Szegedi Tudományegyetem

Természettudományi és Informatikai Kar

Számítógépes Optimalizálás Tanszék

Informatika Doktori Iskola

Módszerek valós hálózatokon játszódó folyamatok leírására és elemzésére

Doktori értekezés tézisei

Bóta András

Témavezetők:
Dr. Krész Miklós
Dr. Pluhár András

Szeged, 2014.

1. Bevezető

A gráfelmélet gyökerei a Königsbergi hidak rejtvényéhez vezetnek vissza. A rejtvényt és a megoldását egy 1736-os cikkében publikálta Leonhardt Euler. Azóta sok minden kiderült a gráfok matematikai tulajdonságairól és alkalmazhatóságáról más tudományterületeken, mint például a szociológia, a biológia vagy akár az optimalizálás és az operációkutatás. A dolgok menete azonban megváltozott, amikor a számítógépek elérhetővé és megfizethetővé váltak a legtöbb kutató számára. Ezek az eszközök lehetővé tették nekik, hogy összegyűjtsék, tárolják, megosszák és tanulmányozzák a valós életben megfigyelt hálózatokat és a hozzájuk kapcsolódó nagyméretű adathalmazokat. Ez a változás vezetett el a hálózatkutatás nevű interdiszciplináris tudományterület kialakulásához, amelyet alapítói a valós hálózatok megfigyelésének és elemzésének szenteltek, a módszertanát pedig a matematika, fizika, szociológia és az informatika területéről kölcsönözték. A hálózatkutatás témái közé tartozik a hálózat szerveződési elveinek, például a fokszám eloszlásnak és a csoportok születésének vizsgálata, valamint a hálózatokon játszódó folyamatok leírása.

A disszertáció célja, hogy ismertesse a szerző munkásságát a hálózatkutatás három nagy témakörében: az átfedő közösségkeresés, a dinamikus közösségkeresés és a fertőzési folyamatok területén. Az ismert közösségkereső algoritmusokkal kapcsolatos tapasztalataink és Csizmadia et al. munkája alapján kifejlesztettük a hub perkolációs közösségkereső algoritmust, amely képes különböző szerkezetű valós hálózatok kezelésére, illetve egy adott hálózatban a közösségek különböző rétegeinek felfedezésére. A módszerünket többek között Newman ismert benchmark hálózatain és Lancichinetti gráf generátora segítségével teszteltük. Két esettanulmányt is bemutattunk. Az egyik magyar cégek tulajdonosi hálózatát vizsgálja, a másik egy magyar és egy angol szóasszociációs hálózat szerkezetét hasonlítja össze.

A közösségi és gazdasági hálózatokkal kapcsolatos tapasztalataink vezettek minket a közösségkeresés egy másik változatához, a dinamikus közösségkereséshez. Palla et al. munkája alapján létrehoztunk egy módszert, amely képes két időben szomszédos gráf közösségeinek egymáshoz rendelésére. A módszerünk tizenegy esemény azonosítására képes, és működése független a használt statikus közösségkereső tulajdonságaitól. A módszerünket két valós életből származó példán teszteltük.

A disszertáció utolsó témaköre a gráfokon játszódó fertőzési folyamatok vizsgálata különös tekintettel ezek gazdasági alkalmazásaira. Gyakori probléma, hogy a fertőzési folyamatok alapjául szolgáló él fertőzési valószínűségek nem állnak rendelkezésre, ilyenkor ezeket becslik vagy valamilyen konstans értéket használnak helyettük. Ennek a feladatnak a hatékony megoldására hoztuk létre az inverz fertőzési problémát, melynek feladata, hogy kiszámítsa az él fertőzési valószínűségeket a fertőzési folyamat bemeneteinek és eredményének ismeretében. Bemutattuk az inverz fertőzési probléma alapjául szolgáló általánosított kaszkád modellt, annak heurisztikáit, és megadtunk egy tanuló algoritmust a probléma megoldására. A módszerünket először mesterséges fertőzési feladatokon teszteltük, majd egy részletes esettanulmányt is publikáltunk, melynek célja a csőd események előrejelzése volt banki tranzakciós hálózatokon.

2. Alapfogalmak

A dolgozatban lévő algoritmusok és módszerek gráfokon alapulnak. Jelölje G gráf pont- és élhalmazait $V(G)$ és $E(G)$. Majdnem minden esetben *irányítatlan* gráfokat használunk. A gráf élein *élsúlyok* formájában, a pontjain és élein *attribútumok* formájában tárolhatunk numerikus információt. A gráfok minden esetben *összefüggőek*, ez lehet valamilyen, a dolgozatban nem említett szűrési folyamat eredménye.

