Ohranjanje lastnosti pri zmanjševanju družbenih omrežij

Neli Blagus, Lovro Šubelj, Aljaž Zrnec, Marko Janković, Marko Bajec
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
{ime.priimek}@fri.uni-lj.si

Izvleček

Z razvojem svetovnega spleta in možnostjo shranjevanja velikih količin podatkov so začeli z omrežji opisovati vedno večje sisteme. Posebno družbena omrežja so vse kompleksnejša zaradi različnih storitev družabnega mreženja na spletu. Analiza velikih omrežij je časovno in prostorsko zahtevna, prav tako je na njih težje opazovati različne dinamične procese. Raziskovalci so za rešitev teh problemov predlagali različne pristope zmanjševanja omrežij. V prispevku predstavimo metode zmanjševanja družbenih omrežij z vzorčenjem in združevanjem, ki so namenjene lažjemu razumevanju ter hitrejši in učinkovitejši analizi. Osredinimo se na raziskovanje podobnosti med osnovnimi in zmanjšanimi omrežji ter na konkretnih primerih družbenih omrežij ovrednotimo uspešnost različnih algoritmov pri ohranjanju pomembnejših lastnosti omrežij.
Ključne besede: analiza omrežij, družbena omrežja, zmanjševanje, vzorčenje, združevanje, ohranjanje lastnosti.

Abstract

Preserving Properties in the Simplification of Social Networks

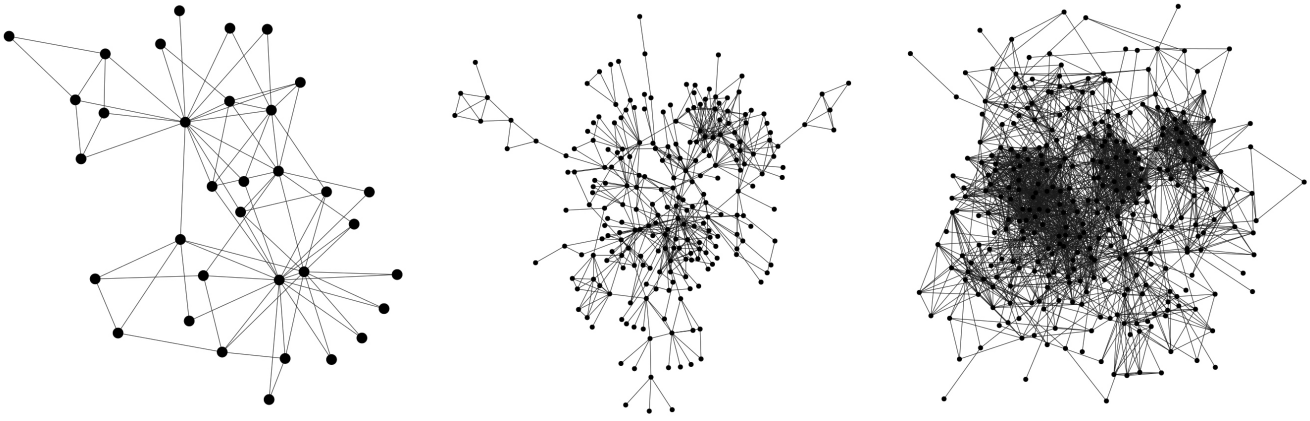
In the past decade, the capability of storing large amounts of data and the evolution of the internet, particularly different services for social networking, increase the size and complexity of networked systems. Their investigation presents a great challenge, since the algorithms for analysis can be temporally or spatially inappropriate, and furthermore observing dynamical processes on large networks can be too expensive. In this paper, we present different existing methods for simplifying social networks in order to provide for easier understanding and more efficient analysis of large networks. We focus on observing similarities between original and simplified networks. Moreover, we analyze several social networks and study the quality of simplification methods based on how well they preserve fundamental network properties.
Key words: network analysis, social networks, simplification, sampling, merging, property preserving.

1 UVOD

Analiza omrežij (Cohen & Havlin, 2010; Newman, 2010) se ukvarja z raziskovanjem sistemov, ki jih predstavimo z omrežji, tj. matematičnimi grafi, sestavljenimi iz vozlišč in povezav med njimi, pri čemer imamo o vozliščih in povezavah dodatno znanje. Poznamo več tipov omrežij, med drugim družbena, informacijska, tehnološka in biološka (Newman, 2003); omrežja pa lahko razdelimo tudi na več vrst glede na usmerjenost (usmerjena in neusmerjena) ali uteženost povezav (utežena in neutežena). V zadnjih letih so velik razmah doživela družbena omrežja (Knoke, Yang & Kuklinski, 2008) predvsem s popularnostjo raznih storitev družabnega mreženja na spletu. Z omrežji so začeli opisovati vedno večje sisteme, kar je povečalo njihovo kompleksnost in otežilo analizo.

Družbeno omrežje je sestavljeno iz vozlišč, ki predstavljajo ljudi, ter povezav, ki pomenijo različne interakcije med njimi, na primer poznanstvo, sorodstvo, soavtorstvo ali kakšno drugo obliko sodelovanja (slika 1). Vozlišča so lahko opisana z različnimi atributi, kot na primer spol, starost, kraj bivanja, področje raziskovanja. Tudi povezave lahko vsebujejo dodatne informacije, kot so vrsta poznanstva ali sorodstva ter oblika sodelovanja ali soavtorstva. Analiza družbenih omrežij se ukvarja z raziskovanjem takšnih sistemov, z opazovanjem zgradbe, globalnih značilnosti ter lastnosti posameznih vozlišč in povezav, s predvidevanjem razvoja v prihodnosti ali z napovedovanjem manjkajočih podatkov.

