

[21] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11, 03 2008.

[22] C. Lee, F. Reid, and A. M. abd N. Hurley. Detecting highly overlapping community structure by greedy clique expansion. *arXiv:1002.1827*, 2010.

[23] D. Lipovac, L. Hajdu, S. Wie, and A. Q. Nyrud. Improving mental wellbeing in organizations with targeted psychosocial interventions. *Business Systems Research*, 11:86–98, 2020.

[24] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 01 2004.

[25] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1*, 1967.

[26] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: An overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 09 2017.

[27] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69:026113, 03 2004.

[28] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 07 2005.

[29] G. Palla, I. Farkas, P. Pollner, I. Derenyi, and T. Vicsek. Directed network modules. *New Journal of Physics*, 9, 04 2007.

[30] M. Rosvall and C. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105:1118–1123, 02 2008.

[31] WHO. Wellbeing measures in primary health care/the depcare project. 2020.