UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO
FAKULTETA ZA MATEMATIKO IN FIZIKO

ODKRIVANJE GOLJUFIJ NA OSNOVI ANALIZE SOCIALNIH MREŽ

Lovro Šubelj

Delo je pripravljeno v skladu s Pravilnikom o podeljevanju
Prešernovih nagrad študentom, pod mentorstvom
izr. prof. dr. Marka Bajca in somentorstvom doc. dr. Matjaža Kukarja

Ljubljana, 2008

Povzetek

V nalogi predstavimo nov sistem za odkrivanje goljufij v avtomobilskem zavarovalništvu. S tem mislimo na izsiljene nesreče, pri katerih posamezniki kasneje z izmišljenimi oziroma pretiranimi škodnimi zahtevki neupravičeno pridobijo sredstva od zavarovalnice. Posebno zanimive so organizirane skupine goljufov, sestavljene iz voznikov, odvetnikov, kiropraktikov, avtomehanikov, policistov ter drugih. Pri razvoju sistema se zato v večini osredotočimo na odkrivanje prav teh.

Za razliko od nekaterih drugih rešitev uporabimo pri predstavitvi podatkov grafe ali natančneje mreže. Slednje so najnaravnejša predstavitev takih podatkov, poleg tega pa omogočajo formulacijo kompleksnih relacij med entitetami, kar je ključno pri odkrivanju takih vrst goljufij. Poleg odkrivanja ključnih entitet je pomemben del sistema tudi predstavitev končnih rezultatov domenskemu ekspertu, saj je popolnoma avtomatsko odkrivanje takih goljufij v praksi nemogoče. Mreže se izkažejo tudi v tem delu sistema, saj omogočijo jasno, a hkrati razmeroma enostavno predstavitev. Sistem sicer ne zahteva označenega začetnega nabora podatkov, je razmeroma enostaven za implementacijo, dopušča vključitev poljubnega znanja o entitetah ter je odprt za številne razširitve in izboljšave.

Sistem preizkusimo na realnih podatkih, pri čemer so doseženi rezultati zelo dobri. Ker je bil vzorec nekoliko manjši in nereprezentativen, ne moremo sklepati o uspešnosti v splošnem, vendar pa smo z doseženim zelo zadovoljni. Podoben odziv smo dobili tudi s strani domenskih ekspertov slovenske zavarovalnice, tako glede rezultatov kot tudi same predstavitve znanja.

Ključne besede: odkrivanje goljufij, socialne mreže, teorija grafov, zavarovalništvo.

Abstract

In this work we present a new system for detecting automobile insurance fraud. Here fraudsters stage car accidents and than issue fake claims so they can gain funds from insurance company. Specially interesting are groups of organized individuals made up of fraudulent drivers, chiropractics, garage mechanics, police officers and others. Our system focuses mainly on detecting such groups.

In contrast to many other solutions, we use networks for the representation of data. Networks are possibly the most natural representation of accidents data, moreover, they provide straightforward formulation of complex relations between different entities. The latter is crucial for detecting such frauds. Besides detecting key entities the system also provides visualization of final results to the domain expert, as it is believed, that fully automatic automobile fraud detection is impossible. We use networks for that part of the system as well, as the representation is clear and also relatively simple. Otherwise the system doesn't require labeled data set, it is relatively simple to implement, allows imputation of arbitrary domain knowledge and can be extended and improved in many ways.

System was tested on real world data and the results were very good. We cannot estimate the efficiency in general, as the sample was smaller and unrepresentative, but we are satisfied with the results. We also got a similar response from the analytics of slovenian insurance company.

Key-words: fraud detection, social networks, graph theory, insurance.

Kazalo

Uvod

Današnji svet omogoča iznajdljivim posameznikom nemalo priložnosti ogoljufati različne institucije. Sem spadajo na primer goljufije v telekomunikacijah, bančništvu, zdravstvu, zlasti veliko pa se jih pojavlja tudi v zavarovalništvu. Pri teh posamezniki z izmišljenimi oziroma pretiranimi škodnimi zahtevki od zavarovalnice neupravičeno pridobijo sredstva, in to pogosto popolnoma neopaženo. Razlog za to je predvsem v neprimerni predstavitvi podatkov ter količini le-teh, saj se odkrivanje goljufij navadno opravlja brez učinkovite programske podpore. Slednje nas motivira k razvoju sistema za odkrivanje zavarovalniških goljufij, ali natančneje t. i. avtomobilskih goljufij. Tu gre večinoma za izsiljene prometne nesreče, kjer goljufi nato hlinijo poškodbe ter tako pretentajo lastno zdravstveno zavarovalnico ali pridobijo sredstva iz obveznega avtomobilskega zavarovanja žrtve. Take goljufije so navadno zelo drage, poleg tega pa tudi ogrožajo življenja nedolžnih udeležencev v prometu. S stališča zavarovalnic so še prav posebno zanimive organizirane skupine in sodelujoči posamezniki, saj ti predstavljajo največjo izgubo. Nemalokrat je v take goljufive skupine vpletenih tudi veliko število različnih osebkov. Poleg posameznikov, ki uprizarjajo nesreče, najdemo tudi odvetnike, kiropraktike, avtomehanike, voznike rešilnih avtomobilov, policiste ter druge, ki so med seboj povezani na različne načine. Primerna predstavitev podatkov je zato ključna za odkrivanje takih vrst goljufij.

V nalogi tako razvijemo sistem za odkrivanje sodelujočih posameznikov v avtomobilskih goljufijah. Pri tem takoj poudarimo, da sistem ni namenjen odkrivanju posameznih goljufivih nesreč ali entitet, temveč skupinam entitet, povezanih v ponavljajoče se nesreče. Poleg odkrivanja goljufivih posameznikov je del sistema tudi predstavitev končnih rezultatov domenskemu ekspertu, ki opravi nadaljnjo raziskavo. Slednje je izjemno pomembno, saj se izkaže, da je popolnoma avtomatsko odkrivanje goljufij v praksi nemogoče. Sistem, za razliko od drugih rešitev, podatke predstavi z grafi, kot ta pojem dojemamo v matematični teoriji, ali natančneje z mrežami. Te predstavljajo najnaravnejšo predstavitev, saj je temelj našega zanimanja predvsem v odnosih med različnimi entitetami, hkrati pa omogočajo jasen prikaz končnih rezultatov ekspertu.

Kot je pogosto pri realnih problemih, gre tudi v tem primeru za občutljive podatke. Le-te je navadno težko ali celo nemogoče pridobiti, sploh v velikem številu. Poleg

vsega pa so pogosto še neoznačeni[1]. Razvoj sistema za avtomatsko odkrivanje goljufij je zatorej močno otežen, saj je omejen zgolj na določene entitete ter na odkrivanje odstopanj, kar ne predstavlja vedno nujno goljufije. Kot pa bomo videli, je kljub vsemu mogoče s primerno analizo ter predstavitvijo podatkov odkriti povezave, ki bi sicer verjetno ostale neopažene.

V nadaljevanju naloge najprej v razdelku 1.1 podrobneje opišemo delovno domeno ter izpostavimo sam problem, ki ga želimo rešiti. Sledi natančnejši opis ciljev naloge (razdelek 1.2), v razdelku 1.3 pa predstavimo sorodno delo. Razdelek 2 podaja teoretično podlago socialnim mrežam ter predstavi bistvene dosežke pri analizah le-teh. Nekatere izmed njih nato uporabimo v razdelku 3, kjer opišemo predlagan sistem za odkrivanje goljufij. Sistem preizkusimo na realnem naboru podatkov, rezultati so podani v razdelku 4. Na koncu naloge sledijo še kritična ocena ter predlogi za nadaljnje delo (razdelek 5).

[1] Z izrazom označeni podatki v strojnem učenju pojmujemo tiste, pri katerih za vsak primer poznamo tudi njegovo klasifikacijo (razred). V nasprotnem primeru so podatki neoznačeni.

Poglavje 1

Goljufije v avtomobilskem zavarovanju

1.1 Problem

Avtomobilske goljufije[1] oziroma izsiljene nesreče so v zadnjem času vse bolj pogoste tudi na slovenskih cestah. Zavarovalnice opažajo, da je velik del prejetih škodnih zahtevkov pretiranih (med 20 in 30 odstotkov [20]), nekateri med njimi tudi popolnoma izmišljeni (okoli 10 odstotkov [20]). V večini primerov gre za "nesreče", ki temeljijo na naslednjem delu člena Zakona o varnosti v cestnem prometu (29. člen ZVCP-1):

"[V]arnostna razdalja mora ne glede na vozne razmere omogočati, da lahko voznik zmanjša hitrost ali ustavi in s tem prepreči trčenje, če voznik, ki vozi pred njim, zmanjša hitrost ali ustavi".

Goljufi tako iščejo situacije, v katerih lahko upravičeno sunkovito zaustavijo svoje vozilo z željo, da vozniku za njimi to ne bo uspelo. Gmotna škoda na vozilih je navadno zelo majhna, kljub temu pa goljuf ob nesreči domnevno utrpi take poškodbe, da bi morali biti obe vozili popolnoma uničeni. Značilni primeri so zvin vratne hrbtenice, udarnina glave ali morda zgolj poškodba zapestja. Ker za tak tip poškodb enostavno ni mogoče preveriti, ali poškodba zares obstaja, goljuf tako enostavno pretenta zavarovalnico. Pogosto je pri takih nesrečah v goljufovem vozilu tudi nenavadno veliko število sopotnikov, ki so podobno kot on "poškodovani". V nekaterih primerih je sopotnikov celo več kot sedežev v vozilu. Skupna cena nesreče je za zavarovalnico tako lahko ogromna, ocenjuje se, da tudi do 25000 evrov [20].

V nalogi pa se ne omejimo zgolj na take vrste goljufij. Izkaže se, da goljufi ne hlinijo venomer poškodb, saj se včasih želijo zgolj okoristiti iz sredstev za popravilo svojega vozila. Ker pri goljufijah pogosto sodelujejo tudi avtoservisi, so izplačana sredstva lahko tudi do nekajkrat višja kot sama škoda na vozilu. Velik del teh vrst nesreč pa ni

[1] Goljufija je kriminalno dejanje, ki ga nekdo namerno naredi, da bi se okoristil, bodisi z oškodovanjem ali zavajanjem koga v zmoto.

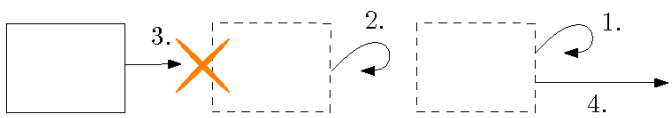
izsiljen, temveč uprizorjen. Pri njih tako sodelujeta dva goljufiva voznika, ki se med seboj zaletita. Navadno je eno od vozil starejše in manjvredno ter drugo skorajda novo in tudi dražje. V večini primerov je slednje v nesreči popolnoma uničeno, dočim na prvem praktično ni sledi. Cena popravila je v takih primerih lahko ogromna. Dogaja se tudi, da je neko vozilo udeleženo v več nesrečah s popolnoma enakimi poškodbami – goljuf tako večkrat prejme sredstva za popravilo neke poškodbe. Prav poseben sklop pa predstavljajo nesreče, ki to dejansko sploh niso. Goljufi zgolj nastavijo svoja vozila na cesto ter nato trdijo, da je prišlo do trčenja.

Kljub raznolikosti vseh takih nesreč v nadaljevanju vse označujemo z izrazom izsiljene nesreče oziroma kar goljufije.

Goljufije pestijo predvsem zavarovalnice, saj imajo le-te zaradi njih ogromne izgube. Seveda to pomeni tudi višje zavarovalnine za vse ostale ljudi. Poleg tega imajo žrtve navadno obilo dela s popravilom svojega vozila, z uveljavljanjem zavarovalnine in v nekaterih primerih celo s tožbami. Posebej je treba omeniti tudi, da predstavljajo uprizorjene nesreče veliko nevarnost za vse ostale udeležence v prometu, predvsem za starejše ljudi. Znani so posamezni primeri, kjer se take nesreče izjalovijo ter končajo s smrtjo nedolžnih udeležencev.