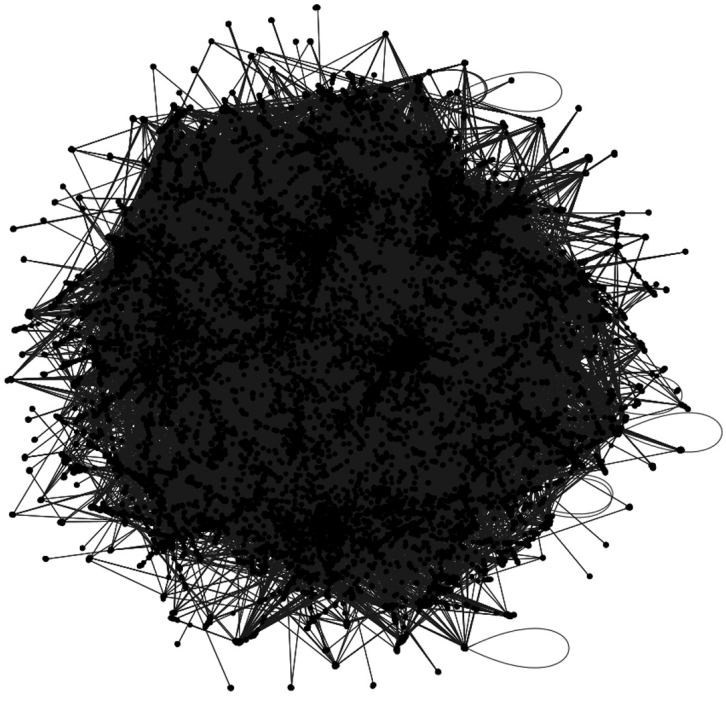
Za velika omrežja štejemo omrežja z več tisoč vozlišči ali več sto tisoč povezavami. Takšna omrežja je težko prikazati in proučevati s prostim očesom (slika 2). Prav tako je časovno zahtevna njihova analiza in opazovanje različnih dinamičnih procesov na



Slika 1: **Primeri družbenih omrežij (od leve proti desni): prijateljstva med člani karate kluba (Zachary, 1977), sodelovanja med slovenskimi znanstveniki na področju računalništva in informatike (Blagus, Šubelj & Bajec, 2012), prijateljstva v družabnem omrežju Facebook (Blagus, Šubelj & Bajec, 2012)**

omrežju. Za rešitev teh problemov so raziskovalci predlagali zmanjševanje omrežij, ki ima izvor v teoriji grafov (npr. razbitje (angl. partitioning) (Feder & Motwani, 1991; Karypis & Kumar, 1998) ali bločno modeliranje (angl. blockmodeling) (Batagelj, 1997)). Z razvojem spleta so začeli podrobneje raziskovati velika omrežja, kot na primer omrežje spletnih strani in internetno omrežje. Za obvladovanje velikosti teh omrežij so raziskovalci uporabljali pristope za zmanjševanje (Krishnamurthy idr., 2005). Poudarek raziskav je bil predvsem na stiskanju (angl. compression) za učinkovito shranjevanje omrežij (Deo & Litow, 1998; Adler & Mitzenmacher, 2001). Z večanjem kompleksnosti omrežij pa so se razvijali tudi pristopi za poenostavitev oziroma zmanjševanje omrežij (angl. simplification), namenjeni preglednejšemu prikazu (Hennessey, Brooks, Fridman & Breen, 2008; Gilbert & Levchenko, 2004) ter učinkovitejši analizi (De Nooy, Mrvar & Batagelj, 2005; Leskovec & Faloutsos, 2006; Hübler, Kriegel, Borgwardt & Ghahramani, 2008).

Uporabna vrednost zmanjševanja pa se ne kaže le v hitrejši analizi. Zmanjšano omrežje prikazuje bolj resnično sliko sistema kot na primer naključno ge-



Slika 2: **Primer velikega družbenega omrežja z 21.363 vozlišči in 182.628 povezavami (Leskovec, Kleinberg & Faloutsos, 2007)**

nerirani približki realnih omrežij. Izkaže se tudi, da so zbrani podatki o sistemu, opisanem z omrežjem, pogosto nepopolni. V tem smislu je že samo omrežje vzorčena slika celotnega sistema, pomembno pa je razumeti, kako sta si resnični in vzorčni sistem podobna. V prispevku se osredinimo na vzorčne sisteme s poudarkom na ohranjanju različnih lastnosti omrežij med zmanjševanjem.

Med procesom zmanjševanja spremenimo omrežje in vplivamo na njegove lastnosti. Pri tem želimo, da se ohranijo lastnosti osnovnega omrežja, kako in katere pa je odvisno od vrste preučevanega omrežja ter namena raziskave. Na primer, pri splošni analizi družbenih omrežij praviloma preučujemo vpliv posameznikov v družbi, zato se pri zmanjševanju omrežij osredinimo na ohranjanje lastnosti posameznih vozlišč, kot so stopnje (angl. degree) in središčnosti (Freeman, 1979) (angl. centrality). Na drugi strani pa so različna biološka omrežja, ki jih zadnje čase pogosto preučujejo v bioinformatiki, sestavljena iz manjših vzorcev vozlišč, kot so motivi (Milo idr., 2001) (angl. motif) in grafki (Pržulj, Wigle & Jurisica, 2004) (angl. graphlet). Pri zmanjševanju je tako zaželeno ohranjanje predvsem zadnjih. V okviru spletnega rudarjenja (angl. Web mining) raziskujemo velika spletna in družabna omrežja, v katerih navadno najdemo različne vrste karakterističnih skupin vozlišč, kot so skupnosti (Girvan & Newman, 2002) (angl. community) in moduli (Šubelj & Bajec, 2012a) (angl. module). Zmanjševanje spletnih omrežij se zato osredinja na ohranjanje omenjenih skupin ter lokalne okolice vozlišč, kar lahko merimo prek gostote ali nakopičenosti vozlišč (Watts & Strogatz, 1998) (angl. clustering). Nazadnje pa je pri zmanjševanju različnih tehnoloških omrežij pomembno predvsem ohranjanje globalnih strukturnih lastnosti, kot so značilne porazdelitve stopenj vozlišč (Barabási & Albert, 1999) ali mešanje med stopnjami (Newman, 2002) (angl. mixing). Prav to ima močan vpliv na različne dinamične procese v tehnoloških omrežjih, kot je npr. prenos podatkovnih paketov po internetnem omrežju (Cohen, Erez, Ben-Avraham & Havlin, 2000) ali širjenje informacij v programskih omrežjih (Šubelj & Bajec, 2012b), kar je predmet številnih raziskav v fiziki. Poudarimo, da zgornja delitev ni stroga, saj je pri zmanjševanju nekaterih omrežij smiselno ohranjanje različnih lastnosti. Tako nas v okviru analize omrežij citiranj v bibliometriki zanimajo predvsem strukturne lastnosti celotnega omrežja (Price, 1965), ocenjevanje znanstvene odličnosti v scientometriki pa temelji na lastnostih posameznih vozlišč v omrežjih citiranj (Walker, Xie, Yan & Maslov, 2007).