A disszertációban szereplő gráfok mindegyike *komplex hálózat*, és a valós életből vett megfigyeléseken alapszik. Ezek, ellentétben az egyszerűbb gráf osztályokkal, mint a fák vagy a rácsok, gyakran nem triviális topológiai tulajdonságokkal rendelkeznek. A komplex hálózatokat a forrásuk szerint többféle csoportba sorolhatjuk: ismertségi hálózatok, információs hálózatok, technológiai hálózatok, gazdasági hálózatok, stb. A hálózatkutatás legfontosabb eredményei közé tartozik az a felismerés, hogy a forrásuktól függetlenül, ezeknek a gráfoknak a viselkedése hasonló.

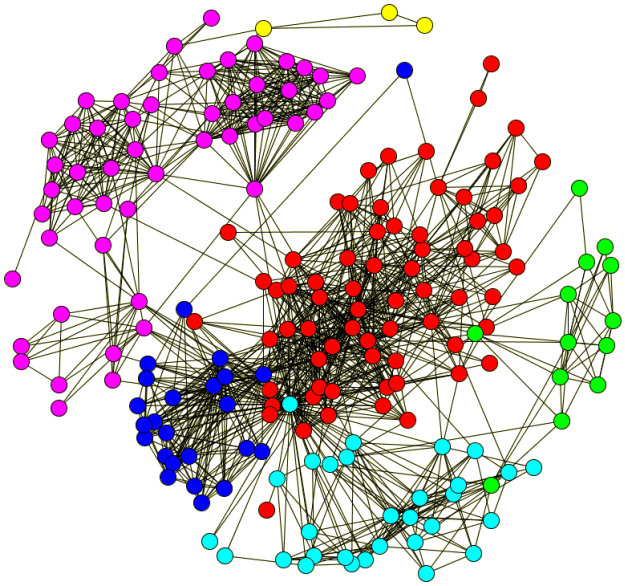
Közösségkeresés

A társadalmi kapcsolatokban megfigyelhető az a fajta viselkedés, hogy az emberek hajlamosak csoportokba szerveződni az érdeklődésük, munkájuk vagy hobbijaik szerint. Az ismertségi hálózatokon ez a viselkedés úgy jelenik meg, hogy a gráf pontjai olyan halmazokba rendeződnek, melyeken belül sűrűk a kapcsolatok, köztük pedig viszonylag kevés. Ezt a jelenséget *közösségstruktúrának* nevezzük, a közösségstruktúra kiszámítását pedig *közösségkeresésnek*. A jelenség azonban nem csak az ismertségi hálózatok jellemzője, más komplex hálózatokban is megtalálható.

A közösségek jelensége ismert, azonban pontos definíciót nehéz rájuk találni. A hagyományos közösségkereső módszerek a közösségeket diszjunkt ponthalmazokként értelmezik, és a következő intuíciót használják. A ponthalmaz olyan partícióit keresik, melyek maximalizálják a halmazokon belül futó élek számát és minimalizálják a közöttük lévő élek mennyiségét. További cél, hogy értelmes közösségeket találjanak, például elvetik azt a triviális megoldást, mely csak egy minden pontot magába foglaló halmazt tartalmaz. A közösségkeresésről egy átfogó összefoglalót itt olvashatunk [14].

A hagyományos közösségkeresők csak diszjunkt ponthalmazokat engednek meg. A valós életben tapasztaltak szerint viszont egy pont több közösségnek is tagja lehet, így eljuthatunk az *átfedő közösségek* fogalmához. Ezzel a fogalommal komolyabban először Palla et al. foglalkozott munkájában [29]. Ugyanitt definiálták a klikk perkolációs átfedő közösségkereső algoritmust is. A maximális klikkek keresése és összefűzése más közösségkereső alapjául is szolgál, de sok más megközelítési mód is szerepel a szakirodalomban [17, 18, 25].

A közösségkeresés egy másik módja a közösségek időbeni változásának követése, ez a *dinamikus közösségkeresés*. A dinamikus közösségkeresés gráfok egy $\{G_i\}_{i\in\mathcal{T}}$ sorozatán alapul, ahol $\mathcal{T} = \{1, \ldots, \ell\}$ egy diszkrét időskálát jelez, a cél pedig annak a felderítése, hogy egy gráf közösségstruktúrája milyen kapcsolatban van a sorozatban szereplő többi gráf struktúrájával. A dinamikus közösségkereső algoritmusok az időben szomszédos gráfok párjait hasonlítják össze, illetve definiálják a közösségek közti lehetséges *események* vagy történések halmazát. A definíció után a szomszédos grá-



1. ábra. Egy közösségi háló egy szeletének közösségstruktúrája. A közösségeket Girvan és Newman módszerével találtuk meg [26].

fok közösségeit egy algoritmus egymáshoz rendeli, és mindegyikhez hozzákapcsolja az események egyikét. A legismertebb algoritmusok itt olvashatóak [1, 28].

