Javni razpis za (so)financiranje raziskovalnih projektov za leto 2020

PREDSTAVITEV RAZISKOVALNEGA PROJEKTA

27.1. **Znanstvena izhodišča ter predstavitev problema in ciljev raziskav**

Bibliografske storitve, kot so Web of Science/Knowledge, Scopus, CiteSeer, Zentralblatt Math, Google Scholar, DBLP, Math Sci, COBISS in druge, nudijo podatke o znanstvenih delih (članki, knjige, poročila, itd.). Posameznik jih običajno uporablja za iskanje del na izbrano tematiko, ustanove pa jih uporabljajo za ovrednotenje in načrtovanje raziskovalnega dela. Uporabljajo se tudi kot vir podatkov za bibliometrične in scientometrične raziskave. V ta namen se podatki na izbrano tematiko pogosto pretvorijo v zbirko sklopljenih bibliografskih omrežij, ki povezujejo raznovrstne enote (dela, avtorje, urednike, revije, ključne besede, ustanove, države, jezike, itd.).

Bistven korak pri izgradnji bibliografskih omrežij je določitev enot (entity resolution) (razrešitev sinonimnih / homonimnih imen/oznak enot). To vprašanje je zelo pomembno tudi pri združevanju podatkov iz različnih virov. Visoka natančnost pri določanju enot je predpogoj za izgradnjo visoko kakovostnih omrežij. Razvili bomo nove, zelo natančne postopke določanja enot za posebne vrste enot, ki upoštevajo medsebojne odnose med enotami. Izdelali bomo tudi programsko podporo za pretvorbo med različnimi zapisi bibliografskih podatkov.

Ustvarjena bibliografska omrežja so pogosto velika (na tisoče ali tudi milijone enot). Za njih analizo je potrebno razviti zelo učinkovite (podkvadratične) algoritme, ki običajno temeljijo na dejstvu, da je večina velikih omrežij redka (število povezav je istega reda kot število vozlišč).

Pomembno orodje pri analizi zbirk sklopljenih omrežij so izpeljana omrežja, ki jih dobimo s prepletanjem normalizacije (deležni (fractional) pristop) in množenja usklajenih omrežij.
Lani smo razdelali teoretično ozadje deležnega pristopa (Batagelj 2020) in pokazali kako lahko časovna omrežja, ki temeljijo na časovnih količinah (Batagelj in Praprotnik 2016), uporabimo v biblimetričnih analizah (Batagelj and Maltseva 2020). V projektu nameravamo raziskati nove možnosti, ki jih odpirata oba pristopa. Novorazvite metode bomo uporabili pri analizi izbranih bibliografskih podatkovij.

Podatke dostopne v bibliografskih podatkovnih bazah bi lahko uporabili tudi za izgradnjo višjestopenjskih storitev za različne vrste uporabnikov. Poiskali bomo nekaj primerov tovrstnih storitev in izdelali zanje prototipne rešitve.

27.2. **Pregled in analiza dosedanjih raziskav in relevantne literature**

Iz posebnih bibliografij (BibTeX, EndNote) in bibliografskih podatkovnih baz lahko pridobimo podatke o delih (članki, knjige, poročila, itd.) za izbrano tematiko. Tipični opis dela vsebuje naslednje podatke: avtorji; naslov; založnik/revija; leto objave; strani. V nekaterih virih so dostopni še drugi podatki, kot so: jezik; klasifikacija; ključne besede; avtorjeva ustanova/država; seznam sklicevanj; povzetek. Te podatke lahko pretvorimo v zbirko usklajenih dvovrstnih omrežij o izbrani tematiki: dela × avtorji; dela × ključne besede; dela × države, in drugi pari množic enot. Poleg tega dobimo še razbitje del glede na leto objave, vektor števila strain, in včasih tudi (enovrstno) omrežje sklicevanj.

Pri izgradnji teh omrežij moramo najprej razrešiti problem mej omrežja (Marsden 1990) – natančno določiti kaj so enote – vozlišča omrežja in katera relacija jih povezuje. Odločiti se moramo ali je

omrežje eno ali dvo-vrstno, katere lastnosti vozlišč in povezav so pomembne za nameravane analize. Pri določitvi povezav moramo odgovoriti na vprašanja:
(1) Ali so povezave usmerjene?
(2) Ali obstajajo različne vrste povezav (relacij), ki jih je potrebno upoštevati?
(3) Ali je lahko par vozlišč povezan z več povezavami?
(4) Kakšne so uteži na povezavah?
(5) Ali se omrežje spreminja skozi čas?

Naslednji problem, ki se pogosto pojavi pri določitvi množice vozlišč je prepoznavanje vozlišč. Enota – vozlišče ima lahko v podatkih več različnih imen (sinonimi) ali pa lahko isto ime označuje različne enote (homonimi, dvoumnost). Na primer, v BibTeXovi bibliografiji področja računske geometrije (Jones 2002) isti avtor nastopa s 7 različnimi imeni: R.S. Drysdale, Robert L. Drysdale, Robert L. Scot Drysdale, R.L. Drysdale, S. Drysdale, R. Drysdale, and R.L.S. Drysdale. Notranja informacija je potrebna, da se odločimo, da sta Otfried Schwarzkopf in Otfried Cheong ista oseba. Na drugi strani v matematični bazi MathSciNet obstaja 57 različnih matematikov z istim imenom Wang, Li (TePaske-King in Richert 2001). Uredniki te baze si že od leta 1985 prizadevajo sproti razreševati problem prepoznavanja avtorjev že pri vnosu podatkov. V prihodnosti se bomo lahko temu problemu izognili s splošno uveljavitvijo pobud kot sta ResearcherID ali ORCID.