V prispevku se osredinimo na družbena omrežja in predstavimo uveljavljene pristope za njihovo zmanjševanje. Ogledamo si tri primere velikih družbenih omrežij ter jih zmanjšamo z različnimi pristopi. Nato opazujemo, kako so si podobna osnovna in zmanjšana omrežja, ter ovrednotimo uspešnost metod glede na ohranjanje različne lastnosti omrežij med zmanjševanjem.

V nadaljevanju najprej predstavimo pristope za zmanjševanje omrežij. Nato si v razdelku 3 ogledamo primere velikih družbenih omrežij ter možnosti za njihovo zmanjševanje z različnimi načini. Prikažemo učinkovitost ohranjanja nekaterih lastnosti omrežij med zmanjševanjem ter razpravljamo o rezultatih analize. V zadnjem razdelku sledi sklep.

2 PRISTOPI ZA ZMANJŠEVANJE OMREŽIJ

Raziskovalci so v zadnjih letih predlagali različne načine zmanjševanja velikih omrežij za namene hitrejše in preprostejše analize. Nekateri so se osredinili le na določen tip omrežij (Biedl, Brejová & Vinar, 2000; Kudelka, Horak, Snasel & Abraham, 2010), drugi so poleg strukture omrežja pri zmanjševanju upoštevali tudi atribute vozlišč in povezav (Tian, Hankins & Patel, 2008; Zhou, Cheng & Yu, 2009), nekateri pa so raziskovali ohranjanje lastnosti omrežij med zmanjševanjem (Leskovec & Faloutsos, 2006; Lee, Kim & Jeong, 2006).

V grobem lahko načine zmanjševanja razdelimo v dve skupini. V prvo spadajo metode vzorčenja omrežja (angl. network sampling), kar pomeni, da iz osnovnega omrežja naključno izberemo vozlišča ali povezave. V drugo skupino spadajo metode združevanja vozlišč oziroma povezav na podlagi njihovih značilnosti; združimo enaka oziroma podobna vozlišča ali povezave v velevozlišča (angl. supernode) ali velepovezave (angl. superedge). V nadaljevanju podrobneje predstavimo metode zmanjševanja iz obeh skupin, s poudarkom na načinih zmanjševanja, ki jih bomo uporabili pri analizi v naslednjem razdelku.

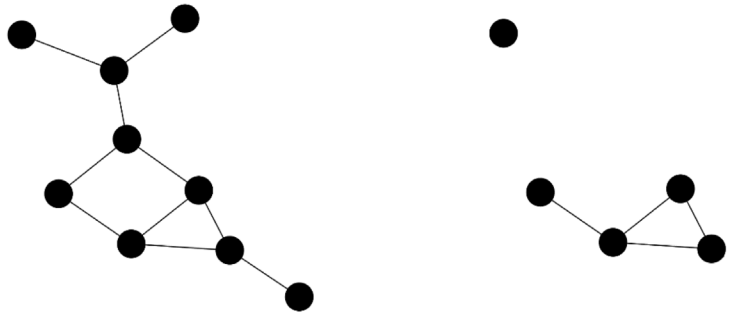
2.1 Zmanjševanje z vzorčenjem

Omrežje lahko vzorčimo na več načinov. Osredinimo se na takšne, ki so preprosti in časovno nezahtevni, a hkrati dobro ohranijo pomembne lastnosti velikih

omrežij: naključno izbiranje vozlišč, naključno izbiranje vozlišč glede na stopnjo, naključno izbiranje povezav (Leskovec & Faloutsos, 2006) ter preiskovanje v širino (Illenberger & Flötteröd, 2011).

Pri naključnem izbiranju vozlišč enakomerno naključno izberemo določeno število vozlišč in vse povezave, ki potekajo med njimi (slika 3). Izkazalo se je, da metoda slabo ohranja porazdelitev stopenj vozlišč (Stumpf, Wiuf & May, 2005). Vozlišča lahko izbiramo tudi sorazmerno glede na izbrano lastnost (npr. stopnja ali mera PageRank (Page, Brin, Motwani & Winograd, 1999)). Za tako generirana omrežja je značilno, da so gostejša, zaradi večje verjetnosti izbire vozlišč z visoko stopnjo pa se spet slabše ohrani porazdelitev stopenj vozlišč. Podobno kot vozlišča lahko v vzorec enakomerno naključno izbiramo povezave. Tako vzorčena omrežja so redkejša z značilno večjim diametrom od osnovnih omrežij. Pri preiskovanju v širino (angl. snowball sampling) enakomerno naključno izberemo začetno vozlišče skupaj z vsemi njegovimi sosedi. Nato v vzorec vzamemo sosednja vozlišča teh sosedov in nadaljujemo postopek, dokler ne dobimo vzorca z želenim številom vozlišč. V smislu povezanosti je tako zmanjšano omrežje najbolj podobno osnovnemu, zaradi česar lahko pričakujemo dobro ohranjanje lastnosti med procesom zmanjševanja.

