



Razvoj algoritmov za analizo omrežij v velikem podjetju

Poročilo operacije Raziskovalci na začetku kariere 2014-15

Lovro Šubelj

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana

lovro.subelj@fri.uni-lj.si

Dokument predstavlja zaključno poročilo operacije Raziskovalci na začetku kariere 2014-15 med raziskovalno organizacijo Univerza v Ljubljani, Fakulteto za računalništvo in informatiko, in podjetjem Petrol d.d., Ljubljana. V nadaljevanju si v zaporednih razdelkih sledijo namen in cilji operacije, opis izvedenih aktivnosti in faz operacije, predstavitev rezultatov in ocena operacije, izbrani primeri uporabe ter zaključki operacije.

Namen in cilji operacije

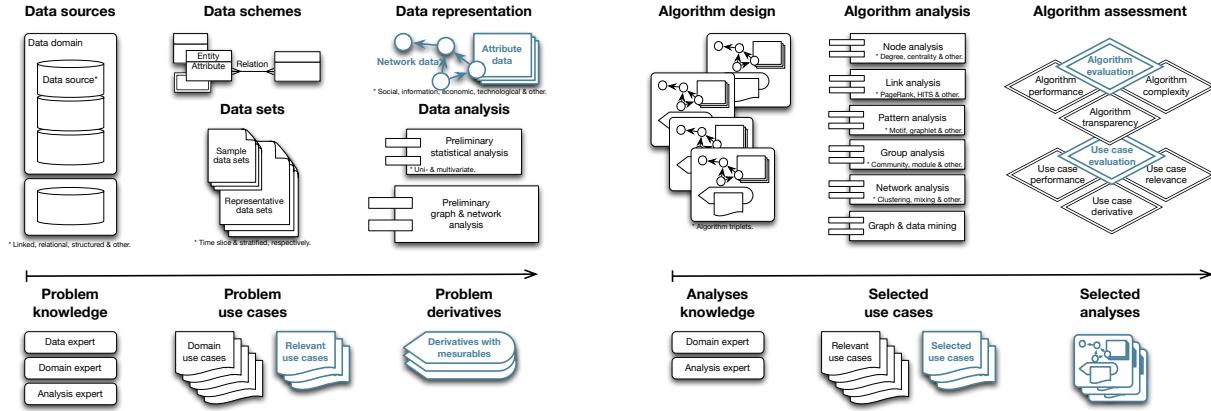
Odnosi med posamezniki, pripadnost strank, svetovni splet, finančne transakcije in energetska infrastruktura so vse primeri kompleksnih omrežij, sestavljenih iz velikega števila med seboj povezanih entitet. Pri tem pa je v različnih realnih omrežjih moč opaziti karakteristične vzorce povezovanja, ki jih značilno ločijo od regularnega ali naključnega sveta. S proučevanjem zgradbe velikih realnih omrežij se ukvarja omrežna znanost (angl. *network science*), ki je trenutno izjemno aktivno znanstveno področje ter temelj številnih uspešnih podjetij. Na drugi strani pa je uporaba omrežne znanosti v Sloveniji omejena predvsem na raziskovalno sfero, dočim je prenos znanja v gospodarstvo manj izrazit. Slednje je vsaj delno posledica dejstva, da je večina razvitih algoritmov namenjena analizi izbranih značilnosti določene vrste omrežij ter ne rešuje dejanskih izzivov podjetij v Sloveniji in ni neposredno uporabna nad podatkovnimi viri, ki so navadno na voljo. Pristopi tako ne odgovarjajo na realne potrebe gospodarstva ter so posledično manj uporabni pri procesih odločanja.

Cilj operacije je bil razvoj algoritmov za analizo omrežij (angl. *network analysis*) zgrajenih na podlagi podatkovnih virov dostopnih podjetjem, z namenom podpore dejanskih primerov uporabe v velikih podjetjih v Sloveniji. Zastavljeni cilji operacije so obsegali pregled dostopnih podatkovnih virov v velikem podjetju, preizkus in oceno obstoječih algoritmov analize omrežij za podporo relevantnih primerov uporabe, razvoj prilagojenih algoritmov za podporo izbranih primerov uporabe ter evalvacijo razvitih algoritmov na izbranih primerih uporabe v velikem podjetju. Namen operacije je bil tako razvoj temeljnega znanja na področju omrežne znanosti, dočim pa je ustvarjeno znanje dejanski odraz potreb gospodarstva v Sloveniji. Posledično je ključen cilj operacije predstavljal objava ustvarjenih rezultatov in znanj v priznanih mednarodnih znanstvenih revijah ter predstavitev na pomembnejših domačih in mednarodnih znanstvenih konferencah.

Aktivnosti in faze operacije

Operacija je potekala v štirih zaporednih fazah skladno z zastavljenim terminskim planom. Aktivnosti prvih dveh faz so predstavljene na Sliki 1, dočim so aktivnosti preostalih dveh faz opisane v nadaljevanju.

Razvoj algoritmov za analizo omrežij v velikem podjetju



Slika 1. Pregled aktivnosti in faz izvajanja operacije. Časovni diagram aktivnosti prve (levo) in druge (desno) faze izvajanja operacije (za opis glej tekst).

Prva faza operacije je vključevala pregled dostopnih podatkovnih virov sodelujočega podjetja ter pridobivanje podatkovnih naborov iz izbranih podatkovnih virov (glej Sliko 1, levo). Le-te so služili za gradnjo omrežij uporabljenih v nadaljevanju operacije. Med izbrane podatkovne vire sodijo zgodovina poslovanja s strankami in poslovnimi partnerji podjetja, demografski in drugi podatki o strankah in poslovnih partnerjih, navodila in pravilniki, ki urejajo poslovanje različnih enot podjetja, celoten nabor spletnih mest podjetja, popis cestne, električne in plinovodne infrastrukture potrebne za delovanje podjetja, prosto dostopne programske knjižnice primerne za izvedbo operacije ter nekateri drugi podporni viri.

Na podlagi preliminarne statistične in omrežne analize izbranih podatkovnih virov so bili v sodelovanju z domenskimi in podatkovnimi eksperti podjetja izbrani relevantni primeri uporabe. Med slednje sodijo napovedovanje kreditnih tveganj oziroma dolgov poslovnih partnerjev z uporabo ekonomskih omrežij (angl. *economic network*) zgrajenih na podlagi denarnega toka, trgovanj ali pripadnosti partnerjev, analiza družbenih omrežij (angl. *social network*) zgrajenih na podlagi odnosov in vzorcev obnašanja strank za namene direktnega oziroma virusnega trženja, ocena delovanja podjetja preko informacijskih omrežij (angl. *information network*) zgrajenih na podlagi sklicevanj med internimi dokumenti, optimizacija spletiča podjetja na osnovi informacijskih omrežij zgrajenih na podlagi povezav med spletnimi mesti, učinkovito napovedovanje v tehnoloških omrežjih (angl. *technological network*) zgrajenih na podlagi umetne podporne infrastrukture podjetja, odpravljanje napak v internih informacijsko-komunikacijskih storitvah in drugih programskih rešitvah z uporabo informacijsko-tehnoloških omrežij ter nekateri drugi podporni primeri uporabe.