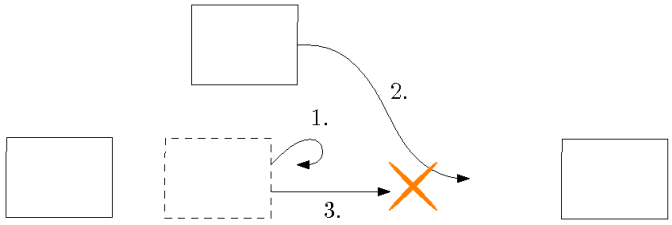
Načinov, kako izsiliti prometno nesrečo, je zelo veliko, predvsem v križiščih z gostim prometom ali v času prometnih konic. Obstajajo pa tudi znane sheme, ki se jih goljufi pogosto poslužujejo. Poleg sunkovitega zaustavljanja vozila zaradi različnih razlogov (maček na cesti, kolesar iz ozadja . . .), poznamo tudi bolj elegantne načine – trije taki so predstavljeni v nadaljevanju (povzeto po [40]).

Napadi in zbeži (*swoop and squat*): Pri tej shemi (slika 1.1) dejansko sodelujeta dva goljufa, ki vozita en za drugim. Ko se za njima pojavi neko vozilo na nezadostni varnostni razdalji, prvi goljuf sunkovito zaustavi svoje vozilo. Goljuf v drugem vozilu manever pričakuje, zato uspe zaustaviti svoje vozilo dovolj hitro, kar pa navadno ne uspe žrtvi za njima. Sedaj pride na vrsto glavna ideja sheme: prvi goljuf pobegne s kraja nesreče, pri čemer pa si drugi zaradi šoka ne zapomni registrske številke ali tipa vozila. Vsa krivda tako pade na voznika za njima, saj le-ta ni vozil na zadostni varnosti razdalji. V nekaterih izvedbah pri shemi sodelujejo tudi drugi goljufi, z namenom, da spravijo žrtev za prej omenjeni vozili.



Slika 1.1: Shema napadi in zbeži. Pravokotniki predstavljajo vozila (goljufa sta narisana s črtkano črto), s puščicami je nakazana smer vožnje (zanka pomeni zaustavljanje), mesto trčenja pa je označeno s križcem. Shema sicer poteka tako kot nakazujejo zaporedne številke.

Spusti in stisni (*drive down*): Goljuf vozi po voznem pasu, dokler ga ne prehiti neko vozilo, ki se želi vključiti pred njim. Goljuf tedaj zmanjša hitrost in navadno še nakaže vozniku, naj se vključi, nakar sunkovito pospeši ter se zaleti vanj. Goljuf nato trdi, da voznika ni želel spustiti pred seboj, oziroma da se je le-ta "vrinil" (glej sliko 1.2).



Slika 1.2: Shema spusti in stisni.

Spelji in ustavi (*start and stop*): Gre za verjetno najbolj znano shemo, kako izsiliti nesrečo. Goljuf čaka pred semaforjem z namenom, da zavije desno. Prav tako tudi žrtev za njim. Ko se na semaforju prižge zelena luč, goljuf še nekoliko počaka – na primer tako dolgo, da mu voznik za njim potrobi. S tem poskrbi, da bo le-ta nato takoj za njim (na premajhni varnostni razdalji). Goljuf spelje ter, v kolikor je kjer koli v bližini prehoda pešec oziroma kolesar, tik pred prehodom sunkovito ustavi. Vozniku za njim to navadno ne uspe, predvsem zaradi premajhne varnostne razdalje. Podobno kot pri prvi shemi tudi tu goljuf izkorišča dejstvo, da voznik ni nikoli kriv, v kolikor se nekdo zaleti v zadnji del njegovega vozila (razen, seveda, če neupravičeno zaustavi svoje vozilo).

Vse izsiljene nesreče imajo določene skupne značilnosti. Navadno se te zgodijo v poznih večernih urah ter izven večjih naselij, saj se tako močno zmanjša možnost prič. Praviloma se vedno pokliče policijo, predvsem zaradi lažjega uveljavljanja zahtevkov pri zavarovalnici. Goljufi so navadno mlajše osebe, večinoma moški, ki niso nikoli pod vplivom alkohola. Kot smo že omenili, je v takih vozilih nenavadno veliko sopotnikov, med katerimi ni otrok. Pogosto je sumljivo tudi razmerje med gmotno škodo in pa domnevnimi telesnimi poškodbami oziroma razmerje med gmotnima škodama obeh vozil.

Podajmo še delitev avtomobilskih goljufij na delno in polno goljufive (povzeto po [20]). Pri delno goljufivih gre v večini primerov zgolj za pretirane škodne zahtevke, pri čemer same nesreče niso načrtovane. Tako goljuf izkoristi priložnost (nesrečo), da uveljavi zahtevke za poškodbe, ki jih dejansko ni utrpel oziroma so plod nekih prejšnjih nesreč ali drugih dogodkov. Z izrazom polno goljufive pa označujemo tiste nesreče, kjer goljufi izsilijo ali uprizorijo nesrečo prav z namenom ogoljufati zavarovalnico. Te so navadno za samo zavarovalnico tudi dražje, vendar vseeno lažje ulovljive kot prve.

1.2 Cilj

Kot smo že omenili, so goljufi pogosto del večjih organiziranih skupin. Te poleg voznikov sestavljajo tudi goljufivi kiropraktiki, zdravniki, odvetniki, avtomehaniki, vlečne službe, zavarovalniški delavci, vozniki rešilnih avtomobilov, policisti ter drugi. Velja pa, da je take skupine zelo težko odkriti, predvsem zaradi nedostopnosti podatkov. V nalogi se zato omejimo na iskanje goljufivih skupin, sestavljenih iz osnovnih entitet, kot so vozniki, sopotniki ter do neke mere policisti. To so tudi entitete, ki jih je moč najti v policijskem zapisniku o nesreči. Sam sistem v veliki meri prilagodimo predvsem odkrivanju povezanih polno goljufivih nesreč. Odkrivanje delno goljufivih je navadno precej težje, saj gre za naključne dogodke – take goljufije bi bilo potrebno odkrivati na nivoju posamezne nesreče. Še enkrat tako poudarimo, da je cilj našega sistema odkrivati posameznike oziroma skupine, ki so udeležene v večje število sumljivih nesreč ter med seboj povezane na različne načine. Poleg tega želimo tudi, da sistem izpostavi ključne "povezave" med temi posamezniki (izpostavi ključne nesreče), še pomembneje pa je, da rezultate jasno in smiselno prikaže domenskemu ekspertu ter mu tako omogoči nadaljnjo raziskavo.

1.3 Sorodno delo

Naloga spada v širše domensko področje odkrivanja goljufij. Te se pojavljajo na mnogih področjih, še posebej v telekomunikacijah, bančništvu, internetnih storitvah ter zdravstvenem in splošnem zavarovalništvu. V literaturi tako najdemo rešitve za odkrivanje ter preprečevanje goljufij iz zelo različnih področij. Predlagana je bila uporaba nekaterih osnovnih metod strojnega učenja, nevronskih mrež, podpornih vektorjev, logistične regresije, združenih dreves (*consolidated trees*), različnih statističnih metod ter drugih [2, 3, 5, 8, 14, 21, 31, 36, 37, 38]. Analize pokažejo, da v praksi navadno nobena od njih ni značilno boljša oziroma slabša od drugih [5, 36]. Metode imajo v večini tudi dve slabi lastnosti. Primerne so predvsem za večje nabore podatkov ter zahtevajo označen začetni nabor, kar v tej domeni navadno predstavlja problem [32]. Pristop, ki ga prestavimo v nalogi, ne zahteva označenih podatkov ter je primeren za okrnjen nabor podatkov.

Kot smo nekoliko že nakazali, naša rešitev uporablja mreže. Te so bile v preteklosti plod mnogih raziskav, tej nalogi pa so sorodni predvsem pristopi k odkrivanju različnih vrst odstopanj v mrežah. Izraz odstopanje ali izstopanje v tem kontekstu ponazarja kakršno koli odstopanje od običajnega, z ozirom predvsem na strukturne lastnosti mrež. Za odkrivanje izstopajočih vozlišč so bile predlagane mere centralnosti [12], metode z naključnimi sprehodi [35] ter različni pristopi k odkrivanju osamelcev (*outliers*). Slednji se izkažejo kot uporabni v začetnih fazah sistema, vendar so za samo odkrivanje goljufij manj primerni, saj ne upoštevajo tudi statičnih[2] lastnosti vozlišč, oziroma ustreznih

[2]Z izrazom statične lastnosti vozlišča označimo tiste, ki so neodvisne od same mreže (na primer spol

entitet. Podobno velja tudi za metode razvrščanja vozlišč [6, 11] (*vertex clustering*) ter odkrivanje skupnosti v mrežah [22, 26] (*community structure*).

Našemu delu se še najbolj približajo metode, ki združijo strukturne[2] lastnosti vozlišč z njihovimi statičnimi lastnostmi. V [30] Noble in sodelavci iščejo odstopanja v mrežah z različnimi tipi vozlišč, vendar se osredotočajo na iskanje izstopajočih struktur v mreži in ne izstopajočih vozlišč. Pristop je tudi nekoliko bolj primeren za večje mreže. Neville in sodelavci [25] predlagajo iterativno klasifikacijo vozlišč, vendar se podobno kot pri mnogih drugih pristopih zahteva označen nabor podatkov. Poleg tega se strukturne lastnosti (povezave) na nek način uporabljajo zgolj za prenos rezultata klasifikacije, pri čemer ta temelji na statičnih lastnostih. Omenimo še [18] s področja ustnega marketinga (*viral marketing*), kjer je predstavljena metoda podobna našemu pristopu. V obeh primerih gre za formulacijo vpliva po mreži, a je ta v vsakem primeru nekoliko drugačna. Metoda v [18] zahteva tudi označen začetni nabor podatkov.

Zaključimo, da je večina obstoječih rešitev odkrivanja goljufij manj primerna za avtomobilske goljufije, saj imajo močne zahteve o podatkih oziroma le-te neprimerno predstavijo. Naša rešitev je drugačna, saj ne zahteva označenih podatkov. Te predstavi z mrežami, ki so bile kljub močnim temeljem teorije mrež za odkrivanje (avtomobilskih) goljufij manj uporabljane.

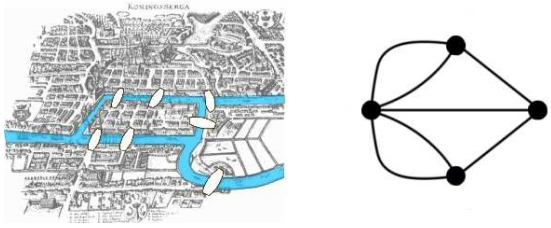
osebe, ki jo predstavlja vozlišče). Nasprotne tem so relacijske ali strukturne lastnosti (povezave vozlišča, njegova centralnost in podobno).

Poglavje 2

Teorija mrež

V nadaljevanju podamo splošen pregled analize mrež s poudarkom na tistih značilnostih in dosežkih, katere nato uporabimo pri zasnovi našega sistema.

Teorija mrež je aplikativna veda diskretne matematike, del teorije grafov. Sami začetki segajo vse do leta 1735, ko je Leonard Euler podal rešitev znanega problema sedmih mostov Königsberga (glej sliko 2.1). Slednje velja za prvi dosežek oziroma prvi pravi dokaz v teoriji grafov. Tudi sama teorija mrež se je do danes že močno razvila, tako na teoretičnem kot na praktičnem področju. Mreže se izkažejo kot nepogrešljive povsod tam, kjer nas zanimajo odnosi med različnimi entitetami oziroma še bolj natančno vzorci v teh odnosih. Primere uporabe tako najdemo v fiziki, kemiji, biologiji, sociologiji, računalništvu in informatiki ter še na mnogih drugih področjih.