Avtorji so predlagali številne druge načine vzorčenja omrežja, kot so na primer različni načini preiskovanja omrežja (Leskovec & Faloutsos, 2006), pri čemer je glavna ideja raziskovanje okolice vozlišča: naključna izbira soseda, naključni sprehod, naključno obiskovanje vozlišč (angl. random jump) ali delno preiskovanje v širino (angl. forest fire). Vozlišča oziroma povezave pa lahko vzorčimo tudi glede na določeno lastnost, pri čemer v vzorec vzamemo vozlišča



Slika 3: **Primer zmanjševanja omrežja z vzorčenjem. Levo je prikazano osnovno, desno pa zmanjšano omrežje z enakomernim naključnim izbiranjem vozlišč.**

ali povezave, ki imajo vrednost določene lastnosti nad nekim pragom, na primer izberemo vozlišča glede na stopnjo ali vmesno središčnost (Hennessey idr., 2008) oziroma povezave glede na utež (Toivonen, Mahler & Zhou, 2010).

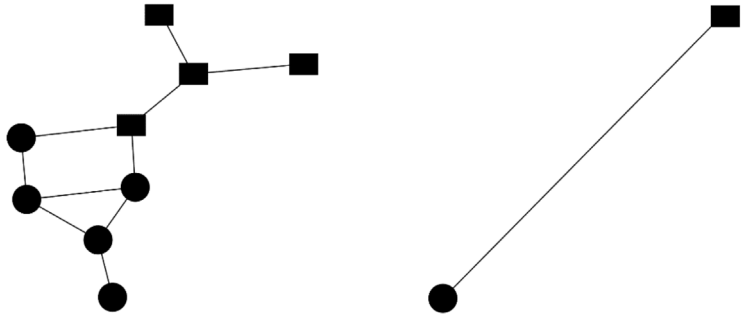
2.2 Zmanjševanje z združevanjem

Pri zmanjševanju omrežja z združevanjem združimo vozlišča v velevozlišča na podlagi njihovih enakih ali podobnih značilnosti. Velevozlišča so povezana z velepovezavami tako, da sta dve velevozlišči povezani, če so med seboj povezana pripadajoča vozlišča osnovnega omrežja. Za zmanjševanje z združevanjem je značilno, da težje kontroliramo velikost zmanjšanega omrežja, saj ne moremo vnaprej določiti, v koliko velevozlišč bodo združena vozlišča.

Veliko realnih omrežij je sestavljenih iz skupnosti (Girvan & Newman, 2002), ki so sestavljene iz tesno povezanih vozlišč, vozlišča iz različnih skupnosti pa so med seboj povezana šibko. Skupnosti v družbenih omrežjih predstavljajo osebe s podobnimi interesi, na primer v omrežjih soavtorstva osebe, ki raziskujejo na sorodnih področjih. Iskanje skupnosti v omrežjih lahko uporabimo za zmanjševanje, tako da združimo vozlišča iz iste skupnosti v velevozlišča (Blagus, Šubelj & Bajec, 2012; Guimerà idr., 2003).

Drugi način zmanjševanja z združevanjem je grobo zrnjenje (angl. coarse-graining) ali renormalizacija, ki je bila v analizo omrežij prenesena iz fizike (Song, Havlin & Makse, 2005). Renormalizacija združi vozlišča v velevozlišča glede na oddaljenost med njimi, tako da za naključno izbrano vozlišče v velevozlišče združi njemu sosednja vozlišča na neki določeni razdalji. Osnovni namen renormalizacije je iskanje samopodobnosti v omrežjih, tj. ohranjanje porazdelitve stopenj vozlišč med procesom

zmanjševanja (Rozenfeld, Gallos, Song & Makse, 2008). Raziskovalci so s spremembo osnovnega načina renormalizacije predlagali veliko drugih metod (Song, Gallos, Havlin & Makse, 2007), ki na primer združujejo povezave namesto vozlišč (W. X. Zhou, Jiang & Sornette, 2007), so primerne samo za določen tip (Itzkovitz idr., 2005) ali za ohranjanje drugih lastnosti omrežja (Blagus, Šubelj & Bajec, 2012). Slika 4 prikazuje zmanjševanje z združevanjem, pri katerem so v velevozlišča združena vozlišča enake oblike.



Slika 4: **Primer zmanjševanja omrežja z združevanjem. Levo je prikazano osnovno, desno pa zmanjšano omrežje, pri katerem so v velevozlišča združena vozlišča enake oblike.**

3 ZMANJŠEVANJE DRUŽBENIH OMREŽIJ

3.1 Primeri družbenih omrežij

V literaturi je bilo raziskovanih veliko različnih družbenih omrežij. Za namene prikaza zmanjševanja na konkretnih primerih smo izbrali tri omrežja (Leskovec, 2009; Kunegis, 2013). Tabela 1 prikazuje njihove osnovne lastnosti.

Tabela 1: **Lastnosti obravnavanih omrežij: ime omrežja, število vozlišč in povezav.**

Omrežje	Vozlišča	Povezave
Facebook	46.952	876.993
DBLP	317.080	1,049.866
Twitter	465.017	835.423

Omrežje Facebook (Viswanath, Mislove, Cha & Gummadi, 2009) je komunikacijsko omrežje med uporabniki spletne storitve Facebook. Omrežje je usmerjeno, vozlišča predstavljajo uporabnike, povezave pa pomenijo pošiljanje sporočil med njimi.

Omrežje DBLP (Yang & Leskovec, 2012) je omrežje sodelovanja med znanstveniki na področju računalništva. Vozlišča predstavljajo raziskovalci, ki so med seboj povezani, če so soavtorji vsaj enega članka.