Druga faza operacije je vključevala pregled in oceno obstoječih algoritmov in pristopov analize omrežij za reševanje izbranih primerov uporabe sodelujočega podjetja (glej Sliko 1, desno). V ta namen so bili najprej obravnavani pristopi za določanje pomembnosti vozlišč (tj. entitet) in povezav v omrežju. Sem sodijo različne stopnje (angl. *degree*), dostopna in vmesna središčnost (angl. *closeness, betweenness centrality*), različne mere nakopičenosti (angl. *clustering*), lastni vektorji (angl. *eigenvector*) ter pristopi analize povezav kot sta algoritma PageRank in HITS. Poleg tega so bili preizkušeni algoritmi za odkrivanje manjših pogostih vzorcev vozlišč kot so motivi (angl. *motif*) in grafki (angl. *graphlet*). Nadalje so bili obravnavani pristopi za odkrivanje večjih karakterističnih skupin vozlišč kot so skupnosti (angl. *community*), moduli (angl. *module*) in mešanice (angl. *mixture*). Preizkušeni so bili različni algoritmi spektralne analize (angl. *spectrum*), optimizacije modularnosti (angl. *modularity*), stiskanja omrežij (angl. *map equation*), izmenjave oznak (angl. *label propagation*), gručenja povezav (angl. *link clustering*), statističnega sklepanja (angl. *inference*), določanja sredic (angl. *core*) ter nekateri drugi pristopi. Nazadnje pa so bili obravnavani še pristopi abstrakcije velikih omrežij kot je metuljčna zgradba (angl. *bow-tie*), odpornost (angl. *robustness*), porazdelitve stopenj in

nakopičenosti, odvisnosti oziroma mešanja med vozlišči (angl. *mixing*) ter premeri (angl. *diameter*). Slednje imajo močan vpliv na različne stohastične in epidemične procese nad omrežji (angl. *dynamics*). Za podporo omenjenim pristopom so bili preizkušeni modeli gradnje velikih omrežij kot je prednostna povezanost (angl. *preferential attachment*) oziroma brezlestvični model (angl. *scale-free*), prevezovanje povezav (angl. *rewiring*) oziroma mali svet (angl. *small-world*), model kopiranj (angl. *copying*), naključni sprehodi (angl. *random walk*), goreči gozd (angl. *forest fire*) ter nekateri drugi. Podobno so bili preizkušeni tudi modeli zmanjševanje ali vzorčenja velikih omrežij kot je naključno ali prednostno izbiranje (angl. *random, preferential selection*), preiskovanje v širino oziroma nenadzorovano širjenje (angl. *snowball sampling*), preiskovanje v globino oziroma sledenje kontaktov (angl. *contact tracing*), naključni sprehodi ter združevanje skupnosti ali particij (angl. *partition*).

Posamezni algoritmi in pristopi so navadno smiseln zgolj v določenih vrstah omrežij, dočim pa rezultati kažejo, da je s pristopi analize omrežij moč učinkovito reševati večino izmed izbranih primerov uporabe. Kljub temu pa obstajajo še številne pomanjkljivosti oziroma možnosti za izboljšave obstoječih pristopov. Ocena primernosti posameznih pristopov za različne vrste omrežij je podana v naslednjem razdelku, ocena primernosti za izbrane primere uporabe pa v nadaljevanju poročila.

Tretja faza operacije je vključevala razvoj prilagojenih oziroma izboljšanih algoritmov in pristopov analize omrežij za podporo izbranih primerov uporabe sodelujočega podjetja. Sočasno je tako potekal razvoj različnih novih pristopov, ki odpravljam posamezne pomanjkljivosti obstoječih. Delo je obsegalo predvsem razvoj in prototipno implementacijo algoritmov, primerjava z obstoječimi pristopi analize omrežij, osnovnimi statističnimi pristopi, standardnimi metodami strojnega učenja (angl. *machine learning*) in podatkovnega rudarjenja (angl. *data mining*) ter preliminarno evalvacijo pristopov na izbranih primerih uporabe. Omenjene prototipne implementacije so bile namenjene izključno testiranju in evalvaciji ter jih kot take ni (bilo) moč tržiti kot produkt podjetja ali podobno. Skladno s ključnim ciljem operacije so bili vsi razviti algoritmi in pristopi objavljeni v priznanih znanstvenih revijah oziroma predstavljeni na mednarodnih znanstvenih konferencah. Predstavitev razvitih pristopov s pripadajočimi objavami je podana v naslednjem razdelku, dočim podrobna evalvacija pristopov na izbranih primerih uporabe sledi v nadaljevanju poročila.

Četrta faza operacije je vključevala testiranje in evalvacijo razvitih algoritmov in pristopov nad izbranimi primeri uporabe sodelujočega podjetja. Delo je potekalo skladno z razvojem pristopov v prejšnji fazi, dočim so podrobni rezultati evalvacije ter pa primerjava z obstoječimi pristopi podani v nadaljevanju poročila. Rezultati sicer kažejo, da kljub relativni učinkovitosti obstoječih pristopov, novo razviti pristopi predstavljajo občutno izboljšanje v različnih primerih uporabe. Podobno kot zgoraj so bili skladno s ključnim ciljem operacije vsi pomembnejši rezultati in znanja objavljeni v priznanih znanstvenih revijah oziroma predstavljeni na mednarodnih znanstvenih konferencah.

Vse štiri faze operacije so potekale skladno z zastavljenimi časovnimi okvirji. Tako je prva faza potekala med januarjem in marcem, druga faza pa med aprilom in septembrom leta 2014. Tretja faza je potekala med oktobrom leta 2014 in marcem leta 2015, četrta faza pa med aprilom in junijem leta 2015.

Rezultati in ocena operacije

Rezultati dela na operaciji v prvi vrsti kažejo, da različne vrste omrežij niso enako zanesljive in dostopne podjetjem. Kot najzanesljivejša se izkažejo tehnološka omrežja kot so cestna (angl. *road network*), plinovodna (angl. *pipeline*) in električna omrežja (angl. *electric grid*), pri čimer so potrebni podatkovni viri večinoma tudi prosti dostopni. Podobno velja za informacijska omrežja kot so svetovni splet (angl. *web graph*), dokumentna omrežja sklicevanj (angl. *document, citation network*) ter programska omrežja (angl. *software network*). Le-ta je moč avtomatsko zgraditi na podlagi internih podatkovnih virov dostopnih podjetjem. Vzorce obnašanja strank ter poslovnih partnerjev podjetja je moč preučevati preko korelačijskih tokovnih omrežij (angl. *correlation network*) ter omrežij sodelovanj in pripadnosti (angl. *affiliation, collaboration network*). Na drugi strani pa standardna spletarna družbena omrežja odnosov med strankami (angl. *online social network*) in ekonomska omrežja trgovanj med partnerji (angl. *trade network*) niso neposredno dostopna podjetjem.

Dočim je na primer moč pridobiti del spletne družabne omrežja preko anket in podobnih inštrumentov, pa je zgradba takih omrežij izjemno nezanesljiva, pridobivanje pa neprimerljivo drag.

