UNIVERSITATEA BABEŞ–BOLYAI
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Teză de Doctorat

Rezumat

Modele de Inteligență Computațională Pentru Probleme de Influență

CONDUCĂTOR ȘTIINȚIFIC:
PROF. DR. HORIA F. POP

STUDENT:
TAMÁS-ZSOLT KÉPES

2024

Cuprins

Abstract

Teza se concentrează în principal pe grafuri și pe componente importante ale grafurilor. Importanța în acest context este definită în mai multe moduri, o abordare a importanței fiind influența sau capacitatea unei componente de a răspândi informații. Căutarea nodurilor influente într-un graf ne-a condus la pregătirea a mai multor articole despre acest subiect. Aceste articole au format ulterior prima jumătate a acestei teze, concentrându-se pe călătoria în timp pe care am făcut-o în investigarea nodurilor influente și a influenței în general. Atenția noastră a fost direcționată în multiple direcții, de la lucrări introductive privind abordările care utilizează Optimizarea Extremală, Cascade, valoarea Shapley și o abordare Monte Carlo, până la combinarea acestor factori și algoritmul rezultat.

A doua definiție a importanței pe care am investigat-o a fost problema componentelor critice. Nodurile și muchiile critice, împreună cu componentele critice ale hipergrafului, au fost toate investigate. A doua parte a tezei s-a concentrat pe problema componentelor critice în toate tipurile de rețele, având o structură care s-a axat mai mult pe rezultate individuale decât pe călătoria în timp. La final, au fost prezentate mai multe aplicații și cazuri de utilizare relevante.

Cuvinte cheie

Influență, Maximizarea Influenței, Componente Critice, Detectarea Nodurilor Critice, Optimizare Extremală, Cascade, Algoritmi Genetici, Valoarea Shapley, Grafuri, Hipergrafuri

Capitolul 1.

Introducere

1.1 Introducere

Cercetarea asumată de această teză este una bazată pe ani de lucrări și publicații anterioare. Ideea metricilor de rețea și analiza rețelelor este un subiect interesant, motivat de utilitatea potențială în domenii care se referă la orice tip de rețea, cum ar fi rețelele rutiere. [36], rețelele sociale [29] chiar și rețelele de comunicare [20] cum ar fi Internetul în sine. Lucrarea se va concentra pe două subiecte separate, care au fost ambele legate de măsurile de rețea, dar care au fost suficient de distincte, deoarece diferă în obiectivele lor. Cele două părți majore ale acestei teze se concentrează pe problema Maximizării Influenței, prezentate în detaliu în Capitolul 3, respectiv problema componentelor critice și variantele acestei probleme, prezentate în Capitolul 4.

Un exemplu bun pentru analiza influenței poate fi o firmă de publicitate în căutarea celor mai influente personalități de pe rețelele sociale sau analiza răspândirii virusului. În ceea ce privește componentele critice, un exemplu bun poate fi analiza eșecurilor rețelelor sau analiza interacțiunilor între deținuți pentru a prezice deținuții critici și a preveni revoltele.

1.2 Obiective

Obiectivul principal asumat de această lucrare a fost asamblarea completă a tuturor lucrărilor anterioare efectuate în timpul procesului de cercetare doctorală, bazate pe ani de publicații, cu mai multe teme de cercetare diferite, dar totuși legate între ele. Părțile separate ale cercetării, care se concentrează pe cele două familii majore de probleme, aveau obiective distincte, aceste obiective fiind prezentate mai jos.

Problema Maximizării Influenței se concentrează pe găsirea nodurilor cu cea mai mare

importanță într-o rețea, importanța în contextul influenței poate fi definită ca abilitatea de difuzare a informațiilor a unui nod, cu alte cuvinte, capacitatea unui nod de a răspândi informații. Obiectivul principal al Maximizării Influenței a evoluat pe parcursul procesului de cercetare. Obiectivele inițiale au fost introducerea conceptelor de bază ale Teoriei Jocurilor în problema Maximizării Influenței și identificarea unui algoritm de optimizare de bază și a unui model de propagare pentru a testa conceptele noastre din Teoria Jocurilor. Odată cu introducerea acestor blocuri de bază ale cercetării noastre asupra influenței, introducerea ideelor noi a fost graduală, cu experimentarea diferitelor modele de difuzie, diferite concepte de teoria jocurilor și calcule matematice. În cele din urmă, am ajuns la obiectivul principal al cercetării noastre asupra Maximizării Influenței, care constă în introducerea unui algoritm complet, care ia în considerare fiecare detaliu al cercetării noastre din domeniul influenței și le aplică unui algoritm complet și complex. Această călătorie și culminarea proceselor sunt prezentate în partea dedicată influenței a tezei în Capitolul 3.

Problema Componentelor Critice a început ca problema Detectării Nodurilor Critice, o problemă similară cu cea a maximizării influenței, cu obiectivul principal al Detectării Nodurilor Critice fiind capacitatea de a determina noduri importante, de data aceasta importanța fiind definită ca fiind criticitatea, unde nodurile critice sunt nodurile care, atunci când sunt eliminate, ar degrada maxim rețeaua. Explorările inițiale ale acestei probleme își aveau rădăcinile în problema Detectării Nodurilor Critice, dar s-au transformat ulterior într-o problemă mai generală a Componentelor Critice, deoarece nu mai vorbeam despre rețele simple sau chiar doar despre noduri în cercetarea noastră, ci despre o abordare din ce în ce mai generalizată a problemei Criticității. Două întrebări simple pe care le-am pus: Ce poate fi critic într-o rețea? și: Ce efect au diferitele tipuri de rețele asupra conceptului de criticitate? Obiectivul cercetării noastre a fost să răspundem la aceste întrebări, fiind dezvoltate mai multe variante ale problemei Componentelor Critice de-a lungul anilor. Obiectivul principal al tezei este colectarea și organizarea acestor cercetări separate, prezentând similarități și evoluția gândurilor noastre despre Criticitate, toate prezentate în Capitolul 4.

1.3 Contribuții Originale

Contribuțiile originale pe care acest proces de cercetare doctorală le-a propus au fost numeroase. În primul rând, în general, principala contribuție a cercetării noastre a fost o multitudine de algoritmi de optimizare variate pentru problemele de influență și criticitate, toate construite una pe cealaltă și toate oferind rezultate din ce în ce mai bune, comparabile sau chiar depășind rezultatele algoritmilor de ultimă generație din literatură.

Separând cele două capitole, principalele noastre contribuții în domeniul Maximizării Influenței au fost introducerea și elaborarea conceptelor din Teoria Jocurilor împreună cu algoritmul de Optimizare Extremală, pentru a redefini și reformula problema Influenței ca un joc cooperativ, cu noduri ca jucători. Această explorare a conceptului nou a fost foarte limitată în literatură, motiv pentru care am încercat să inovăm în această direcție. O altă inovație a venit odată cu introducerea valorii Shapley, o valoare mai specializată în teoria jocurilor care poate calcula contribuția individuală la un joc, și traducerea valorii Shapley în limbajul Influenței. Mai târziu, introducerea unei aproximări a valorii Shapley a fost principala inovație, deoarece, în măsura în care puteam să observăm, procesul în care am aproximat valoarea Shapley era unic. În cele din urmă, introducerea câtorva noi îmbunătățiri la procesul nostru a condus la dezvoltarea algoritmului nostru final, care a fost principala contribuție pe care cercetarea noastră a adus-o domeniului Influenței.

Pentru problemele legate de Criticitate, contribuțiile noastre nu au urmat o abordare liniară; în schimb, am încercat să inovăm în toate domeniile detectării a Componentelor Critice, cu mai multe abordări diferite, cum ar fi Algoritmii Genetici, Abordările Lacome (Greedy), și chiar variante ale algoritmului de Optimizare Extremală fiind propuse. Cea mai originală contribuție a noastră în domeniu trebuie să fie explorarea Hipergrafurilor în legătură cu Criticitatea, și mai specific, explorarea metricilor de centralitate legate de Hipergrafuri în legătură cu Criticitatea, care a fost o idee extrem de nouă.

Am propus o serie de aplicații pentru ambele probleme, cum ar fi analiza rețelelor de citare pentru problema Maximizării Influenței, și analiza pieței de valori, inflației și rețelelor politice pentru problema Detectării Criticității. Pe lângă analizele prezentate pe diferitele rețele, am propus o nouă metrică de robustețe a rețelelor, o abordare practică, care folosește unul dintre algoritmii noștri de criticitate pentru a determina robustețea unei rețele; această nouă metrică se comportă foarte bine în comparație cu alte metrici robustețe cunoscute, prezentându-se ca o alternativă bună la măsurările existente.

Capitolul 2.

Noțiuni teoretice de bază

2.1 Definiții legate de Grafuri

Definiție 2.1.1 (Grafuri). Un Graf este reprezentat a fiind cuplul $G = (V, E)$, unde $V = \{v_1, v_2, ..., v_n\}$ este setul de noduri sau vârfuri, iar $E = \{e_1, e_2, ..., e_n\}$ este setul de muchii. În grafurile tradiționale, setul E conține perechi de noduri. Valorile n și m denotă numărul de noduri, respectiv de muchii.