Podobno v sklicevanjih iz WoSa najdemo naslednja imena revij: NUCLEIC ACIDS RES, NUCL ACIDS RES, NUCLEIC ACIDS RES S, NUCLEIC ACIDS RES S2, NUCL ACID RES, NUCL ACIDS RES S2, NUCL ACIDS S SER, NUCL ACIDS RES S, NUCL AC RES, NUCLEIC ACIDS RES S1, Nucleic Acids Res, NUCL ACIDS RES S1 ali Q J R MET SOC, Q J R METEOROL SOC, Q J ROY METEOR SO S1, Q J ROY METEOR SOC, Q J ROY METEOR SOC B, QUART J ROY METEOR S, QUART J ROY METEOROL, QUART J ROY METEOROL SOC, QUART J ROYAL METEOR. Vprašanje, ki se nemudoma zastavi je, ali ta imena določajo isto revijo? Obstaja Standard Serial Number (ISSN 2020), mednarodni system za enolično določitev revij in drugih serijskih virov. Težava je, ker se v WoSu ta dogovor pri sklicevanjih ne uporablja. Pri prepoznavanju revij lahko uporabimo še Global serials directory (Ulrichsweb 2020) in Journal Abbreviation Sources (JAS 2020) ter še nekaj drugih storitev in virov.

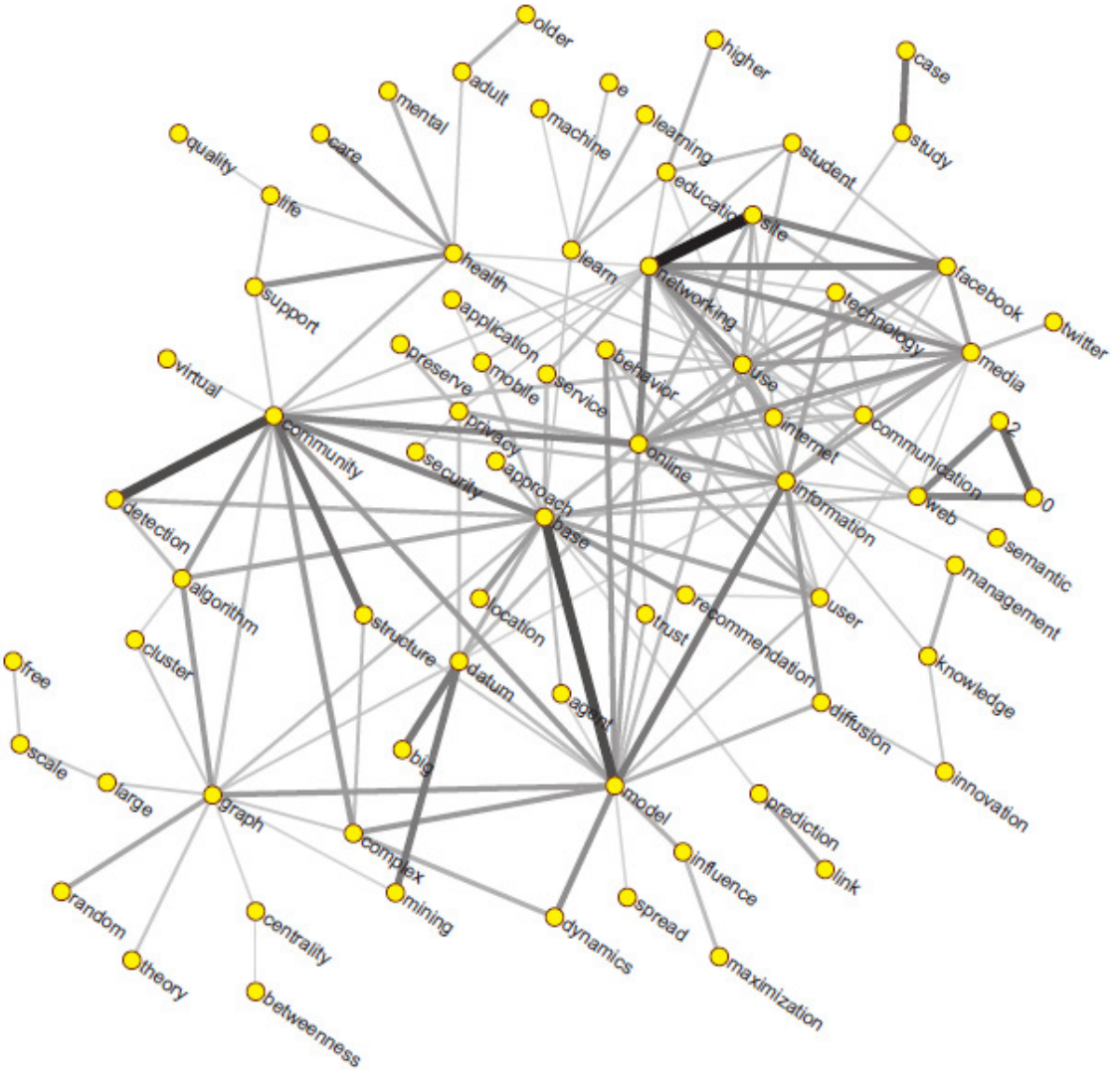
Problem prepoznavanja se pojavi tudi pri luščenju enot iz besedil v podatkovnih poljih. Pri pridobivanju ključnih besed iz naslova ali povzetka najprej odstranimo nepomembne besede (stopwords). Preostale zanimive besede (ali fraze) običajno standardiziramo, tako da jih zamenjamo s kanonskimi predstavniki. Na primer, termine ‘function’, ‘map’, ‘mapping’ in ‘transformation’ lahko v matematični literature obravnavamo kot enakovredne. Podobno je z besedami v večjezičnih virih. Pri razreševanju the problemov si lahko pomagamo s seznami enakovrednih terminov ali slovarji.