Fertőzési modellek

Gráfokon sokféle folyamatot lehet értelmezni. A disszertáció egy nagyobb része gazdasági események terjedésével foglalkozik, ez pedig a *fertőzési* vagy *diffúziós modellek* területe. Ezek a modellek sokféle jelenség terjedésével foglalkoznak: viselkedési minták, információ, betegségek, stb. Ezeket a modelleket ugyancsak sokféle tudományterület használja. Számunkra a legfontosabb modell a független kaszkád (Independent Cascade) modell, amit Domingos és Richardson vezetett be [12, 19].

A fertőzési modellek állapotokat rendelnek a gráf pontjaihoz, amelyek a fertőzési bizonyos fázisait jelzik. A független kaszkád modell állapotai a következők. Egy pont lehet *fertőzhető*, ekkor nem fertőzött de más pontok megfertőzhetik. Egy pont lehet *fertőzött*, ekkor más pontokat is meg tud fertőzni, végül egy pont lehet *eltávolított*, ekkor fertőzött de már nem képes más pontokat megfertőzni. Ezekre az állapotokra *inaktív*, *újonnan aktivált* és *aktív* pontokként is fogunk hivatkozni.

Egy fertőzési modellt olyan folyamatként lehet felírni, melynek két bemenete van. Az első egy súlyozott gráf, melyek élein 0 és 1 közötti valós számok vannak definiálva: $\forall e \in E(G), 0 \leq w_{u,v} \leq 1$ ezek az él fertőzési valószínűségek. A második bemenet a kezdeti fertőzött pontok $A_0 \subset V(G)$ halmaza. Ezek a pontok a folyamatot fertőzött állapotban kezdik. Maga a folyamat lépésekben vagy iterációkban történik, és t iterációban fejeződik be amennyiben $A_t = \emptyset$, az eredménye pedig a folyamat során megfertőződött $A = \bigcup_{i=0}^{t} A_i$ pontok halmaza. A modellt eredetileg irányított gráfokra fogalmazták meg, de könnyű általánosítani irányítatlan esetre is, ekkor az él fertőzési valószínűségek szimmetrikusak $w_{u,v} = w_{v,u}$.

A fertőzési modell meghatározza, hogy a pontok hogyan fertőzhetik meg egymást. Legyen az i-edik iterációban újonnan aktivált pontok halmaza $A_i \subseteq V(G)$. Az $i+1$ iterációban minden $u \in A_i$ pont megpróbálja aktiválni az $v \in V \setminus \cup_{0 \leq j \leq i} A_j$ inaktív szomszédait a köztük lévő $w_{u,v}$ él fertőzési valószínűség szerint. Amennyiben ez sikeres a következő iterációban v újonnan aktivált lesz. Ha több mint egy pont akarja ugyanazt a pontot aktiválni, a próbálkozások egymástól függetlenül bármilyen sorrendben történhetnek. Ha $A_t = \emptyset$ akkor a folyamat véget ért a t iterációban. Könnyű belátni, hogy a folyamat véges számú lépésben véget ér.

3. A disszertáció új tudományos eredményei

A disszertáció eredményei három tézispontba sorolhatóak:

1. Nagy felbontású rugalmasan paraméterezhető átfedő közösségkereső algoritmus kifejlesztése és alkalmazásai.

2. Dinamikus közösségkereső algoritmus kifejlesztése, ami alkalmas nagy méretű valós hálózatok kezelésére.

3. Él fertőzési valószínűségek becslésére szolgáló módszertan és fertőzési modell fejlesztése és banki alkalmazásai.

Ebben a fejezetben mindhárom pontot ismertetni fogjuk.

3.1. A hub perkolációs módszer

A *hub perkolációs* módszert [4] azzal a céllal hoztuk létre, hogy rugalmasságával minél több alkalmazás feltételeinek képes legyen megfelelni. Ez egy klikk alapú átfedő közösségkereső algoritmus komplex hálózatokra, melynek paraméterei képesek a megtalált közösségstruktúra finomhangolására. Ez lehetővé teszi, hogy közösségeket többféle felbontásban is megfigyeljük, legyenek ezek nagy, kevéssé átfedő csoportok, vagy akár kis, szorosan összekapcsolódó közösségek. Az algoritmus fejlesztésére nagy hatással voltak azok a benyomások, melyeket más algoritmusok, mint a klikk perkolációs vagy az N^{++} módszer [3] használata során szereztünk. Ezen felül segítették a munkánkat Newman [15, 27] és Zachary [30] ismert benchmark hálózataival kapcsolatos tapasztalataink, melyeknek nyomai az algoritmus sok részleteiben megtalálhatóak.