Un graf orientat este un tip de graf în care fiecare muchie are o direcție, cu un nod de pornire și un nod de sosire. O muchie orientată este de obicei reprezentată folosind o pereche ordonată de două noduri.

Definiție 2.1.2 (Vecini). Având două noduri $v, w \in V$, v și w sunt considerate vecini dacă și numai dacă $\exists e \in E$ muchie astfel încât $e = (v, w)$.

Definiție 2.1.3 (Grad, Grad de intrare, Grad de ieșire). Gradul unui nod $v \in V$ poate fi dat ca $|\{w \in V$, unde v și w sunt vecini$\}|$, în esență, cardinalitatea mulțimii care conține fiecare nod care este vecin cu v.

Gradul de intrare și gradul de ieșire sunt definite doar pentru grafurile orientate, gradul de intrare al unui nod v este numărul de muchii care se îndreaptă spre nodul respectiv, în timp ce gradul de ieșire al lui v este numărul de muchii care pleacă din nodul v.

Definiție 2.1.4 (Drumuri). Un drum într-un graf G poate fi definit ca o secvență de vârfuri $v_1, v_2, v_3, ..., v_n$, unde fiecare pereche de vârfuri adiacente (v_i, v_{i+1}) este conectată printr-o muchie în graf. Cu alte cuvinte, pentru fiecare $i = 1..n - 1$, există o muchie (v_i, v_{i+1}) în graf.

Definiție 2.1.5 (Cicluri). Un cerc într-un graf G poate fi definit ca un drum $v_1, v_2, v_3, ..., v_n$, unde $n \geq 3$, iar vârfurile v_1 și v_n sunt conectate printr-o muchie (v_n, v_1).

Definiție 2.1.6 (Arbori). Un arbore este un graf neorientat, conectat și aciclic. Un arbore este perechea $T = (V, E)$, unde V este mulțimea de noduri și E este mulțimea de muchii.

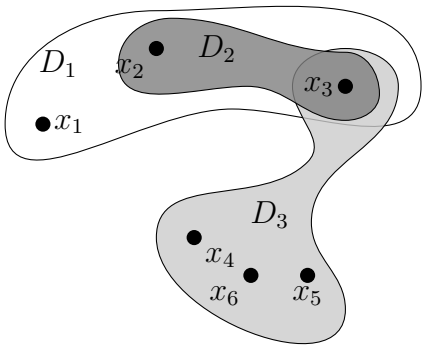

Figura 2.2.1: Un exemplu simplu de un hipergraf cu șase noduri și trei hiper-muchii

Definiție 2.1.7 (Componente conectate). Dat fiind un graf $G = (V, E)$, unde V este mulțimea vârfurilor și E este mulțimea muchiilor, o componentă conectată C este un subansamblu de vârfuri $C \subseteq V$astfel încât pentru fiecare pereche de vârfuri $(u, v) \in C$,există un drum în G care leagă u și v.

2.2 Definiții legate de Hypergrafuri

Hipergrafurile, introduse și formalizate în [7], pot fi considerate generalizări ale grafurilor simple, cu o definiție similară cu cea a grafurilor simple, prezentată în Definiția 2.2.1.

Definiție 2.2.1 (Hipergrafuri). Un hipergraf poate fi definit ca o pereche $\mathcal{H} = (X, \mathcal{D})$, unde $X = \{x_1, x_2, ..., x_n\}$ este mulțimea de noduri, $\mathcal{D} = \{D_1, D_2, ..., D_m\}$ este mulțimea de hiper-muchii, constând din submulțimi ale lui X, iar n și m se referă la numărul de noduri și hiper-muchii, respectiv.

Exemplu 2.2.1. Un exemplu pentru un hipergraf poate fi observat în Figura 2.2.1. Hipergraful are șase noduri, $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, și trei hiper-muchii $D = \{D_1, D_2, D_3\}$, $D_1 = \{x_1, x_2, x_3\}$, $D_2 = \{x_2, x_3\}$, $D_3 = \{x_3, x_4, x_5, x_6\}$.

Unele definiții clasice ale grafurilor trebuie redefinite într-un mediu de hipergraf.

Definiție 2.2.2 (Vecini). Având două noduri $v, w \in V$, v și w sunt considerate vecini dacă și numai dacă $\exists D$ hiper-muchie astfel încât $u \in D$ și $v \in D$.

Definiție 2.2.3 (Grad). Gradul unui nod $u \in X$ poate fi exprimat ca $|\{v \in V$, unde u și v sunt vecini$\}|$, în esență, cardinalitatea mulțimii care conține fiecare nod care este vecin cu u, Dacă mai multe noduri împărtășesc aceeași hiper-muchie cu u, fiecare nod din aceeași hiper-muchie se numără ca un grad separat.

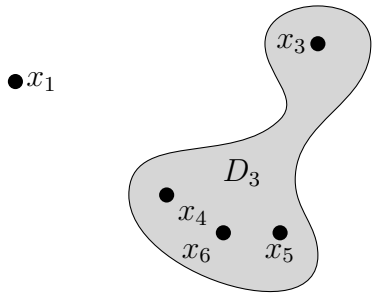

Figura 2.2.2: Ștergere puternică a nodului x_2 din hipergraful prezentat în Figura 2.2.1

Definiție 2.2.4 (Drumuri). Un drum într-un hipergraf $\mathcal{H}$ poate fi definit ca o secvență de vârfuri $v_1, v_2, v_3, ..., v_n$, unde fiecare pereche adiacentă de vârfuri (v_i, v_{i+1}) împărtășesc o hiper-muchie în hipergraf. Cu alte cuvinte, pentru fiecare $i = 1..n-1$, există o hiper-muchie D astfel încât $v_i \in D$ și $v_{i+1} \in D$.

Definiție 2.2.5 (Componente conectate). Dat fiind un hipergraf $\mathcal{H} = (X, D)$, unde X este mulțimea vârfurilor și D este mulțimea hiper-muchiilor, o componentă conectată C este un subansamblu de vârfuri $C \subseteq X$ astfel încât pentru fiecare pereche de vârfuri $(u, v) \in C$, există un drum în hipergraf $\mathcal{H}$ care leagă u și v.

Definiție 2.2.6 (Ștergerea puternică a nodurilor). Dat fiind un hipergraf $\mathcal{H} = (X, D)$, unde X este mulțimea vârfurilor și D este mulțimea hiper-muchiilor, ștergerea puternică a unui nod constă în eliminarea unui nod v din mulțimea de noduri X, împreună cu fiecare hiper-muchie la care nodul v este conectat. Alte noduri din aceleași hiper-muchii nu sunt luate în considerare la ștergere și rămân în loc.

Exemplu 2.2.2. Un exemplu de ștergere puternică a unui nod din hipergraful dat în Exemplul 2.2.1 este prezentat în Figura 2.2.2.

Definiție 2.2.7 (Ștergerea slabă a nodurilor). Dat fiind un hipergraf $\mathcal{H} = (X, D)$, unde X este mulțimea vârfurilor și D este mulțimea hiper-muchiilor, ștergerea slabă a unui nod constă în eliminarea unui nod v din mulțimea de noduri X. Hiper-muchia care conține nodul v este șters numai dacă nodul v este singurul nod din acea hiper-muchie.

Exemplu 2.2.3. Un exemplu de ștergere slabă a unui nod din hipergraful dat în Exemplul 2.2.1 este prezentat în Figura 2.2.2.

2.3 Probleme de optimizare

Problemele de optimizare sunt o clasă de probleme în care, începând de la un stadiu de bază, trebuie să se ajungă la un stadiu nou, îmbunătățit, ajungând la o soluție pentru problema dată.

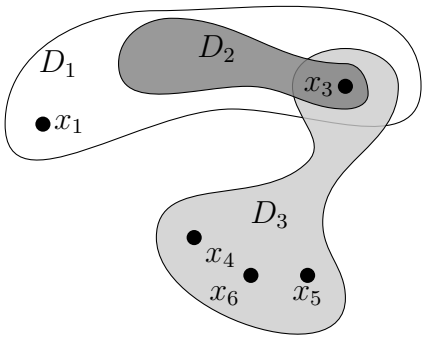

Figura 2.2.3: Ștergere slabă a nodului x_2 din hipergraful prezentat în Figura 2.2.1

Această soluție trebuie să fie una care oferă o cantitate crescută de satisfacție pentru o anumită metrică arbitrară.

Definiție 2.3.1 (Probleme de optimizare). Problemele de optimizare pot fi definite astfel:

$$P = (X, f, \omega)$$

unde, P este problema de optimizare, X este spațiul de căutare, f este funcția obiectiv și ω este mulțimea de constrângeri.

Definiție 2.3.2 (Funcția obiectiv). Funcția obiectiv f este funcția pe care dorim să o optimizăm și poate fi definită astfel: $f : X \to Y$, unde X este spațiul de căutare, în timp ce $Y \in \mathbb{R}$ este un număr real.

Definiție 2.3.3 (Minimum global). $x^* \in X$ este un minim global pentru funcția obiectiv $f : X \to Y$ if $f(x^*) \leq f(x), \forall x \in X$.