Dodatni vir težav pri prepoznavanju izhaja iz slovnice uporabljenega jezika. Na primer v angleščini se lahko glagol ‘go’ pojavi v besedilu v različnih oblikah, kot so ‘go’, ‘goes’, ‘gone’, ‘going’ in ‘went’. Za razreševanje tovrstnih problemov lahko uporabimo postopke krnjenja ali lematizacije, ki jih najdemo v knjžnicah za obdelavo naravnega jezika, kot sta NLTK (Bird et al. 2009; Perkins 2010) in MontyLingua (Liu 2004).

Splošni problem prepoznavanja enot (entity resolution) je dobro razdelan v literaturi iz rudarjenja v besedilih (text mining) (Buscaldi & Rosso, 2008; Talburt, 2011; Christen, 2012; Reitz & Hoffmann, 2013; Momeni & Mayr, 2016; Windham, 2018; Yadav & Bethard, 2019). Pri bibliometričnih analizah potrebujemo rešitve z visoko natančnostjo. Poskusili jih bomo razviti z upoštevanjem posebnosti zgradbe bibliografskih podatkov.

Iz podatkov pridobljenih iz bibliografskih baz lahko ustvarimo različna bibliografska omrežja. Na primer, s programom WoS2Pajek lahko iz podatkov pridobljenih iz WoS ustvarimo naslednja

dvovrstna omrežja: omrežje avtorstev WA na dela × avtorji, omrežje revij WJ na dela × revije, omrežje ključnih besed WK na dela × ključne besede, in (enovrstno) omrežje sklicevanj Cite na delih. Poleg tega dobimo še lastnosti del: razbitje year del glede na leto objave; razbitje DC, ki razlikuje dela na dela s polnim opisom (DC[w] = 1) in dela, ne katera se samo sklicuje (DC[w] = 0); in vektor števila strani NP. Pri analizi the omrežij lahko določimo frekvenčne porazdelitve raznih enot (avtorji, revije, ključne besede), ki opisujejo celostne lastnosti omrežij. Določimo lahko tudi najpomembnejše enote (Cerinšek & Batagelj, 2015). Pomembno orodje pri analizi zbirk sklopljenih omrežij je množenje omrežij, ki ustvari nova, izpeljana omrežja, ki opisujejo povezanosti med spočetka neposreno nepovezanimi množicami vozlišč. Na primer, omrežje AK = t(WA) · WK (t(A) je transponirano omrežje A) poveže avtorje s ključnimi besedami (Batagelj & Cerinšek, 2013).



Slika 1

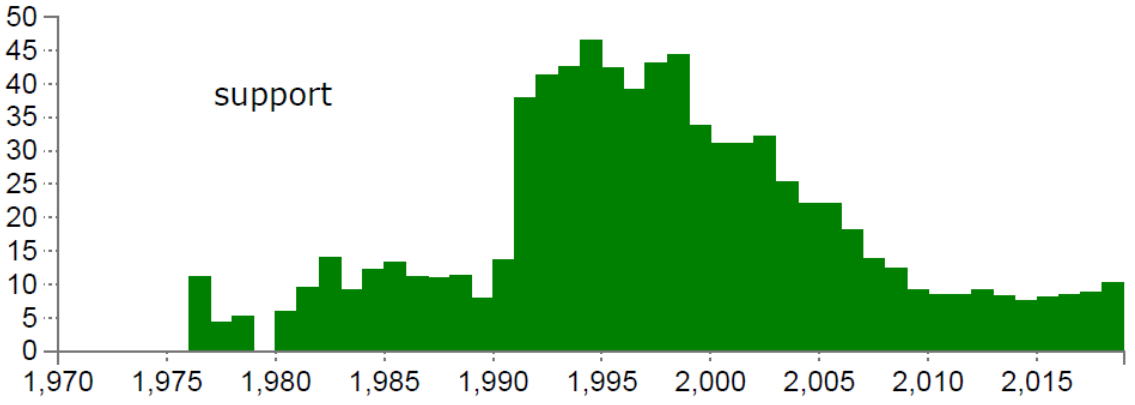
Deležni (fractional) pristop je predlagal Lindsey (1980). Na primer, pri analizi soavtorstev mora biti skupni prispevek vseh soavtorjev posameznega dela enak 1. Ker ne poznamo dejanskih prispevkov, ocenimo prispevek posameznega avtorja kot 1 deljeno s številom soavtorjev. Normalizacijo omrežja A dobljeno na ta način označimo z n(A). Alternativno pravilo, Newmanova normalizacija, ki izključuje samosodelovanje, je predlagal Newman (2001, 2004) – delimo s število soavtorjev – 1. Pred kratkim

je vprašanje normalizacije zopet postalo “vroče” (Batagelj & Cerinšek, 2013; Cerinšek & Batagelj, 2015; Perianes-Rodriguez et al., 2016; Prathap & Mukherjee, 2016; Leydesdorff & Park, 2017; Gauffriau, 2017) so se ponovno poglobili v ozadje deležnega pristopa. V članku (Batagelj 2020) smo predstavili teoretični okvir, ki temelji na razcepu produkta matrik na vsoto zunanjih produktov ustreznih vektorjev. Ta nam omogoča vpogled v zgradbo bibliografskih omrežij dobljenih z normalizacijo in množenjem omrežij.