A hub perkolációs algoritmus két fogalmon alapul: a *klikkeken* és a *hub-okon*. A klikkek adott k számú ponttal rendelkező teljesen összefüggő részgráfok. Egy klikk maximális, ha nem részhalmaza egyetlen más klikknek sem. A maximális klikkek halmazának megadása általános esetben nagyon nehéz feladat, komplex hálózatokon azonban a Bron-Kerbosch algoritmus változatai [10, 13] elfogadható teljesítményt nyújtanak. A klikkeken belül minden pont kapcsolatban van egymással, így ezek a közösségek legtisztább formájának tekinthetők. Ezt az intuíciót a legtöbb klikk alapú közösségkereső használja.

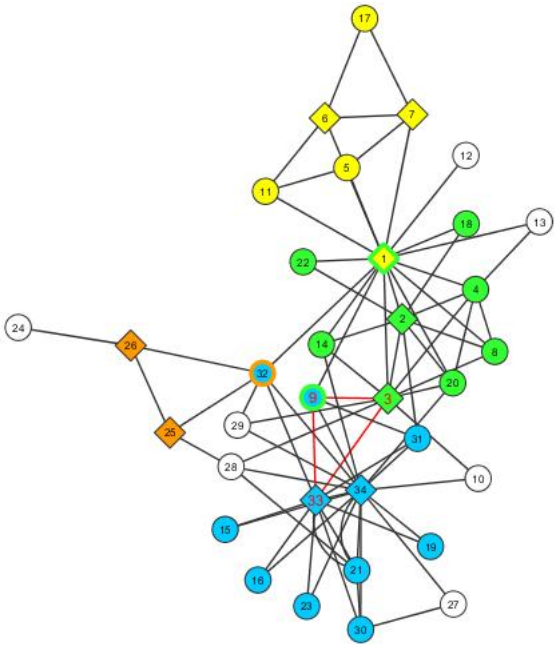
A valós életből származó példák alapján egy másik megfigyelést is tehetünk. A közösségeken belül nem mindegyik pont egyenrangú. A csoportok összetartásában

egyes pontoknak nagyobb szerep jut másoknál. Ezeket a pontokat hub-oknak nevezzük és segítségükkel kapcsoljuk össze és terjesztjük ki a klikkeket. A hub-ok azonosítása a *hub kiválasztási stratégia* feladata. A disszertációban több stratégiát is megadtunk.

A módszert a következő lépésekre lehet felosztani:

1. Keressük meg a 3-nál nagyobb maximális klikkek halmazát.
2. Válasszuk ki a hub-okat a hub kiválasztási stratégia szerint. Ennek a stratégiának lehet egy q paramétere.
3. Találjuk meg a hub-okból álló k-klikkeket, és terjesszük ki őket egy korlátozott perkolációs szabály szerint.
4. Olvasszuk össze azokat a közösségeket, melyekben ugyanazok a hub-ok szerepelnek.

A módszer működését kétféle módon befolyásolhatjuk. Az egyik a k szűrési paraméter, a másik a hub kiválasztási stratégia, és annak lehetséges q paramétere. Ezek a paraméterek teszik lehetővé módszerünk sokrétű felhasználhatóságát. Azt, hogy a paraméterek milyen hatással vannak a megtalált közösségstruktúrára, sokféleképpen mértük: a közösségek számával, a közösségek méretének eloszlásával, a közösségek közti átlagos átfedéssel és a közösségen kívüli pontok számával. Erre a célra Newman benchmark hálózatait [15, 27] használtuk.



2. ábra. Zachary karate klub hálózatának közösségei [30]. A hub-okat rombuszok jelölik. A többszínű pontok több közösséghez is tartoznak. A medián hub kiválasztási stratégiát használtuk $k = 2$ paraméterrel. A $9, 3, 33$ pontok egy külön közösséget alkotnak, és a 9-es pont három közösséghez is tartozik egyszerre.

Eredmények és esettanulmányok

A módszerünk hatékonyságát többféle módon mértük. Lancichinetti és Fortunato közösség alapú gráf generátorát [25] használtuk arra, hogy összemérjük módszerünk eredményeit a népszerű klikk perkolációs módszer [29] eredményeivel. A mutual information [24] nevű mérőszámot használtuk arra, hogy összehasonlítsuk a gráfok természetes közösség struktúráját[1] az említett két módszer által találtakkal. Megmutattuk, hogy a módszerünk jobban közelíti a természetes közösségeket, amennyiben ezek között nagyok az átfedések.

Két esettanulmányt is megadtunk. Megvizsgáltuk a magyar céginformációs adatbázisból készített tulajdonosi hálózat közösségstruktúráját, és megállapítottuk a vállalatok csoportjainak földrajzi elhelyezkedését, ipari tevékenységeit és korát. A megfigyelt cégfajtára (Kft) jellemző, hogy cégeinek közösségei földrajzilag egymáshoz közel helyezkednek el, és jellemzően ugyanabban a gazdasági szektorban tevékenykednek. A cégek közösségeinek kora azonban szórást mutat.