Definiție 2.3.4 (Problemă de optimizare cu un singur obiectiv). Problema de optimizare cu un singur obiectiv, care minimizează, poate fi descrisă astfel:

$$\min_{x \in \mathbb{R}^n} f(x)$$

,

astfel ca $h_j(x) = 0, j = 1, ..., J$ și $g_k(x) \leq 0, k = 1, ..., K$, unde h_j și g_k sunt orice număr de funcții de constrângere, în timp ce $x = (x_1, ..., x_n)$.

Este important să observăm că definițiile de mai sus se referă doar la minimizare, dar definiții similare pot fi date și pentru problemele de maximizare.

Capitolul 3.

Maximizarea influenței

3.1 Introducere în problema de maximizare a influenței

Problema de Maximizare a Influenței (IMP) este problema care descrie răspândirea informațiilor într-o rețea, nodurile care răspândesc informația la rate mai mari sunt numite noduri influente. Scopul principal al IMP este de a găsi un set de noduri dintr-o rețea, care pot răspândi informația la un grad maxim, maximizând astfel influența setului găsit. Răspândirea informațiilor este simulată folosind modele de difuzie, care sunt numite și modele de propagare. Utilizând aceste modele, putem defini funcția de influență, care este folosită pentru a estima numărul de noduri activate, după aplicarea modelului specific de difuzie.

Definiție 3.1.1 (Funcția de Influență). Dat fiind un graf $G = (V, E)$,funcția de influență $f_D(S) : 2^V \rightarrow \mathbb{R}^+$ poate fi definită ca fiind numărul mediu de noduri activate de D folosind nodurile din mulțimea S a noduri semință, unde $S \subseteq V$ este o mulțime de noduri, iar D este un model de propagare (difuzie).

Definiție 3.1.2 (IMP). Dat fiind un graf $G = (V, E)$, folosind D ca model de propagare, MP poate fi definită ca problema de a găsi o mulțime S de noduri semință care va maximiza funcția de influență:

$$\max_{S \subseteq V, |S| \leq k} f_D(S),$$

unde $|.|$ reprezintă cardinalitatea unei mulțimi, iar $k \in \mathbb{N}^*$ este un parametru care stabilește dimensiunea mulțimii de noduri semință la o valoare fixă.

IMP este o problemă de optimizare complexă și bine studiată, NP-hard, cu mai multe algoritmi propuși în literatură, precum în [27], [22], [11], [19], etc.

3.2 Modele de Propagare

Problema de maximizare a influenței folosește modele de propagare pentru a simula răspândirea informațiilor într-un mediu [17]. Există două clase principale de algoritmi de propagare: deterministe și probabilistice.

În modelele deterministe, putem calcula mulțimea de noduri care au fost activate de procesul de propagare. Rezultatele acestor calcule au fost folosite în diverse aplicații în literatură, o utilizare principală pentru aceste tipuri de modele fiind rețelele sociale și analiza efectuată asupra acestora, cum ar fi în [18].

Modelele de propagare probabilistice își propun să estimeze răspândirea informațiilor și în general oferă rezultate mai realiste. Ele realizează acest lucru folosind o variabilă probabilistică sau alte metode stochastice pentru a include variabilitatea în rezultate.

Unele dintre modelele de propagare mai populare investigate în prezent în literatură sunt modelele cascade, cum ar fi Modelul de Cascadă Independentă (ICM) [22], Modelul de Cascadă Ponderată (WCM) [22] sau modelul prag liniar [31]. Pentru cea mai mare parte a cercetării noastre legate de influență, variantele algoritmilor de cascadă au fost utilizate ca principală formă de simulare a propagării.

3.2.1 Cascada

Algoritmul de cascadă este un model de propagare folosit în simularea difuziei informațiilor într-o rețea socială. Este larg utilizat în literatură pentru multe aplicații diferite [22]. Algoritmul principal este extrem de similar cu o căutare pe lățime într-o rețea, cu trei factori principali de diferențiere: creșterea numărului de noduri de pornire și creșterea numărului de noduri investigate într-o iterație dată, împreună cu un proces de traversare diferit, probabilistic. Acest proces de traversare diferit permite algoritmului de traversare, care ar putea fi considerat mult prea simplistic, să simuleze difuzia informațiilor. Pasul de activare este prezent în procesul de selecție a vecinilor, ideea fiind că fiecare vecin al nodurilor investigate în mod curent ar trebui să fie luat în considerare, dar nu fiecare vecin ar trebui să fie vizitat. Selecția vecinului care urmează să fie vizitat (sau activat) se face folosind o variabilă probabilistică p ceea ce înseamnă că pentru fiecare vecin al setului de noduri investigate în mod curent, există o probabilitate de p ca vecinul investigat să fie activat.

Conectând funcționalitățile de mai sus, algoritmul de cascadă, indiferent de varianta aleasă, va oferi întotdeauna un număr, care este cardinalitatea mulțimii de noduri activate cu succes. O

descriere matematică mai detaliată ar fi:

$$\sigma(A_0) = |A|$$

unde A_0 reprezintă mulțimea inițială de noduri, în timp ce A reprezintă mulțimea de noduri activate.

Trebuie făcută o observație importantă: mulțimea de noduri de pornire este întotdeauna garantată să fie activată, ceea ce înseamnă că cardinalul mulțimii de noduri activate va fi întotdeauna cel puțin egal cu cardinalul mulțimii de noduri de pornire.

$$\sigma(A_0) \geq |A_0|$$

O altă observație importantă este că, având în vedere natura probabilistică a algoritmului de cascadă, o singură rulare nu furnizează rezultate statistic consistente; este necesară media a mai multor rulări ale algoritmului de cascadă pentru a echilibra orice anomalii și a oferi rezultate concluzive. Acest lucru înseamnă că trebuie dat un rezultat mediu al mai multor rulări al algoritmului de cascadă, rezultatul fiind astfel un număr real, în ciuda faptului că reprezintă o cardinalitate.

Cascada Independentă

Modelul de Cascadă Independentă (ICM) [22] este o variantă larg utilizată a algoritmului de cascadă, și este modelul de propagare ales pe parcursul celei mai mari părți a procesului de cercetare.

Cascada Independentă folosește o probabilitate globală, statică de p (valori cel mai des întâlnite variază de la 1% la 5%).

Cascada Ponderată

Modelul de Cascadă Ponderată (WCM), așa cum este propus în [22], este un model de difuzie conținut în familia mai largă a algoritmului de cascadă. Funcționalitatea sa este similară cu cea a Modelului de Cascadă Independentă, cu principala diferență observată în probabilitatea de propagare $p$folosită în algoritmul de cascadă. Cascada Ponderată folosește o probabilitate per nod $p_{w'}$ care dă probabilitatea de activare a vecinului w'. Această probabilitate este calculată ca:

$$p_{w'} = \frac{1}{in\_degree(w')}$$

3.3 Optimizarea Extremală

Natura și multe fenomene fizice prezintă caracteristici puternice de auto-optimizare [5], iar multe medii naturale co-dependente sunt optimizate prin selecția indivizilor nedoriți sau "răi" și înlocuirea lor în sistem.

Cercetarea principală efectuată pentru această teză s-a concentrat pe interpretarea IMP ca un mediu co-dependent, iar selecția indivizilor menționată a condus la utilizarea algoritmilor din familia de variații ale algoritmului de Optimizare Extremală (EO) [9] [8]. EO este definit în Definiția 3.3.1.

Definiție 3.3.1 (Optimizarea Extremală). Optimizarea Extremală este un algoritm de optimizare, cu idea de a împărți soluția unei probleme propuse în mai multe componente mici. Fiecare componentă trebuie să contribuie calculabil la calitatea oricărei soluții propuse. Mai mult, în orice instanță dată a algoritmului EO, există mai mulți candidați de soluții care sunt analizați în același timp.

Principala provocare a conversiei de la probleme bazate pe grafuri, cum ar fi problema de maximizare a influenței, la lumea și terminologia algoritmilor de optimizare, mai precis EO, a fost interpretarea datelor și traducerea terminologiei între cele două lumi. O altă problemă a fost legarea terminologiei modelului de propagare cu terminologia atât a grafurilor, cât și a EO.

O soluție s și cea mai bună soluție s_{best} sunt ambele mulțimi de noduri, care sunt submulțimi ale lui V. Aceste seturi sunt indivizii din terminologia EO și au o mărime de k, care este numărul care limitează mărimea lui s din terminologia IMP. Fiecare individ s este compus din noduri, care sunt componentele din terminologia EO.

Atât indivizii, cât și componentele unui individ, necesită o funcție de fitness. Aceste funcții de fitness au evoluat de-a lungul drumului nostru de cercetare. Funcția de fitness a fiecărui individ ar trebui să fie legată de influența deținută de mulțimea de noduri din care este compus fiecare individ. Această influență este calculată folosind modelul de propagare ales. Dacă luăm în considerare metoda noastră principală de propagare, Metoda de Cascadă Independentă, funcția noastră de influență și, prin urmare, funcția de fitness a fiecărui individ poate fi reprezentată ca

$$f_{ICM}(s) = \bar{\sigma}(s),$$

unde $\bar{\sigma}(s)$ este media a mai multor rulări ale modelului de propagare.