Na Sliki 1 je prikazan glavni povezavni otok (najmočneje povezana vozlišča) v omrežju nKK sopojavljanja ključnih besed v omrežju za področje analize družbenih omrežij (SNA). Ključne beside so lematizirane. nKK = t(n(WK)).n(WK), |W| = 70792, |K| = 32409, |E(nKK)| = 2799530.

Podrobnejši vpogled v razvoj bibliografskih omrežij je omogočen z upoštevanjem časa. V časovnem omrežju se prisotnost in dejavnost vozlišč in povezav spreminja s časom, prav tako se spreminjajo vrednosti njih lastnosti. V opis časovnega omrežja moramo vključiti to spreminjanje. Zgodnje uporabe časovnih omrežij srečamo pri razporejanju opravil (CPM, Pert) v operacijskih raziskavah (Moder & Phillips, 1970); pri analizi transportnih omrežij analysis (Bell & Iida, 1997); in kot omrežja omejitev v umetni inteligenci (Dechter, 2003). Različni pristopi so bili predlagani tudi za analizo podatkov (Holme, 2015). Najpogosteje se uporablja prečni (cross-sectional) pristop, pri katerem čas sestavlja končno število časovnih točk (trenutkov ali intervalov). Časovna rezina v dani časovni točki je navadno omrežje, ki ga sestavljajo vsa v tej časovni točki dejavna vozlišča in povezave. Omrežje analiziramo tako, da analiziramo vsako rezino posebej in rezultate združimo v “časovno vrsto”. Zanimiv je tudi pristop časovno-spreminjajočih se grafov TVG (Casteigts, Flocchini, & Quattrociocchi, 2012).

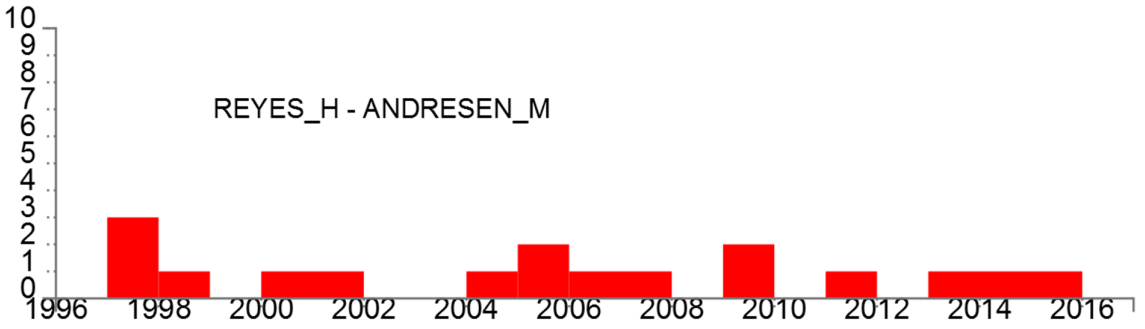
V članku Batagelj and Praprotnik (2016) smo predlagali vzdolžni (longitudinal) pristop k opisu in analizi časovnih omrežij, ki temelji na časovnih količinah. Pristop je alternativa običajnemu prečnemu pristopu. Časovna količina opisuje kako se pripadajoča lastnost spreminja skozi čas. Predlagani pristop ima naslednje odlike: (1) deluje tako za diskretni kot za zvezni čas (2) se sam (znotraj operacij) prilagaja razdrobljenosti podatkov (3) rezultati metod so običajno spet časovna omrežja ali seznami časovnih količin. Pristop je mogoče uporabiti tudi za analizo časovnih bibliografskih omrežij. Lahko ga uporabimo tudi v podobnih kontekstih. Opis tekočega stanja na področju časovnih omrežij je mogoče najti v novi knjigi Holme & Saramäki (2019). V pravkar objavljenem članku (Batagelj & Maltseva, 2020) smo pokazali, kako lahko tradicionalna bibliografska omrežja in podatke o letu objave del predelamo v ustrezna časovna (trenutna ali nakopičena) omrežja.



Slika 2

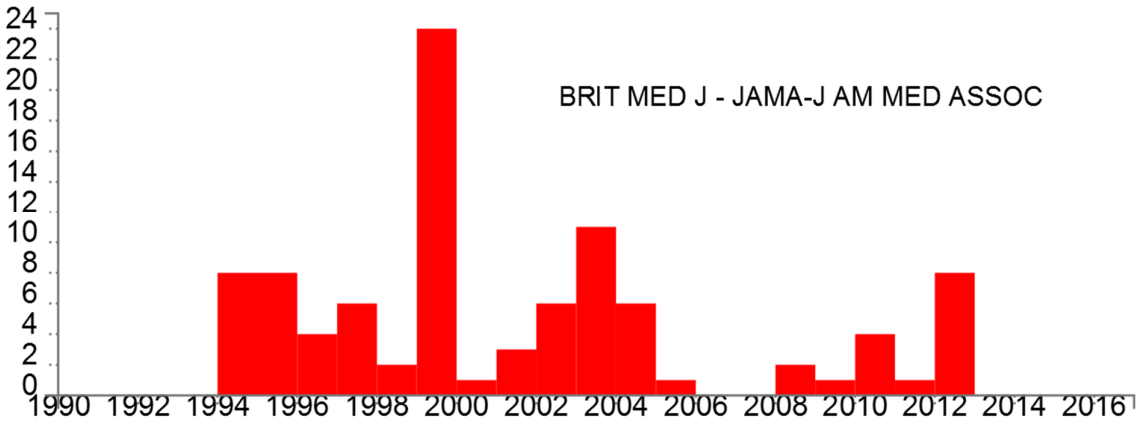
Na primer, za izbrano leto smo v trenutnem časovnem omrežju WKi izračunali razmerje med številom vseh pojavitev dane ključne besede in številom pojavitev najpogostejše ključne besede. To razmerje

