Hálózatok rekonstrukciója és felhasználásuk az optimalizálásban

Doktori disszertáció

Homolya Viktor

Témavezető: Dr. Vinkó Tamás

Informatika Doktori Iskola
Informatikai Intézet
Természettudományi és Informatikai Kar
Szegedi Tudományegyetem

Szeged
2024

1. Bevezető

Az optimalizálás története kiterjedt és számos területet felölel, beleértve a matematikát, a mérnöki tudományokat, a közgazdaságtant és a számítástechnikát. Az optimalizálás célja, egy probléma lehetséges megoldásainak halmazából megtalálni a legjobb megoldást.

A gráfelmélet és a hálózattudomány kezdetét Leonhard Euler königsbergi hidak problémájának megoldásától számítjuk. Az évek során hatalmas fejlődésen esett át köszönhetően más tudomány területekben való felhasználhatóságának. A számítógépek nagy adatfeldolgozási képességei miatt a gyakorlatban felmerülő problémákat kezdték vizsgálni. Ennek köszönhetően alakult ki a hálózattudomány.

Ma az optimalizálás, gráfelmélet és a hálózattudomány rendkívül interdiszciplináris terület. Matematika, informatika, fizika, biológia és más területek kutatói együttműködnek a komplex rendszerek vizsgálatánál és problémák megoldásának megtalálásáért, melyekhez az optimalizálás és gráfelmélet eszközeit használják.

2. Gráf rekonstrukció

A 2. fejezet a szakirodalomban régóta ismert gráf rekonstrukció azon változatával foglalkozik, ahol az ismert adataink a csúcsok Közelség centralitás (Closeness Centrality) értékei és a rekonstruálandó gráf fa. A fejezet cikk formája a [2] publikáció.

A probléma azért érdekes, mert felhasználható korlátozott adatokból való hálózatok helyreállításához, adattömörítéshez vagy generatív modellek készítéséhez, melyek teszteléseket segíthetik.

Az első gráf rekonstrukciós probléma a fokszám értékekből való gráf építése volt. Ehhez megfogalmazott feltételről és algoritmusról a [5, 6, 7] cikkekben olvashatunk. Más centralitási mértékekre készítettek algoritmusokat, például a [9] cikkben található.

Ezen munkában a jelenleg elterjedt populáció-alapú, sztochasztikus algoritmusoktól eltérő eszközt szerettünk volna alkotni. A Közelség központiságból való rekonstrukciót tudomásunk szerint nem vizsgálták a munkánk befejeztéig. Célunk volt ezen specializált feladat megértése és olyan összefüggések megtalálása, melyekkel segíthetjünk az általános esetek megoldását.

A probléma nehézségét növeli, hogy egynél több gráf is rendelkezhet a megadott szekvenciával megegyező Közelségi centralitás értékekkel. Ezen gráfok nem feltétlen izomorfak.

Több algoritmust alkottunk és teszteltünk. Munkánk során extra információt használtuk fel, tudtuk mely csúcsok levelek. Teszteket végeztünk ezen extra információ elhagyásáról, mellyel algoritmusaink futási ideje nőtt, de a megoldott példák száma nem csökkent. Észrevételeink egyike volt a gráfban szomszédos csúcsok értékei közötti összefüggés. Ebből megfogalmazott tételünk:

2.1. Tétel. *Adott $G = (V, E)$ gráf és $(v_1, v_2) \in E$ él. Felhasználva a*

$$V_G(v_i, v_j) := \{v : v \in V,\ d_G(v, v_i) \leq d_G(v, v_j)\} \tag{1}$$

jelölést teljesül, hogy

$$ds[v_1] - ds[v_2] = |V_G(v_2, v_1)| - |V_G(v_1, v_2)|. \tag{2}$$

Empirikus erdeményekből azt tapasztaltuk, hogy a fák fokszámsorozat és a rekonstruálási feladat nehézsége (ezen algoritmusokkal) között kapcsolat van.