Pentru fitnessul fiecărei componente, trebuie să calculăm contribuția fiecărui nod la fitnessul individului complet. Principala idee a cercetării noastre a fost să considerăm problema influenței ca un joc cooperativ coalizat din teoria jocurilor, și apoi să folosim această înțelegere a teoriei

jocurilor pentru a calcula contribuțiile fiecărui nod ca jucători ai jocului. Primul nostru algoritm folosește o funcție mai simplistă. Valoarea de fitness a unui nod activ i din s este calculată ca fiind contribuția sa marginală la $\bar{\sigma}$:

$$f_i(s) = \bar{\sigma}(s) - \bar{\sigma}(s \setminus i) - 1,$$

unde $s \setminus i$ denotă mulțimea nodurilor active din s fără nodul i.

Optimizarea Extremală cu Variație Temporală

Principalul dezavantaj al algoritmului de Optimizare Extremală este dificultatea sa în a evita optimii locali. Una dintre soluțiile propuse ca și remediere la această problemă a fost utilizarea unei versiuni cu Variație Temporală a EO, care își propune să schimbe un set de componente minimale din orice individ în loc de un singur component la un moment dat. Numărul de componente schimbate într-o iterație dată este denumit q. Au fost luate în considerare trei versiuni ale procesului de modificare a parametrilor:

Versiunea liniară de bază:

$$q = max(1, \lfloor \frac{k * (T_{Max} - T)}{T_{Max}} \rfloor),$$

unde q este numărul de noduri proiectate să fie înlocuite în generația curentă, k este mărimea setului s, numărul de noduri din orice individ dat din algoritmul EO, T este numărul curent de generații, T_{Max} este numărul total de generații, iar $\lfloor \cdot \rfloor$ reprezintă partea întreagă a unui număr real.

Versiunea liniară alternativă:

$$q = max(1, \lfloor \frac{k * (T_{Max} - 2 * T)}{T_{Max}} \rfloor),$$

Versiunea exponențială:

$$q = \lfloor max(1, \frac{1}{2} * k * (k-1)^{\frac{-T}{T_{Max}}} \rfloor$$

Versiunea liniară alternativă a fost aleasă deoarece a oferit cele mai bune rezultate pentru scopurile noastre, cu o mică modificare, pentru a ne asigura că numărul de noduri înlocuite nu depășea jumătate din totalul nodurilor.

3.4 Valoarea Shapley

Valoarea Shapley a fost propusă în [32], și este utilizată pentru a calcula contribuția fiecărui jucător la un joc cooperativ în teoria jocurilor, făcând-o un concept de soluție popular pentru un joc. Are multe descrieri, dar pentru scopurile noastre, valoarea Shapley poate fi descrisă ca metoda de diviziune echitabilă a câștigurilor unui joc cooperativ, unde câștigul poate fi împărțit. Această diviziune echitabilă ar trebui să se relaționeze la contribuția individuală a fiecărui jucător.

Definiție 3.4.1 (Joc cooperativ coalitional). Un joc cooperativ coalitional $\Delta = (N, v)$ conține două elemente:

- Jucătorii jocului, conținuți în setul N;
- O funcție caracteristică $v : 2^N \to \mathbb{R}$ care atribuie valori reale submulțimilor de jucători.

Definiție 3.4.2 (Valoarea Shapley). Valoarea Shapley ϕ_i poate fi definită în contextul unui joc cooperativ coalitional, unde măsoară contribuția medie a jucătorului i în toate coalițiile jocului și este calculată astfel:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(v(S \cup \{i\}) - v(S))$$

unde $|\cdot|$ reprezintă cardinalitatea unui set.

Deoarece valoarea Shapley poate fi utilizată pentru a calcula măsura contribuției fiecărui jucător la valoarea finală a jocului, conceptul ar trebui să fie convertabil în limbajul optimizării extremale și ar trebui să putem calcula contribuția fiecărei componente, adică fitness-ul lor, la valoarea finală a fitness-ului individului.

3.4.1 Aproximarea Valorii Shapley

În timp ce valoarea Shapley oferă rezultate excelente, în principal pentru că aproximează contribuția fiecărui nod la rezultatul final într-un mod realistic, făcând astfel trivială alegerea nodurilor cu valoare minimă, există o problemă majoră cu calculul valorii Shapley. În fiecare instanță a calculului valorii Shapley, fiecare coaliție de noduri este luată în considerare. Acest lucru face ca calculul să fie extrem de costisitor în orice aplicație practică.

Pentru a depăși această problemă, am folosit o metodă de aproximare, care s-a dovedit a apropia suficient de bine rezultatele unui calcul complet al valorii Shapley. Această aproximare folosește un număr de ordonări distincte de dimensiune k, conținând aceleași noduri ca și s, dar

în permutări diferite. Apoi, calculăm contribuția nodurilor bazându-ne doar pe aceste ordonări diferite și pe coalițiile din aceste ordonări. Deoarece calculul nostru depinde de ordinea nodurilor în coaliție, asigurarea unui număr suficient de ordonări luate în considerare a fost una dintre principalele noastre provocări.

3.5 Abordarea Monte Carlo

Abordarea obișnuită a algoritmului folosește întreaga rețea originală dată ca parametru, pentru a calcula rezultatele rulării ICM, ceea ce necesită efectuarea a mai multor rulări ICM. Deoarece modelul de propagare are o natură probabilistică, dacă dorim să oferim un nivel de certitudine rezultatelor, ar trebui să facem media a mai multor rulări de propagare. Acest lucru este atât lent, cât și imprevizibil, timpul de rulare crescând odată cu creșterea nivelului de certitudine dorit. Noi am propus o îmbunătățire, folosind o abordare Monte Carlo pentru a introduce o fază inițială de generare a rețelei.

Schema funcționează prin crearea unui număr n_G de rețele noi pornind de la rețeaua originală, acest n_G fiind similar ca mărime cu numărul de rulări Cascade care ar fi fost necesare pentru rezultate precise în scenariul original. Fiecare rețea nou creată $G_i = (V, E_i)$ conține fiecare nod al rețelei originale V și un procent mic de muchii din rețeaua originală $E_i \subset E$. Muchiile păstrate pentru o rețea nouă sunt selectate aleator din E, iar probabilitatea de a selecta o muchie pentru a fi păstrată este p, aceeași probabilitate folosită în cazul cascadelor independente originale.

Utilizând aceste rețele nou generate, complexitatea și timpul de rulare scad semnificativ, deoarece am modificat spațiul de căutare. Algoritmul de cascadă folosit pentru aceste rețele noi este redus la o versiune non-probabilistică, cu natura probabilistică a căutării fiind incorporată în generarea rețelelor. Această schimbare de perspectivă ne permite să pregătim un număr mare de rețele generate în faza de configurare a algoritmului, iar apoi să folosim ulterior media rulărilor pe fiecare rețea pentru a obține un rezultat final pentru un anumit set de semințe. Putem să folosim totuși același set de semințe deoarece nodurile din rețelele generate nu s-au schimbat.

Aproximarea valorii Shapley este, de asemenea, afectată de modificarea algoritmului de cascadă. Revizuirea introducerii noului Monte Carlo Cascade în algoritmul de aproximare a valorii Shapley nu a modificat validitatea sau eficacitatea aproximării.

3.6 Algoritmul SIM-EO

Pe parcursul cercetării noastre asupra problemei de maximizare a influenței, intențiile au fost clare în a rezuma toate posibilele îmbunătățiri propuse în timpul pașilor iterați făcuți pe algoritmi, și un algoritm final a fost dezvoltat folosind aceste îmbunătățiri. Algoritmul propus a fost numit Maximizarea Influentei Shapley - Optimizarea Extremală (SIM-EO), recunoscând atât tipul de algoritm, care este o variantă de Optimizare Extremală, cât și folosirea specială a fitness-ului, aceea de a folosi valoarea Shapley ca fitness al componentelor.

Descrierea algoritmului SIM-EO Algoritmul se inspiră din toate iterațiile anterioare. Inițial, se bazează pe versiunea simplificată a primului nostru algoritm bazat pe EO, renunțând la selecția mai complexă a lui s_{best} prezentă acolo, dar păstrând un cadru similar al algoritmului.

SIM-EO începe cu algoritmul de generare a rețelelor Monte Carlo, utilizând îmbunătățirile propuse Monte Carlo, creând mai multe rețele folosind probabilitatea Cascade ca bază pentru probabilitatea de generare a rețelei și un parametru n_G pentru a da numărul necesar de rețele.

Fitnessul fiecărui individ este calculat folosind o variantă a algoritmului ICM, care este modificat prin înlocuirea naturii probabilistice din ICM, și necesitatea mai multor rulări pe aceeași rețea, cu abordarea folosirii mai multor rețele, fiecare rețea fiind diferită, fiecare fiind generată folosind metoda Monte Carlo.

Algoritmul folosește apoi algoritmul Shapley Value, mai precis algoritmul de aproximare a valorii Shapley folosind rețelele Monte Carlo, ca valoare de fitness pentru fiecare componentă.

Numărul de componente schimbate este reprezentat de numărul m și este calculat folosind ecuația îmbunătățită din optimizarea extremală cu variație temporală, care se bazează pe versiunea liniară alternativă din ecuația 3.3.