Omrežje Twitter (De Choudhury idr., 2010) je omrežje sledilcev, pridobljeno iz spletne storitve Twitter. Omrežje je usmerjeno in vsebuje prijateljske povezave med uporabniki.

3.2 Zmanjševanje

Za prikaz zmanjševanja omrežij uporabimo pet pristopov, ki so podrobneje predstavljeni v razdelku 2. Med metodami zmanjševanja z vzorčenjem izberemo enakomerno naključno izbiranje vozlišč, naključno izbiranje vozlišč glede na stopnjo, enakomerno naključno izbiranje povezav ter vzorčenje s preiskovanjem v širino. Vsa zmanjšana omrežja so velikosti 15 odstotkov števila vozlišč osnovnega omrežja, kar se je v literaturi izkazalo za dovolj primerno (Leskovec & Faloutsos, 2006) za ohranitev določenih lastnosti osnovnega omrežja. Med metodami združevanja izberemo združevanje glede na skupnosti z izmenjavo oznak. Velikost zmanjšanega omrežja je pri združevanju odvisna od števila skupnosti osnovnega omrežja. Posledično so zmanjšana omrežja lahko različno velika. V našem primeru sta omrežji Facebook in DBLP zmanjšani na 5 do 15 odstotkov velikosti osnovnega omrežja, medtem ko je omrežje Twitter zmanjšano na približno 0,5 odstotka velikosti osnovnega omrežja.

V prispevku je poudarek na raziskovanju družbenih omrežij, zato za prikaz ohranjanja lastnosti omrežij pri zmanjševanju izberemo lastnosti, ki bi nas zanimale pri njihovi splošni ali lokalni analizi. Opazujemo porazdelitve petih lastnosti, ki so v podobnih raziskavah prav tako pogosto analizirane (Leskovec & Faloutsos, 2006; Lee, Kim & Jeong, 2006):

- vhodna stopnja: v usmerjenem omrežju pove, koliko povezav kaže v vozlišče;

- izhodna stopnja: v usmerjenem omrežju pove, koliko povezav kaže iz vozlišča;
- stopnja vozlišča: število sosednih povezav (v usmerjenem omrežju seštevek vhodne in izhodne stopnje);
- nakopičenost: gostota ali tranzitivnost omrežja v okolici določenega vozlišča;
- vmesna središčnost (angl. betweenness centrality): število najkrajših poti med vsemi pari vozlišč, ki gredo skozi določeno vozlišče.

Posamezno lastnost osnovnega in zmanjšanega omrežja primerjamo z *D*-statistiko Kolmogorov-Smirnov testa, ki nam pove, koliko sta si podobni dve porazdelitvi (manjša kot je vrednost statistike, večja je podobnost med porazdelitvama). Zaradi nedeterminističnosti metod smo za vse metode izvedli za vsako omrežje pet ponovitev ter izračunali povprečje rezultatov. Rezultate analize prikazuje tabela 2, v kateri so s krepko pisavo označene dobro ohranjene lastnosti (vrednost *D*-statistike pod 0,2), na sliki 5 pa so podrobneje prikazane porazdelitve posameznih lastnosti osnovnih omrežij v primerjavi z zmanjšanimi.

Tabela 2: **Rezultati ohranjanja lastnosti pri zmanjševanju treh družbenih omrežij z različnimi algoritmi**

Omrežje	Stopnja	Izhodna stopnja	Vhodna stopnja	Nakopičenost	Vmesna središčnost
Enakomerno naključno izbiranje vozlišč					
Facebook	0,418	0,426	0,399	**0,029**	0,443
DBLP	0,422	0,351	0,624	0,703	0,277
Twitter	0,297	**0,004**	0,790	**0,031**	**0,005**
Naključno izbiranje vozlišč glede na stopnjo					
Facebook	0,414	**0,093**	**0,101**	**0,019**	0,227
DBLP	0,412	**0,066**	**0,187**	0,304	**0,080**
Twitter	0,430	**0,029**	0,245	**0,028**	**0,029**
Preiskovanje v širino					
Facebook	**0,104**	0,230	0,228	**0,018**	0,226
DBLP	**0,096**	**0,094**	**0,171**	**0,092**	**0,136**
Twitter	**0,072**	**0,018**	0,265	**0,031**	**0,016**
Enakomerno naključno izbiranje povezav					
Facebook	0,414	0,137	0,132	0,034	0,224
DBLP	0,426	0,198	0,502	0,767	0,251
Twitter	0,690	0,027	0,068	0,238	0,005
Združevanje glede na skupnosti z izmenjavo oznak					
Facebook	0,451	**0,145**	**0,168**	**0,113**	0,424
DBLP	0,574	0,401	0,219	0,244	**0,154**
Twitter	0,647	0,995	0,807	**0,145**	0,948