A Maximum Presolve algoritmusunk gyorsan elkészülő és közel jó megoldásival más algoritmusok kezdő értékeinek kiválasztását segíthetjük. A populáció-alapú algoritmusokban való felhasználása jövőbeli tervünk.

A populáció-alapú algoritmusoktól eltérő heurisztikus algoritmust készítettünk Corrector néven, mely A Maximum Prsolve algoritmusunk eredményét használja fel. Ezen algoritmus különböző változatait teszteltük.

A Recursive algoritmusunk pontos megoldások keresésre lett fejlesztve. Az 1000 méretű gráfok tesztelésénél azt tapasztaltuk, hogy bizonyos fokszámeloszlás tulajdonsággal rendelkező gráfok esetén nem hatékony, akárcsak a Maximum Presolve és a Corrector együttes használata. További összefüggéseket szeretnénk keresni, melyekkel a futási idő csökkenne ilyen gráfokon, illetve az általános egyszerű, súlyozatlan gráf esetének megoldását is segítjük.

3. Gráf realizáció

A 3. fejezet a rekonstrukciónál egyszerűbb feladatot tárgyal, a realizációt. A feladat csak az, hogy eldöntsük az adott adatokból építhető-e gráf. Az előzőhöz fejezethez hasonlóan közelségi centralitás értékekből dolgozunk.

A munka során fa és nem-fa gráfokra sikerült 4-4 feltételt alkotni. Ezen feltételek csak szükségesek, így a feltételeket teljesítő adatsorból nem garantált a pontos egyezésű gráf alkotása. A feltételeket bizonyítottuk is.

A 2. fejezetben szereplő probléma gyors input tesztelésére használhatóak ezek a feltételek. Így elkerülhető, hogy költségesebb algoritmusokat kelljen használni nem megfelelő adatokon.

A megfogalmazott feltételek a következők:

3.1. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, fa gráf esetén, ha $|V| = N$, akkor $2 \mid \sum_{v \in V} ds(v)$. Ha $2 \mid N$, akkor $\forall v : \ (2 \mid ds(v) \vee 2 \nmid ds(v))$.*

3.2. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, fa gráf minden v csúcsára igaz, hogy $N - 1 \leq ds(v) \leq \frac{N(N+1)}{2}$, ahol $|V(G)| = N$.*

3.3. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, fa gráfra igaz, ha δ a G gráf ds értékeinek vektora és $|V(G)| = N$, akkor $N - 2 \leq \max(\delta) - \min(\delta)$.*
$2 \mid N$ esetén $\max(\delta) - \min(\delta) \leq \frac{N^2}{4} - \frac{N}{2}$, különben $\max(\delta) - \min(\delta) \leq \frac{N^2}{4} - \frac{N}{2} + \frac{1}{4}$.

3.4. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, fa gráfra igaz, ha δ a G ds értékeinek vektora és $|V(G)| = N$, akkor*

$$2(N-1)^2 \leq \sum_i \delta[i] \leq a_N,$$

ahol $a_{n+1} = a_n + n(n+1)$, $a_1 = 0, a_2 = 2$

3.5. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, összefüggő, nem-fa gráf esetén, ha $|V| = N$, akkor $2 \mid \sum_{v \in V} ds(v)$.*

3.6. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, nem-fa gráf minden v csúcsára igaz, ha $|V(G)| = N$, akkor $N - 1 \leq ds(v) \leq \frac{N(N+1)}{2} - 1$.*

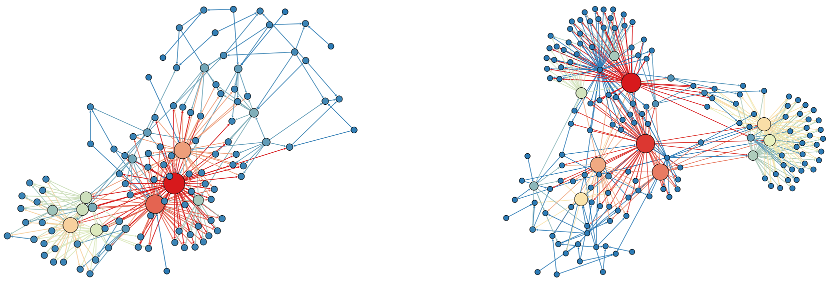
3.7. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, összefüggő, nem-fa gráfra igaz, ha δ a G ds értékeinek vektora és $|V(G)| = N$, akkor $2 \mid N$ esetén $\max(\delta) - \min(\delta) \leq \frac{N^2}{4} - \frac{N}{2}$, különben $\max(\delta) - \min(\delta) \leq \frac{N^2}{4} - \frac{N}{2} + \frac{1}{4}$.*