În ceea ce privește parametrii algoritmului, facem distincție între parametrii, care se referă la problema maximizării influenței, și parametrii care sunt specifici algoritmului în discuție. Parametrii includ n_G, $MaxIterations$, care reprezintă numărul maxim de generații pentru care algoritmul va rula, și m împreuna cu rețeaua inițială G, și valorile k și p.

3.7 Experimentele numerice și o aplicație propusă

Deși experimentele numerice sunt prea ample pentru a se potrivi acestui rezumat cu scop mai mic, se poate oferi o scurtă descriere a rezultatelor. Algoritmul SIM-EO și variantele sale anterioare au oferit rezultate remarcabile, comparabile cu, sau chiar mai bune decât, algoritmii

de ultimă generație investigați. Această îmbunătățire a fost în termeni de rezultate, în ceea ce privește complexitatea temporală, algoritmii noștri au fost toți mai lenți și mult mai intensivi din punct de vedere computațional. Acest lucru oferă un compromis între abordările euristice actuale cele mai bune, care oferă rezultate rapide, dar inexacte sau variantele noastre de EO, care oferă rezultate mai bune, deși mai lente. Au fost efectuate teste atât pe rețele inspirate din lumea reală, cât și pe rețele sintetice, iar aproape fiecare tip de rețea a arătat aceleași relații între algoritmi, care au fost descrise anterior.

Aplicație: Analiza rețelelor de jurnale cu citare ridicată

Un singur caz de testare bazat pe lumea reală a fost propus pentru varianta InfEO a algoritmului, deoarece restul cercetărilor bazate pe influență au fost în mare parte teoretice. Pentru testarea în lumea reală, a fost construită o rețea folosind date din baza de date a articolelor Web of Science (WoS) [1], mai exact, a fost creată o rețea de citare folosind articole din domeniul informaticii. Articolele au fost selectate folosind o interogare specifică de căutare, care include categoria, anul și eticheta *highly cited*.[2].

În rețeaua rezultată din aceste date, nodurile erau reviste, în timp ce legăturile erau citări între articole din diferite reviste; o legătură direcționată exista între două noduri dacă nodul de pornire avea un articol care referenția nodul de destinație. Astfel, rețeaua a fost construită din 606 articole, rezultând un număr de 7482 de noduri, din care 131 aveau un grad de ieșire pozitiv. Rețeaua mai conținea și 14479 de legături. Cele 131 de reviste care aveau un grad de ieșire pozitiv erau candidații principali pentru a fi identificați ca noduri influente de către algoritmul nostru.

Analiza efectuată asupra rețelei de reviste foarte citate a oferit o perspectivă interesantă asupra unei posibile aplicații a problemei de maximizare a influenței și a arătat utilitatea algoritmului InfEO într-un scenariu pseudo-real.

1. (https://apps.webofknowledge.com/), ultima accesare 29.06.2023
2. https://images.webofknowledge.com/images/help/WOS/hs_citation_applications.html, ultima accesare 29.06.2023

Capitolul 4.

Componente critice în rețele

4.1 Problema Criticității și Detectarea Nodurilor Critice

Unul dintre principalele domenii de cercetare care preocupă analiza rețelelor este domeniul care investighează problema criticității în rețele. Problema Detectării Nodurilor Critice (CNDP) este descrisă în [26], și în sondajul efectuat de [24], și este cea mai comună problemă de criticitate investigată în literatură. CNDP poate fi simplu descrisă ca o problemă de identificare. Trebuie să identificăm nodurile importante într-o rețea. Importanța este definită în funcție de o metrică specifică. Este crucial să distanțăm această definiție de definiția influenței, unde nodurile trebuie să aibă capacitatea maximă de difuzare a informațiilor. Aici, nodurile trebuie să fie importante în sensul integrității rețelei.

Descrierea de mai sus este vagă în mod intenționat, deoarece aproape fiecare aspect al problemei de detectare a nodurilor critice poate fi modificat. Putem căuta unul sau mai multe noduri sau chiar alte componente ale rețelei, cum ar fi în sondajul [37], unde autorii descriu alte tipuri de elemente critice și printre acestea, nodurile critice. Putem defini un nod important conform mai multor metrici, putem chiar utiliza diferite tipuri de rețele.

Această variabilitate este una dintre cauzele pentru care CNDP este folosit într-o mare varietate de studii în literatură. A fost folosit pentru analiza rețelelor sociale în [10], [15], studii privind vulnerabilitatea rețelelor în [14] și managementul riscului rețelelor în [3].

Unul dintre principalele componente ale CNDP este măsura folosită pentru detectarea nodurilor critice. În literatură au fost propuse mai multe măsuri, iar mai multe sunt posibile, principala întrebare adresată de cercetători fiind: De ce este un nod critic? Ce face un nod critic?

În [2] sunt discutate trei versiuni ca răspuns la întrebările noastre, acestea fiind cele mai populare variante ale CNDP în literatură și sunt următoarele: Problema *kMaxComp*, conectivitatea în perechi (pairwise connectivity) și problema *MinMaxC*. Acestea vor fi descrise în detaliu în

curând.

În ceea ce privește complexitatea, s-au analizat mai multe măsuri de conectivitate, iar CNDP a fost dovedit a fi NP-hard pentru toate măsurile investigate în [33]. Plecând de la această dilemă, au fost propuse diferite metode de rezolvare, însă problema *kMaxComp* nu a fost investigată în detaliu.

Definiție 4.1.1 (Problema Detectării Nodurilor Critice (CNDP)). Problema detectării nodurilor critice poate fi definită ca fiind problema găsirii unui set de noduri, setul având o dimensiune fixă de k, în orice graf dat, astfel încât, după eliminarea setului selectat, graful să se deterioreze maxim, conform unei măsuri arbitrare σ.

Împreună cu problema *kMaxComp*, au fost studiate trei forme distincte ale lui σ pentru rețelele tradiționale pe parcursul procesului de cercetare prezentat în această teză. Aceste probleme definite de aceste măsuri σ sunt următoarele:

Definiție 4.1.2 (Problema kMaxComp). Problema *kMaxComp* constă în eliminarea unui set de noduri, care ar duce la numărul maximal de componente conexe rămase într-un graf deteriorat. Formal, dacă S denotă setul de noduri șterse, având o dimensiune de k, și $\mathcal{H}(G[V \setminus S])$ denotă setul de componente conexe ale grafului G după eliminarea setului selectat de noduri, în esență graful deteriorat, atunci problema *kMaxComp* poate fi descrisă folosind următoarea ecuație:

$$\max_{S \subset V} |\mathcal{H}(G[V \setminus S])|,$$
$$\text{astfel ca } |S| \leq k,$$

unde $|\cdot|$ reprezintă cardinalitatea mulțimii.

Problema kMaxComp a fost forma principală a posibilelor variante ale CNDP pe care le-am folosit în cercetările de bază privind detectarea criticității, cum ar fi algoritmul CN-EO, algoritmul MAXC-GA, oferind totodată baza pentru algoritmul nostru Hyp-GA axat pe hipergrafuri. Este probabil cea mai utilizată metrică din cele trei metrici principale găsite în literatură și poate fi folosită pentru orice tip de detectare a criticității, indiferent de rețeaua în discuție.

Definiție 4.1.3 (CNP - Pairwise Connectivity (Măsurarea conectivității în perechi)). În acest caz, trebuie să minimizăm următoarea funcție obiectiv:

$$f(A) = \sum_{C_i \in G[V \setminus A]} \frac{|C_i|(|C_i| - 1)}{2},$$

, unde C_i este mulțimea de noduri care se află într-un component conectat concret după deteriorarea grafului, în timp ce $|\cdot|$ indică cardinalul unei mulțimi, adică dimensiunea componentei.

Pentru această ecuație, luăm în considerare componentele conectate după degradarea grafului, prin eliminarea nodurilor selectate de problema CNDP.

Măsurarea conectivității în perechi a fost folosită ca metrică în interpretarea noastră a problemei combinate de detectare a nodurilor și a muchiilor critice, deoarece poate fi utilizată în mod semnificativ pentru detectarea deteriorării rețelei. Focalizarea noastră pentru acea cercetare a fost utilizarea problemei combinate ca mijloc de detectare a punctelor critice de deteriorare a rețelei, oferind o posibilă metrică de securitate a rețelei. Utilizarea conectivității în perechi în ceea ce privește criticitatea este explicată în detaliu în paragraful 4.2.1, unde este prezentat un exemplu pentru calculele necesare în acest caz.

Definiție 4.1.4 (MinMaxC). Această problemă constă în minimizarea dimensiunii celei mai mari componente după eliminarea nodurilor selectate de algoritm ca fiind potențial critice. Formal:

$$\min |(\max_{C \in \mathcal{H}} C)|,$$

unde $\mathcal{H}$ este mulțimea care conține componentele conectate ale grafului, în timp ce $|\cdot|$ indică cardinalul unei mulțimi.

Deși s-au efectuat teste folosind problema MinMaxC, nici o cercetare concretă nu a fost complet bazată pe această variantă a CNDP, deoarece am constatat că, pentru scopurile noastre, MinMaxC nu a oferit rezultatele dorite, iar combinat cu faptul că această problemă era intens computațională, cel puțin implementarea inițială a acesteia, s-a luat decizia de a nu se explora în profunzime această variantă. Cu toate acestea, este încă una dintre abordările mai utilizate pentru problema CNDP, ceea ce înseamnă că a fost necesară o definiție.