izraža pomembnost dane ključne besede v izbranem letu in ima vrednosti med 0 in 100%. Na Sliki 2 so prikazane letne spremembe pomembnosti ključne besede "support" na področju analize družbenih omrežij.



Slika 3

Trenutno soavtorsko omrežje Coi dobimo tako Coi = t(WAi) · WAi, kjer je WAi trenutna časovna različica omrežja WA. Utež Coi(a,b) je časovna količina, ki šteje število del, ki sta jih avtor a in avtor b napisala skupaj vsako leto. Na Sliki 3 je prikazano soavtorstvo po letih za avtorja Reyes H. in Andersen M. na področju strokovnih pregledov (peer-review) (podatki iz Batagelj, Ferligoj, Squazzoni, 2017).



Slika 4

Izpeljano trenutno časovno omrežje, ki opisuje sklicevanja med revijami dobimo kot JCJ = t(WJi) · CiteI · WJc. Pozor, prvi dve omrežji sta trenutni, tretje je nakopičeno. Utež JCJ(i,j) šteje število sklicevanj v posameznem letu iz del objavljenih v reviji I na dela objavljena v reviji j. V posebnem primeru i=j dobimo časovno količino ki šteje samosklicevanja za revijo i. Na Sliki 4 je prikazana časovna količina JCJ(BRIT MED J, JAMA-J AM MED ASSOC) za področje strokovnih pregledov.

Za določitev ključnih besed specifičnih za skupino avtorjev ali del lahko uporabimo pristop TF-IDF (Robertson, 2004). V Tabeli 1 so podane najbolj specifične ključne besede za področje analize družbenih omrežij glede na uteži TF-IDF za izbrane tri revije (skupine del).

	SOC NETWORKS		LNCS		PHYSICA A	
Rank	Value	Id	Value	Id	Value	Id
1	0.1389	graph	0.1464	graph	0.3674	complex
2	0.1375	model	0.1407	base	0.2318	dynamics
3	0.1350	**structure**	0.1218	user	0.1761	model
4	0.1199	tie	0.1172	privacy	0.1659	spread
5	0.1015	centrality	0.1038	web	0.1208	rumor
6	0.1002	random	0.1016	online	0.1126	evolution
7	0.0965	**structural**	0.0995	**network**	0.1114	world
8	0.0912	personal	0.0994	datum	0.1099	epidemic
9	0.0899	**network**	0.0934	information	0.1084	**structure**
10	0.0809	exponential	0.0902	model	0.1071	free
11	0.0808	p	0.0888	analysis	0.0978	community
12	0.0780	power	0.0867	algorithm	0.0966	small
13	0.0768	equivalence	0.0777	detection	0.0931	node
14	0.0755	analysis	0.0735	recommendation	0.0913	detection
15	0.0740	friendship	0.0713	community	0.0881	base
16	0.0730	accuracy	0.0710	**social**	0.0871	scale
17	0.0729	exchange	0.0696	semantic	0.0849	diffusion
18	0.0713	datum	0.0690	learn	0.0844	opinion
19	0.0691	measure	0.0679	mining	0.0824	game
20	0.0682	blockmodel	0.0654	use	0.0806	**network**
21	0.0678	organization	0.0630	mobile	0.0754	propagation
22	0.0643	asterisk	0.0624	trust	0.0741	graph
23	0.0629	dynamics	0.0623	collaborative	0.0712	agent
24	0.0591	status	0.0592	visualization	0.0701	sir
25	0.0584	informant	0.0586	application	0.0700	algorithm
26	0.0573	mode	0.0575	service	0.0655	spreader
27	0.0569	generator	0.0561	search	0.0641	evolutionary
28	0.0535	core	0.0560	query	0.0640	emergence
29	0.0526	markov	0.0554	twitter	0.0612	information
30	0.0502	effect	0.0553	design	0.0602	distribution
Total:	18.6443		19.5058		14.8126	

Tabela 1

Vse tri pristope, normalizacija, množenje omrežij in časovna omrežja, lahko med seboj prepletamo in razpenjajo velik, v veliki meri neraziskan proctor za razvoj novih metod.