Slika 2.1: Problem sedmih mostov Königsberga – ali se je moč sprehoditi po mestu na levi strani slike tako, da pri tem prečkamo vsak most natanko enkrat? Problem rešimo tako, da narišemo ustrezen graf (slika desno), in ugotovimo, da v njem ne obstaja Eulerjeva pot. [41]

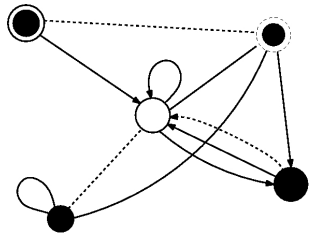
Podajmo sedaj nekoliko bolj formalno definicijo mrež[1]. Mreža N je določena z dvema množicama V in E, $N := (V, E)$. V je množica vozlišč ter E množica povezav med njimi (povezave so zaenkrat še neusmerjene).

$$V = \{v_1, v_2, \ldots v_n\} \tag{2.1}$$

[1]V literaturi se za izraz *network* uporabljata prevoda omrežje in mreža. Zaradi enotnosti v nadaljevanju striktno uporabljamo slednjega, tudi v primerih, ko bi bilo dejansko bolj primerno uporabiti izraz graf.

$$E \subseteq \{\{v_i, v_j\} | \, v_i, v_j \in V\} \tag{2.2}$$

Slednje se sklada tudi z definicijo grafov v matematični teoriji s to razliko, da tu vozlišča ter povezave opremimo še z dodatnimi lastnostmi. Tako so vozlišča in povezave lahko različnih tipov oziroma nosijo še dodatne vrednosti ali uteži (glej sliko 2.2). Kot primer navedimo mrežo prometnih nesreč, kjer vozlišča predstavljajo osebe in povezave soudeleženost pri neki prometni nesreči. Naravno bi si želeli poleg vozlišč hraniti tudi spol ali starost udeležene osebe oziroma vrsto soudeleženosti poleg povezav. Ravno to pa nam omogoča uporaba mrež.



Slika 2.2: Primer enostavne mreže z različnimi tipi vozlišč in povezav.

Zgornjo definicijo navadno še razširimo, tako da dopuščamo usmerjenost povezav. Vsaka povezava je tako urejen par vozlišč (in ne množica tako kot prej).

$$E \subseteq \{(v_i, v_j) | \, v_i, v_j \in V\} \tag{2.3}$$

Dovoljujemo tudi, da je med dvema vozliščema več paralelnih povezav ter da se povezava začne in konča v istem vozlišču – zanke (glej sliko 2.2). V teoriji grafov poznamo tudi pojem hiperpovezav, kjer povezava ni nujno dva-elementarna podmnožica vozlišč – ena povezava lahko med seboj povezuje poljubno število vozlišč. Zadnja posplošitev je sicer manj pogosta, čeprav se lahko v nekaterih primerih izkaže kot zelo koristna.

Preučevanih je bilo kar nekaj različnih mrež realnega sveta. V grobem jih delimo v naslednje štiri sklope [28]:

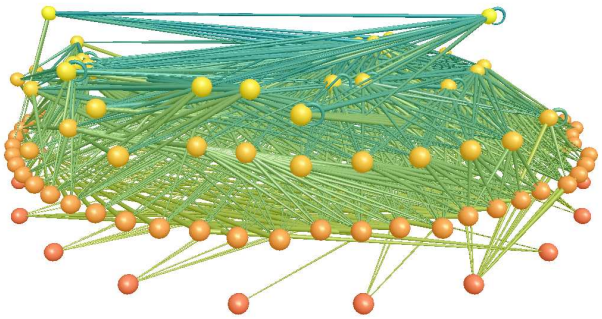
Socialne mreže: so vse tiste mreže, ki opisujejo poljubno skupino ljudi z nekimi relacijami ali interakcijami med seboj. Mednje tako spadajo tiste, ki opisujejo prijateljstva med posamezniki, partnerske odnose med podjetji, soavtorstvo člankov ter tudi mreže prometnih nesreč, ki jih obravnavamo v tej nalogi. Ta vrsta mrež je bila poleg informacijskih v preteklosti najbolj raziskovana.

Informacijske mreže: tipičen primer te vrste so mreže citiranj med znanstvenimi članki. Vozlišča predstavljajo različne članke, med katerimi so usmerjene povezave, v kolikor prvi citira drugega. Ker lahko članki citirajo zgolj pretekle članke, gre v tem primeru očitno za aciklično mrežo, kjer njena topologija oziroma struktura nazorno predstavlja informacijo, shranjeno v vozliščih (člankih). Od tod tudi

izraz informacijske mreže. Primera te vrste sta tudi mreža spletnih hiper-strani ter mreža posameznikov v omrežju vsak z vsakim (*peer to peer network*).

Tehnološke mreže: so vse mreže, ki predstavljajo umetno narejena omrežja, predvsem za oskrbo z določeno dobrino. To so na primer električno ali telefonsko omrežje, internet (fizične povezave), cestno, železniško in letalsko omrežje. Omenimo še, da ima pri takih mrežah pomembno vlogo tudi sama geografska lokacija vozlišč oziroma ustreznih entitet, kar za ostale vrste večinoma ne velja.

Biološke mreže: mreže so zelo uporabne tudi v biologiji, predvsem na področjih molekularne biologije, genetike ter nevrologije. Zanimiv primer je tudi prehrambena mreža (veriga), sestavljena iz različnih živih bitij, ki so med seboj povezana, v kolikor ustrezajo relaciji *plenilec-plen*. Primer take mreže je viden na sliki 2.3.



Slika 2.3: Prehrambena mreža živali iz jezera Little Rock, Wisconsin. Dve vozlišči (živali) sta povezani, v kolikor ustrezata relaciji *plenilec-plen*. [39]

Kljub veliki raznolikosti mrež ter področij uporabe se izkaže, da imajo vse mreže kar nekaj skupnih, večinoma strukturnih značilnosti. Te predstavljamo v naslednjem razdelku.

2.1 Skupne značilnosti

Eden najbolj znanih pojavov v teoriji mrež je gotovo t. i. učinek majhnega sveta (*small-world effect*). Nanj je med prvimi opozoril Stanley Milgram [24], ko je v šestdesetih letih prejšnjega stoletja opravil naslednji eksperiment. Opazovanim osebam je zadal nalogo, da prek pisem navežejo stik z neko drugo osebo na svetu, pri tem pa jim ni podal nobenih drugih informacij. Rezultat je bil presenetljiv, saj je večina teh pisem prišla na cilj v samo približno šestih korakih (velik del se jih je sicer izgubilo na poti).

Eksperiment tako nakaže, da je razdalja med poljubnima dvema osebama očitno zelo majhna, veliko manjša, kot bi sprva pričakovali. Iz eksperimenta se je razvil tudi izraz šest prostostnih stopenj (*six degrees of freedom*), ki se pogosto pojavlja v literaturi.

Pojav navadno merimo tako, da opazujemo povprečno razdaljo med poljubnima dvema vozliščema. Naj bo n število vozlišč v mreži ter l povprečna dolžina geodetke med poljubnim parom vozlišč v mreži (geodetka je najkrajša pot med dvema vozliščema). Velja

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d_{ij}, \tag{2.4}$$

kjer je d_{ij} dolžina geodetke med i in j oziroma 0, kadar vozlišči i in j nista povezani (mreža je tedaj sestavljena iz več komponent). Vrednost l nam poda zelo dobro oceno, ali je pojav prisoten pri določeni vrsti mrež. Večina objavljenih rezultatov to tudi potrjuje, saj l navadno ni nič večji od 6, tudi pri mrežah z nekaj tisoč oziroma nekaj milijoni vozlišč.

Pojav je z matematičnega stališča pravzaprav očiten. V večini mrež velja, da število vozlišč na razdalji d od nekega centralnega vozlišča narašča eksponentno glede na d. Iz enačbe (2.4) tedaj nemudoma sledi, da l narašča kot $\log n$. Pojav tako ni nepričakovan, v zadnjih letih pa se pojavljajo dosežki, ki nakazujejo, da je vrednost l verjetno še manjša.

Raziskovalci so veliko zanimanja posvetili tudi porazdelitvi stopenj vozlišč v realnih mrežah (stopnja vozlišča je število njegovih povezav). Izkaže se, da je ta precej drugačna od porazdelitve stopenj naključnih ali regularnih mrež (oziroma grafov). Naj bo p_k verjetnost, da je stopnja nekega naključno izbranega vozlišča enaka k. Porazdelitev stopenj vozlišč v naključni mreži, kjer je vsaka od $\frac{1}{2}n(n-1)$ možnih povezav prisotna z verjetnostjo p, je očitno binomska

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{2.5}$$

oziroma Poissonova v limiti ko $n \to \infty$. To se močno razlikuje od realnih mrež, saj se izkaže, da je v tem primeru porazdelitev navadno močno raztegnjena v desno (*right skewed*). Slednje pomeni, da obstajajo v realnih mrežah vozlišča, katerih stopnja je močno nad povprečjem, kljub vsemu pa je le-teh navadno premalo za dobro statistično ocenjevanje porazdelitve pri velikih k.

Že sama definicija verjetnosti nam zagotavlja, da p_k pada vsaj tako hitro kot k^{-1} ko $k \to \infty$. V nasprotnem primeru porazdelitev očitno ne bi bila integrabilna. V večini primerov se p_k zmanjšuje še hitreje, kot

$$p_k \sim k^{-\alpha} \tag{2.6}$$

za neko fiksno vrednost $\alpha > 1$. Take mreže se navadno označuje z izrazom *scale-free* mreže, pogosto pa rečemo tudi, da porazdelitev stopenj vozlišč zadošča potenčnemu

zakonu (*power law*) oziroma Paretovemu zakonu. Omenimo še, da se zakon kaže tudi v porazdelitveni funkciji[2] stopenj $F_K(k)$, vendar je eksponent v tem primeru enak $-(\alpha-1)$.

$$F_K(k) = \sum_{i=k}^{\infty} p_i \sim k^{-(\alpha-1)} \tag{2.7}$$

K je naključna spremenljivka, ki meri stopnjo vozlišča k. V literaturi se pojavljajo tudi mreže, kjer p_k pada še hitreje, eksponentno [1, 29]. Obstajajo tudi mreže, kjer se porazdelitev p_k ne podreja potenčnemu zakonu, temveč je le-ta na primer Gaussova ali eksponentna [1].

V primeru usmerjenih grafov se obravnava nekoliko zaplete, saj sedaj verjetnosti ocenjujemo v dvo dimenzionalnem prostoru vhodnih ter izhodnih stopenj. Zaradi enostavnosti se navadno v tem primeru ocenjuje zgolj robne verjetnosti za vhodne oziroma izhodne stopnje.

Pogosto nas zanima tudi največja stopnja vozlišča v mreži k_{max}. Za znano porazdelitev stopenj p_k velja, da je pričakovana vrednost največje stopnje enaka

$$\mathbb{E}[K_{max}] = \sum_{k=0}^{\infty}((p_k + 1 - F_K(k))^n - (1 - F_K(k))^n)k. \tag{2.8}$$

Kadar porazdelitev p_k zadošča potenčnemu zakonu, lahko k_{max} ocenimo kot

$$k_{max} \sim n^{\frac{1}{\alpha-1}}. \tag{2.9}$$

V velikem številu socialnih mrež je moč opaziti tudi pojav tranzitivnosti. Tu nas zanima, ali predpostavka, da je vozlišče v_i povezano z vozliščem v_j ter vozlišče v_j z vozliščem v_k, poveča verjetnost, da je v_i povezan tudi z v_k. Lep primer slednjega je mreža prijateljstev med osebami, saj je prijatelj prijatelja pogosto tudi naš prijatelj. Tranzitivnost navadno merimo s koeficientom razvrščanja[3] (*clustering coefficient*), definiranim kot

$$C = 3\frac{\text{število trikotnikov}}{\text{število povezanih trojic vozlišč}}, \tag{2.10}$$

kjer pa ne upoštevamo vzporednih povezav ali zank.

Opazimo lahko, da predstavlja C ravno prej omenjeno verjetnost oziroma verjetnost, da sta dve vozlišči, ki imata skupnega soseda, povezani. Izkaže se, da je koeficient v realnih mrežah navadno precej višji kot v naključno konstruiranih mrežah, sumi se celo, da se C približuje neničelni limitni vrednosti, ko $n \to \infty$ (z nekaterimi izjemami).

Kot primer povejmo, da koeficient razvrščanja za mrežo na sliki 2.2 znaša $C = \frac{9}{15}$.

Obstaja še vrsta drugih značilnosti mrež, vendar so v večini manj primerne za naše namene. Omenimo zgolj še elastičnost mrež (*resilience*), kjer preučujemo odpornost mrež na odstranitev vozlišč, ter odkrivanje skupnosti v mrežah. Področjema namenjamo nekaj pozornosti v naslednjih razdelkih, najprej pa se posvetimo konstrukciji naključnih mrež.

[2]Porazdelitveno funkcijo definiramo kot $F_K(k) = P(K \geq k)$ in ne $F_K(k) = P(K < k)$, kot je sicer v navadi.

[3]Pojma ne smemo mešati z razvrščanjem, kot ga poznamo v analizi podatkov.