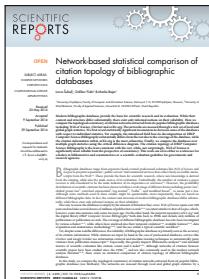
Ocena primernosti obstoječih pristopov analize omrežij za reševanje izbranih primerov uporabe sodelujočega podjetja kaže, da so različni pristopi uspešni v različnih omrežjih. Za določanje pomembnosti posameznih vozlišč se v primeru redkejših tehnoloških in informacijsko-tehnoloških omrežij izkažejo mere središčnosti, dočim so v gostejših informacijskih, družbenih in ekonomskih omrežjih primernejše različne stopnje in lastni vektorji. V obeh primerih obstajajo še različne možnosti za izboljšave. Pri analizi pogostih vzorcev vozlišč so uspešni pristopi na osnovi grafkov, dočim pa motivi zaradi inducirane zgradbe niso primerni. Grafki se izkažejo predvsem v primeru tehnoloških in nekaterih korelacijskih omrežij. Goste skupine vozlišč ozziroma skupnosti so prisotne v večini omrežij, za njihovo odkrivanje pa so najprimernejše metode stiskanja omrežij, optimizacije modularnosti in pa spektralne analize v primeru, da je število skupin znano. Kljub vsemu pa se omenjeni pristopi osredotočajo zgolj na skupnosti ter zanemarjajo druge karakteristične skupine vozlišč kot so na primer moduli in mešanice. Za namene abstrakcije velikih omrežij so v primeru tehnoloških in informacijskih omrežij najprimernejše različne dekompozicije in mešanja, dočim imajo v družbenih in ekonomskih omrežjih večji pomen porazdelitve stopenj in nakopičenosti. Predvsem v prvem primeru pa obstoječi pristopi ne zagotavljajo zadovoljivega vpogleda v zgradbo omrežja.

Zgradbo družbenih in ekonomskih omrežij sicer najnatančneje simulirajo model kopiranj, naključni sprehodi in goreči gozd. Nasprotno pa obstoječi modeli niso primerni za gradnjo usmerjenih informacijskih in tehnoloških omrežij. Pri vzorčenju omrežij se še najbolj izkaže sledenje kontaktov, primerljivo natančnost pa dosežeta tudi nenadzorovano širjenje in naključni sprehodi. Kljub temu pa noben od pristopov dejansko ne daje zadovoljivih rezultatov v različnih omrežjih, dočim obstaja tudi veliko pomanjkanje metod za primerjavo in ocenjevanje zanesljivosti tako dobljenih omrežij. Za razliko od vsega zgoraj je napovedovanje povezav razmeroma neodvisno od vrste omrežja, posebej uspešni pa so pristopi na podlagi stopenj, sosedov, grafkov in skupnosti. Nenazadnje pa se napovedovanje lastnosti vozlišč izkaže kot pomemben pristop za reševanje izbranih primerov uporabe, ki še pa ni zadovoljivo rešen.

Kot odziv na zgoraj izpostavljene pomanjkljivosti obstoječih algoritmov in pristopov analize omrežij, so bili tekom izvajanja operacije razviti številni novi pristopi (glej Sliko 2). Predlagana je bila nadgradnja obstoječih pristopov določanja pomembnosti vozlišč na osnovi stopenj in dinamičnih procesov [8] (angl. *dynamic process*). Poleg tega so bili razviti izboljšani pristopi za odkrivanje skupin vozlišč na osnovi ekstremne optimizacije [2] (angl. *extremal optimisation*) in hierarhične izmenjave oznak [3, 14] (angl. *hierarchical propagation*), ki so za razliko od vseh ostalih pristopov primerni za odkrivanje splošnih skupin brez vsakršnega predznanja. Za namene abstrakcije informacijskih in tehnoloških omrežij so bili predstavljeni novi pristopi dekompozicije stopenj [1] (angl. *degree bow-tie*) in pa mešanja nakopičenosti [2] (angl. *clustering mixing*). Razvit je bil nov model gradnje informacijskih in družbenih omrežij na osnovi dinamike citiranj [11] (angl. *citation dynamics*), ki simulira zgradbo omrežij bolj natančno kot vsi obstoječi modeli. Nadalje so bili predlagani izboljšani pristopi vzorčenja omrežij na osnovi induciranih podgrafov [10, 13] (angl. *induced subgraph*) ter novi statistični pristopi za primerjavo vzorčenih omrežij [4] in ocenjevanje zanesljivosti posameznih omrežij [6, 17]. Pri tem je bila pojasnjena tudi evolucija skupin vozlišč med vzorčenjem omrežja [7] ter razvoj časovnih informacijskih omrežij [9]. Za namene analize ekonomskih omrežij so bili predlagani novi pristopi tokovnih omrežij [12] kot so graf vidljivosti (angl. *visibility graph*), korelacijska omrežja in omrežja prehodov (angl. *transition network*) ter razvit eksperimentni sistem za ruderjanje podatkovnih tokov [5, 15] (angl. *stream mining*). Poleg vsega omenjenega pa je bila razvita tudi metodologija napovedovanja lastnosti vozlišč na osnovi sosesčin [18] (angl. *neighborhood*) in skupnosti vozlišč [16], ki podpira izjemno širok spekter primerov uporabe.

Skladno s ključnim ciljem operacije so bili razviti algoritmi in pristopi najprej objavljeni v osmih originalnih znanstvenih prispevkih v uglednih mednarodnih revijah, pri čimer sta še dva prispevka v procesu recenzije ozziroma oddaje (glej Sliko 2, panel A). Vse omenjene revije so vključene v indeks SCI ter vključujejo tudi prosto dostopne revije kot sta *Scientific Reports* in *PLoS ONE*. Poudarimo, da je zaradi pogojev varovanja osebnih podatkov sodelujočega podjetja, evalvacija razvitih pristopov nad izbranimi primeri uporabe podana v tem poročilu, dočim pa so bili v objavljenih prispevkih pristopi preizkušeni nad primerljivimi prosto

Razvoj algoritmov za analizo omrežij v velikem podjetju



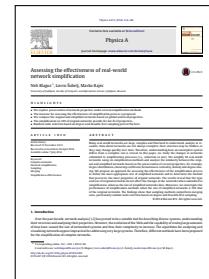
[1] *Sci. Rep.* (2014)



[2] *Comp. Sys.* (2014)



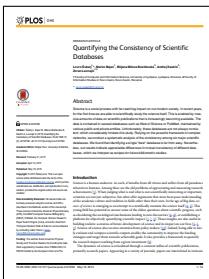
[3] *Physica A* (2014)



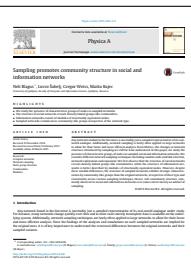
[4] *Physica A* (2014)



[5] *PLoS ONE* (2014)



[6] *PLoS ONE* (2015)



[7] *Physica A* (2015)



[8] *J. Informetr.* (2015)

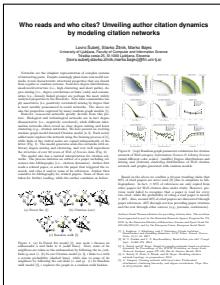


[9] *JASIST* (recenzija)



[10] *arXiv* (oddaja)