Ena od značk prirejenih posameznemu avtorju področje raziskav (RI). Na primer, v bazi Sicris/Cobiss je področje raziskav opisano hierarhično – značka 1.01.05 zaobjema (1) naravoslovje, (01) matematika, (05) teorija grafov. Žal v več primerih ta podatek manjka ali pa ne opisuje ustrezno področje raziskav danega raziskovalca. V nekaterih primerih je prišlo do napake pri vnosu podatkov. Po drugi strani, baza Scopus vsebuje za revije podatke o znanstvenih področjih (SF), ki jih revija pokriva. Tako je revija "Ars Mathematica Contemporanea" razvrščena kot Mathematical journal, ki pokriva štiri matematična področja: Algebra and Number Theory, Discrete Mathematics and Combinatorics, Geometry and Topology, in Theoretical Computer Science. Radi bi raziskali, kako je RI povezano s SF in nato to zvezo uporabili za razkritje osamelcev in napak v podatkih. Podatke bomo preverili še z drugimi neodvisnimi količinami, kot so sodelovalna razdalja in ujemanje v ključnih besedah. Pričakujemo, da bodo soavtorji imeli podobna področja raziskav.

Raziskati nameravamo tudi povezanosti med prerezi, otoki in sredicami ter področji raziskav avtorjev glede na omrežje soavtorstev. Ustvariti nameravamo nove omrežne mere, ki merijo odstopanje danega porojenega podomrežja od enobarvnega, pri čemer so barve področja raziskav.

27.3. **Podroben opis vsebine in programa dela raziskovalnega projekta**

WP1. Izboljšave programske podpore za pretvorbo bibliografskih podatkov v omrežja
Nekaj programov za pretvorbo bibliografskih podatkov smo že razvili (WoS2Pajek, DBLP, Zbmath, BiBTeX2Pajek). Najobsežnejši med njimi je WoS2Pajek. V projektu nameravamo razširiti zmogljivosti programa WoS2Pajek z novimi možnostmi (država, jezik, ustanova). Dodatno nameravamo razviti program za pretvorbo podatkov v drugih oblikah zapisa (BibTeX, Zbmath, DBLP, RIS) v obliko WoS. To nam bo omogočilo, da uporabimo WoS2Pajek tudi za pretvorbo podatkov v drugih oblikah. Na ta način omogočimo tudi združevanje podatkov iz različnih virov.

WP2. Metode in orodja za prepoznavanje enot (entity resolution)
Izgradnja bibliografskih omrežij na izbrano temo je iterativen process. Začnemo z zadetki za poizvedbo s ključnimi besedami značilnimi za izbrano temo. Ustvarimo ustrezno zbirko bibliografskih omrežij in na njih opravimo začetno analizo, s katero razkrijemo manjkajoča pomembna dela (pogosto citirana dela, ki niso zadetki) in jih dodamo med začetne podatke – zapolnjevanje podatkov. Pri čiščenju podatkov poskušamo čim bolje prepoznati enote – včasih, npr. za razrešitev dvoumnosti, s popravljanjem našega izvoda začetnih podatkov. Problem prepoznavanja je dobro razdelan v rudarjenju v podatkih, a splošne metode ne zagotavljajo visoke natančnosti. V našem primeru lahko uporabimo posebne lastnosti zgradbe bibliografskih podatkov za pripravo izboljšanih algoritmov prepoznavanja za posamezne vrste enot.

WP3. Teoretične raziskave v analizi bibliografskih omrežij
Pomembno orodje v analizi sklopljenih omrežij (bibliografska omrežja so poseben primer) je množenje omrežij (Batagelj & Cerinšek, 2013), ki nam omogoča izračun izpeljanih omrežij. Da zagotovimo enakovredno obravnavo vseh enot uporabimo deležni (fractional) pristop z normalizacijo omrežij. Njegovo teoretično ozadje je bilo razdelano v pravkar objavljenem članku (Batagelj 2020). V člankih Batagelj & Praprotnik (2016) in Batagelj &Maltseva (2020) smo ponudili vzdolžni pristop k analizi časovnih omrežij in pokazali, kako ga lahko uporabimo za analizo časovnih bibliografskih omrežij. Nadaljevali bomo z raziskavami uporabe the treh pristopov v analizi bibliografskih omrežij.

WP4. Programska izvedba novih metod za analizo bibliografskih omrežij in njih uporaba na izbranih podatkovjih
Bibliografska omrežja so lahko velika (nekaj tisoč ali celo nekaj milijonov enot - vozlišč). Razvita programska podpora mora ponuditi rešitve, ki lahko tudi take podatke učinkovito obdelajo – v nekaj sekundah ali minutah. Osnovna podpora bo dana v nekaj knjižnicah v Pythonu ali Rju. Mogoče jih je tudi ponuditi združene v obliki prijaznega orodja – nekakšno računalo z bibliografskimi omrežji (podobno Pajku). Za prikaz zmogljivosti analize (časovnih) bibliografskih omrežij bomo ustvarili nekaj večjih zbirk za izbrana znanstvena področja in jih analizirali, da bi dobili vpogled v njihov razvoj in zgradbo.

WP5. Višjestopenjske bibliografske storitve
Analizo bibliografskih omrežij lahko uporabimo za preverjanje skladnosti bibliografskih podatkov in storitev in za avtomatsko pripravo predlogov za manjkajoče podatke. Uporabimo jo lahko tudi za razvoj višjestopenjskih bibliografskih storitev za razne vrste uporabnikov. Na primer:
Uredniki revij: izbira primernega recenzenta za dani članek; ovrednotenje recenzentov.
Vzdrževanje baze: avtomatični predlogi za manjkajoče vrednosti; preverjanje usklajenosti podatkov.

Avtorji: v katero revijo naj pošljem ta članek? Ključne besede za ta članek?
Raziskovalci: kdo so primerni partnerji za ta projekt?
Študent: rad bi seznanil s tem področjem. Katere članke naj preberem?