Egy részletes esettanulmányban megvizsgáltuk és összehasonlítottuk egy magyar és egy angol szó-asszociációs hálózat közösségstruktúráját [9]. Azonosítottuk azokat a szavakat, melyek a közösségek középpontjaiban állnak, ezek tipikusan kategórianevek, gyakori melléknevek vagy kollektív főnevek. A szavak egy része mindkét hálózatban szerepel, ezek jellemzően alapvető igényekhez vagy mindennapi eseményekhez kapcsolódnak, de nagyok a különbségek a két gráf között.

3.2. Dinamikus közösségkeresés

A *dinamikus közösségkereső* algoritmusunk [6] kifejlesztéséhez két valós életből származó feladat vezetett el minket. Ezeknek a feladatoknak több olyan követelményük is volt, amelyeket a meglévő módszerek nem tudtak egyidejűleg teljesíteni. Az első a gyorsaság volt. A vizsgált feladatokban nagy gráfok szerepeltek, amelyek rendelkeztek ugyan a komplex hálózatok tulajdonságaival, de nagyon sok ponttal és éllel rendelkeztek.

A második probléma az általunk kiválasztott statikus közösségkereső algoritmus volt. Mindenképpen átfedő közösségkereső algoritmust szerettünk volna használni, de a disszertációban említett okok miatt nem voltunk megelégedve a klikk perkolációs módszerrel. Palla et al. [28] dinamikus közösségkereső módszere azonban ennek az algoritmusnak az egyik különleges tulajdonságát használja. Végül az N^{++} módszert [3] választottuk közösségkeresőnek, és elvetettük Palla módszerének használatát.

Az utolsó követelmény a közösségekkel kapcsolatos *események* száma és jellege volt. A Palla által javasoltakat nem találtuk alkalmasnak arra, hogy leírják a vizsgált gráfjaink dinamikáját.

Az általunk használt megközelítés Palla et al. munkáján alapszik, de ezt sok helyen módosítottuk, hogy használhatóbbá tegyük a módszert. Kibővítettük a közösség-események halmazát, és átalakítottuk az összehasonlítást végző algoritmust, így az már nem függ a felhasznált közösségkereső különleges tulajdonságaitól.

[1] Ezeket a gráf generátor adta meg.

A módszerünk egy gráfsorozat két szomszédos G_1 és G_2 elemét hasonlítja össze a következő módon.

1. Határozzuk meg G_1 és G_2 közösségeit egy közösségkereső segítségével.
2. Számoljuk ki G_1 és G_2 unióját G_U-t, majd futtassuk le a közösségkeresőt ezen is.
3. Rendeljük hozzá G_U közösségeit G_1 és G_2 közösségeihez.
4. A hozzárendelések összegzésével megkapjuk a G_1 és G_2 közötti eseményeket.

A módszerünkkel két valós életből származó hálózat közösség struktúrájának dinamikáját vizsgáltuk. Az eredményeink azt mutatják, hogy lényeges különbség van a két hálózat viselkedése között. Az egyik hálózat stabil, de lassan és folyamatosan veszít a közösségeiből, a másik gyorsan változik, de megtartja közösségeinek számát.

3.3. Inverz fertőzés

Az inverz fertőzés vizsgálatának ötlete Csernenszky et al. 2009-ben publikált munkájából [11] származik. A dolgozatuk célja a fertőzési modellek használhatóságának vizsgálata volt banki tranzakciós hálózatokon, az eredményeik pedig azt mutatták, hogy csőd előrejelzésre használt módszerek pontosságát javítja, ha figyelembe veszik azt a hatást, melyet a hálózat pontjai egymásra fejtenek ki. Ez az úgynevezett hálózati hatás, annak a valószínűsége, hogy egy pont viselkedése hat a szomszédjaira is. Ez a jelenség adta az ötletet, hogy kifejlesszünk egy módszert, amely képes ezeknek a valószínűségeknek a becslésére, ez pedig az *inverz fertőzési problémához* [8] vezetett el minket.

A mi módszerünk lényegesen különbözik a szakirodalomban meglévőktől [16, 22], bár a terület kevéssé kutatott. A meglévő módszerekkel ellentétben a mienknek nem előfeltétele a fertőzési folyamat lépéseinek ismerete. Ehelyett a banki tranzakciós hálózatokkal kapcsolatos tapasztalatainkra épít, mint a vállalati csőd-valószínűségek becsléseinek használata, vagy a pont és él attribútumok jelenléte.

Az inverz fertőzési probléma létrehozásához szükséges volt további algoritmusok és módszerek kifejlesztése. Az egyik legfontosabb ezek közül az *általánosított kaszkád* fertőzési modell [7]. Ez a modell a független kaszkád modellre épül, és egy valószínűségeken alapuló keretrendszert használ a fertőzési folyamat bemeneteinek és kimeneteinek a megadására.