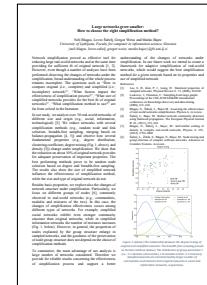
A Mednarodne revije



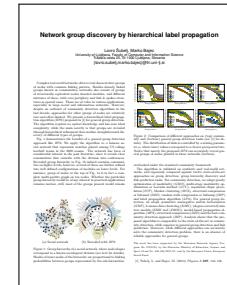
[11] *NetSci '14*



[12] *NetSci '14*



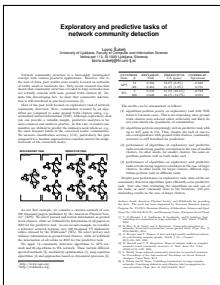
[13] *NetSci '14*



[14] *EUSN '14*



[15] *CAiSE '14*



[16] *NetSci '15*



[17] *NetSci '15*



[18] *ARS '15*



Mreženje slovenskih netvorkašev *MSN* '15

B Konference in dogodki

Slika 2. Znanstveni prispevki objavljeni tekom izvajanja operacije. Panel A prikazuje prispevke objavljene v mednarodnih znanstvenih revijah, panel B pa prispevke predstavljene na mednarodnih znanstvenih konferencah ter izvedene dogodke (za opis glej tekst).

dostopnimi podatkovnimi viri. Nadalje so bili razviti pristopi predstavljeni v osmih prispevkih na priznanih mednarodnih konferencah kot so International Conference on Network Science *NetSci '14/15*, European Conference on Social Networks *EUSN '14* in International Workshop on Social Network Analysis *ARS '15* (glej Sliko 2, panel **B**). Slednje trenutno veljajo za najprestižnejše konference na področju analize omrežij in so potekale na Berkleyu, Kalifornija, na Capriju, Italija, ter v Barceloni in Zaragozi, Španija. Rezultati operacije so bili predstavljeni tudi v okviru seminarjev na različnih članicah Univerze v Ljubljani kot sta Fakulteta za računalništvo in informatiko ter Fakulteta za matematiko in fiziko, na Univerzi v Novem mestu ter na University of West Bohemia v Plznu, Češka. Nenazadnje pa je bilo delo predstavljeno tudi v okviru prvega Mreženja slovenskih netvorkašev *MSN '15*, ki ga je soorganiziral raziskovalec na operaciji.

Algoritmi in pristopi razviti tekom izvajanja operacije predstavljajo občutno izboljšanje obstoječih pristopov na izbranih primerih uporabe, pri čimer je podrobna evalvacija podana v nadaljevanju poročila. Rezultati operacije se tako direktno skladajo z zastavljenimi cilji, zatorej jih ocenujemo kot uspešne.

Odstopanja in skladnost operacije

Operacija je potekala skladno z zastavljenim terminskim planom, dočim tekom izvajanja operacije ni prišlo do nikakršnih odstopanj od zastavljenih ciljev oziroma do tveganj za uspešen zaključek operacije.

Operacija je potekala skladno z namenom, cilji in predmetom javnega razpisa ter širšimi cilji trajnostnega razvoja. V obdobju izvajanja operacije se je okreplila povezava in izmenjava znanja med različnimi sektorji sodelujočega podjetja, podizvajalci podjetja ter sodelujočo raziskovalno organizacijo. Operacija je vzbudila veliko zanimanja, kar je pripomoglo k prenosu znanja v obe smeri. Izvajanje operacije je tako omogočilo prenos analitičnega znanja iz raziskovalne organizacije v podjetje ter pa domenskega znanja iz podjetja v raziskovalno organizacijo, kar bo pripomoglo k trajnostnemu razvoju obeh entitet. Poleg tega so bili rezultati operacije predstavljeni tudi na različnih mednarodnih znanstvenih konferencah ter objavljeni v več prispevkih v priznanih znanstvenih revijah iz področja analize omrežij (glej Sliko 2). Le-to bo pripomoglo k razpoznavnosti in trajnostnemu razvoju sodelujoče raziskovalne organizacije ter pa slovenske znanosti nasploh. Kjerkoli mogoče so bili prispevki objavljeni v prosto dostopnih znanstvenih revijah, kar bo pripomoglo k še lažji diseminaciji pridobljenega znanja ter posledično trajnostnemu razvoju širše družbe. Poleg vsega omenjenega pa so bili rezultati operacije predstavljeni tudi v okviru domačih in mednarodnih seminarjev ter prvega Mreženja slovenskih netvorkašev *MSN '15*, ki ga je soorganiziral raziskovalec na operaciji. Dogodek je zbral skoraj petdeset slovenskih raziskovalcev s širšega področja omrežne znanosti, kar je že pripomoglo k dvigu sodelovanja in izmenjavi med slovenskimi raziskovalnimi organizacijami.

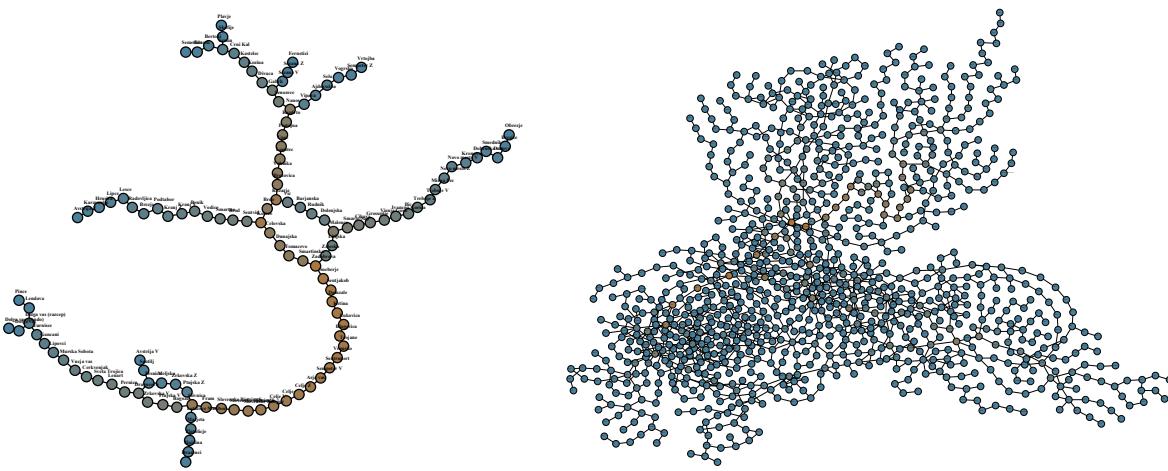
V okviru operacije je bila zagotovljena enakost možnosti ter preprečena vsakršna diskriminacija med predstavniki sodelujočega podjetja in podizvajalcev oziroma sodelujoče raziskovalne organizacije. Pri izvajaju operacije so bili upoštevani pogoji zaščite in varovanja osebnih podatkov, kot jih narekuje zakonodaja.

Primeri uporabe operacije

V nadaljevanju je podana evalvacija algoritmov in pristopov analize omrežij za reševanje izbranih primerov uporabe v sodelujočem podjetju. V zaporednih razdelkih so podani primeri uporabe nad tehnološkimi, informacijskimi, družbenimi in ekonomskimi omrežji, dočim pa je posamezne primere uporabe moč direktno prenesti na druga omrežja enake vrste. Zaradi enostavnosti najuspešnejši obstoječi pristopi analize omrežij ter pa novi pristopi razviti tekom izvajanja operacije niso obravnavani ločeno.

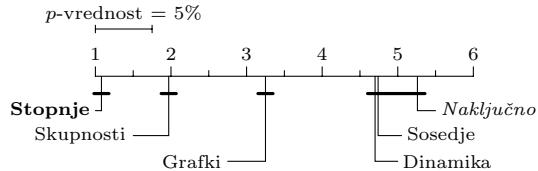
Tehnološka omrežja

Tehnološka omrežja kot sta cestno in plinovodno omrežje navadno predstavljajo neko umetno infrastrukturo, pri čimer so vozlišča omrežja podrejena določenim tehnološkim omejitvam (npr. število cest v križišču).