În afară de aceste trei cazuri principale de utilizare, am propus și o a patra abordare sub formă de utilizare a metricii Centralității de centralitatea ponderată a gradului nodului (Weighted Node Degree Centrality (WNDC)) specific hipergrafului, în cercetarea care a dus la problema WNDC CNDP, prezentată în secțiunea 4.2.2. Acesta a fost o abordare nouă, și ca atare, o explicație detaliată a fost oferită în secțiunea corespunzătoare.

Algoritmul CN-EO

După cum sugerează și numele, CN-EO este un algoritm bazat în principal pe algoritmul de Optimizare Extremală, care a constituit baza cercetării noastre în domeniul maximizării influenței. Cu toate acestea, am încercat să inovăm în abordările noastre anterioare prin introducerea unei variante noi de Optimizare Extremală ca bază pentru CN-EO, și anume algoritmul NoisyEO din [28]. În acea lucrare, NoisyEO a fost utilizat cu succes pentru problema detectării comunităților și s-a dovedit a fi adaptabil la problema criticității, cu câteva modificări notabile. NoisyEO

funcționează prin introducerea unei proceduri de schimbare în algoritm. Odată ce rezultatele par să stagneze pentru un număr extins de iterații, procedura de schimbare este inițiată.

Procedura de schimbare funcționează prin modificarea aleatorie a rețelei $G = (V, E)$ creând $G' = (V', E')$, unde $V' \subseteq V$ și $E' \subseteq E$ sunt setul de noi noduri și muchii, respectiv. Aceste seturi noi sunt obținute prin eliminarea aleatorie a nodurilor și muchiilor din V și E originale, eliminarea fiind realizată cu o probabilitate de p_{shift} care este un parametru al algoritmului nostru. Noua rețea G' este acum folosită ca rețea în algoritmul EO pentru un număr fix de iterații nrG. În timpul acestor iterații, cea mai bună soluție de până acum poate fi modificată, important este că nu reținem rezultatul, ci doar setul de noduri potențial critice. După aceste iterații, rețeaua originală este readusă, dar soluția propusă rămâne neschimbată față de soluția obținută în procedura de schimbare, iar în acest fel căutarea poate scăpa dintr-un optim local.

Algoritmul a fost testat pe unele dintre rețelele de referință introduse anterior și a furnizat rezultate bune comparativ cu rezultatele din literatură.

Algoritmul MAXC-GA

O altă abordare inițială pentru rezolvarea CNDP a fost algoritmul MAXC-GA, care poate fi descris ca un algoritm genetic simplu. Scopul specific al acestei cercetări a fost de a crea o soluție pentru CNDP utilizând cea mai mică cantitate de informații specifice problemei în timpul fazei de căutare. Deoarece algoritmul propus este un algoritm genetic, există câteva informații generale despre structura și utilizarea algoritmului care trebuie elaborate. Am folosit o reprezentare binară, schemă de încrucișare cu două puncte, o schemă de mutație uniformă și o selecție bazată pe turneu. Valoarea de fitness a unui individ a fost calculată folosind numărul de componente rămase după deteriorarea grafului cauzată de ștergerea nodurilor marcate ale individului.

Deși testarea pe benchmark-uri a fost realizată cu succes pe MAXC-GA, demonstrând că este un bun concurent pentru algoritmul CNDP, pentru lucrarea care a creat algoritmul MAXC-GA, nu a fost propus un caz de utilizare real similar cu rezultatele propuse pentru abordarea EO. În schimb, lucrarea a fost propusă ca o lucrare de tip proof-of-concept.

4.2 Generalizări ale Problemei Detectării Nodurilor Critice

4.2.1 Problema Combinată de Detectare a Nodurilor și Muchiilor Critice

Criticitatea într-o rețea poate fi legată de orice formă a componentelor rețelei, nu doar de noduri. Una dintre componentele care sunt de asemenea investigate în literatură și ar trebui să ofere

rezultate interesante sunt muchiile unui graf. În afară de problema detectării nodurilor critice, putem defini problema detectării muchiilor critice.

Definiție 4.2.1 (Problema Detectării Muchiilor Critice (CEDP)). Într-un graf $G = (V, E)$, obiectivul problemei este de a găsi un set de l muchii, care pot fi considerate critice conform unei metrice date, care măsoară degradarea grafurilor după ce muchiile din l sunt eliminate.

Problema de detectare a nodurilor critice (CNDP) și problema de detectare a muchiilor critice (CEDP) coexistă în literatură. Am propus o abordare inovatoare pentru problema de criticitate, combinând cele două probleme într-una singură, obținând problema combinată de detectare a nodurilor și a muchiilor critice (CNEDP), care este o abordare mult mai puțin studiată a familiei de probleme legate de criticitate. CNEDP poate fi utilizat pentru a simula scenarii din lumea reală, ceea ce înseamnă că poate fi util în diverse aplicații (de exemplu, rețele de drumuri, rețele de calculatoare etc.). Poate funcționa prin ștergerea nu numai a nodurilor sau a muchiilor, ci a unei combinații între cele două.

Problema (k, l)-CNEDP

Definiție 4.2.2 (Problema de Detectare a Nodurilor și a Muchiilor Critice (CNEDP)). Dat fiind graful $G = (V, E)$, CNEDP constă în găsirea unei colecții W de mărime k, care conține noduri din rețeaua originală, și o colecție F e mărime l, care conține muchii din rețeaua originală, astfel încât, atunci când sunt șterse simultan, să degradeze graful într-un mod maximal, conform unei măsuri date σ. Problema introdusă este denumită (k, l)-CNEDP.

Putem identifica o problemă interesantă care apare atunci când ștergem două componente separate ale grafului: nodurile și muchiile. Este evident că eliminarea unui nod va elimina toate muchiile conectate la el, deoarece nu poate exista o muchie între noduri inexistente, a trebuit să luăm în considerare dacă eliminarea unei muchii ar elimina și nodurile conectate la ea. Această posibilitate a fost rapid exclusă, deoarece eliminarea nodurilor ar elimina și muchiile, și am ajunge rapid la un graf gol.

Măsura de conectivitate a rețelei folosită pentru această cercetare a fost pairwise connectivity. În cazul nostru, prin urmare, am trebuit să minimizăm următoarea funcție obiectiv:

$$f(A) = \sum_{C_i \in G[V \setminus A]} \frac{\delta_i(\delta_i - 1)}{2},$$

unde $A \subseteq V$, $C_i$este o colecție care conține toate componentele conectate în graficul deteriorat, cu δ_i reprezentând dimensiunea componentei conectate C_i.

Putem observa că găsirea unui set ideal pentru componente critice ar duce la dezintegrarea completă a grafului, cu dimensiunile componentelor fiind reduse toate la 1, ceea ce ar rezulta

într-o valoare a funcției obiectiv de 0. De asemenea, putem identifica faptul că stabilirea valorii lui k respectiv l la 0 ar reduce problema noastră la CNDP sau CEDP, respectiv. În cele din urmă, putem observa că CNEDP ar trebui să fie NP-completă, deoarece în [4], o variantă a CNP a fost dovedită a fi NP-completă, cu CNP fiind o sub-problemă a problemei noastre curente.

Abordarea greedy pentru (k, l)-CNEDP

În cercetarea propusă în [2] au fost prezentate trei soluții non-evolutive pentru CNDP. O adaptare a celui de-al doilea abordaj este folosită de noi în timpul acestei cercetări, adaptată la (k, l)-CNEDP. Am folosit conectivitatea pairwise în procesul greedy, proces care se bazează pe următoarea funcție:

$$GR2(S_X) = argmax\{(f(S_X) - f(S_X \cup \{t\}) : t \in X \setminus S_X\}$$

unde S_X poate fi unul dintre seturile de soluții propuse pentru noduri sau muchii, ceea ce înseamnă că X reprezintă fie setul original de noduri, fie setul original de muchii ale rețelei. Această funcție este descrisă în [2] și a fost adoptată de noi.

Algoritm genetic pentru (k, l)-CNEDP

Ca o abordare suplimentară, am creat un algoritm genetic simplu pentru a aborda problema CNEDP. Operatorii principali sunt următorii: codificare bazată pe listă, fitness-ul este calculat ca conectivitatea pairwise după deteriorare, selecția pe baza unui turneu, două variante de mutație, ambele bazate pe înlocuire, fără nevoia unui operator de reparare. De asemenea, s-a folosit un scheme de selecție $(\mu + \lambda)$.

Compararea între cele două metode Comparările inițiale s-au făcut folosind un număr fix de calculări de fitness ca limită, dar pentru a asigura comparații corecte, trebuia să permitem ambelor algoritmi să ruleze pe deplin, totuși, algoritmul greedy nu a putut procesa în mod realist rețele mai mari datorită scalării algoritmului fiind exponențială. Au urmat rezultate de comparații mai concrete, unde calculările greedy nu au fost limitate și rezultatele au fost derivate din rulări complete în cazurile în care algoritmul greedy a putut termina. Rezultatele au arătat că în majoritatea cazurilor GA era mai potrivit, dar pentru unele cazuri specifice abordarea greedy a fost mai bună.