Az általánosított kaszkád modell

Az általánosított kaszkád (Generalized Cascade) modell minden egyes ponthoz egy 0 és 1 közötti valós p_v értéket rendel, amely azt a valószínűséget jelzi, amivel az adott pont még a folyamat kezdete előtt megfertőződik. Ezeket az értékeket *a priori eloszlásnak* nevezzük. A hálózat pontjai még a fertőzési folyamat kezdete előtt egymástól függetlenül megfertőződhetnek ezekkel a valószínűségekkel. Az a priori eloszlás fog megfelelni a független kaszkád modell kezdeti fertőző pontjainak ebben a modellben. Az általánosított kaszkád modell képes az a priori fertőzéseknek és ezek

hálózati hatásainak a kiszámítására. Az a priori eloszláshoz hasonlóan a modell kimenetét egy *a posteriori eloszlás* adja meg, melynek p'_v értékei jelzik, hogy a pontok mekkora valószínűséggel fertőződnek meg a folyamat során. A pontok közötti fertőzés terjedésének módja ugyanaz, mint a független kaszkád modellnél megismert, bár az általánosított modell képes más fertőzési modellek leírására is.

A következő módon definiálhatjuk az általánosított kaszkád modellt:

Az általánosított kaszkád modelll: Legyen adott egy megfelelően súlyozott G gráf és a p_v a priori fertőzési értékek, a modell ezek alapján kiszámolja az a posteriori eloszlás p'_v értékeit minden $v \in V(G)$-re.

Az általánosított kaszkád modell kiszámítása #P-nehéz, ezért négy heurisztikus módszert adtunk meg a felgyorsítására [5].

- A teljes szimulációs módszer (Complete Simulation) Kempe et al. [20] módszerének átalakítása az általánosított kaszkád modell feltételeinek megfelelően.
- Az élszimulációs módszer (Edge Simulation) a szimuláció és az algebrai számítások olyan kombinációja, mely csökkenti a szimulációkra annyira jellemző szórás értékét.
- Ha az él fertőzési valószínűségek kicsik, akkor a fertőzések nem terjednek túl messze a kiindulópontjaiktól. A korlátozott szomszédság heurisztika (Neighborhood Bound Heuristics) ezt a jelenséget használja ki.
- A kaszkád fertőzési folyamatot le lehet cserélni egy könnyebben kiszámolható modellre, ez az ALE heurisztika.

Az inverz fertőzési probléma

Az inverz fertőzési probléma (Inverse Infection Problem) [8] definíciója hasonló az általánosított kaszkád modell definíciójához:

Az inverz fertőzési probléma: Legyen adott egy súlyozatlan G gráf, a p_v a priori és a p'_v a posteriori fertőzési értékek. Számoljuk ki ezekből a w_e él fertőzési valószínűségeket minden $e \in E(G)$-re.

Minden egyes élvalószínűséget kiszámolni egyrészt nehéz, másrészt a probléma maga aluldefiniált, még akkor is, ha a hálózat nagyon kicsi. Ehelyett feltételezzük, hogy az éleken lévő valószínűségek a pontokon és éleken levő más attribútumoknak egy (normalizált) függvényeként állnak elő. Így csak ezeknek a függvényeknek az együtthatóit kell kiszámolnunk, és mivel az attribútumok és együtthatók száma korlátozott, a probléma megoldhatóvá válik.

A fentiek alapján definiálhatunk egy tanuló módszert.

- A definíció szerint az a posteriori eloszlás a probléma bemeneteihez tartozik. Ezt felhasználhatjuk egy tanuló módszer referencia adathalmazaként.
- Az attribútum függvények kezdeti együtthatóit véletlenszerűen választjuk észszerű korlátok közül.

- Az attribútum függvények, az együtthatók és az a priori eloszlás segítségével kiszámolhatunk egy, az együtthatókhoz kapcsolható a posteriori eloszlást.

- Egy hibafüggvény segítségével meghatározzuk az eltérést a referencia eloszlás és az újonnan számolt a posteriori eloszlás között.

Ez a feladat a globális optimalizálás tipikus esete, a megoldás módszerének pedig néhány egyéb próbálkozás után Kennedy és Mendes FPSO algoritmusát [21] választottuk.

A módszerünk korlátait mesterséges fertőzési feladatokkal fedeztük fel. Megállapítottuk az optimalizálás stabilitását és pontosságát, egy általános megközelítési módot adtunk a megfelelő attribútum függvény kiválasztására, megvizsgáltuk milyen hatásai vannak a használt heurisztikáknak a módszer eredményeire, és kipróbáltuk hogyan viselkedik a módszerünk, ha a bemenetei hiányosak vagy hibásak. A módszer alkalmas arra, hogy kevés lépésszámmal is pontos becslést adjon az él fertőzési valószínűségekre, ezt az attribútumok száma és az attribútum függvények alakja csak kevéssé befolyásolja még pontatlan bemenetek esetében is.