A Slovenske in evropske avtoceste

	Korelacija ρ_S	Točnost AUC
Stopnje	0.355	70.4%
Skupnosti	0.327	64.3%
Graffi	0.120	56.8%
Dinamika	0.136	52.2%
Sosedje	0.136	52.2%
<i>Naključno</i>	<i>0.055</i>	<i>49.7%</i>

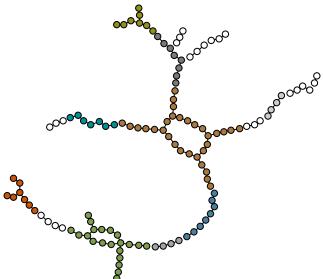


B Napovedovanje odsekov

	Korelacija ρ	Korelacija ρ_S
Središčnosti	0.682	0.769
Skupnosti	0.175	0.345
Stopnje	0.308	0.287
<i>Naključno</i>	<i>0.092</i>	<i>0.117</i>
Graffi	0.030	0.077
Sosedje	0.000	0.000

	Korelacija ρ	Korelacija ρ_S
Središčnosti	0.708	0.772
Stopnje	0.308	0.359
Vektorji	0.010	0.308
Povezave	0.133	0.131
<i>Naključno</i>	<i>0.084</i>	<i>0.092</i>
Nakopičenost	0.000	0.000

C Napovedovanje obremenitev



	Ujemanje NMI	Točnost CA
Sosedje	-	86.9%
Izmenjava	0.775	82.9%
Stiskanje	0.763	65.1%
Spekter	0.757	63.6%
Modularnost	0.736	60.5%
Povezave	0.692	12.1%

D Napovedovanje postajališč

Slika 3. Slovensko in evropsko avtocestno omrežje. Panel A prikazuje obe avtocestni omrežji, panel B točnost napovedovanja manjkajočih cestnih odsekov, panel C natančnost napovedovanja prometnih obremenitev ter panel D točnost napovedovanja postajališč sodelujočega podjetja (za opis glej tekst).

Cestna omrežja

Cestno omrežje predstavlja pomembno infrastrukturo za delovanje sodelujočega podjetja. Slika 3, panel **A** prikazuje slovensko in evropsko avtocestno omrežje iz leta 2012, kjer barve vozlišč ustrezajo povprečnim dnevnim prometnim obremenitvam. Cestna omrežja so pogosto pomanjkljiva zato je smiselno napovedovanje manjkajočih cestnih odsekov z uporabo pristopov napovedovanja povezav. Slika 3, panel **B** prikazuje točnost napovedovanja za evropsko avtocestno omrežje in pa kritični diagram razlik za izbrane pristope. Najbolje se izkažejo inverzne stopnje vozlišč, ki omogočajo napovedovanje z več kot 70% točnostjo, pri čimer pa je pristop statistično značilno boljši od ostalih pri stopnji tveganja 5%. Zadovoljive rezultate dosegajo še pristopi na osnovi skupnosti in grafkov, dočim pa so ostali pristopi primerljivi z naključnim napovedovanjem.

Dnevne prometne obremenitve predstavljajo pomemben vir informacij za optimizacijo poslovanja ter tudi delovanja sodelujočega podjetja. Na drugi strani pa je avtomatsko štetje prometa v Sloveniji realizirano le deloma in razmeroma drago. Slika 3, panel **C** prikazuje natančnost napovedovanja prometnih obremenitev za posamezne cestne odseke (levo) ter pa vozlišča slovenskega avtocestnega omrežja (desno). Mere središčnosti dosežejo izjemno visoko stopnjo korelacije preko 0.75, delno zadovoljive rezultate pa kažejo še pristopi na osnovi stopenj in skupnosti vozlišč. Podobno kot prej so ostali pristopi primerljivi z naključnim napovedovanjem.

Različna obcestna in druga postajališča predstavljajo razbitje cestnega omrežja. Slika 3, panel **D** prikazuje razbitje slovenskega avtocestnega omrežja glede na bencinska postajališča sodelujočega podjetja (levo), ki jih je moč napovedovati z uporabo pristopov odkrivanja skupnosti (desno). Najboljše ujemanje kažejo pristopi izmenjave oznak, ki dosežejo 83% točnost napovedovanja bencinskih postajališč, dočim pristop na osnovi soseščin doseže 87% točnost. Delno zadovoljive rezultate kažejo še pristopi stiskanja omrežij, spektralne analize in optimizacije modularnosti.

Plinovodna omrežja

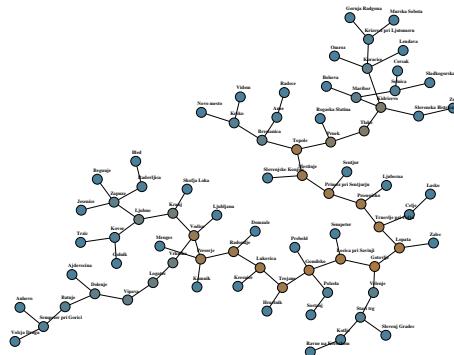
Podobno kot zgoraj tudi plinovodno omrežje predstavlja pomembno distribucijsko infrastrukturo sodelujočega podjetja. Slika 4, panel **A** prikazuje hrbtenico slovenskega plinovodnega omrežja iz leta 2015, kjer barve vozlišč ustrezajo ocjenjenim povprečnim dnevnim obremenitvam. Slika 4, panel **B** prikazuje točnost napovedovanja manjkajočih plinovodov in pa kritični diagram razlik za izbrane pristope. Najbolje se izkažejo pristopi na osnovi grafkov, ki dosežejo več kot 70% točnost, zadovoljive rezultate pa kaže še pristop na osnovi inverznih stopenj vozlišč. Oba omenjena pristopa sta statistično značilno boljša od ostalih pri stopnji tveganja 5%, dočim so slednji primerljivi z naključnim napovedovanjem.

Informacijska omrežja

V informacijskih omrežjih kot so svetovni splet ter omrežja sklicevanj med dokumentni in programi vozlišča predstavljajo delčke informacije (npr. spletna mesta), usmerjene povezave pa ustrezajo toku informacij skozi preučevan sistem.

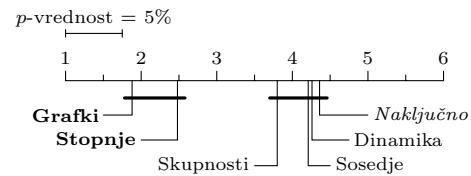
Dokumentna omrežja

Slika 5, panel **A** prikazuje omrežje sklicevanj med prosto dostopnimi navodili in pravilniki delovanja sodelujočega podjetja iz leta 2015, kjer barve vozlišč označujejo vrsto dokumenta. Dinamika avtorjev dokumentov ima neposreden vpliv na zgradbo takih omrežij kar je moč preučevati s pomočjo modelov gradnje omrežij. V okviru izvajanja operacije je bila tako potrjena 15 let stara teza, da avtorji preberejo zgolj okrog 20% dokumentov na katere se sklicujejo. Slika 5, panel **B** prikazuje natančnost ujemanja večjega dokumentnega omrežja in pa kritični diagram razlik za izbrane modele gradnje omrežij. Daleč najboljše ujemanje kaže model citiranj, ki je statistično značilno boljši od modela kopiranj pri stopnji tveganja 5%. Delno zadovoljivo natančnost dosežeta še model naključnih sprehodov in goreči gozd, dočim pa razlike med ostalimi modeli niso statistično značilne.