4.2.2 Detectarea Nodurilor Critice în Hipergrafuri

Au fost făcute încercări pentru a combina conceptul de hipergrafuri și CNDP. Două abordări principale au fost propuse pentru CNDP bazat pe hipergrafuri.

Definiție 4.2.3 (Detectarea Nodurilor Critice în Hipergrafuri). Dat fiind un hipergraf $\mathcal{H} = (X, \mathcal{D})$, CNDP pentru hipergrafuri constă în eliminarea slabă a unei colecții de k noduri pentru a maximiza numărul de componente conectate rămase în graful deteriorat.

Hyp-GA

Cercetările inițiale pe tema hipergrafurilor în nodurile critice au dus la algoritmul Hyp-GA. Așa cum sugerează și numele, este încă o implementare a algoritmului genetic, de data aceasta având accentul pe hipergrafuri în loc de noi tipuri de algoritmi. Acest algoritm funcționează prin utilizarea unei reprezentări simplificate a unui hipergraf, ca un graf care conține sub-grafuri complete, sau formal o *reprezentare a clicilor*, și apoi lucrând pe această reprezentare ca un simplu graf. Algoritmul folosește aceleași componente și parametri ca și abordarea noastră anterioară bazată pe GA. Cu toate că oferă rezultate bune, reprezentarea ineficientă împreună cu algoritmul simplu nu au oferit rezultate remarcabile.

Nodurile critice în hipergrafuri cu centralitatea ponderată a gradului nodului

Încă o abordare GA axată pe hipergrafuri și criticitate, cu rezultate mult mai interesante. Au fost introduse câteva tactici evolutive principale care lipseau în abordarea anterioară: o reprezentare mai bună folosind biblioteci de hipergrafuri dedicate; o schemă algoritmică mai bună în general, cu un GA îmbunătățit; și introducerea centralității ponderate a gradului nodului ca metrică pe baza căreia am putut calcula criticitatea.

Centralitatea Ponderată a Gradului Nodului (WNDC) Principalul scop al acestei cercetări și principalul factor de diferențiere față de lucrările anterioare este introducerea unei măsuri de centralitate specifice hipergrafurilor ca măsură de criticitate pentru problema noastră. Această nouă metrică este numită Centralitatea Ponderată a Gradului Nodului (WNDC), care a fost prezentată în [21] ca o încercare de a extinde măsurile tradiționale de centralitate, obișnuite în cercetarea grafurilor, în domeniul hipergrafurilor.

În general, greutățile w ale unei hiper-muchii iau în considerare doi metrici care descriu muchia dată. Multiplicitatea (m_j), care descrie frecvența apariției unei hiper-muchii date în rețea, și cardinalitatea (c_j), care se referă la numărul de noduri care sunt incluse în acea hiper-muchie. În timp ce rezultatele testelor de referință au fost interesante, deoarece nu exista alte

cercetări cu care să le comparăm, principala atracție a fost comparația cu o abordare euristică propusă de noi și aplicația în lumea reală.

Comparații între euristică și GA Validarea rezultatelor obținute de GA ar trebui să fie un pas important în validarea utilității algoritmului. Am propus o euristică care, similar, a folosit WNDC, prin eliminarea nodurilor cu cel mai mare WNDC. Această rezultare, la rândul ei, a oferit o bază bună cu care am putut compara rezultatele algoritmului genetic, deoarece intuiția ar spune că, dacă eliminăm cele k noduri cu cel mai mare WNDC, atunci rezultatele ar fi mai bune decât orice altă combinație de k noduri eliminate. În realitate, algoritmul nostru GA a obținut combinații mai bune de noduri eliminate, arătând încă o dată că întreaga mulțime de noduri are un impact mai mare decât suma părților sale.

4.3 Aplicații propuse pentru variantele Problemei de Detectare a Nodurilor Critice

4.3.1 Utilizare practică pentru CNDP: analiza piețelor

Articolul care a prezentat CN-EO a fost una dintre cele mai interesante cercetări realizate în timpul acestei perioade doctorale, cu un rezultat practic care s-a dovedit a fi interesant. Utilizarea practică găsită pentru CNDP și algoritmul CN-EO a fost o analiză a piețelor. Pentru analiza rețelelor economice, există câteva exemple în literatură, rețelele bancare fiind unul dintre acestea, unde nodurile reprezintă totul de la bănci la persoanele influente din acele bănci și multe altele. Analiza pieței de valori este un aspect important al lucrărilor legate de rețele economice, unul dintre primele astfel de articole fiind [12], iar pentru noi [16] fiind de asemenea important, deoarece a analizat piața de valori din China dintr-o perspectivă de influență.

Am utilizat o versiune neponderată și simplificată a unei rețele de piețe de valori din [25], care a fost obținută din analiza corelațiilor dintre acțiunile de pe piața de valori din New York pe o perioadă de doi ani. Utilizând CN-EO, am calculat cele mai critice noduri. Dimensiunea setului de noduri critice a început de la 3 și a ajuns până la 8.

4.3.2 Aplicație pentru CNEDP: o nouă metrică de robustețe a rețelelor

A fost propusă o aplicație pentru CNEDP combinat, o nouă metrică care ar putea fi folosită pentru a măsura robustețea rețelei și care, dacă ar produce rezultate promițătoare, ar oferi o alternativă bună pentru măsurile deja prezente în literatură. În literatură, există mai multe măsuri de robustețe, care în mod obișnuit încearcă să calculeze robustețea prin examinarea diferitelor

seturi de proprietăți ale rețelei.

Am testat noua noastră metrică propusă folosind un set de rețele din lumea reală de dimensiuni medii și mari. Acestea includ rețele de infrastructură [30], [35], [23], rețele cerebrale [1], rețele de energie electrică [30], rețele de interacțiune [34] și o rețea informatică [23].

Noua măsură de robustețe a rețelei a fost denumită $NE_{k,l}$, care se baza pe algoritmul nostru (k,l)-CNEDP care folosea conectivitatea pairwise ca criteriu de criticitate. Noua metrică are forma următoare:

$$NE_{k,l} = \frac{2 \cdot (k,l)\text{-CNEDP}}{(n-k-1)(n-k-2)} \in [0,1]$$

.

O observație interesantă poate fi făcută în ceea ce privește termenii acestei ecuații. Ecuația conține rezultatele cel mai nefavorabile ale conectivității pairwise după eliminarea a k noduri, ceea ce înseamnă că valoarea va fi întotdeauna între 0 și 1, deoarece CNEDP poate numai să se apropie doar de rezultatul maxim.

$NE_{k,l}$ poate fi considerată o metrică bună de robustețe datorită rezultatelor, deoarece oferă rezultate independente față de alte metrici și funcționează și pe grafuri neconectate. De asemenea, poate fi configurată, datorită celor două parametri k și l. Aceste fapte au făcut ca această realizare să fie una dintre cele mai importante contribuții ale procesului nostru de cercetare de-a lungul anilor.

4.3.3 Aplicație pentru CNDP pe hipergrafuri: O analiză a inflației pe hipergrafuri

Propunem ca aplicație în lumea reală utilizarea algoritmului nostru pe un hipergraf construit din date reale privind inflația. Datele pentru aproximativ 123 de țări au fost disponibile public[1] și conțineau informații despre rata inflației din 1960 până în 2019. De atunci, site-ul a fost închis, dar rezultatele rămân valabile.

A fost realizată o analiză pe ultimii zece ani de date, din 2010 până în 2019. Țările au fost eliminate dacă aveau date incomplete, astfel rămânând cu 98 de țări. Un hipergraf este apoi construit din aceste date prin considerarea țărilor ca noduri, în timp ce hiper-muchii reprezintă diferite intervale de inflație într-un an dat, ceea ce înseamnă că există un hiper-muchie care conține țări care au avut, de exemplu, o rată negativă a inflației într-un an studiat, etc. De fapt, au fost obținute patru hipergrafuri pentru patru seturi de ani din intervalul marcat (2010-2012, 2013-2015, 2015-2017, 2017-2019). Această reprezentare poate fi utilă, deoarece reprezintă dinamica inflației pentru țări.

1. https://dice.ifo.de/en/node/358439, ultimul accesat la 20/09/2021

Este aleasă o valoare de 10 pentru k, ceea ce înseamnă că dorim să identificăm cele 10 noduri cele mai critice din rețea. Rezultatele pot oferi o privire asupra posibilelor regiuni instabile sau regiuni cu rate similare de schimbare de-a lungul anilor, deoarece un nod critic în această configurație ar însemna țări care au schimbat ratele de inflație de mai multe ori în perioada investigată.

4.3.4 Aplicație pentru CNDP pe hipergrafuri: Analiza datelor comitetelor congresionale și senatoriale din SUA

Ca o aplicație a CNDP pe hipergrafuri, am investigat două rețele din lumea reală, ambele derivate din datele congresionale din SUA, create de Charles Stewart și Jonathan Woon. Aceste rețele au fost folosite inițial în [13] și le-am implementat pentru aplicația noastră din lumea reală și pentru a dovedi corectitudinea algoritmului nostru. Aceste rețele agregă membrii în comitetele congresului SUA, fie în Camera Reprezentanților, fie în Senat.