A dolgozat végén bemutattuk az inverz fertőzi probléma alkalmazását egy banki tranzakciós hálózaton [2]. Az alkalmazás célja a csődesemények előrejelezetőségének javítása volt. Mivel az inverz fertőzési probléma fejlesztésekor figyelembe vettünk banki alkalmazásokkal kapcsolatos követelményeket, így az esettanulmányt könnyű volt megvalósítani. A módszerünk jobb becsléseket nyújtott, mint a hagyományos módszerek, képes volt arra, hogy azonosítsa a csődre leginkább hajlamos cégeket, és 2013-ban hozzávették az OTP Bank eszközrendszeréhez.

4. Tézispontok

A dolgozat eredményeit a következő tézispontokban összegezzük.

4.1. Átfedő közösségkeresés

Az első pontunkban egy *új nagy felbontású rugalmasan paraméterezhető átfedő közösségkereső algoritmust és alkalmazásait* ismertettünk.

1. Részletes leírását adtuk a hub perkolációs közösségkereső algoritmusnak és több hub kiválasztási stratégiának beleértve egy olyat is, mely képes súlyozott hálózatok kezelésére. Megvizsgáltuk, hogy a módszer paraméterei milyen hatással vannak a megtalált közösségstruktúrára, és ezt a hatást Newman [15,27] ismert hálózatain demonstráltuk.

2. Lancichinetti közösség alapú gráf generátorát [25] használtuk arra, hogy összemérjük módszerünk eredményeit a népszerű klikk perkolációs módszer [29] eredményeivel.

3. Módszerünk segítségével megvizsgáltuk a magyar céginformációs adatbázisból készített tulajdonosi hálózat közösségstruktúráját.

4. Egy részletes esettanulmányban megvizsgáltuk és összehasonlítottuk egy magyar és egy angol szó-asszociációs hálózat közösségstruktúráját [9].

A hub perkolációs módszert magát, a paraméterek hatásának demonstrációját Newman hálózatain, valamint a gazdasági esettanulmányt (1-3) egy publikáció foglalja össze, melyet egy nemzetközi folyóirathoz nyújtottuk be [4]. Elfogadásáról még nem kaptunk visszajelzést. A szó-asszociációs hálózatokkal kapcsolatos esettanulmány (4) egy nemzetközi konferencia kiadványában fog megjelenni [9].

4.2. Dinamikus közösségkeresés

A disszertáció második tézispontja egy *dinamikus közösségkereső algoritmus kifejlesztése* volt, ami *alkalmas nagy méretű valós hálózatok kezelésére.*

1. Bevezettük a módszerünk által értelmezett tizenegy dinamikus közösségi eseményt. Részletes leírást adtunk a dinamikus közösségkereső algoritmushoz, és ismertettük az időbonyolultságát is.

2. A módszerünkkel két valós életből származó hálózat közösségstruktúrájának dinamikáját vizsgáltuk.

A tézisponttal kapcsolatos eredményeink egy nemzetközi folyóiratban jelentek meg [6].

4.3. Inverz fertőzés

Az utolsó tézispont egy *él fertőzési valószínűségek becslésére szolgáló módszertan és fertőzési modell kifejlesztése* volt valamint ennek *banki alkalmazásai.*

1. Definiáltuk az általánosított kaszkád fertőzési modellt, amely az inverz fertőzési probléma alapjául szolgál.

2. Az általánosított kaszkád modell kiszámítása nehéz, ezért létrehoztunk négy különböző heurisztikus módszert a számítások felgyorsítására: teljes szimuláció, élszimuláció, korlátozott szomszédság heurisztika és az ALE heurisztika.

3. Definiáltuk az inverz fertőzési problémát. Legyen adott egy súlyozatlan G gráf, a p_v a priori és a p'_v a posteriori fertőzési értékek. Számoljuk ki ezekből a w_e él fertőzési valószínűségeket minden $e \in E(G)$-re.

4. Megadtunk egy tanuló módszert, ami a problémát egy optimalizálási feladatra egyszerűsíti.

5. A módszerünk korlátait mesterséges fertőzési feladatokkal fedeztük fel. Megállapítottuk az optimalizálás stabilitását és pontosságát, egy általános megközelítési módot adtunk a megfelelő attribútum függvény kiválasztására, megvizsgáltuk milyen hatásai vannak a használt heurisztikáknak a módszer eredményeire, és kipróbáltuk hogyan viselkedik a módszerünk, ha a bemenetei hiányosak vagy hibásak.

6. Végül ismertettük az inverz fertőzési probléma alkalmazását egy banki tranzakciós hálózaton.

Az inverz fertőzéssel és az általánosított kaszkád modellel kapcsolatos előzetes eredményeinket (1, 3) egy rövid publikációban közöltük egy nemzetközi folyóiratban [7]. Az általánosított kaszkád modell részletes leírását, elemzését, valamint a kapcsolódó heurisztikákat (1, 2) egy nemzetközi folyóiratban jelentettük meg [5]. Az inverz fertőzési probléma, a tanuló módszer és ezeknek a kiértékelése mesterséges fertőzési feladatokon (3-5) egy nemzetközi konferencia kiadványában található meg [8]. A banki alkalmazás (6) egy ismert nemzetközi folyóiratban fog megjelenni [2].