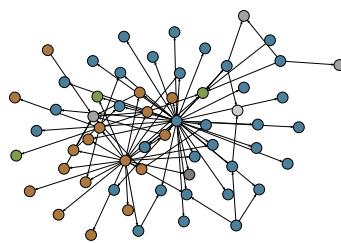
A Slovenski plinovod

	Korelacija ρ_S	Točnost AUC
Graffi	0.386	71.5%
Stopnje	0.328	66.4%
Skupnosti	0.203	54.2%
Sosedje	0.000	51.7%
Dinamika	0.000	51.7%
Naključno	0.222	48.8%



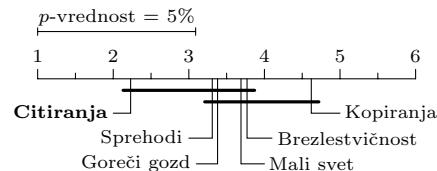
B Napovedovanje vodov

Slika 4. Slovensko plinovodno omrežje. Panel **A** prikazuje plinovodno omrežje z barvno označenimi ocenami obremenitev, panel **B** pa točnost napovedovanja manjkajočih plinovodov (za opis glej tekst).



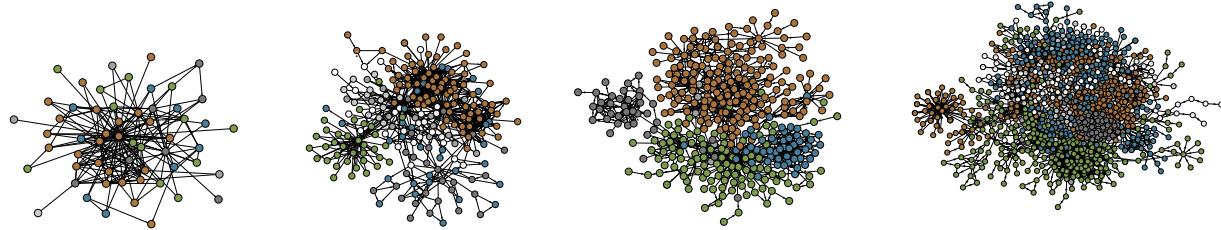
A Navodila in pravilniki

	Ujemanje F	Rang R
Citiranja	0.759	2.23
Sprehodi	0.570	3.31
Goreči gozd	0.504	3.38
Mali svet	0.000	3.69
Brezlestvičnost	0.382	3.77
Kopiranja	0.147	4.62



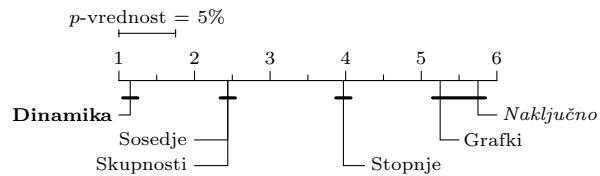
B Modeliranje omrežij

Slika 5. Omrežje sklicevanj med dokumenti. Panel **A** prikazuje dokumentno omrežje sodelujočega podjetja z barvno označenimi vrstami, panel **B** pa natančnost modelov gradnje omrežij (za opis glej tekst).



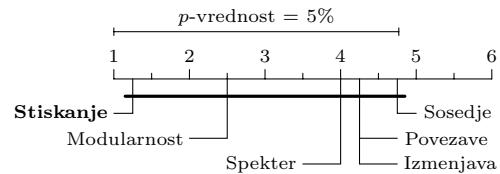
A Programske knjižnice

	Korelacija ρ_S	Točnost AUC
Dinamika	0.584	79.0%
Sosedje	0.532	76.4%
Skupnosti	0.490	75.7%
Stopnje	0.301	67.3%
Grafki	0.082	54.2%
<i>Naključno</i>	<i>0.053</i>	<i>50.5%</i>



B Napovedovanje odvisnosti

	Ujemanje NMI	Točnost CA
Stiskanje	0.475	89.8%
Izmenjava	0.449	88.5%
Modularnost	0.492	86.0%
Povezave	0.453	83.8%
Sosedje	-	81.8%
Spekter	0.452	76.6%

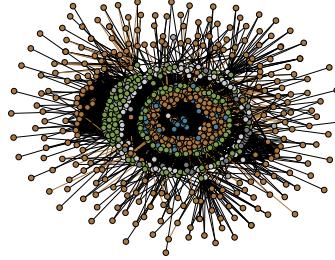


C Napovedovanje lastnosti

Slika 6. Omrežja odvisnosti med programskimi razredi. Panel **A** prikazuje omrežja programskih knjižnic potrebnih za izvajanje operacije, kjer barve označujejo visokonivojske pakete knjižnic, panel **B** prikazuje točnost napovedovanja manjkajočih programskih odvisnosti in panel **C** pa prikazuje točnost napovedovanja lastnosti programskih razredov (za opis glej tekst).

Programska omrežja

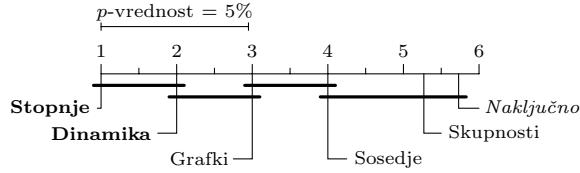
Slika 6, panel **A** prikazuje omrežja odvisnosti med razredi različnih programskih knjižnic, kjer barve označujejo visokonivojske pakete knjižnic. Zaporedomo si sledijo programska knjižnica razvita tekom izvajanja operacije ter podporne knjižnice za delo z grafi, statistične analize in podatkovno ruderjanje. Za namene odpravljanja napak Slika 6, panel **B** prikazuje točnost napovedovanja manjkajočih programskih odvisnosti in pa kritični diagram razlik za izbrane pristope. Najbolje se izkažejo pristopi na osnovi dinamičnih procesov, ki dosežejo 79% točnost in so statistično značilno boljši od vseh ostalih pristopov pri stopnji tveganja 5%. Delno zadovoljive rezultate dosegata še pristopa na osnovi sosedčin in skupnosti vozlišč. Slika 6, panel **C** prikazuje točnost napovedovanja lastnosti razredov kot so paketi z uporabo odkrivanja skupnosti ter kritični diagram razlik za izbrane pristope. Najvišjo točnost kar 80% doseže pristop na osnovi stiskanja omrežij, dočim pa razlike med različnimi pristopi niso statistično značilne pri stopnji tveganja 5%.



	Ujemanje NMI	Točnost CA
Spekter	0.401	62.7%
Modularnost	0.369	49.1%
Izmenjava	0.295	48.5%
Stiskanje	0.286	48.4%
Sredice	0.023	42.6%
Sosedje	-	15.7%

A Napovedovanje domen

	Korelacija ρ_S	Točnost AUC
Stopnje	0.741	92.7%
Dinamika	0.683	88.1%
Grafki	0.638	82.3%
Sosedje	0.443	74.7%
Skupnosti	0.017	50.2%
<i>Naključno</i>	<i>0.001</i>	<i>50.0%</i>



B Napovedovanje povezav

Slika 7. Omrežje povezav med spletnim mestom. Panel **A** prikazuje spletno omrežje sodelujočega podjetja z barvno označenimi domenami (levo) ter točnost napovedovanja domen (desno), panel **B** pa točnost napovedovanja manjkajočih povezav (za opis glej tekst).