WP6. Objava rezultatov
Delo na projektu bomo nadzorovali na seminarji, na katerem bodo sodelavci sproti poročali o napredku in morebitnih težavah. Dobljene rezultate bomo predstavili na mednarodnih znanstvenih srečanjih in objavili v znanstvenih revijah. Razvita programska oprema skupaj z ustrezno dokumentacijo in podatki bo dostopna na Githubu kot odprto-kodna. Letna in zaključno poročilo so tudi del tega WP.

27.4. **Razpoložljiva raziskovalna oprema (nad 5.000 €) potrebna za izvedbo projekta**

Za analizo večine podatkovij zadostuje boljši prenosnik z 16 GB pomnilnika. Za zelo velika podatkovja bomo po potrebi uporabili računalniške zmogljivosti dostopne na IAM.

27.5. **Upravljane projekta: podroben načrt uresničevanja in časovna razporeditev**

Projekt je razdeljen na šest paketov (WP) naprej razdeljenih na opravila:

WP1. Izboljšave programske opreme za pretvorbo bibliografskih podatkov v omrežja
   a. Izboljšave programa WoS2Pajek (dodati: jezik, država, ...), izvedba v Python 3
   b. Pretvorbe iz drugih oblik zapisa (Scopus, DBPL, RIS, ...) v obliko WoS

WP2. Metode in orodja za prepoznavanje enot (entity resolution)
   a. Prepoznavanje del
   b. Prepoznavanje oseb (avtorjev)
   c. Prepoznavanje revij
   d. Prepoznavanje ključnih besed
   e. Prepoznavanje držav

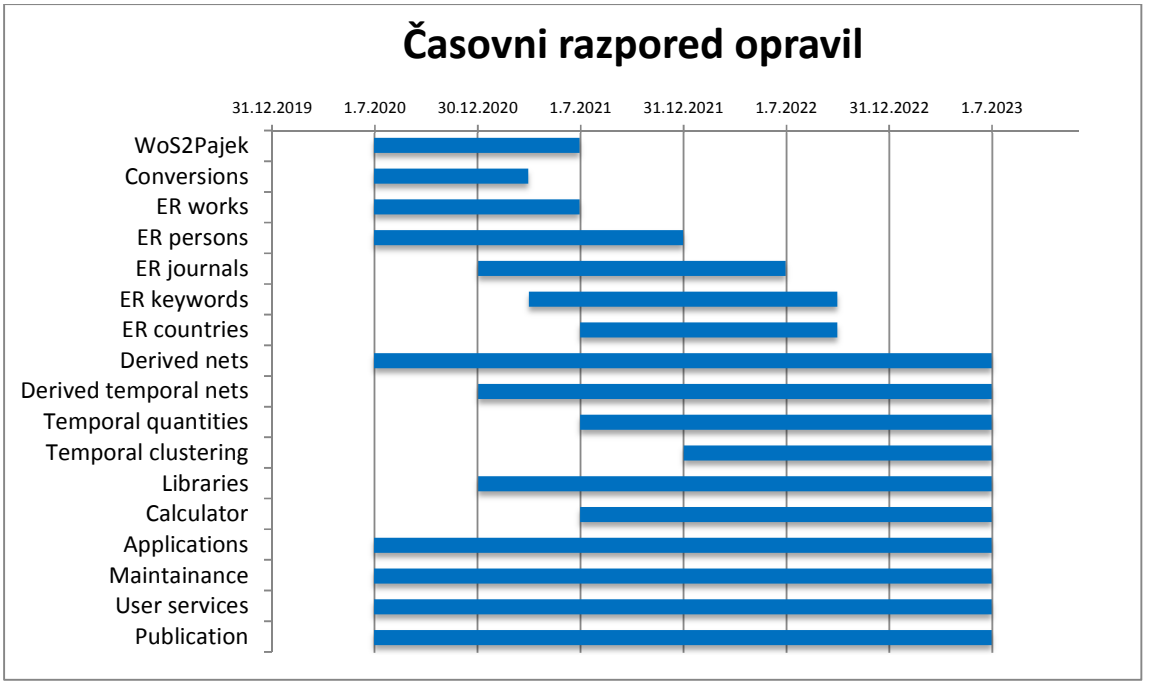
WP3. Teoretične raziskave v analizi bibliografskih omrežij
   a. Nova izpeljana omrežja z normalizacijo in množenjem
   b. Časovne različice izpeljanih omrežij
   c. Nove časovne količine za opis lastnosti časovnih bibliografskih omrežij
   d. Razvrščanje v časovnih bibliografskih omrežjih

WP4. Programska izvedba novih metod za analizo bibliografskih omrežij in uporabe na izbranih podatkih
   a. Vključitev novih algoritmov v knjižnice Nets, TQ in Biblio ali samostojna
   b. Načrt in izvedba »računala« z bibliografskimi omrežji
   c. Izgradnja zbirk bibliografskih omrežij na izbrane teme in njih analiza

WP5. Višjestopenjske storitve
   a. Iskanje zanimivih storitev in razvoj metod za podporo vzdrževanja bibliografskih baz
   b. Iskanje zanimivih storitev in razvoj metod za podporo višjestopenjskih storitev

WP6. Publication (seminars, conferences, papers) of the results

Posamezni WPji so dinamično soodvisni – napredek v posameznem osnovnem opravilu razširi možnosti, ki jih lahko uporabimo v analizah ali poveča natančnost rezultatov. Tematiko projekta poznamo in se lahko nemudoma lotimo dela.