3.8. Tétel. *Minden $G = (V, E)$ egyszerű, irányítatlan, súlyozatlan, összefüggő, nem-fa gráfra igaz, ha δ a G ds értékeinek vektora és $|V(G)| = N$, akkor $N(N-1) \leq \sum_i \delta[i] \leq a_N - 2(N-2)$, ahol $a_{n+1} = a_n + n(n+1)$, $a_1 = 0, a_2 = 2$.*

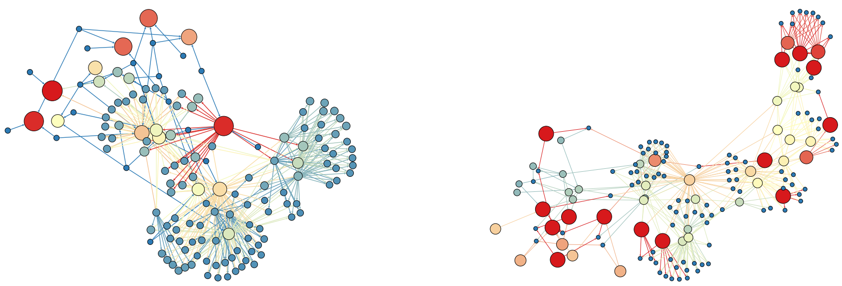
4. Memetic Differential Evolution hálózattudományi eszközökkel

A 4. fejezet, mely a [3, 4] cikkekből készült, a Memetic Differential Evolution (MDE) [10] populáció-alapú algoritmus vizsgálatáról és fejlesztéséről szól. Néhány közismert tesztfüggvény globális optimumát kerestük.

Az eredeti algoritmusban szereplő szabályt megváltoztattuk. A megtalált lokális optimumokból épített hálózatbeli centralitási értékeket használtuk az egyedek kiválasztására. A különböző szabályokkal különböző hálózatok készültek. Példa látható a 1. és 2. ábrákon.



1. ábra. *MDE LON-ok: fokszám centralitást (balra) és közöttiség centralitást (jobbra) használva. A csúcsméret a megfelelő centralitási értéket jelenti.*



2. ábra. *MDE LON-ok: PageRank centralitást (balra) és közelség centralitást (jobbra) használva. A csúcsméret a megfelelő centralitási értéket jelenti.*

Az eredeti algoritmust kezdtük vizsgálni és az általa gyártott hálózatokat. A további teszteléseink során nagyobb dimenzió méretet, eltolást és forgatást használtunk, hogy a tesztelendő függvények globális optimumát nehezebb legyen megtalálni. A globális optimum értékeket ismertük.

A közismert nyolc MDE változat futtatása közben épített hálózatokat felhasználva egyszerű újraindítási szabályt fogalmaztunk meg a korai konvergencia okozta hiba ellen, amivel bonyolultabb problémáknál javulást értünk el.

Az újraindítási szabály használatával elért eredményekről példa látható a 1. és 2. táblázatokon, amiken a jelölések a következők:

- S a sikeres futtatások százaléka, azaz hányszor sikerült elérni a globális minimumot;
- 'Best' a legjobb megtalált függvényérték a K futásból;
- 'Avg' a futások átlagos függvényértéke;
- 'Adf' (Average difference failure) az átlagos eltérés a megtalált függvényérték és a globális optimum érték között azokból a futásokból, amelyek nem voltak sikeresek;
- 'LS' (Local Search) az átlagos lokális keresések száma sikeres futásonként;
- 'SP' (Success Performance) a siker teljesítmény, melyet úgy számítanak, hogy

$$\text{átlag(lokális keresések száma a sikeres futások során)} \times \frac{K}{\text{sikeres futások száma}}.$$