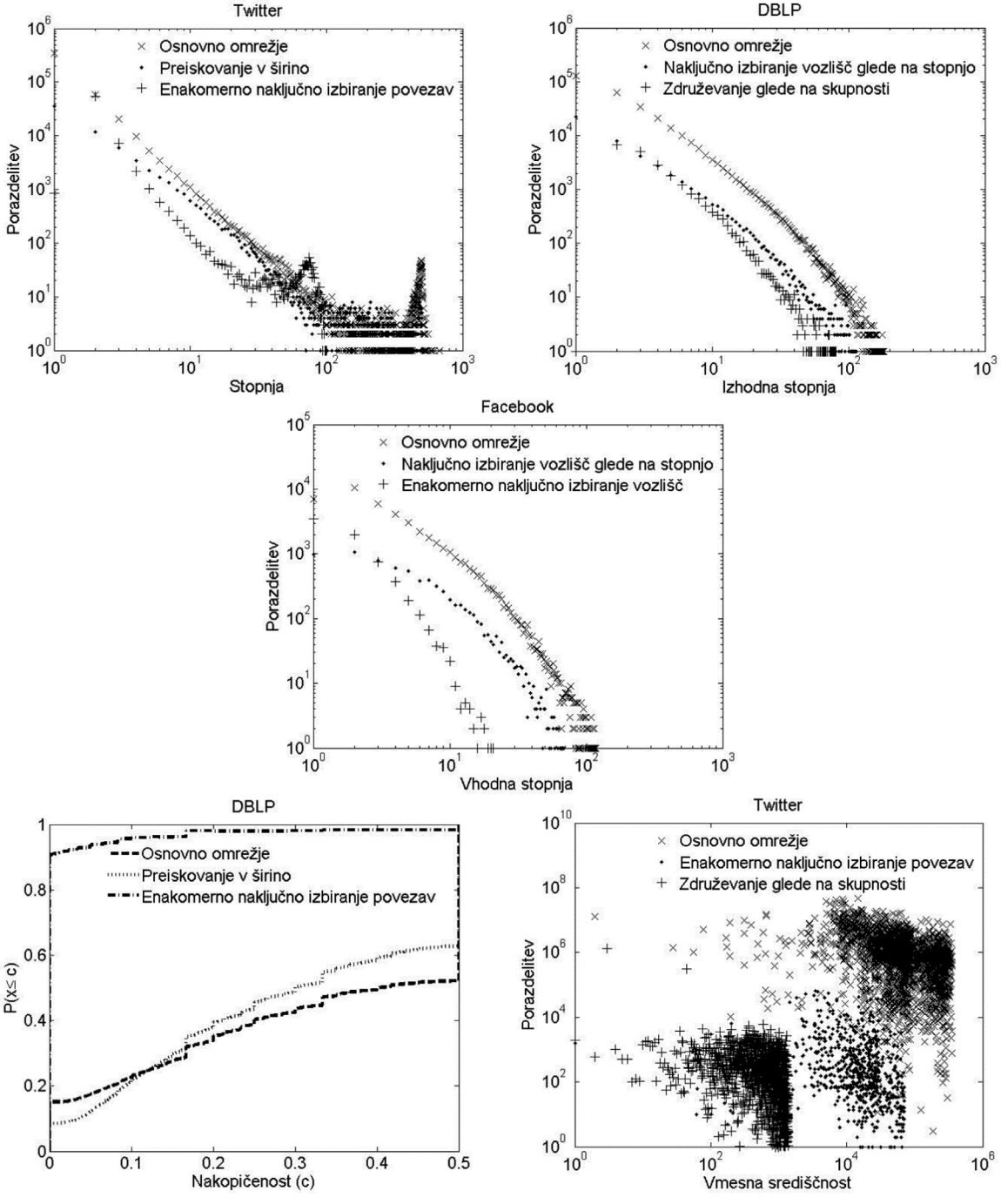
3.3 Razprava

Iz tabele 2 razberemo, da je pri vseh metodah zmanjševanja najbolje ohranjena nakopičenost omrežja. Izbrane metode najslabše ohranijo stopnje vozlišč, po drugi strani pa se izhodne stopnje vozlišč ohranijo dobro pri vseh metodah razen pri združevanju glede na skupnosti.

Posamezne lastnosti se pri različnih metodah ohranijo različno dobro (slika 5). Stopnjo vozlišč najbolje ohrani preiskovanje v širino, isto velja za nakopičenost. Vmesna središčnost ter vhodna in izhodna stopnja se najmanj spremenijo pri naključnem izbiranju vozlišč glede na stopnjo. Omenimo še, da pri opazovanju ohranjanja lastnosti najbolj izstopa stopnja vozlišč, ki je dobro ohranjena le pri preiskovanju v širino. Ta pristop se izkaže za najboljšega tudi po uspešnosti ohranjanja vseh drugih lastnosti (večina vrednosti pod 0,2). Njegova posebnost v primerjavi z drugimi opazovanimi metodami je v tem, da v vzorec ne zajame več nepovezanih delov omrežja, temveč neki naključno izbrani povezani del omrežja. To

očitno igra pomembno vlogo pri uspešnosti ohranjanja lastnosti med zmanjševanjem.

Malo manj uspešni sta metodi naključno izbiranje vozlišč glede na stopnjo in enakomerno naključno izbiranje povezav, medtem ko se enakomerno naključno izbiranje vozlišč in združevanje glede na skupnosti z izmenjavo oznak izkažeta za najmanj uspešni. Zmanjšano omrežje z enakomernim naključnim izbiranjem vozlišč je precej nepovezano, kar očitno slabo vpliva na ohranjanje lastnosti. Metoda



Slika 5: **Primeri ohranjanja lastnosti; prikazane so porazdelitve posameznih lastnosti osnovnega ter zmanjšanih omrežij z najboljšo (druga po vrsti) in najslabšo (tretja po vrsti) metodo (od zgoraj navzdol, od leve proti desni): porazdelitev stopnje vozlišč za omrežje Twitter, porazdelitev izhodnih stopenj vozlišč za omrežje DBLP, porazdelitev vhodnih stopenj vozlišč za omrežje Facebook, kumulativna porazdelitev nakopičenosti za omrežje DBLP ter porazdelitev vmesne središčnosti za omrežje Twitter**

združevanja najbolj med vsemi upošteva strukturo omrežja in je zato tudi časovno zahtevnejša. Prav tako se bolj osredinja na globalno zgradbo omrežja in manj na lokalno, kar pri ohranjanju vseh opazovanih lastnosti pomeni slabost.

Pri (ne)uspešnosti metode združevanja na omrežju Twitter poudarimo, da ima to omrežje velike skupnosti, zato je zmanjšano omrežje majhno (0,5 % velikosti osnovnega omrežja). Posledično so lastnosti slabše ohranjene v primerjavi z drugimi metodami, pri katerih je velikost zmanjšanega 15 odstotkov velikosti osnovnega omrežja.