Slika 5

Razpored opravil na Sliki 5 je narejen s predpostavko, da bo projekt začel 1. julija 2020. Pravi razpored dobimo z ustreznim pomikom na dejanski datum začetka.

27.6 Viri

Batagelj, V, Cerinšek, M: On bibliographic networks. Scientometrics 96 (2013) 3, 845-864.

Batagelj, V., Doreian, P., Ferligoj, A., Kejžar, N.: Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley Series in Computational and Quantitative Social Science. Wiley, 2014.

Batagelj, V., Ferligoj, A., & Squazzoni, F. (2017). The emergence of a field: A network analysis of research on peer review. Scientometrics, 113(1), 503–532.

Batagelj, V., Praprotnik, S.: An algebraic approach to temporal network analysis based on temporal quantities. Social Network Analysis and Mining, 6(2016)1, 1-22

Batagelj, V.: On Fractional Approach to Analysis of Linked Networks. Scientometrics, 2020.

Batagelj, V., Maltseva, D.: Temporal Bibliographic Networks. Journal of Informetrics, 2020.

Bell, M. G. H., & Iida, Y. (1997). Transportation network analysis. Chichester: Wiley.

Bird, S., Klein, E., Loper, E.: NLTK - Natural Language Processing with Python. 2020. https://www.nltk.org/book/

Buscaldi, D., Rosso, P.: A conceptual density-based approach for the disambiguation of toponyms. Int. J. Geogr. Inf. Sci. 22, 3 (January 2008), 301–313.

Casteigts, A., Flocchini, P., Quattrociocchi, W., et al. (2012). Time-varying graphs and dynamic networks. International Journal of Parallel, Emergent andDistributed Systems, 27(5), 387–408.

Cerinšek, M., Batagelj, V.: Network analysis of Zentralblatt MATH data. Scientometrics, 102(2015)1, 977-1001.

Christen, P.: Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer-Verlag, Berlin, Heidelberg, 2012

Dechter, R. (2003). Constraint processing. San Francisco: Morgan Kaufmann.

Doreian, P., Batagelj, V., Ferligoj, A. (eds.): Advances in Network Clustering and Blockmodeling. Wiley, 2020. ISBN: 978-1-119-22470-9

Gauffriau, M.: A categorization of arguments for counting methods for publication and citation indicators. Journal of Informetrics, 11(2017)3, 672-684.

Holme, P. (2015). Modern temporal network theory: A colloquium. European Physical Journal B, 88, 234.

Holme, P., & Saramäki, J. (Eds.). (2019). Temporal network theory. Springer.

ISSN: 2020. https://www.issn.org/ , https://www.nuk.uni-lj.si/informacije/ISSN

JAS: Journal Abbreviation Sources. 2020. https://www.abbreviations.com/jas.php

Jones, B.: Computational Geometry Database. 2002. ftp://ftp.cs.usask.ca/pub/geometry/

Leydesdorff, L., Park, H.W.: Full and Fractional Counting in Bibliometric Networks. Journal of Informetrics Volume 11, Issue 1, February 2017, Pages 117–120.

Lindsey, D.: Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship. Social Studies of Science, 10(1980)2, 145–162.

Maltseva, D., Batagelj, V.: Social network analysis as a field of invasions: bibliographic approach to study SNA development. Scientometrics, November 2019, Volume 121, Issue 2, pp 1085–1128

Maltseva, D., Batagelj, V.: Towards a Systematic Description of the Field Using Keywords Analysis: Main Topics in Social Networks. Scientometrics, 2020

Marsden, P.V.: Network Data and Measurement. Annual Review of Sociology, Vol. 16 (1990), pp. 435-463

Moder, J. J., & Phillips, C. R. (1970). Project management with CPM and PERT. Second edition. New York: Van Nostrand Reinhold Company.

Momeni, F., Mayr, P.: Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names. In: Fuhr N., Kovács L., Risse T., Nejdl W. (eds) Research and Advanced Technology for Digital Libraries. TPDL 2016. Lecture Notes in Computer Science, vol 9819. Springer, (2016, p. 386-391

MontyLingua: A Free, Commonsense-Enriched Natural Language Understander for English. 2004/2020. http://alumni.media.mit.edu/~hugo/montylingua/

Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 64(2001)1, 016132.

Perianes-Rodriguez, A., Waltman, L., Van Eck, N.J.: Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics, 10(2016)4, 1178-1195.

Prathap, G., Mukherjee, S.: A conservation rule for constructing bibliometric network matrices. 2016. https://arxiv.org/abs/1611.08592

Reitz, F., Hoffmann, O.: Learning from the Past: An Analysis of Person Name Corrections in the DBLP Collection and Social Network Properties of Affected Entities. In Özyer, T., Rokne, J., Wagner, G., Reuser, A.H.P. (Eds.): The influence of technology on social network analysis and mining. Springer, Vienna 2013, p. 427-453

Robertson, S.: Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60(2004)5, 503–520.

Talburt, J. (Eds.): Entity Resolution and Information Quality. Morgan Kaufmann, 2011

TePaske-King, B., Richert, N.: The Identification of Authors in the Mathematical Reviews Database. Issues in Science and Technology Librarianship No. 31 (Summer 2001), http://www.istl.org/01-summer/databases.html

Ulrichsweb: 2020. http://ulrichsweb.serialssolutions.com/

Yadav, V., Bethard, S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. October 2019, https://arxiv.org/abs/1910.11470

Windham, M.: Unstructured Data Analysis: Entity Resolution and Regular Expressions in SAS. SAS Institute, 2018