Spletne omrežja

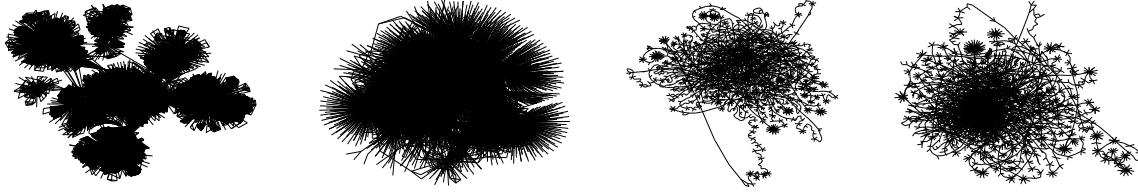
Spletne mesta sodelujočega podjetja predstavljajo pomemben vir informacij poslovnim partnerjem in strankam. Zato je smiselna optimizacija domen in povezav spletiča za namene enostavnnejše in učinkovitejše navigacije. Slika 7, panel **A** prikazuje del spletnega omrežja sodelujočega podjetja iz leta 2015 z barvno označenimi spletnimi domenami (levo) ter točnost napovedovanja domen z uporabo odkrivanja skupnosti vozlišč (desno). Najbolje se izkažejo pristopi spektralne analize, ki dosežejo 63% točnost, dočim delno zadovoljive rezultate dosegajo še pristopi optimizacije modularnosti, izmenjave oznak in stiskanja omrežij. Slika 7, panel **B** prikazuje še točnost napovedovanja spletnih povezav in pa kritični diagram razlik za izbrane pristope. Najvišjo točnost kar 93% dosežejo stopnje vozlišč, dočim zadovoljive rezultate dosega tudi pristop na osnovi dinamičnih procesov. Oba omenjena pristopa sta statistično značilno bolj zanesljiva od večine preostalih pri stopnji tveganja 5%, pri čimer so slednji primerljivi z naključnim napovedovanjem.

Družbena omrežja

V družbenih ali socialnih omrežjih kot je spletna storitev Facebook vozlišča ustrezajo posameznikom ali skupinam, neusmerjene povezave pa predstavljajo odnose ali interakcije med njimi (npr. prijateljstvo).

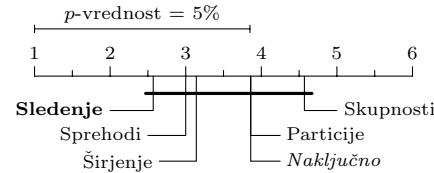
Spletne družbena omrežja

Spletna storitev Facebook predstavlja bogat nabor odnosov med strankami sodelujočega podjetja, ki jih je moč uporabiti za namene virusnega trženja. Na drugi strani pa spletno družbeno omrežje Facebook ni prosto dostopno ter preveliko za direktno analizo. Zatorej je potrebno vzorčenje. Slika 8, panel **A** prikazuje omrežja Facebook iz leta 2011 dobljena z različnimi pristopi vzorčenja omrežij, ki simulirajo procese v



A Facebook omrežja

	Ujemanje F	Rang R
Sledenje	0.549	2.57
Sprehodi	0.214	3.00
Širjenje	0.235	3.14
<i>Naključno</i>	0.223	3.86
Particije	0.057	3.86
Skupnosti	0.000	4.57



B Vzorčenje omrežij

Slika 8. Spletna družbena omrežja Facebook. Panel **A** prikazuje družbena omrežja Facebook dobljena z različnimi pristopi vzorčenja, panel **A** pa natančnost ujemanja z originalnim (za opis glej tekst).

praksi. S prostim očesom lahko opazimo, da imajo omrežja značilno različno zgradbo. Slika 8, panel **B** prikazuje ujemanje vzorčenih omrežij z originalnim ter kritični diagram razlik za izbrane pristope. Najboljše ujemanje doseže sledenje kontaktom, dočim zadovoljive rezultate kažeta še pristopa naključnih sprehodov in nenadzorovanega širjenja. Vseeno pa med pristopi ni statistično značilnih razlik pri stopnji tveganja 5%.

Ekonomski omrežji

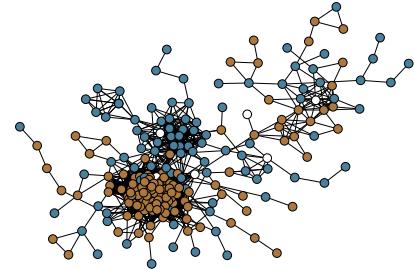
V ekonomskih omrežjih kot so korelacijska tokovna omrežja in pa omrežja trgovanj vozlišča ustrezajo posameznikom ali ustanovam, (ne)usmerjene povezave pa predstavljajo finančne interakcije, ekonomsko pripadnost ali sodelovanje (npr. upravni odbori).

Korelacijska tokovna omrežja

Slika 9, panel **A** prikazuje korelacijsko tokovno omrežje poslovnih partnerjev sodelujočega podjetja za izbrano slovensko občino, kjer barve vozlišč ponazarjajo partnerje z več kot 10% neporavnanih dolgov ob koncu izbranega leta. Vsakemu partnerju je sicer prirejen dvanajstmestni vektor, ki predstavlja poslovanje partnerja po mesecih izbranega leta. Dva partnerja sta povezana, v kolikor si delita podoben vzorec poslovanja merjeno preko korelacijskega koeficiente pripadajočih vektorjev. Opazimo, da imajo partnerji z oziroma brez neporavnanih obveznosti očitno podoben vzorec poslovanja ter tako tvorijo skupnosti v omrežju. Slika 9, panel **B** prikazuje točnost napovedovanja dolgov partnerjev z uporabo pristopov za določanje pomembnosti vozlišč (levo) in pristopov za odkrivanje skupnosti v omrežju (desno). V obeh primerih pristopi dosežejo kar 77% točnost, dočim se med pristopi pomembnosti vozlišč najbolje izkažejo grafki in lastni vektorji, med pristopi odkrivanja skupnosti pa izmenjava oznak in stiskanje omrežij.

Omrežja pripadnosti in sodelovanj

Slika 10, panel **A** prikazuje dvodelno omrežje pripadnosti poslovnih partnerjev sodelujočega podjetja za izbrano slovensko občino (levo) ter pripadajoče enodelno omrežje sodelovanj (desno), kjer barve vozlišč

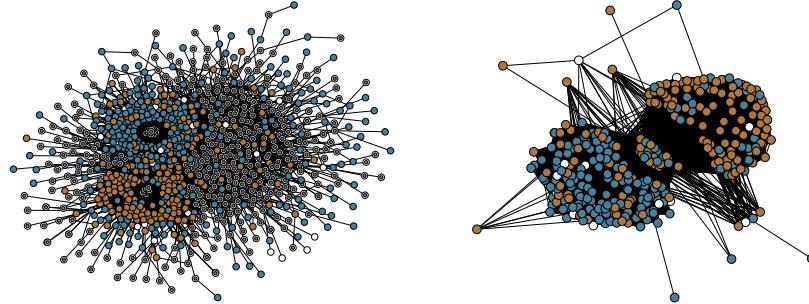


A Korelacijsko tokovno omrežje partnerjev

	Korelacija ρ_S	Točnost AUC		Ujemanje NMI	Točnost CA
Grafki	0.493	76.7%	Izmenjave	0.178	76.9%
Vektorji	0.455	76.7%	Stiskanje	0.176	76.3%
Stopnje	0.358	70.9%	Sosedje	-	75.1%
Središčnosti	0.370	70.3%	Modularnost	0.175	73.7%
<i>Naključno</i>	<i>0.044</i>	<i>50.2%</i>	Spekter	0.147	73.6%
Nakopičenost	0.037	47.7%	Sredice	0.223	71.3%

B Napovedovanje dolgov

Slika 9. Korelacijsko tokovno omrežje partnerjev. Panel **A** prikazuje tokovno omrežje partnerjev z barvno označenimi neporavnanimi dolgovi, panel **B** pa točnost napovedovanja dolgov (za opis glej tekst).