2.2 Naključne mreže

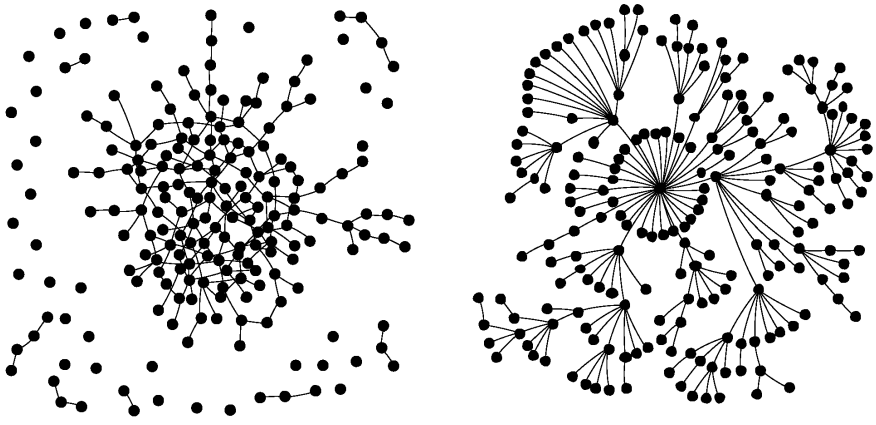
Študij naključnih mrež je predvsem pomemben s stališča razumevanja obnašanja mrež. Poleg tega modeli, ki dobro opisujejo realne mreže, ponujajo alternativni vir izgradnje naključnih podatkov. V nadaljevanju zato opišemo tipične modele ter predstavimo nekatere njihove lastnosti.

Poissonove naključne mreže

Najenostavnejši model naključnih mrež so Poissonove mreže (slika 2.4), katere so neodvisno odkrili Erdös in Rényi [10] ter Solomonoff in Rapoport [33]. Kot smo že omenili, jih dobimo tako, da med vsakim parom vozlišč postavimo povezavo z neko verjetnostjo p (neodvisno med seboj). Porazdelitev stopenj vozlišč je očitno binomska (glej enačbo (2.5)) oziroma Poissonova v limiti ko $n \to \infty$,

$$p_k \approx \frac{\lambda^k e^{-\lambda}}{k!}, \tag{2.11}$$

kjer je $\lambda = p(n-1)$. Očitno je, da se pri takem modelu vsaka izmed možnih mrež ne pojavi z enako verjetnostjo. Mreža z m povezavami se pojavi z verjetnostjo $p^m(1-p)^{\binom{n}{2}-m}$, in tako porazdelitev ni enakomerna. Obstajajo modeli, ki popravijo slednjo slabost (vsaka od mrež se pojavi z enako verjetnostjo), vendar porazdelitev stopenj še vedno ostaja Poissonova. To je tudi glavna pomanjkljivost takih modelov, saj se, kot smo videli v razdelku 2.1, porazdelitev stopenj navadno podreja potenčnemu zakonu - $p_k \sim k^{-\alpha}$ za nek $\alpha > 1$.



Slika 2.4: Dva primera naključnih mrež. Levo je mreža s Poissonovo porazdelitvijo stopenj ter desno mreža s porazdelitvijo stopenj po potenčnemu zakonu. Slednja je konstruirana po načelu prednostne povezanosti. [34]

Koeficient razvrščanja (glej enačbo (2.10)) je zaradi neodvisnosti povezav enak $C = p$, kar je navadno precej manj kot pri realnih mrežah. Poissonov model je tako neprimeren za mreže z visoko stopnjo tranzitivnosti, izkaže pa se, da model dobro posnema učinek majhnega sveta realnih mrež.

Naključne mreže po potenčnemu zakonu

Zaradi nerealistične porazdelitve stopenj prejšnjega modela je bilo veliko raziskovanja usmerjenega v iskanje naključnih mrež, ki se podrejajo potenčnemu zakonu. V nadaljevanju predstavimo dva taka modela. Oba temeljita na načelu prednostne povezanosti (*preferential attachment*), po katerem imajo vozlišča z višjo stopnjo večjo verjetnost, da bodo v nadaljevanju povezana (oziroma jim bo dodana povezava). Omenimo, da je načelo ena od razlag, zakaj se realne mreže podrejajo potenčnemu zakonu.

Prvi model (*steady-state model*) sta predlagala Eppstein in Wang [9]. Začnemo s poljubno naključno mrežo, navadno kar s Poissonovo iz prejšnjega razdelka, in r-krat ponovimo naslednje štiri korake:

1. (enakomerno) naključno izberemo vozlišče v ter njegovo povezavo (v, u) (neodvisno med seboj),
2. (enakomerno) naključno izberemo vozlišče w,
3. naključno izberemo vozlišče y, proporcionalno stopnji vozlišč (načelo prednostne povezanosti),
4. v kolikor (w, y) ni povezava v mreži, odstranimo (v, u) ter dodamo (w, y).

Opazimo, da predstavlja zgornji postopek (aperiodično) Markovsko verigo z neko limitno porazdelitvijo. V kolikor je parameter r dovolj velik, bo dobljena mreža blizu te porazdelitve, ne glede na to, s kakšno mrežo smo začeli. Empirično je bilo pokazano [9], da se tako zgrajene mreže podrejajo potenčnemu zakonu, vendar formalni dokaz ni znan.

Drug model je še nekoliko bolj enostaven, sicer zelo znan model Barabásija in Alberta [4]. Dejansko gre za naključno mrežo, ki se razvija skozi čas, vendar je ta vidik za nas manj pomemben. Začnemo z nekim manjšim številom vozlišč n_0 ter nato na vsakem koraku dodamo novo vozlišče, ki ga povežemo z $m \leq n_0$ naključno izbranimi vozlišči. Slednja izbiramo po načelu prednostne povezanosti – vsako vozlišče je izbrano z verjetnostjo, proporcionalno njegovi stopnji. Po r korakih tako dobimo naključno mrežo z $n_0 + r$ vozlišči in mr povezavami. Možno je pokazati, da se tako konstruirana mreža podreja potenčnemu zakonu oziroma natančneje $p_k \sim k^{-3}$.

Primer mreže s porazdelitvijo stopenj po potenčnemu zakonu je viden na sliki 2.4.

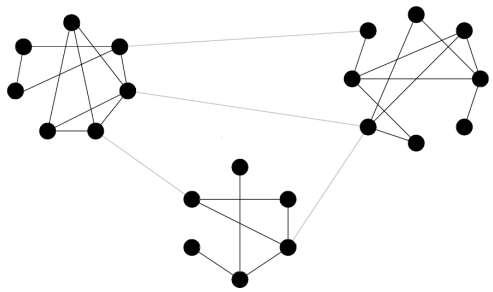
Drugi modeli

Poleg predstavljenih obstaja še cela vrsta drugih modelov. Od takih, ki konstruirajo mreže po potenčnemu zakonu (glej prejšnji razdelek), do bolj splošnih za poljubno porazdelitev stopenj vozlišč. Obstajajo tudi različni modeli za usmerjene in dvodelne mreže ter modeli, ki se razvijajo skozi čas. Omenimo še model majhnega sveta (*small-world model*), primeren za mreže, pri katerih ima pomembno vlogo geografska lokacija vozlišč.

Zanimive so tudi eksponentne naključne mreže oziroma p^* modeli v bolj splošni obliki. Pri slednjih opišemo mrežo z nekimi merljivimi lastnostmi (na primer Hamiltonskost mreže) in jo naključno konstruiramo z verjetnostjo, proporcionalno neki linearni kombinaciji omenjenih lastnosti. Želja je, da bi bilo s takimi modeli moč bolje razumeti pojave, kot je tranzitivnost mrež, saj le-te v tem trenutku še ne znamo vključiti v splošne naključne mreže.

2.3 Odkrivanje skupnosti

Domneva se, da večina predvsem socialnih mrež vsebuje t. i. skupnosti (*communities*). Z izrazom označujemo množice vozlišč z velikim številom povezav med njimi (znotraj skupnosti) ter malo povezavami med samimi skupnostmi (glej sliko 2.5). Slednje je za nas še posebej zanimivo, saj lahko brez škode predpostavimo, da bodo v mrežah nesreč te skupnosti ustrezale ravno skupinam goljufivih posameznikov. Seveda ne velja, da bodo vsa vozlišča (posamezniki) v takih "goljufivih" skupnostih goljufiva. Goljufiva skupnost bo v večini primerov nadmnožica goljufive skupine.

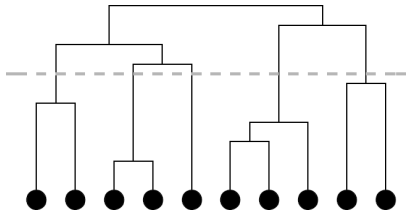


Slika 2.5: Primer mreže s tremi skupnostmi (povezave med skupnostmi so narisane svetleje). [26]

Kadar je obravnavana mreža dovolj redka (nepovezana), bodo skupnosti pogosto ustrezale že kar povezanim komponentam mreže. Seveda pa je to manj pogosto in na nek način nakazuje na pomanjkanje podatkov.

Standardna metoda iskanja skupnosti se imenuje hierarhično razvrščanje (*hierarchical clustering*). Pri tem vsaki izmed $\binom{n}{2}$ možnih povezav najprej priredimo neko utež oziroma jakost povezanosti (če povezave sicer ni v mreži, ji priredimo jakost 0). Nato

iterativno dodajamo po eno povezavo tako (začnemo z vsemi vozlišči brez povezav), da zmanjšujemo skupno jakost povezanosti med vozlišči. Postopek lahko ustavimo na poljubnem koraku oziroma tedaj, ko je trenutna mreža sestavljena iz želenega števila skupnosti (povezanih komponent). Navadno predstavimo rezultat z dendrogramom, ki je zgolj grafičen prikaz združevanja v zgornjem postopku. Primer dendrograma je viden na sliki 2.6. Opazimo, da lahko sedaj dobimo poljubno število skupnosti tako, da zgolj prerežemo dendogram na ustrezni višini (na sliki 2.6 tako dobimo 5 skupnosti).



Slika 2.6: Primer dendrograma, ki prikazuje rezultat hierarhičnega razvrščanja. Višina povezave med dvema vozliščema (oziroma skupnostima) nam pove, na katerem koraku smo ju združili – izraža njuno podobnost. [28]

Jakost povezanosti je seveda lahko poljubna metrika. Večinoma se uporabi kakšna utežena razdalja med vozlišči, število različnih poti ali velikost najmanjšega prereza. V zadnjem času je bilo veliko uspeha tudi z uporabo vmesnosti povezav (*edge betweenness*). Vmesnost povezave e je definirana kot delež najkrajših poti med vsemi pari vozlišč, ki gredo skozi e, oziroma

$$B(e) := \frac{|\{(v_i, v_j)|v_i, v_j \in V \wedge i < j \wedge e \in g(v_i, v_j)\}|}{\binom{n}{2}}, \tag{2.12}$$

kjer je $g(v_i, v_j)$ geodetka med vozliščema v_i in v_j. Ker je navadno med skupnostmi zgolj nekaj povezav, gre skozi te velik del najkrajših poti. Povezave med skupnostmi imajo tako visoko vmesnost $B(e)$. To nam omogoči naslednji preprost postopek za odkrivanje skupnosti [15]: dokler obstaja kakšna povezava v mreži, odstrani tisto, ki ima največjo vmesnost. Podobno kot pri hierarhičnem razvrščanju predstavimo rezultat z dendogramom, ki ga nato prerežemo na ustrezni višini.

Obstaja še kar nekaj v večini kompleksnejših metod za iskanje skupnosti. Sem spadajo metode spektralne delitve (*spectral partitioning*), metode na osnovi pretoka oziroma prereza, informacijsko-teoretične metode ter druge metode razvrščanja vozlišč. Poleg omenjenega lahko skupnosti iščemo tudi z razvrščanjem, kot ta pojem dojemamo v analizi podatkov – iskanje skupin (gruč) v nekem k-dimenzionalnem prostoru. Seveda je potrebno v tem primeru vsa vozlišča predstaviti v atributnem jeziku, oziroma je potrebno vsako izmed vozlišč opisati z nekim določenim naborom atributov. Pogosto se izkaže, da zadnja preslikava ni enostavna, zato raje uporabimo metode, razvite prav za mreže.