1. táblázat. *Mutatók és gráf mértékek a forgatott $Schwefel_{20}$-nál*

szabály	S	Best	Avg	Adf	LS	SP
best/1/bin	0	-7905.9	-7371.6	1007.9	∞	∞
best/1/exp	0	-8142.7	-7204.9	1174.6	∞	∞
best/2/bin	0	-8261.2	-7886.8	492.8	∞	∞
best/2/exp	0	-8024.3	-7629.7	749.9	∞	∞
rand/1/bin	2	-8379.6	-7875.2	514.6	3520	176000
rand/1/exp	0	-8024.3	-7639.5	740.1	∞	∞
rand/2/bin	**20**	-8379.6	**-8202.2**	**221.7**	15408	**77040**
rand/2/exp	4	-8379.6	-8114.2	276.4	5530	138250

Az első vizsgálatainkhoz hasonlóan a futás közben épített hálózatban lévő centralitás értékeket használtuk új szabályokhoz. Ezeket a szabályokat hasonlítottuk össze az eredeti klasszikussal ugyanazon bonyolultabb tesztfüggvényeken, melyeket a fejezett korábbi részében használtunk.

A felhasznált centralitási mértékek a közelség, a közöttiség, a fokszám, a harmonikus centralitás és a PageRank volt. Az adott paraméterekkel a klasszikus MDE bizonyult az egyik leghatékonyabbnak.

2. táblázat. *Mutatók és gráf mértékek a forgatott* $Schwefel_{20}$*-nál az Above-Below újraindítási feltétellel*

szabály	S	Best	Avg	Adf	LS	SP
best/1/bin	0	-8142.7	-7685.6	693.9	∞	∞
best/1/exp	0	-8024.3	-7464.8	914.8	∞	∞
best/2/bin	0	-8261.2	-7993.3	386.2	∞	∞
best/2/exp	0	-8261.2	-7834.0	545.6	∞	∞
rand/1/bin	0	-8142.7	-7899.7	479.8	∞	∞
rand/1/exp	0	-8261.2	-7639.2	740.4	∞	∞
rand/2/bin	**26**	-8379.6	**-8188.6**	**258.1**	13524	**52017**
rand/2/exp	4	-8379.6	-8111.9	278.8	5850	146250

5. Befolyás terjedés maximalizálás probléma terének vizsgálata

Az 5. fejezet egy gráf-alapú optimalizálási probléma vizsgálatáról szól, melynek neve Befolyás terjedés maximalizálás [8]. A probléma lokális optimumait derítettük fel és alkottunk belőlük hálózatot, hasonlóan az előző fejezetbelihez. A fejezet az [1] publikációként jelent meg.

Célunk a probléma megértése volt és olyan összefüggések találása, melyekkel a problémához tudunk előfeldolgozóknak vagy egyéb módszereknek segítő észrevételeket megfogalmazni.

A közismert Monte Carlo-módszert alkalmazó terjedési modellek közül a Független Kaszkádokat (Independent Cascades) vizsgáltuk. Megfogalmaztunk lokális optimumokat ebben a diszkrét problémában, melyekből építettünk egy az előző fejezetben szereplő Lokális Optimumok Hálózatától eltérőt. Ezen hálózat élei súlyozottak és írányítottak. A definíciójához igazodva alkottunk egy hegymászó-jellegű algoritmust.

Az algoritmusunk nem a probléma megoldását kereste. Egy-egy optimális megoldás megtalálásához a szakirodalomban ismert a Befolyás terjedés Maximalizálás problémához tartozó Mohó algoritmust használtuk. Az algoritmusunk terjedési modellt ismételte az előző iteráció eredményétől függő kezdőértékkel az általunk megfogalmazott feltételek teljesüléséig.