Deoarece rezultatele prezentate aici sunt bazate pe date din lumea reală, o interpretare a rezultatelor ar putea fi utilă.

Un rezultat interesant al acestei cercetări arată cât de mult mai puțin rigidă este structura Camerei Reprezentanților comparativ cu structura Senatului. În rețeaua Camerei Reprezentanților, există un total de 1290 de noduri, din care 51% sunt prezente cel puțin o dată în lista critică. Dacă ne concentrăm în schimb pe nodurile care apar în cel puțin jumătate din totalul seturilor critice, acest număr scade la doar 4.5%, în timp ce dacă ne-am uita la 75% din totalul seturilor critice, am obține doar 2 noduri critice, ceea ce corespunde unui procent incredibil de mic de 0.015% din toate nodurile. Putem face calcule similare pentru rețeaua Senatului. Din cele 282 de noduri, doar 28.7% apar o dată, 9.9% apar în jumătate din listele critice, iar 3.2% apar în 75% din totalul seturilor critice.

Putem trage unele deducții generale despre funcționarea Congresului, cum ar fi faptul că membrii Senatului sunt mai critici și există mai mulți membri critici ai Senatului, comparativ cu Camera Reprezentanților. Aceste date nu sunt deloc definitive, dar ideea principală din spatele utilizării acestor date poate duce la identificarea unor figuri cheie printre elitele politice ale Americii.

Capitolul 5.

Concluzii și viitorul

Maximizarea Influenței Pentru Maximizarea Influenței, introducerea atâtor inovații, începând de la algoritmul EO, până la combinarea EO cu Cascada și Teoria Jocurilor, introducerea valorii Shapley și îmbunătățirile ulterioare ale procesului nostru, au condus la un algoritm final SIM-EO, care poate fi considerat complet. Încă mai este mult de lucrat, perspectivele viitoare potențiale includ analiza posibilei reintroduceri a stilului Inf-EO, îmbunătățirea EO, care a fost abandonată devreme în cercetarea noastră din cauza constrângerilor computaționale. O altă opțiune ar fi să investigăm diferite valori ale teoriei jocurilor, cum ar fi valoarea Banzhaf [6] sau diferite modele de propagare, cum ar fi modelul pragului liniar. Am putea, de asemenea, să investigăm diferite tipuri de algoritmi, cum ar fi abordarea GA pentru problema influenței. În cele din urmă, problema maximizării influenței online ar putea fi, de asemenea, investigată, practic, maximizarea influenței pe o rețea în evoluție, cu valori k non-stabile.

În concluzie, maximizarea influenței a fost un subiect care a dus la rezultate interesante și ar trebui să merite revizitarea în viitor.

Detectarea Nodurilor Critice În contrast cu problema de maximizare a influenței, problema detectării nodurilor critice nu a avut propus un algoritm final unificator. Ne-am concentrat pe multe variante ale problemei de criticitate, de la variante bine cunoscute precum CNDP până la abordări complet noi precum WNDC-CNDP pe hipergrafuri. Viitorul poate fi interesant pentru problema de criticitate. În primul rând, am putea să continuăm să creăm o colecție completă din îmbunătățirile propuse, similar cu partea de influență, de asemenea, am putea combina părțile deja existente ale cercetării noastre privind criticitatea, sau am putea investiga concepte noi, cum ar fi alte componente critice sau noi tipuri de algoritmi pentru problemele existente.

CNDP a fost și rămâne un subiect extrem de important de cercetat, datorită aspectului de securitate al acestei probleme. Rezultatele obținute până acum ne oferă suficientă motivație pentru a continua cercetarea în acest domeniu.

Bibliografie

[1] K. Amunts, C. Lepage, L. Borgeat, H. Mohlberg, T. Dickscheid, M.-É. Rousseau, S. Bludau, P.-L. Bazin, L. B. Lewis, A.-M. Oros-Peusquens, N. J. Shah, T. Lippert, K. Zilles, and A. C. Evans. Bigbrain: An ultrahigh-resolution 3d human brain model. *Science*, 340 (6139):1472–1475, 2013.

[2] R. Aringhieri, A. Grosso, P. Hosteins, and R. Scatamacchia. A general evolutionary framework for different classes of critical node problems. *Engineering Applications of Artificial Intelligence*, 55:128–145, 2016.

[3] A. Arulselvan, C. W. Commander, P. M. Pardalos, and O. Shylo. Managing network risk via critical node identification. *Risk management in telecommunication networks, Springer*, 2007.

[4] A. Arulselvan, C. W. Commander, L. Elefteriadou, and P. M. Pardalos. Detecting critical nodes in sparse graphs. *Computers & Operations Research*, 36(7):2193–2200, 2009.

[5] P. Bak. The discovery of self-organized criticality. In *How nature works: The science of self-organized criticality*, pages 33–48. Springer, 1996.

[6] J. F. Banzhaf. Weighted voting does not work: A mathematical analysis. In *Rutgers Law Review*, 1965.

[7] C. Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

[8] S. Boettcher and A. Percus. Nature's way of optimizing. *Artificial Intelligence*, 119(1-2): 275–286, 2000.

[9] S. Boettcher and A. G. Percus. Optimization with extremal dynamics. *Complexity*, 8(2): 57–62, 2002.

[10] S. P. Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1):21–34, 2006.

[11] D. Bucur and G. Iacca. Influence maximization in social networks with genetic algorithms. In *European conference on the applications of evolutionary computation*, pages 379–392. Springer, 2016.

[12] K. T. Chi, J. Liu, and F. C. Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 17(4):659–667, 2010.

[13] P. S. Chodrow, N. Veldt, and A. R. Benson. Hypergraph clustering: from blockmodels to modularity. *Science Advances*, 2021.

[14] T. N. Dinh and M. T. Thai. Precise structural vulnerability assessment via mathematical programming. In *2011-MILCOM 2011 Military Communications Conference*, pages 1351–1356. IEEE, 2011.

[15] N. Fan and P. M. Pardalos. Robust optimization of graph partitioning and critical node detection in analyzing networks. In *International Conference on Combinatorial Optimization and Applications*, pages 170–183. Springer, 2010.

[16] Y.-C. Gao, Y. Zeng, and S.-M. Cai. Influence network in the chinese stock market. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(3):P03017, 2015.

[17] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.

[18] F. Gursoy and D. Gunnec. Influence maximization in social networks under deterministic linear threshold model. *Knowledge-Based Systems*, 161:111–123, 2018.

[19] K. Hussain, M. N. Mohd Salleh, S. Cheng, and Y. Shi. Metaheuristic research: a comprehensive survey. *Artificial intelligence review*, 52(4):2191–2233, 2019.

[20] W. Jiang. Graph-based deep learning for communication networks: A survey. *Computer Communications*, 185:40–54, 2022.

[21] K. Kapoor, D. Sharma, and J. Srivastava. Weighted node degree centrality for hypergraphs. In *2013 IEEE 2nd Network Science Workshop (NSW)*, pages 152–155, 2013. doi: 10.1109/NSW.2013.6609212.

[22] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[23] J. Kunegis. Konect: The koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 1343–1350, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320382. doi: 10.1145/2487788.2488173. URL https://doi.org/10.1145/2487788.2488173.

[24] M. Lalou, M. A. Tahraoui, and H. Kheddouci. The critical node detection problem in networks: A survey. *Computer Science Review*, 28:92–117, 2018.

[25] V. Latora, V. Nicosia, and G. Russo. *Complex networks: principles, methods and applications*. Cambridge University Press, 2017.

[26] J. M. Lewis and M. Yannakakis. The node-deletion problem for hereditary properties is np-complete. *Journal of Computer and System Sciences*, 20(2):219 – 230, 1980. ISSN 0022-0000.

[27] Y. Li, J. Fan, Y. Wang, and K.-L. Tan. Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1852–1872, 2018.

[28] R. I. Lung, M. Suciu, and N. Gaskó. Noisy extremal optimization. *Soft Computing*, 21(5): 1253–1270, 2017.

[29] J. C. Mitchell. Social networks. *Annual review of anthropology*, 3(1):279–299, 1974.

[30] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015. URL http://networkrepository.com.

[31] P. Shakarian, A. Bhatnagar, A. Aleali, E. Shaabani, and R. Guo. The independent cascade and linear threshold models. In *Diffusion in Social Networks*, pages 35–48. Springer, 2015.

[32] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317, 1953.

[33] S. Shen, J. C. Smith, and R. Goli. Exact interdiction models and algorithms for disconnecting networks via node deletions. *Discrete Optimization*, 9(3):172–188, 2012.

[34] SocioPatterns. Infectious contact networks. URL http://www.sociopatterns.org/datasets/.

[35] L. Šubelj and M. Bajec. Robust network community detection using balanced propagation. *The European Physical Journal B*, 81(3):353–362, 2011.

[36] R. C. Thomson and D. E. Richardson. A graph theory approach to road network generalisation. In *Proceeding of the 17th international cartographic conference*, pages 1871–1880, 1995.

[37] J. L. Walteros and P. M. Pardalos. *Selected Topics in Critical Element Detection*, pages 9–26. Springer New York, New York, NY, 2012. ISBN 978-1-4614-4109-0. doi: 10.1007/978-1-4614-4109-0_2. URL https://doi.org/10.1007/978-1-4614-4109-0_2.