2.4 Odkrivanje odstopanj

Sam naslov razdelka zahteva obrazložitev. Z besedo odstopanje[4] ali izstopanje v tem kontekstu ponazarjamo kakršno koli odstopanje od običajnega. Izstopajoče je tako vsako vozlišče z nenavadno visoko stopnjo, veliko centralnostjo ali vmesnostjo kot tudi vozlišče, ki ima pomembno vlogo pri elastičnosti mreže. Pogosto se izkažejo kot zanimive tudi izstopajoče povezave ter (pod)strukture znotraj mreže, vendar pa se v tem razdelku v večini posvetimo zgolj izstopajočim vozliščem ter nekaterim metodam za njihovo iskanje.

Najprej omenimo pristope, ki temeljijo na naključnih sprehodih oziroma lastnih vrednostih. Sem spada znan algoritem PageRank [7], ki izračuna rang vozlišč na podlagi njihove "pomembnosti" oziroma centralnosti v mreži. Primeren je predvsem za določanje pomembnosti posameznih spletnih strani. Algoritem HITS [19] za vsako vozlišče v usmerjeni mreži iterativno izračuna dve vrednosti (*authority and hub score*), ki predstavljata vlogo vozlišča. Tipičen primer uporabe je zopet mreža spletnih strani. Odkrivanje odstopanj je bilo raziskovano tudi v primeru dvodelnih grafov. V [35] avtorji predlagajo metodo, pri kateri so izstopajoča vozlišča tista, ki med seboj povezujejo nepodobna vozlišča, kjer se podobnost določi s pomočjo naključnih sprehodov po mreži. Podobno idejo predstavlja tudi algoritem SimRank [16].

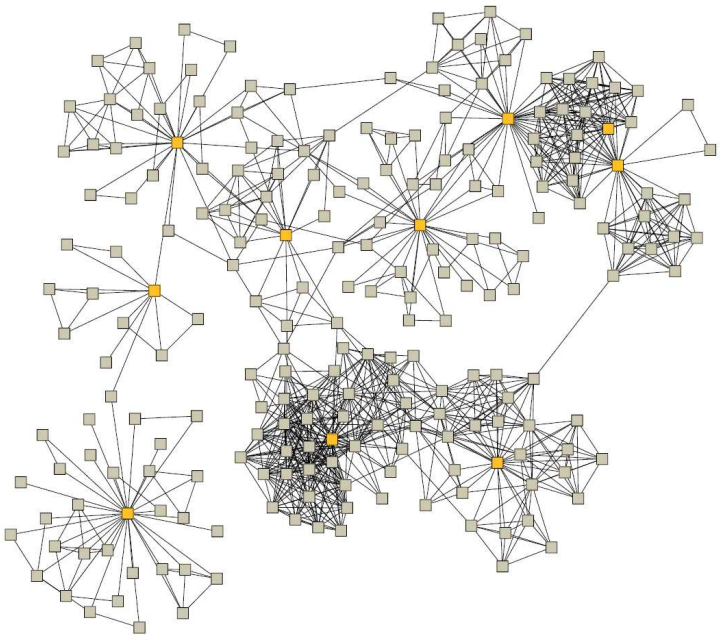
Centralnost oziroma pomembnost vozlišč je moč določati tudi na bolj enostavne načine. Dobra ocena je že kar sama stopnja vozlišča, saj imajo centralna vozlišča pogosto tudi visoko stopnjo. Ker slednje ni pravilo, je morda nekoliko boljša rešitev izpostaviti zgolj vozlišča, ki ležijo v minimalnem pokritju mreže (minimalno pokritje je najmanjša množica vozlišč, za katera velja, da ima vsaka povezava mreže vsaj eno svoje vozlišče v pokritju). Zadnji pristop je seveda nekoliko problematičen, saj je pokritje NP-težak problem.

Veliko število drugih metod za določanje centralnosti temelji na analizi (najkrajših) poti med posameznimi vozlišči. Podobno kot smo definirali vmesnost povezav lahko definiramo tudi vmesno centralnost vozlišč (*vertex betweenness centrality*). Vmesna centralnost vozlišča v je število najkrajših poti med poljubnim parom vozlišč, ki gredo skozi to vozlišče.

$$B(v) := \frac{|\{(v_i, v_j)|v_i, v_j \in V \wedge i < j \wedge v \in g(v_i, v_j)\}|}{\binom{n}{2}} \tag{2.13}$$

Mera da oceno pomembnosti vozlišč, temelji pa na dejstvu, da gre skozi vmesno centralna vozlišča veliko število najkrajših poti, precej več kot skozi neko izolirano vozlišče. Mera ima tudi veliko različic, kot je razdaljna centralnost vozlišča (*vertex distance centrality*), definirana kot povprečna razdalja do vseh vozlišč v mreži. Slednje nam da oceno dejanske (ne vmesne) centralnosti vozlišč. Na sliki 2.7 lahko vidimo mrežo z označenimi vozlišči z visoko centralnostjo.

[4] Anomalija.



Slika 2.7: Vozlišča z visoko centralnostjo so označena oranžno. [27]

Poglavje 3

Odkrivanje goljufij v mrežah nesreč

V razdelku 1.3 smo spoznali različne pristope k odkrivanju goljufij v avtomobilskem zavarovalništvu. Ti v večini predstavijo podatke z običajnim atributnim zapisom, kar močno oteži, če ne celo onemogoči odkrivanje različnih skupin sodelujočih goljufov – goljufi so navadno med seboj povezani na kompleksne načine, ki jih je v splošnem težko opisati v atributnem jeziku. Naš sistem je drugačen od omenjenih, saj za predstavitev podatkov uporablja mreže (razdelek 2). Te predstavljajo najnaravnejši opis različnih povezav med entitetami, kar olajša odkrivanje, poleg tega pri pretvorbi podatkov ne pride do nikakršne izgube informacije. Sistem za razliko od ostalih ne zahteva označenega začetnega nabora podatkov ter je primeren predvsem v primerih, ko imamo na voljo manj podatkov. Kot smo videli v razdelku 1, je to v tej domeni pogosto. Verjame se tudi, da je popolnoma avtomatsko odkrivanje goljufij v praksi nemogoče, zato v sistemu na koncu poleg smiselne predstavitve dobljenih rezultatov ponudimo tudi možnost za usmerjeno nadaljnjo raziskavo.

Predlagan sistem kot vhodne podatke uporablja podatke iz policijskih zapisnikov o nesrečah. Te podrobneje predstavimo v razdelku 3.1, v nadaljevanju pa podajamo oris celotnega sistema, ki ga razdelimo na tri dele.

V prvem delu iz policijskih zapisnikov najprej izločimo posamezne entitete (udeleženci, policisti, nesreče in vozila). Te nato povežemo v mreže glede na povezanost v samih nesrečah oziroma glede na neke skupne lastnosti. Zgradimo več vrst mrež, saj sistem v nadaljevanju za različne namene uporablja različne mreže. Na koncu mreže po potrebi tudi nekoliko poenostavimo tako, da jih razbijemo glede na skupnosti, ki se pojavljajo znotraj njih. Kot bomo videli, slednje storimo brez izgube za splošnost.

Mreže, ki jih dobimo kot rezultat prvega dela sistema, so dejansko sestavljene iz več manjših povezanih komponent. Vsaka taka komponenta opisuje skupino povezanih entitet. Namen drugega dela je identificirati sumljive komponente znotraj mreže, pri čemer se osredotočimo predvsem na strukturne lastnosti posameznih komponent. V ta namen za vsako komponento konstruiramo naključne mreže ter na podlagi njih ocenimo, ali določena komponenta izstopa oziroma je sumljiva. Nesumljive komponente na koncu tega dela sistema zavržemo.

V zadnjem, tretjem, delu sistema za vsako sumljivo komponento poiščemo še ključne entitete znotraj nje – izpostavimo sumljive (goljufive) skupine posameznikov. V ta namen uporabimo preprosto iterativno metodo, ki zna upoštevati tako relacijske kot statične lastnosti entitet. Metoda izračuna za vsako entiteto stopnjo sumljivosti, ki jo lahko nato uporabimo za usmerjeno nadaljnjo raziskavo.

Vsi trije deli sistema so zaporedoma opisani v razdelkih 3.2.1, 3.2.2, 3.2.3, katerim sledi še razdelek o predstavitvi rezultatov in uporabi le-teh pri nadaljnji raziskavi (razdelek 3.3). Kot pa smo omenili že prej, v nadaljevanju najprej povemo nekaj več o samih podatkih v policijskih zapisnikih.

3.1 O podatkih

Sistem osnujemo zgolj na podatkih, ki jih je moč pridobiti iz policijskih zapisnikov o nesrečah. Slednje se v veliko primerih tudi sklada s stanjem v realnem svetu, saj zaradi nedostopnosti podatkov zavarovalnice mnogokrat nakazujejo sredstva osebam, o katerih ne vedo nič drugega kot tisto, kar je vsebovano v samem zapisniku. Sicer so zapisniki pol strukturirana besedila, ki vsebujejo osnovne podatke o udeležencih nesreče, vozilih ter o sami nesreči. V večini primerov so znana tudi imena policistov, ki so nesrečo obravnavali, redko tudi morebitne priče. V nadaljevanju podamo podrobnejši opis podatkov, ki jih poznamo za posamezno entiteto.

Policisti: znana so zgolj imena policistov.

Udeleženci: poleg imena poznamo tudi spol osebe, rojstni datum, stalni naslov in državljanstvo. Seveda je poznana tudi vloga udeleženca v posamezni nesreči (povzročitelj oziroma oškodovanec ter voznik oziroma sopotnik).

Vozila: registrska številka, znamka ter model vozila. Navadno sta znana tudi zavarovalnica, pri kateri je vozilo obvezno zavarovano, ter lastnik vozila.

Dogodek (nesreča): čas in kraj nesreče, opis poteka ter nestrokovna ocena policistov o vrednosti gmotne škode na vozilih. Pogosto zapisnik vsebuje tudi opis poškodb, ki so jih utrpeli udeleženci.

Določeni podatki so za naše namene manj uporabni, predvsem zaradi nekonsistentnosti oziroma pomanjkanja drugih. Tako na primer kraj nesreče navadno ni enolično podan, saj vsak policist lokacijo nesreče opredeli na svoj način. Podobno velja za oceno gmotne škode na vozilih ter morda tudi opise poškodb. Tudi znamka in model vozila sta večinoma manj uporabna, saj bi nas zanimal kvečjemu cenovni razred vozila ali pa število sedežev, kar pa iz teh podatkov ne znamo določiti. Na drugi strani se izkaže tudi, da nekateri podatki ne nosijo nobene informacije[1]. Tako je lastnik vozila največkrat

[1]Glede na naše podatke.

voznik sam, državljanstva udeležencev so skorajda vedno slovenska, stalni naslovi so praviloma različni.

Za namene odkrivanja goljufij je tako smiselno uporabiti zgolj določene podatke oziroma atribute. Nekateri izmed njih so statične lastnosti entitet (na primer spol osebe), drugi predstavljajo relacije med entitetami – relacijske lastnosti (na primer relacija med voznikom ter ustrezno nesrečo). Vsi atributi so predstavljeni v tabeli 3.1, njihov tip je razviden iz konteksta.