A munka során az általunk ebben a részben megfogalmazott Lokális Optimumok Hálózatának diszkrét elemeket tartalmazó csúcsaira definiáltunk távolság függvényt, melyet Vertex Set Distance-nek neveztünk el (röviden VSD). A VSD formális leírása a következő:

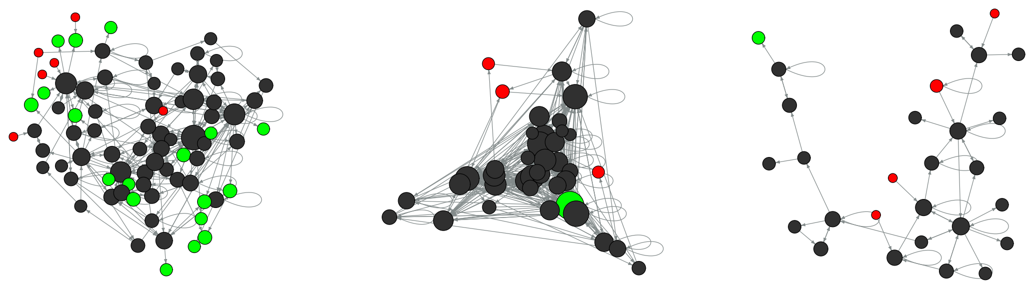
Ha V_1 és V_2 a G gráf csúcshalmazának k-elemű részhalmazai. Legyen

$$VSD(v_1, v_2) = \min_{y \in S_k} \left(\sum_{i=1}^{k} d(V_{1_i}, V_{2_{y_i}}) \right),$$

ahol $d : V(G) \times V(G) \to \mathbb{N}_0$, a gráfon vett távolság, és S_k az első k természetes számból álló permutációk halmaza.

A VSD-ről bizonyítottuk, hogy matematikai értelemben vett távolságfüggvény.

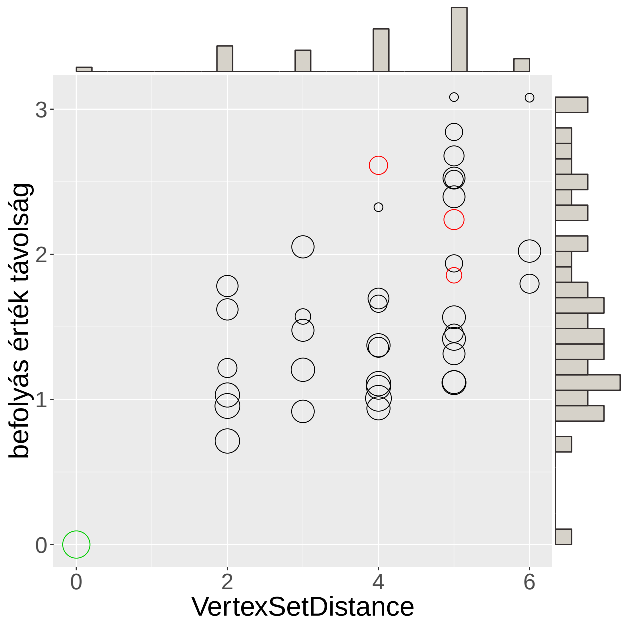
Vizsgálatokat végeztünk különböző generatív modellekkel készített gráf típusokon, melyeket használtuk fel inputként. A hegymászó-jellegű algoritmusunk által adott hálózatokat egyszerűsítettük, bizonyos felesleges elemeket töröltünk. Ezen tesztek eredményei közül láthatunk példát a 3. ábrán.



(a) *Watts-Strogatz* **(b)** *Barabási-Albert* **(c)** *Cooper-Frieze*

3. ábra. *Lokális optimumok hálózatai, szűrt változatok (H gráfok)*

A fejezetben ezen diszkrét probléma terét vizualizáltuk is a VSD és a probléma értékének összefüggésében további kísérletek segítéséért. Példa látható a 4. ábrán.