Glavna uporabna vrednost zmanjševanja omrežij se kaže v tem, da je analiza manjšega omrežja razumljivejša, prav tako pa zanjo porabimo manj časa. Posebno je to lahko uporabno pri analizi velikega omrežja, ki se ne spreminja veliko. Tako omrežje zmanjšamo in ga do naslednje večje spremembe uporabljamo za analizo. To na primer uporabljajo oglaševalci, ki testirajo širjenje novic o novih produktih na zmanjšanem omrežju in učinkovite pristope uporabijo na velikem omrežju (Ebbes, Huang & Rangaswamy, 2012). Podobno velja tudi za analizo drugih dinamičnih procesov (širjenje govoric, virusov in bolezni po omrežju). Simulacije takšnih procesov je preprosteje izvajati na manjših omrežjih, pri čemer je ključnega pomena podobnost med osnovnim in zmanjšanim omrežjem.

4 SKLEP

V zadnjih letih, predvsem z razvojem interneta in različnih družabnih mrež na spletu, sistemi, opisani z omrežji, postajajo vse večji in kompleksnejši. Raziskovalci so zato predlagali številne pristope za zmanjševanje velikih omrežij, ki omogočajo preprostejšo, hitrejšo ter bolj razumljivo analizo. Pomembno pri tem je, da se med zmanjševanjem čim manj spremenijo različne lastnosti omrežja.

V prispevku smo predstavili pristope za zmanjševanje družbenih omrežij. Na treh primerih omrežij iz realnega sveta smo prikazali delovanje različnih pristopov za zmanjševanje in opazovali ohranjanje pomembnih lastnosti omrežij med zmanjševanjem. Rezultati so pokazali, da se med analiziranimi metodami najbolje izkaže preiskovanje v širino, ki dobro ohrani vse opazovane lastnosti. Med lastnostmi se pri večini metod zmanjševanja dobro ohranijo izhodna stopnja, nakopičenost in vmesna središčnost.

Glede na to, da smo se pri raziskavi osredinili le na družbena omrežja, glavno možnost za nadaljnje delo pomeni analiza večjega števila omrežij različnih tipov (poleg družbenih še informacijska, tehnološka, biološka) in velikosti. Razširjena analiza bi omogočila iskanje zakonitosti, ki veljajo pri zmanjševanju omrežij, kot na primer primerjava učinkovitosti metod zmanjševanja v odvisnosti od tipa in velikosti osnovnega omrežja. Rezultati te in nadaljnjih raziskav bodo poleg boljšega razumevanja in hitrejše analize omogočali tudi učinkovitejšo uporabo metod za zmanjševanje omrežij.

VIRI IN LITERATURA

[1] Adler, M. & Mitzenmacher, M. (2001). Towards compressing web graphs. *Proceedings of the Data Compression Conference* (str. 203–212). IEEE.

[2] Barabási, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512.

[3] Batagelj, V. (1997). Notes on blockmodeling. *Social Networks*, 19, 143–155.

[4] Biedl, T., Brejová, B. & Vinar, T. (2000). Simplifying flow networks. *Mathematical Foundations of Computer Science*, 192–201.

[5] Blagus, N., Šubelj, L. & Bajec, M. (2012). Self-similar scaling of density in complex real-world networks. *Physica A*, 391, 2794–2802.

[6] Cohen, R., Erez, K., Ben-Avraham, D. & Havlin, S. (2000). Resilience of the Internet to random breakdowns. *Phys. Rev. Lett.*, 85, 4626.

[7] Cohen, R. & Havlin, S. (2010). *Complex networks: structure, robustness and function*. Cambridge University Press.

[8] De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L. & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media* (str. 34–41).

[9] De Nooy, W., Mrvar, A. & Batagelj, V. (2005). *Exploratory social network analysis with Pajek* (št. 27). Cambridge University Press.

[10] Deo, N. & Litow, B. (1998). A structural approach to graph compression. *Proceedings of the 23th MFCS Workshop on Communications* (str. 91–101). Citeseer.

[11] Doreian, P., Batagelj, V. & Ferligoj, A. (2004). *Generalized blockmodeling* (št. 25). Cambridge University Press.

[12] Ebbes, P., Huang, Z. & Rangaswamy, A. (2012). Subgraph sampling methods for social networks: The good, the bad, and the ugly. *Available at SSRN 1580074*.

[13] Feder, T. & Motwani, R. (1991). Clique partitions, graph compression and speeding-up algorithms. *Proceedings of the 23th annual ACM symposium on Theory of computing* (str. 123–133). ACM.

[14] Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Soc. Networks*, 1, 215–239.

[15] Gilbert, A. C. & Levchenko, K. (2004). Compressing network graphs. *Proceedings of the Link KDD workshop at the 10th ACM Conference on KDD*. Citeseer.

[16] Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99, 7821–7826.

[17] Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. & Arenas, A. (2003). Self-similar community structure in a network of human interactions. *Physical Review E*, 68, 065103.

[18] Hennessey, D., Brooks, D., Fridman, A. & Breen, D. (2008). A simplification algorithm for visualizing the structure of complex graphs. *Information Visualisation 12th International Conference* (str. 616–625). IEEE.

[19] Hübler, C., Kriegel, H. P., Borgwardt, K. & Ghahramani, Z. (2008). Metropolis algorithms for representative subgraph sampling. *Proceedings of the 8th International Conference on Data Mining* (str. 283–292). IEEE.

[20] Illenberger, J. & Flötteröd, G. (2011). *Estimating properties from snowball sampled networks*. VSP Working Paper 11-01, TU Berlin, Transport Systems Planning and Transport Telematics.

[21] Itzkovitz, S., Levitt, R., Kashtan, N., Milo, R., Itzkovitz, M. & Alon, U. (2005). Coarse-graining and self-dissimilarity of complex networks. *Physical Review E*, 71, 016127.

[22] Karypis, G. & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20, 359–392.

[23] Knoke, D., Yang, S. & Kuklinski, J. H. (2008). *Social network analysis* (št. 2). Sage Publications.