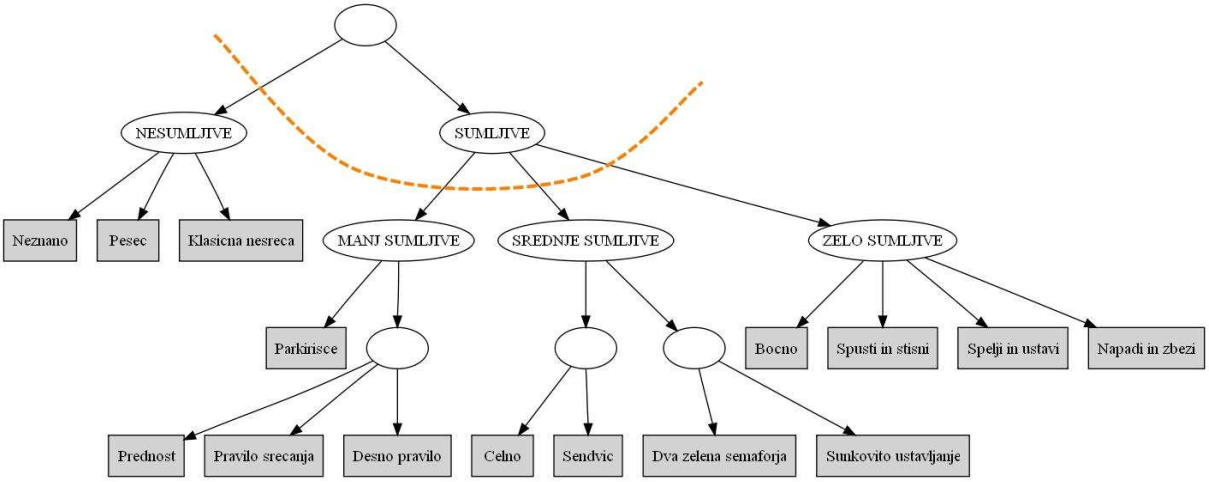
Entiteta	**Ime atributa**	**Opis atributa**
Policist	*ime*	atribut služi zgolj kot identifikacija entitete
Udeleženec	*ime*	identifikacija entitete
	spol	spol udeleženca
	starost	starost udeleženca
Vozilo	*registrska številka*	identifikacija entitete
Nesreča	*številka*	zaporedna številka nesreče (identifikacija)
	čas	čas nesreče
	sumljivost	sumljivost nesreče
	visoka gmotna škoda	ocena gmotne škode na vozilih (binarni atribut)
	sumljive poškodbe	obstoj sumljivih poškodb (binarni atribut)
	voznik povzročitelj	povzročitelj nesreče
	voznik oškodovanec	oškodovanec v nesreči
	sopotniki povzročitelja	sopotniki povzročitelja nesreče
	sopotniki oškodovanca	sopotniki oškodovanca v nesreči
	vozilo povzročitelj	vozilo povzročitelja nesreče
	vozilo oškodovanec	vozilo oškodovanca v nesreči
	policisti	policisti na kraju nesreče

Tabela 3.1: Predstavitev atributov entitet, ki jih uporabimo pri sistemu za odkrivanje goljufij.

Dva izmed atributov zahtevata še dodatno razlago. Za vsako nesrečo predhodno določimo njen tip oziroma vrsto. Nekatere vrste nesreč smo spoznali že v razdelku 1 pri opisu standardnih shem za uprizarjanje nesreč, druge je moč videti na sliki 3.1. Opisa posameznih vrst na tem mestu ne podajamo, povejmo pa, da gre za klasifikacijo nesreč na podlagi opisa policistov, pri čemer upoštevamo tudi nekatere druge dejavnike (na primer sledi zaviranja na cestišču, vidljivost ...). *Sumljivost* nesreče tako določimo na podlagi vrste nesreče – za namene naloge ustvarimo hierarhijo (drevo) sumljivosti različnih vrst nesreč, ki je predstavljena na sliki 3.1. Drevo ustrezno prerežemo (glej sliko 3.1), in tako postane *sumljivost* nesreče atribut s štirimi vrednostmi (vrednosti so zaporedoma enake 0.33, 0.50, 0.75 in 1.00).

V atribut lahko sicer s pomočjo kompleksne hierarhije nesreč vnesemo veliko količino domenskega znanja. Tudi zrnatost atributa bi bila v tem primeru lahko veliko

večja.



Slika 3.1: Hierarhija sumljivosti različnih vrst nesreč. Listi drevesa predstavljajo različne vrste nesreč – najmanj sumljive nesreče so na skrajni levi, najbolj pa na skrajni desni. Drevo prerežemo kot prikazuje slika, s čimer razdelimo nesreče v štiri razrede.

Razlago zahteva še binarni atribut *sumljive poškodbe*. Sem spadajo vse tipične poškodbe pri goljufijah (na primer zvin vratne hrbtenice), kjer praviloma poškodovanec zavrne zdravniško pomoč. Navadno v zapisniku piše, da je udeleženec *"stokal o bolečinah v vratu, a zavrnil zdravniško pomoč"*. V kolikor pri nesreči pride do težjih poškodb oziroma je kateri od udeležencev z reševalnim vozilom odpeljan v bolnišnico, je vrednost tega atributa enaka 0 (neresnično).

Opazimo, da sam model podatkov ni primeren za nesreče, kjer sta kriva oba udeležena voznika oziroma krivde ni bilo moč dokazati nobenemu od njiju. Model zato nekoliko posplošimo in dovoljujemo tudi take nesreče; vsaki nesreči v ta namen določimo še nek dodaten atribut oziroma nekoliko spremenimo predstavitev (uvedemo pojem skupine, ki predstavlja voznika, vozilo ter sopotnike). Kot bomo videli v naslednjem razdelku, krivdo v vsakem primeru ponazorimo s smerjo povezav med vozlišči (entitetami) v mrežah nesreč.

3.2 Sistem

3.2.1 Predstavitev z mrežami

Relacijske atribute iz prejšnjega razdelka najnaravneje predstavimo kot povezave med ustreznimi entitetami. Tako dobimo usmerjene mreže nesreč, kjer vozlišča predstavljajo same entitete, povezave pa različne relacije med njimi. Vsekakor lahko na tak način dobimo veliko število različnih mrež, predvsem odvisno od tega, katere entitete

vključimo v mreže in kako jih dejansko povežemo med seboj. Tudi sistem za odkrivanje goljufij v različnih delih uporablja različne mreže.

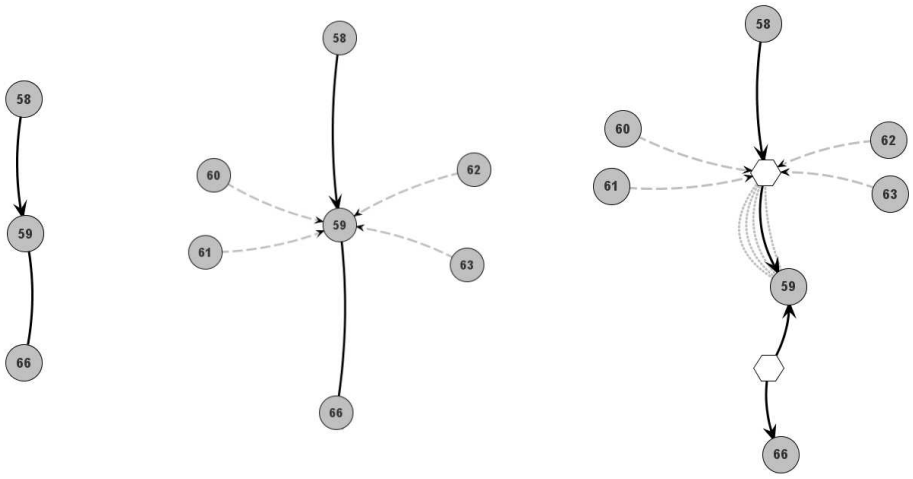
Mreža voznikov. Najpreprostejšo mrežo dobimo tako, da med seboj povežemo zgolj voznika, ki sta bila udeležena v določeni nesreči (slika 3.2). Nesreče so tako ponazorjene s povezavami, kjer njihova usmerjenost predstavlja krivdo – povezava se začne v krivem vozniku in konča v oškodovanem. V primerih, ko krivda ni jasna, je povezava neusmerjena. V taki mreži tako ni udeležencev, ki so bili venomer sopotniki, podobno tudi ne vključimo policistov ter vozil.

Mreža sopotnikov. V mrežo voznikov dodamo še sopotnike, ki jih povežemo z voznikom vozila. Povezave so podobno kot prej usmerjene (začnejo se v sopotniku ter končajo v vozniku), a seveda drugačnega tipa kot povezave med vozniki. Tako mrežo imenujemo mreža sopotnikov (glej sliko 3.2), uporabimo pa jo v drugem delu sistema.

Mreža nesreč. Glavna slabost mreže sopotnikov je, da v večini primerov ni jasno število sopotnikov v posamezni nesreči. Če je nek voznik sodeloval pri dveh nesrečah ter je bil vsakič z njim v vozilu en sopotnik, je mreža sopotnikov popolnoma identična tisti, ki bi jo dobili, če bi bila oba omenjena sopotnika prisotna pri isti nesreči. Slednja je s stališča goljufij seveda bolj zanimiva (glej razdelek 1). V mrežo sopotnikov zato dodamo še dodatna vozlišča, ki predstavljajo nesreče. Tako dobimo mrežo nesreč (slika 3.2), ki je zelo podobna prvotni, a s to razliko, da sta sedaj voznika oziroma sopotnik in voznik povezana preko vozlišča, ki predstavlja ustrezno nesrečo. Usmerjenost in vrsta povezav ostajata enaka kot prej. Med voznika in nesrečo pa postavimo še dodatne neusmerjene povezave, katerih število je enako številu sopotnikov pri nesreči. Kljub tem dodatnim povezavam seveda v večini primerov še vedno ni jasno, h kateremu vozniku spada nek sopotnik, vendar, kot bomo videli v razdelku o zadnjem delu sistema 3.2.3, taka predstavitev zadošča za naše namene.

Mreža zaenkrat še ne vključuje udeleženih vozil. Slednjih pa ne vključimo kot samostojna vozlišča, na primer med voznike in nesreče, saj je naš prvotni namen odkrivati sumljive udeležence in do neke mere tudi nesreče, ne pa sumljiva vozila. Uporaba posebnih vozlišč za sama vozila je zato nepotrebna. Dodaten razlog za to je tudi, da v kolikor se nekdo zaleti večkrat, v splošnem ni nič bolj sumljivo, če to stori z različnimi ali z istim vozilom. Sumljivo se zdi zgolj, v kolikor neko vozilo v več nesrečah vozi več različnih voznikov. Vozila zato vključimo v mrežo tako, da le povežemo nesreče, pri katerih se pojavi isto vozilo, a ga vozi drug voznik. Pri tem še dodatno zahtevamo, da spadata obe nesreči v isto povezano komponento mreže. Slednje nam zagotavlja, da bo vsaki povezani komponenti v mreži sopotnikov ustrezala natanko ena komponenta v mreži nesreč in obratno.

V mrežo bi lahko enostavno vključili tudi policiste tako, da jih dodamo kot posebna vozlišča oziroma zgolj povežemo ustrezne nesreče, kot smo to storili pri vozilih. Vendar



Slika 3.2: Različne vrste mrež, ki opisujejo nesreče – levo je mreža voznikov, na sredini mreža sopotnikov ter desno mreža nesreč. Okrogla vozlišča predstavljajo udeležence, šestkotniki pa nesreče. Povezave, ki ustrezajo voznikom, so narisane s polno črto, tiste, ki ustrezajo sopotnikom, pa s črtkano črto. Pikčaste povezave med voznikom in nesrečo so dodatne povezave, ki smo jih dodali zaradi sopotnikov v vozilu.

tega iz več razlogov ne storimo. Namreč, o policistih ne vemo popolnoma nič drugega kot to, da so bili prisotni pri določenih nesrečah. Pri tem posebej poudarimo, da te nesreče ne opredeljujejo policistov tako kot ostale udeležence. V nadaljevanju bomo tudi videli, da pri odkrivanju goljufij na nek način predpostavimo, da so nam znane vse nesreče, pri katerih je sodelovala določena entiteta. Slednjega seveda ne moremo predpostaviti za policiste. Tudi sama predpostavka o neki močni povezanosti med nesrečama, pri katerih je sodeloval isti policist, se zdi pri takem pomanjkanju ostalih podatkov nekoliko naivna. Zaradi vseh naštetih razlogov tako ne vključimo policistov v mrežo nesreč ter jih ne obravnavamo pri odkrivanju goljufij, seveda pa lahko kljub temu kasneje posredno ocenimo njihovo sumljivost iz samih nesreč oziroma udeležencev.

V prvem delu sistema tako najprej iz podatkov zgradimo ustrezne mreže. Za namene odkrivanja goljufij potrebujemo zgolj mrežo sopotnikov ter mrežo nesreč, za kasnejšo vizualizacijo oziroma predstavitev znanja pa še nekatere druge. Vendar slednjih navadno ne potrebujemo v celoti, zato te mreže po potrebi zgradimo šele na koncu (glej razdelek 3.3).

Vsaka od mrež je dejansko sestavljena iz več povezanih komponent, ki opisujejo skupine povezanih entitet. Kot smo omenili že prej, sama konstrukcija obeh vrst mrež zagotavlja, da obstaja bijektivna relacija med komponentami mreže sopotnikov in komponentami mreže nesreč. To je zelo pomembno, saj sistem v naslednjem delu uporablja eno vrsto mrež in kasneje drugo, v obeh primerih pa dejansko dela z istim

naborom podatkov.