4. ábra. VSD *értékek egy Barabási-Albert algoritmussal készített input gráfon*

6. A szerző hozzájárulásai

A 2. fejezetben szereplő munkából a disszertáció szerzője alkotta meg és implementálta a ismertett algoritmusokat, **Maximum Presolve, Corrector** és **Recursive**. A tesztelések megtervezése és kiértékelése, a fejezetben szereplő eredmények cikk formájú leírása közös munka a témavezetővel.

A 3. fejezetben található gráf realizációhoz tartozó feltételeket és bizonyításaikat a disszertáció szerzője alkotta.

A 4. fejezetben bemutatott algoritmusok implementálását, az utolsó alfejezetben szereplő tesztekhez szükséges kiegészítések kivételével, és a tesztek elvégzését a szerző hajtotta végre.

A tesztek tervezése, az eredmények értelmezése és a fejezetben bemutatott **Above-Below szabály** a témavezető és a szerző közös munkája.

Az 5. fejezetben ismertetett **hegymászó-jellegű algoritmus** megalkotása, a szakirodalomban ismert terjedési modell kivételével, a témavezető és a szerző közös munkája, a speciális Lokális Optimumok Hálózatának megfogalmazásával, a tesztek tervezésével és az eredmények értelmezésével együtt.

A szerző végezte el az implementálásokat, teszteléseket, a fejezetben szereplő **Vertex Set Distance** (VSD) mérték megfogalmazását és bizonyítását arról, hogy matematikai értelemben vett távolság függvény.

Hivatkozások

A szerző folyóirat publikációi

[1] **Homolya V.** és Vinkó T. Befolyás terjedés optimumainak hálózatáról. *Alkalmazott Matematikai Lapok*, 37, 167–179, 2020.

[2] **V. Homolya** and T. Vinkó. Closeness centrality reconstruction of tree graphs. *Central European Journal of Operations Research*, (közlésre elfogadva) 2023.

A szerző konferencia kötetekben megjelent publikációi

[3] **V. Homolya** and T. Vinkó. Memetic differential evolution using network centrality measures. In *Proceedings of the 14th International Global Optimization Workshop,* AIP Publishing, 250–254, 2019.

[4] **V. Homolya** and T. Vinkó. Leveraging local optima network properties for memetic differential evolution. In *Optimization of Complex Systems : Theory, Models, Algorithms and Applications,* Springer-Verlag, 109–118, 2020.

Egyéb hivatkozások

[5] P. Erdős and T. Gallai. Gráfok eloírt fokú pontokkal (graphs with points of prescribed degrees, in hungarian). *Mat. Lapok*, 11:264–274, 1961.

[6] S. L. Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, 1962.

[7] V. Havel. A remark on the existence of finite graphs. *Casopis Pest. Mat.*, 80:477–480, 1955.

[8] D. Kempe, J Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[9] M. Lozano and F. J. Rodriguez. Network reconstruction from betweenness centrality by artificial bee colony. *Swarm and Evolutionary Computation*, 62:100851, 2021.

[10] A.P. Piotrowski. Adaptive memetic differential evolution with global and local neighborhood-based mutation operators. *Information Sciences*, 241:164–194, 2013.

7. Summary

Chapter 2, **Graph reconstruction**, deals with the version of graph reconstruction that has long been known in the literature, where the known data are the Closeness Centrality values of the vertices and the graph to be reconstructed is tree.

In this work, we wanted to create a tool different from the currently widespread population-based, stochastic algorithms. Our goal was to understand this specialized task and to find connections with which we could help solve general cases.

We created and tested several algorithms. One of our observations was the correlation between the values of adjacent vertices in the graph. From empirical results, we found that there is a relationship between the degree sequence of the trees and the difficulty of the reconstruction task.

Chapter 3, **Graph realization**, discusses a simpler task than reconstruction, realization. The only task is to decide whether a graph can be built from the given data. Similar to the previous chapter, we work from Closeness Centrality values.

During the work, it was possible to create 4 conditions for tree and non-tree graphs. These conditions are only necessary. We have also proved the conditions.

These conditions help to avoid using more expensive algorithms on inappropriate data.