Irodalomjegyzék

[1] S. Asur, S. Parthasarathy and D. Ucar, An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data* **3**(4), 2009.

[2] A. Bóta, A. Csernenszky, L. Győrffy, Gy. Kovács, M. Krész, A. Pluhár, Applications of the Inverse Infection Problem on bank transaction networks. Accepted for publication in: *Central European Journal of Operations Research.*

[3] A. Bóta, L. Csizmadia, A. Pluhár, Community detection and its use in Real Graphs. *Proceedings of the 2010 Mini-Conference on Applied Theoretical Computer Science*, 95–99, 2010.

[4] A. Bóta, M. Krész, A high resolution clique-based overlapping community detection algorithm for small-world networks. Submitted.

[5] A. Bóta, M. Krész and A. Pluhár, Approximations of the Generalized Cascade Model. *Acta Cybernetica* **21** 37–51, 2013.

[6] A. Bóta, M. Krész, A. Pluhár, Dynamic Communities and their Detection. *Acta Cybernetica* **20** 35–52, 2011.

[7] A. Bóta, M. Krész and A. Pluhár, Systematic learning of edge probabilities in the Domingos-Richardson model. *Int. J. Complex Systems in Science*, **1(2)** (2011) 115–118.

[8] A. Bóta, M. Krész and A. Pluhár, The inverse infection problem. *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, 75–83, IEEE, 2014. http://dx.doi.org/10.15439/978-83-60810-58-3.

[9] A. Bóta, L. Kovács, The community structure of word association graphs. Accepted for publication in: *The Proceedings of the 9th International Conference on Applied Informatics*, Eger, 2014.

[10] C. Bron, J. Kerbosch, Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* **16** (9): 575–577, 1973. doi:10.1145/362342.362367.

[11] A. Csernenszky, Gy. Kovács, M. Krész, A. Pluhár, T. Tóth, The use of infection models in accounting and crediting. *Challenges for Analysis of the Economy, the Businesses, and Social Progress* Szeged (2009) pp. 617–623.

[12] P. Domingos, M. Richardson, Mining the Network Value of Costumers. *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, ACM (2001) 57–66.

[13] D. Eppstein, D. Strash, Listing all maximal cliques in large sparse real-world graphs. *Experimental Algorithms*, Springer Berlin Heidelberg, 364–375, 2011.

[14] S. Fortunato, Community detection in graphs. *Physics Report*, **486**(3):75–174, 2010.

[15] M. Girvan, M. E. J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, **99**(12):7821–7826, 2002.

[16] A. Goyal, F. Bonchi, L.V.S. Lakshmanan, Learning influence probabilities in social networks. *Proceedings of the third ACM International Conference on Web search and data mining.* ACM (2010) 241–250.

[17] S. Gregory, Finding overlapping communities in networks by label propagation. *New J. Phys.*, **12**(10):103018, 2010.

[18] E. Griechisch, A. Pluhár, Community Detection by using the Extended Modularity. *Acta Cybernetica*, **10**:69–85, 2011.

[19] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the Spread of Influence though a Social Network. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2003) 137–146.

[20] D. Kempe, J. Kleinberg, E. Tardos, Influential Nodes in a Diffusion Model for Social Networks. *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP)*, Springer-Verlag (2005) 1127–1138.

[21] J. Kennedy, R. Mendes, Neighborhood topologies in fully informed and best-of-neighborhood particle swarms. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews.* **36** (4) (2006) 515–519.

[22] M. Kimura, K. Saito, Tractable models for information diffusion in social networks. *Knowledge Discovery in Databases*, Lecture Notes in Computer Science Springer Berlin / Heidelberg, (2006), 259–271.

[23] A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, **80**(1):016118, 2009.

[24] A. Lancichinetti, S. Fortunato, J. Kertész Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, **11**(3):033015, 2009.

[25] A. Lancichietti, F. Radicchi, J. J. Ramasco, S. Fortunato, Finding statistically significant communities in networks. *PLoS One*, **6**(4):e18961, 2011.

[26] M. E. J. Newman, Detecting community structure in networks. *Eur. Phys. J. B* **38**, 321–330 (2004).

[27] M. E. J. Newman, The structure of scientific collaboration networks. *PNAS*, **98**(2):404–409, 2001.

[28] G. Palla, A.-L. Barabási and T. Vicsek, Quantifying social group evolution. *Nature* **446**, 664–667 (2007).

[29] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043):814–818, 2005.

[30] W. W. Zachary, An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 452–473, 1977.