Poleg gradnje mrež je del prvega dela sistema tudi dodatna delitev komponent mreže v manjše, v kolikor je to seveda potrebno. Slednje storimo predvsem z namenom, da olajšamo delo nadaljnjim fazam sistema ter tudi zaradi dejstva, da bodo goljufive skupine posameznikov verjetno sestavljene zgolj iz manjšega števila oseb (glej razdelek 1.1). Mreže poenostavimo na naslednji način. Vsako komponento mreže rekurzivno delimo tako, da na vsakem koraku odstranimo povezavo z največjo vmesnostjo, kot smo le-to definirali v razdelku 2.3. Komponenta tako navadno razpade na dve manjši. Ker imajo vse odstranjene povezave visoko vmesnost, si zagotovimo, da odstranjujemo zgolj povezave med morebitnimi skupnostmi, ne pa tudi povezave znotraj njih – spomnimo se, da gre skozi povezave med skupnostmi veliko večje število najkrajših poti kot čez povezave znotraj skupnosti, vmesnost pa meri ravno število takih poti. Komponente mreže tako poenostavimo brez izgube pri odkrivanju. Sam postopek prekinemo, ko so dobljene komponente dovolj majhne oziroma opisujejo dovolj majhno množico nesreč.

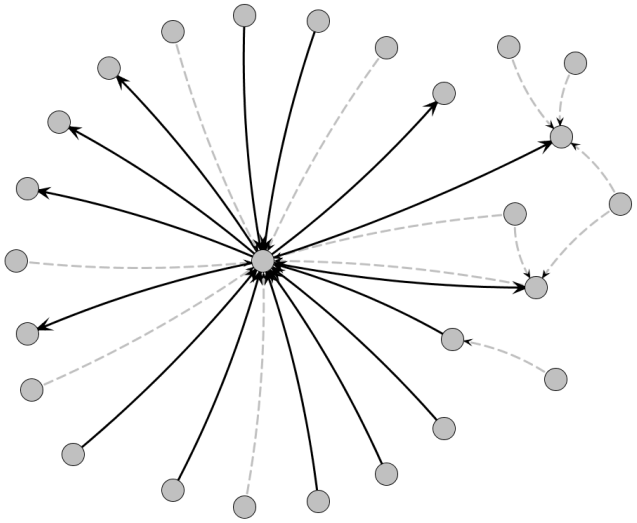
Delitev komponent dejansko opravimo nad mrežo sopotnikov, za zagotavljanje prej omenjene bijektivne relacije pa je potrebno mreži na vsakem koraku tudi nekoliko popraviti. V kolikor smo odstranili povezavo, ki predstavlja sopotnika, storimo podobno tudi v mreži nesreč. V nasprotnem primeru smo odstranili povezavo, ki predstavlja nesrečo. Tedaj je potrebno v mreži sopotnikov odstraniti tudi vse tiste, ki predstavljajo sopotnike pri tej nesreči, ter v mreži nesreč odstraniti vozlišče, ki ustreza nesreči. V tem primeru lahko komponenta razpade tudi na več kot dve manjši. Večini korakov se sicer lahko izognemo tako, da enostavno zgradimo mreže nesreč naknadno.

3.2.2 Identifikacija sumljivih komponent

Razdelek opisuje drugi del sistema, v katerem identificiramo sumljive komponente znotraj mrež, konstruiranih v prejšnjem delu (razdelek 3.2.1). Ta del sistema v celoti uporablja zgolj mrežo sopotnikov, ki je enostavnejša, a nosi vso potrebno informacijo.

Vsaka komponenta mreže predstavlja neko skupino nesreč in udeležencev, pri čimer so nekatere od teh skupin goljufive. Na podlagi analize podatkov ter znanja domenskih ekspertov lahko izpostavimo lastnosti komponent, ki ustrezajo goljufivim skupinam (v nadaljevanju goljufiva komponenta). Take komponente so veliko večje od negoljufivih, tako s stališča vozlišč kot tudi povezav (oziroma udeležencev in nesreč). V večini gre za razmeroma sumljive nesreče, razmerje med številom udeležencev in številom nesreč pa je zelo majhno. Poleg tega obstajajo tudi značilne strukturne lastnosti takih komponent. Na sliki 3.3 lahko vidimo komponento mreže, ki vsebuje večino omenjenih lastnosti.

Vsaka goljufiva komponenta gotovo vsebuje vsaj nekatere od teh lastnosti, zato jih je moč odkrivati na razmeroma preprost način. Natančneje, za vsako komponento lahko zgolj ocenimo, ali se ta glede na opisane lastnosti značilno razlikuje od običajnega (naključnega) oziroma presega neke mejne vrednosti. V kolikor komponenta po večini lastnosti izstopa, jo proglasimo kot sumljivo, v nasprotnem primeru pa zavržemo. Če



Slika 3.3: Komponenta mreže sopotnikov z večino lastnosti, ki so značilne za goljufive komponente.

predpostavimo, da lastnosti dobro opisujejo goljufive komponente oziroma jih dobro ločijo od negoljufivih, jih bomo na ta način gotovo identificirali.

Naj bo K komponenta mreže ter S_i preslikava, ki ustreza neki i-ti lastnosti, definirana kot

$$S_i(K) = \begin{cases} 1 & \text{komponenta } K \text{ glede na } i\text{-to lastnost izstopa} \\ 0 & \text{sicer} \end{cases}.$$

V drugem delu sistema tako znotraj mreže kot sumljive identificiramo tiste komponente K, za katere velja

$$S(K) = \sum_{i=1}^{h} S_i(K) \geq \frac{h}{2}, \tag{3.1}$$

kjer je h število opazovanih lastnosti. Komponente, kjer je $S(K) < \frac{h}{2}$, zavržemo. Na koncu tega dela tako dobimo množico sumljivih komponent, za katere predpostavljamo, da vsebujejo goljufive skupine. Slednje predstavlja tudi vhodne podatke za zadnji del sistema.

Predstavimo sedaj še dejanskih 5 lastnosti oziroma preslikav S_i, ki jih uporabimo za odkrivanje goljufivih komponent.

1) Prva lastnost temelji na dejstvu, da je razmerje med številom udeležencev in številom nesreč v komponenti zelo majhno (označimo s $PPC(K)$ za neko komponento K). Pri popolnoma neodvisnih nesrečah je to razmerje enako 2, v primeru goljufivih

komponent pa se močno približa 1. Ustrezna preslikava za prvo lastnost je

$$S_1(K) = \begin{cases} 1 & PPC(K) \leq \theta_1 \\ 0 & \text{sicer} \end{cases},$$

kjer θ_1 predstavlja prag, ki ga nastavi domenski ekspert glede na to, kakšna vrednost za $PPC(\cdot)$ se šteje za sumljivo. V sistemu nastavimo vrednost na $\theta_1 = 1.25$ (ustreza situaciji ko se en udeleženec zaleti s štirimi drugimi).

2) Ker goljufive komponente navadno vsebujejo veliko število razmeroma sumljivih nesreč, jih lahko identificiramo tudi na podlagi skupne sumljivosti nesreč, kot smo to definirali v razdelku 3.1. Naj bo $CS(K)$ vsota sumljivosti vseh nesreč, ki jih opisuje komponenta K. Tedaj je

$$S_2(K) = \begin{cases} 1 & CS(K) \geq \theta_2 \\ 0 & \text{sicer} \end{cases},$$

kjer prag θ_2 nastavimo na 3.00 (ustreza trem zelo sumljivim nesrečam oziroma večjemu številu manj sumljivih). Izbiro vrednosti za θ_2 sicer zopet prepustimo domenskemu ekspertu oziroma jo nastavimo glede na čas, ki ga ima slednji na voljo za nadaljnjo raziskavo.

3, 4, 5) Preostale tri lastnosti (preslikave) se osredotočajo na strukturne značilnosti goljufivih komponent. Ocenjujemo premer komponente, največjo stopnjo vozlišča ter največjo vmesno centralnost vozlišča (vrednosti označimo z $val_i(K)$, $i = 3, 4, 5$, zaporedoma). Za vsako komponento bi te vrednosti radi primerjali s tistimi, ki bi jih dobili za neke negoljufive komponente oziroma negoljufiv svet. Ker slednjih nimamo,[2] si pomagamo tako, da konstruiramo naključne mreže, čim bolj podobne dejanskemu stanju, ter vrednosti ocenimo glede na te. Naj bo $P(V_i)$ porazdelitev za val_i, ki jo dobimo na podlagi velikega števila naključno konstruiranih mrež (V_i je naključna spremenljivka, ki meri i-to vrednost). Tedaj je

$$S_3(K) = \begin{cases} 1 & val_3(K) \leq \theta_3, \text{ kjer je } P(V_3 \leq \theta_3) = \alpha \\ 0 & \text{sicer} \end{cases}$$

ter

$$S_i(K) = \begin{cases} 1 & val_i(K) \geq \theta_i, \text{ kjer je } P(V_i \geq \theta_i) = \alpha \\ 0 & \text{sicer} \end{cases}$$

za $i = 4, 5$. Opazimo, da so $\theta_3, \theta_4, \theta_5$ vezane v zgornjih izrazih, vrednost α pa v vseh primerih nastavimo na 0.05.

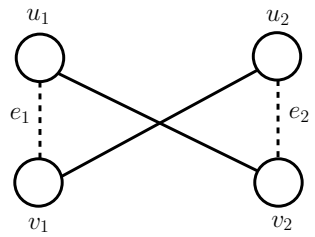
Za ocenjevanje porazdelitve $P(V_i)$ bi lahko uporabili kar ene od naključnih mrež, ki smo jih predstavili v razdelku 2.2. Slednje bi tudi pomenilo, da določene

[2] V razdelku 5 bomo videli, da bi v nasprotnem primeru postopali nekoliko drugače.

porazdelitve poznamo in jih ne bi bilo potrebno statistično ocenjevati. Najprimernejše bi bile sicer verjetno kar Poissonove, saj v primeru, da ni goljufij, lahko predpostavimo, da so nesreče neki naključni dogodki v času. Zdi se tudi, da se porazdelitev stopenj v naših mrežah nesreč verjetno ne podreja potenčnemu zakonu (gotovo ne v mrežah brez goljufij)[3], zato uporaba naključnih mrež po potenčnemu zakonu sploh ni primerna.

Kljub vsemu so verjetno tudi Poissonove mreže zelo daleč od dejanskih mrež nesreč. Naključne mreže zato konstruiramo na nekoliko drugačen način, podobno kot pri modelu majhnega sveta. Začnemo z neko komponento (mrežo) ter naključno prevezujemo povezave v njej. Po določenem številu korakov (prevezav) dobimo naključno mrežo, ki je verjetno, vsaj glede na relacijo takih prevezav, bolj podobna mrežam nesreč kot splošne naključne mreže. Prevezave potekajo na naslednji način:

1. naključno izberi povezavi $e_1 = \{u_1, v_1\}, e_2 = \{u_2, v_2\}$ brez skupnih krajišč,
2. odstrani e_1, e_2 ter dodaj povezavi $\{u_1, v_2\}, \{u_2, v_1\}$ oziroma $\{u_1, u_2\}, \{v_1, v_2\}$ (naključno izberi).



Slika 3.4: Primer prevezave – odstranimo povezavi e_1 in e_2 ter dodamo dve novi (označeni s polno črto).

Primer prevezave je viden na sliki 3.4.

Pri konstrukciji ohranimo sama vozlišča ter povezave v komponenti, tako komponenta opisuje iste nesreče ter udeležence, ki so se sedaj zaleteli na nekoliko drugačen način. Tako dejansko velja, da dobimo neko porazdelitev $P(V_i|K)$ odvisno od začetne komponente K. $P(V_i|K)$ zato ocenimo za vsako komponento posebej. Poudarimo še, da v postopku naredimo zgolj manjše število prevezav (primerljivo s številom povezav), saj bi v nasprotnem primeru dobili popolnoma (splošno) naključno mrežo, brez sledi o tisti, s katero smo začeli.

Opazimo, da postopek nikoli ne spremeni stopnje nekega vozlišča. To je seveda problem, saj je stopnja ravno ena od lastnosti, ki jih ocenjujemo. Nadalje velja tudi,

[3]Temelji na predpostavki, da v realnem svetu skorajda ni oseb, ki se zaletijo res zelo velikokrat. To pa pomeni, da porazdelitev stopenj v mrežah nesreč ne more padati tako počasi kot pravi potenčni zakon.

da zaradi tega nekaterih mrež enostavno ni moč nikakor spremeniti (prevezati). Problem rešimo na sila enostaven način. V mrežo zgolj dodamo neko vozlišče v_e, ki ga povežemo z vsemi ostalimi. Postopek je nato identičen prejšnjemu, s to razliko, da na koncu še odstranimo vozlišče v_e (ter s tem tudi vse njegove povezave). Stopnja vozlišč se sedaj lahko spreminja – vsakič ko v prevezavi sodeluje vozlišče v_e, spremenimo (končno) stopnjo dvema izmed vozlišč (enemu se poveča, drugemu zmanjša).