A Pripadnosti in sodelovanja partnerjev

	Korelacija ρ_S	Točnost AUC		Ujemanje NMI	Točnost CA
Vektorji	0.364	71.0%	Izmenjave	0.163	71.6%
Središčnosti	0.330	68.9%	Modularnost	0.155	71.6%
Stopnje	0.330	68.8%	Sosedje	-	71.6%
Nakopičenost	0.213	62.2%	Sredice	0.134	70.1%
Grafki	0.213	62.1%	Spekter	0.121	69.9%
<i>Naključno</i>	<i>0.035</i>	<i>49.7%</i>	Stiskanje	0.154	69.7%

B Napovedovanje dolgov

Slika 10. Omrežji pripadnosti in sodelovanja partnerjev. Panel **A** prikazuje omrežji pripadnosti in sodelovanja poslovnih partnerjev sodelujočega podjetja z barvno označenimi neporavnanimi dolgovi partnerjev, panel **B** pa točnost napovedovanja dolgov (za opis glej tekst).

ponazarjajo partnerje z več kot 10% neporavnanih dolgov ob koncu izbranega leta. V omrežju pripadnosti so partnerji povezani z računovodskimi temeljnicami na katerih se pojavljajo, pri čimer so slednje prikazane z označenimi vozlišči. Na drugi strani pa so v omrežju sodelovanj partnerji povezani, v kolikor se pojavijo na zadostnem številu skupnih temeljnic. Podobno kot zgoraj partnerji z oziroma brez neporavnanih dolgov tvorijo skupnosti v omrežju. Slika 10, panel **B** prikazuje točnost napovedovanja dolgov partnerjev z uporabo pristopov pomembnosti vozlišč (levo) in odkrivanja skupnosti (desno). V obeh primerih je dosežena 71% točnost, dočim se med pristopi pomembnosti vozlišč najbolje izkažejo lastni vektorji, med pristopi odkrivanja skupnosti pa izmenjava oznak, optimizacija modularnosti in soseščine.

Zaključki operacije

Izvajanje operacije je obsegalo pregled dostopnih podatkovnih virov v sodelujočem podjetju, oceno obstoječih algoritmov analize omrežij za podporo relevantnih primerov uporabe, razvoj prilagojenih algoritmov za podporo izbranih primerov uporabe, evalvacijo razvitih pristopov na izbranih primerih uporabe, objavo razvitih pristopov v priznanih znanstvenih revijah ter predstavitev na mednarodnih konferencah. Rezultati operacije ustrezajo ciljem zastavljenim v prijavni vlogi, pri čimer je operacija potekala skladno s terminskim planom ter namenom, cilji in predmetom javnega razpisa. Operacijo zato ocenujemo kot uspešno.

Univerza v Ljubljani
Kongresni trg 12, 1000 Ljubljana

Petrol d.d., Ljubljana
Dunajska cesta 50, 1527 Ljubljana

prof. dr. Ivan Svetlik, rektor
po pooblastilu prof. dr. Nikolaj Zimic, dekan

Pavel Škerlj, direktor IT

Ljubljana, 10.7.2015

Ljubljana, 9.7.2015

žig

žig

- [1] Šubelj L, Fiala D, Bajec M. Network-based statistical comparison of citation topology of bibliographic databases. *Sci Rep.* 2014;4:6496.
- [2] Šubelj L, Žitnik S, Blagus N, Bajec M. Node mixing and group structure of complex software networks. *Advs Complex Syst.* 2014;17(7-8):1450022.
- [3] Šubelj L, Bajec M. Group detection in complex networks: An algorithm and comparison of the state of the art. *Physica A.* 2014;397:144–156.
- [4] Blagus N, Šubelj L, Bajec M. Assessing the effectiveness of real-world network simplification. *Physica A.* 2014;413:134–146.
- [5] Žitnik S, Šubelj L, Bajec M. SkipCor: Skip-mention coreference resolution using linear-chain conditional random fields. *PLoS ONE.* 2014;9(6):e100101.
- [6] Šubelj L, Bajec M, Boshkoska BM, Kastrin A, Levnajič Z. Quantifying the consistency of scientific databases. *PLoS ONE.* 2015;10(5):e0127390.
- [7] Blagus N, Šubelj L, Weiss G, Bajec M. Sampling promotes community structure in social and information networks. *Physica A.* 2015;432:206–215.
- [8] Fiala D, Šubelj L, Žitnik S, Bajec M. Do PageRank-based author rankings outperform simple citation counts? *J Infometr.* 2015;9(2):334–348.
- [9] Šubelj L, Fiala D. Publication boost in Web of Science journals alters common citation distributions. *sub to JASIST.* 2015;p. 8.
- [10] Blagus N, Šubelj L, Bajec M. Empirical comparison of network sampling techniques. e-print arXiv:150602449v2. 2015;p. 14.
- [11] Šubelj L, Žitnik S, Bajec M. Who reads and who cites? Unveiling author citation dynamics by modeling citation networks. In: Proceedings of the International Conference on Network Science. Berkeley, CA, USA; 2014. p. 1.
- [12] Šubelj L, Weiss G, Blagus N, Bajec M. What coins the bitcoin? Exploratory analysis of bitcoin market value by network group discovery. In: Proceedings of the International Conference on Network Science. Berkeley, CA, USA; 2014. p. 1.
- [13] Blagus N, Šubelj L, Weiss G, Bajec M. Large networks grow smaller: How to choose the right simplification method? In: Proceedings of the International Conference on Network Science. Berkeley, CA, USA; 2014. p. 1.
- [14] Šubelj L, Bajec M. Network group discovery by hierarchical label propagation. In: Proceedings of the European Social Networks Conference. Barcelona, Spain; 2014. p. 284.
- [15] Šubelj L, Bosnič Z, Kukar M, Bajec M. Automatization of the stream mining process. In: Proceedings of the International Conference on Advanced Information Systems Engineering. Thessaloniki, Greece; 2014. p. 409–423.
- [16] Šubelj L. Exploratory and predictive tasks of network community detection. In: Proceedings of the International Conference on Network Science. Zaragoza, Spain; 2015. p. 1.
- [17] Šubelj L, Fiala D, Bajec M. Consistency of citation topology of bibliographic databases. In: Proceedings of the International Conference on Network Science. Zaragoza, Spain; 2015. p. 1.
- [18] Šubelj L. Large network community detection in practical scenarios. In: Proceedings of the International Workshop on Social Network Analysis. Capri, Italy; 2015. p. 78.