[24] Krishnamurthy, V., Faloutsos, M., Chrobak, M., Lao, L., Cui, J.-H. & Percus, A. G. (2005). Reducing large internet topologies for faster simulations. *Proceedings of the 4th International IFIP-TC6 Networking Conference* (str. 328–341). Springer.

[25] Kudelka, M., Horak, Z., Snasel, V. & Abraham, A. (2010). Social Network Reduction Based on Stability. *Computational Aspects of Social Networks* (str. 509–514). IEEE.

[26] Kunegis, J. (2013). KONECT - the Koblenz Network Collection. Retrieved from http://konect.uni-koblenz.de/.

[27] Lee, S. H., Kim, P.-J. & Jeong, H. (2006). Statistical properties of sampled networks. *Physical Review E*, 73, 016102.

[28] Leskovec, J., Kleinberg, J. & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1, 1–40.

[29] Leskovec, J. (2009). Stanford Network Analysis Project. Retrieved from http://snastr.stanford.edu/index.html.

[30] Leskovec, J. & Faloutsos, C. (2006). Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (str. 631–636). ACM.

[31] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2001). Network motifs: Simple building blocks of complex networks. *Science*, 298, 824–827.

[32] Newman, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, 89, 208701.

[33] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45, 167–256.

[34] Newman, Mark E. J. (2010). *Networks: an introduction*. Oxford University Press.

[35] Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab*.

[36] Pržulj, N., Wigle, D. A. & Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics*, 20, 340–348.

[37] Price, D. J. de S. (1965). Networks of scientific papers. *Science*, 149, 510–515.

[38] Rozenfeld, H. D., Gallos, L. K., Song, C. & Makse, H. A. (2008). Fractal and transfractal scale-free networks. *e-print arXiv:08082206v1*.

[39] Song, C., Gallos, L. K., Havlin, S. & Makse, H. A. (2007). How to calculate the fractal dimension of a complex network: The box covering algorithm. *Journal of Statistical Mechanics*, 2007, 03006.

[40] Song, C., Havlin, S. & Makse, H. A. (2005). Self-similarity of complex networks. *Nature*, 433, 392–395.

[41] Stumpf, M. P., Wiuf, C. & May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 4221–4224.

[42] Šubelj, L. & Bajec, M. (2012a). Ubiquitousness of link-density and link-pattern communities in real-world networks. *Eur. Phys. J. B*, 85, 32.

[43] Šubelj, L. & Bajec, M. (2012b). Software systems through complex networks science: Review, analysis and applications. *Proceedings of the KDD Workshop on Software Mining* (str. 9–16). Beijing, China.

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

[45] Tian, Y., Hankins, R. A. & Patel, J. M. (2008). Efficient aggregation for graph summarization. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (str. 567–580). ACM.

[46] Toivonen, H., Mahler, S. & Zhou, F. (2010). A framework for path-oriented network simplification. *Advances in Intelligent Data Analysis IX*, 220–231.

[47] Viswanath, B., Mislove, A., Cha, M. & Gummadi, K. P. (2009). On the evolution of user interaction in facebook. *Proceedings of the 2nd ACM workshop on Online social networks* (str. 37–42). ACM.

[48] Walker, D., Xie, H., Yan, K.-K. & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *J. Stat. Mech.*, 2007, P06010.

[49] Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393, 440–442.

[50] Yang, J. & Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics* (str. 3). ACM.

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

[52] Zhou, W. X., Jiang, Z. Q. & Sornette, D. (2007). Exploring self-similarity of complex cellular networks: The edge-covering method with simulated annealing and log-periodic sampling. *Physica A*, 375, 741–752.

[53] Zhou, Y., Cheng, H. & Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2, 718–729.

■

Neli Blagus je mlada raziskovalka v Laboratoriju za podatkovne tehnologije na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Raziskovalno se ukvarja z analizo omrežij.

■

Lovro Šubelj je asistent na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Poučuje predvsem predmete s področja podatkovnih baz. Raziskovalno se ukvarja z analizo realnih omrežij, natančneje z odkrivanjem značilnih skupin vozlišč v velikih kompleksnih omrežjih. Je avtor ali soavtor številnih prispevkov v strokovnih in znanstvenih publikacijah.

■

Aljaž Zrnec je magistriral leta 2002 na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Leta 2006 je doktoriral s področja konstruiranja metodologij. Zaposlen je v Laboratoriju za podatkovne tehnologije kot asistent za področje podatkovnih baz. Na raziskovalnem področju se ukvarja s konstruiranjem metodologij, podatkovnimi bazami NoSQL in računalništvom v oblaku. Je avtor ali soavtor številnih prispevkov v strokovnih in znanstvenih publikacijah.

■

Marko Janković je mladi raziskovalec v Laboratoriju za podatkovne tehnologije na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Njegova glavna raziskovalna področja obsegajo dogodkovno vodene arhitekture, procesiranje in rudarjenje po podatkovnih tokovih ter internet stvari.

■

Marko Bajec je izredni profesor na Fakulteti za računalništvo in informatiko Univerze v Ljubljani, kjer poučuje dodiplomske in podiplomske predmete s področja razvoja informacijskih sistemov in podatkovnih baz. Raziskovalno se ukvarja z metodami in pristopi k snovanju in razvoju informacijskih sistemov, obvladovanjem informatike ter v zadnjih letih predvsem s podatkovnimi tehnologijami za predstavitev, analizo in vizualizacijo podatkov. Leta 2009 je ustanovil Laboratorij za podatkovne tehnologije ter prevzel njegovo vodenje. Je član številnih domačih in tujih združenj, komisij in odborov. V okviru fakultete je vodil več aplikativnih in raziskovalnih projektov. Svoje raziskovalne rezultate in dosežke iz prakse redno objavlja v domačih in mednarodnih znanstvenih in strokovnih krogih.