Chapter 4, **Memetic Differential Evolution with Network Science tools**, is about the investigation and development of the population-based algorithm Memetic Differential Evolution. We searched for the global optimum of some well-known test functions.

We created new rules based on centrality values from the local optima network of provisional results of MDE.

Using the networks built while running the eight well-known MDE versions, we formulated a simple restart rule against the error caused by early convergence, with which we achieved an improvement in more complicated problems.

The centrality measures closeness, betweenness, degree, harmonic centrality and PageRank were used for further tests.

Chapter 5, **Investigation of the space of Influence Maximization Problem** deals with the investigation of a graph-based optimization problem, Influence Maximization. We discovered the local optima of the problem and created a network from them, similar to the one in the previous chapter.

Our goal was to understand the problem and to find connections with which we can formulate helpful ideas for preprocessors or other methods. We created a hill-climbing algorithm to map the space of problem.

During the work, we defined a distance function for the vertices containing discrete elements of the Local Optima Network formulated in this section, which we named Vertex Set Distance. Finally, we visualized the space of the discrete problem to help further experiments.

Társszerzői nyilatkozat

Homolya Viktor “Hálózatok rekonstrukciója és felhasználásuk az optimalizálásban” című PhD disszertációjában a következő eredményekben Homolya Viktor hozzájárulása volt a meghatározó:

A disszertáció 2. fejezetében és a

1. V. Homolya and T. Vinkó. Closeness centrality reconstruction of tree graphs. *Central European Journal of Operations Research*, (közlésre elfogadva) 2023

cikkben szereplő eredményekre vonatkozóan:

- A disszertáció szerzője alkotta meg és implementálta az algoritmusokat.
- A tesztelések megtervezése, kiértékelése és az eredmények közös munka a témavezetővel.

A disszertáció 3. fejezetében bemutatott eredményekkel kapcsolatosan:

- A disszertáció szerzője fogalmazta meg a fejezetben ismertetett feltételeket és dolgozta ki azok bizonyításait.

A disszertáció 4. fejezetében és a

1. V. Homolya and T. Vinkó. Memetic differential evolution using network centrality measures. In *Proceedings of the 14th International Global Optimization Workshop*, AIP Publishing, 250–254, 2019.
2. V. Homolya and T. Vinkó. Leveraging local optima network properties for memetic differential evolution. In *Optimization of Complex Systems : Theory, Models, Algorithms and Applications*, Springer-Verlag, 109–118, 2020.

cikkekben szereplő eredményekre vonatkozóan:

- Az algoritmusok implementálását, az utolsó alfejezetben szereplő tesztekhez szükséges kiegészítések kivételével, és a tesztek elvégzését a szerző hajtotta végre.
- A tesztek tervezése, az eredmények értelmezése és az Above-Below szabály a témavezető és a szerző közös munkája.

A disszertáció 5. fejezetében és a

1. V. Homolya and T. Vinkó. Befolyás terjedés optimumainak hálózatáról. *Alkalmazott Matematikai Lapok*, 37, 167–179, 2020.

Cikkben szereplő eredményekkel kapcsolatban:

- A hegymászó-jellegű algoritmus megalkotása, a szakirodalomban ismert terjedési modell kivételével, a témavezető és a szerző közös munkája, a speciális Lokális Optimumok Hálózatának megfogalmazásával, a tesztek tervezésével és az eredmények értelmezésével együtt.
- A szerző végezte el az implementálásokat, teszteléseket, a VSD megfogalmazását és bizonyítását arról, hogy valódi távolság fogalom.

Ezek az eredmények Homolya Viktor PhD disszertációján kívül más tudományos fokozat megszerzésére nem használhatók fel.

Szeged, 2024. január 12.

Jelölt

Témavezető

Az Informatika Doktori Iskola vezetője kijelenti, hogy jelen nyilatkozatot minden társszerzőhöz eljuttatta, és azzal szemben egyetlen társszerző sem emelt kifogást.

Szeged, 2024. január 12.

SZEGEDI TUDOMÁNYEGYETEM INFORMATIKA DOKTORI ISKOLA SZEGED

Doktori Iskola vezető