V sistemu ne uporabimo statistike l in koeficienta razvrščanja C, ki smo ju omenili v razdelku 2.1. Prva se zdi neprimerna, ker z njo navadno merimo, da je svet veliko manjši, kot se zdi sprva. V našem primeru pa je ravno nasprotno, saj bi radi merili, da svet še vseeno ni tako majhen, kot se to izkaže v goljufivih komponentah - učinek ne tako majhnega sveta (*not so small world effect*).

Koeficient razvrščanja C se sicer zdi zelo uporaben, saj je kakršna koli tranzitivnost v mreži močno sumljiva ter pogosto pomeni goljufijo. Vendar pa ima večina od omenjenih naključnih mrež previsoko tranzitivnost, da bi z njimi lahko ocenjevali negoljufiv svet, ki je po predpostavki skorajda brez tranzitivnosti – vendar ne popolnoma, zato ne moremo preprosto izpostaviti komponent, za katere velja $C > 0$.

3.2.3 Odkrivanje ključnih entitet

V prejšnjem delu sistema smo v mreži identificirali sumljive komponente, katere v tem delu natančneje raziščemo. Povedano na hitro, entitetam v vsaki (sumljivi) komponenti izračunamo stopnjo sumljivosti, na podlagi katere lahko nato identificiramo ključne udeležence ter nesreče. Rezultat tega zadnjega dela je tako stopnja sumljivosti za vsako entiteto, najbolj sumljivi udeleženci v neki (povezani) komponenti pa verjetno predstavljajo skupino sodelujočih goljufov.

Stopnjo sumljivosti določimo entitetam s pomočjo iterativne metode, ki temelji na opazki, da lahko vsako entiteto dobro opredelimo z njenimi lastnimi lastnostmi, predvsem pa z lastnostmi entitet, s katerimi je povezana. Natančneje, vsak udeleženec je dobro opredeljen z nesrečami, v katerih je sodeloval, vsaka nesreča je dobro opredeljena s svojimi udeleženci. Idejo ponazarja naslednji znani rek:

Povej mi, kdo so tvoji prijatelji, in povedal ti bom, kdo si.

Slednje se lepo sklada tudi z mrežo nesreč, ki jo uporabimo v tem delu sistema. Velja, da je vsako njeno vozlišče dobro opredeljeno s svojimi neposrednimi sosedi – vozlišče, ki ustreza udeležencu, je povezano ravno s svojimi nesrečami in obratno. Sama konstrukcija mreže nesreč je sedaj verjetno nekoliko bolj jasna.

Opazimo, da je ideja na nek način zelo lokalna, saj se pri opredelitvi določene entitete upošteva zgolj njene neposredne sosede. Kot pa bomo videli v nadaljevanju, lahko z iterativnim ocenjevanjem stopnje sumljivosti premagamo to slabost.

Predstavimo sedaj samo metodo za določanje stopnje sumljivosti entitet (v nadaljevanju sumljivost entitet). Ta deluje nad vsako komponento mreže posebej. Naj bodo $u_1, \dots u_s$ udeleženci in $a_1, \dots a_t$ nesreče, ki jih opisuje neka komponenta K, ter naj bo $V_K(\cdot)$ bijektivna preslikava, ki za vsako entiteto (udeleženec ali nesreča) določi ustrezno vozlišče. S $s^i(\cdot)$ označimo še sumljivost neke entitete na i-ti iteraciji. Metoda tedaj poteka v naslednjih treh korakih:

1. inicializiramo sumljivost za vse entitete kot

$$\forall i : s^0(u_i) = \frac{1}{s} \text{ ter } \forall i : s^0(a_i) = \frac{1}{t},$$

2. dokler velja $\sum_{i=1}^{s}(s^k(u_i) - s^{k-1}(u_i))^2 > \epsilon^2$, ponovi:

$$\forall i : s^{k+1}(a_i) = f_{ent}(a_i) \sum_{e=\{v,V_K(a_i)\}\in E(K),\ x=V_K^{-1}(v)} f_e(e, a_i)\, s^k(x) \tag{3.2}$$

$$\forall i : s^{k+1}(u_i) = \gamma s^k(u_i) + (1-\gamma) f_{ent}(u_i) \sum_{e=\{v,V_K(u_i)\}\in E(K),\ a_j=V_K^{-1}(v)} f_e(e, u_i)\, s^{k+1}(a_j) \tag{3.3}$$

$$\forall i : s^{k+1}(u_i) = \frac{s^{k+1}(u_i)}{\sum_{j=1}^{s} s^{k+1}(u_j)}. \tag{3.4}$$

   Na vsaki iteraciji najprej ocenimo sumljivost nesreč kot uteženo linearno kombinacijo sumljivosti vseh sosedov nesreče, ki so lahko tako udeleženci kot nesreče (enačba (3.2)). $f_{ent}(x)$ je faktor, ki predstavlja statične lastnosti entitete x (sumljivost ter čas nesreče, starost udeleženca ..., glej razdelek 3.2.1), $f_e(e, x)$ pa utež, ki je odvisna od tipa povezave e (voznik-nesreča, sopotnik-nesreča ...). Omenimo, da s pomočjo faktorjev formuliramo tudi krivdo v nesreči (usmerjenost povezav).

   Sumljivost nesreč potrebujemo zgolj zato, da lahko sedaj ocenimo sumljivost udeležencev, kar je dejansko cilj tega dela sistema (enačba (3.3)). Sumljivost je enaka linearni kombinaciji stare ter nove vrednosti, katero ocenimo podobno kot prej, s to razliko, da uporabljamo pri izračunu že nove ocene za sumljivost nesreč (sosedi udeležencev so lahko le nesreče). Na koncu iteracije še normaliziramo sumljivosti udeležencev, saj bi se sicer te zgolj povečevale (enačba (3.4)).

3. sumljivost udeležencev na koncu normaliziramo glede na sumljivost komponente $CS(K)$ (razdelek 3.2.2):

$$\forall i : s^{k+1}(u_i) = s^{k+1}(u_i)\, CS(K).$$

Zadnji korak nam zagotavlja, da je moč med seboj primerjati sumljivost entitet, ki prihajajo iz različnih komponent. V kolikor tega ne storimo, bo sumljivost entitete tem manjša, tem večja bo komponenta, v kateri je entiteta vsebovana.

V sistemu nastavimo $\gamma = 0.75$ ter $\epsilon = 10^{-6}$. Natančnih vrednosti faktorjev $f_{ent}(\cdot)$ in $f_e(\cdot,\cdot)$ na tem mestu ne podajamo, povejmo zgolj, da v kolikor je e neka dodatna povezava med voznikom in nesrečo, ki smo jo dodali zaradi sopotnikov v vozilu (glej razdelek 3.2.1), tedaj je $f_e(e, a_i) = 0, \forall i$. Take povezave tako ne upoštevamo pri ocenjevanju sumljivosti nesreč, temveč zgolj pri udeležencih. To je tudi razlog, zakaj pri določanju $f_e(\cdot,\cdot)$ potrebujemo drugi argument.

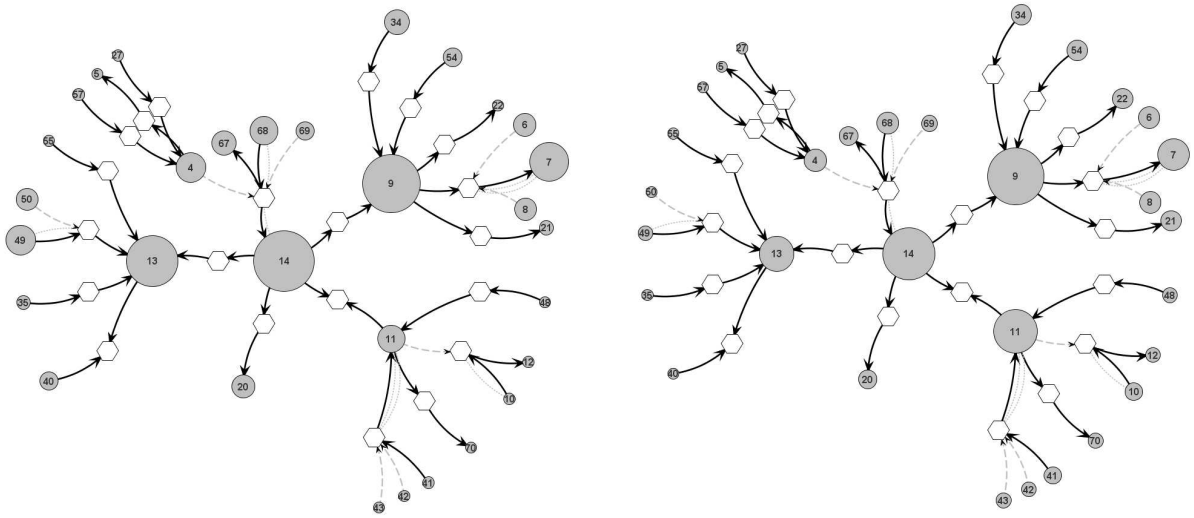
Faktorje sicer nastavimo glede na pomembnost posameznih lastnosti in relacij za vrsto goljufij, ki jih želimo na ta način odkriti. V kolikor nas zanimajo zgolj strukturne lastnosti, enostavno nastavimo $f_{ent}(\cdot) = 1$ (lahko tudi $f_e(\cdot,\cdot) = 1$). Metoda tedaj izpostavi entitete z visoko centralnostjo, stopnjo ter vmesnostjo.

Poudarimo, da je namen metode določiti sumljivost udeležencev in ne tudi nesreč, ki na nek način služijo zgolj za prenos sumljivosti med entitetami. Vendar pa slednje naredimo implicitno v samem postopku. Podobno bi seveda lahko storili tudi za entitete, ki se ne pojavljajo v mrežah nesreč. Na primer, sumljivost vozila lahko ocenimo iz sumljivosti njegovih voznikov ter iz nesreč, v katerih je bilo vozilo udeleženo. Podobno lahko storimo tudi za policiste.

Opazimo, da je sumljivost udeleženca močno odvisna od števila nesreč, v katerih je le-ta sodeloval, saj je od tega odvisno število členov v vsoti iz enačbe (3.3). Slednje je seveda pravilno, saj so goljufi udeleženi v veliko večje število nesreč kot ostali udeleženci. Vendar pa pri tem dejansko predpostavimo, da za posameznega udeleženca poznamo vse njegove nesreče. Ker temu pogosto ni tako, lahko zato nekoliko popravimo izračun sumljivosti. V enačbi (3.3) vsoto nadomestimo s povprečjem njenih členov, ki ga pomnožimo s povprečjem med številom nesreč ustreznega udeleženca in povprečnim številom nesreč na udeleženca. Tako število nesreč nekoliko potegnemo k povprečju, s čimer zmanjšamo pomembnost tega atributa pri ocenjevanju sumljivosti.

Kot smo omenili že na samem začetku, se zdi, da je metoda dokaj lokalna, saj se pri računanju sumljivosti upoštevajo zgolj neposredni sosedi neke entitete. Vendar pa zaradi iterativnega ocenjevanja očitno sledi, da se na k-ti iteraciji metode dejansko upošteva vsa $2k$-okolica neke entitete ($2k$-okolica vozlišča so vsa vozlišča na razdalji manjši ali enaki $2k$). Slednjega seveda ne bi bilo mogoče doseči "na silo", saj bi bila formulacija gotovo prezapletena.

S tem je zaključen zadnji del sistema. Kot rezultat dobimo ocene sumljivosti za posamezne entitete (glej sliko 3.5), kar lahko sedaj uporabimo, da v vsaki sumljivi komponenti izpostavimo ključne udeležence in nesreče – sumljive (goljufive) skupine posameznikov ter povezujoče nesreče (glej sliko 3.6). Zaradi ustrezne normalizacije je moč določiti tudi najbolj sumljivo entiteto med vsemi komponentami – entiteto, katero naj domenski analitik najprej razišče. Več o uporabi rezultatov za nadaljnjo raziskavo povemo v naslednjem razdelku.



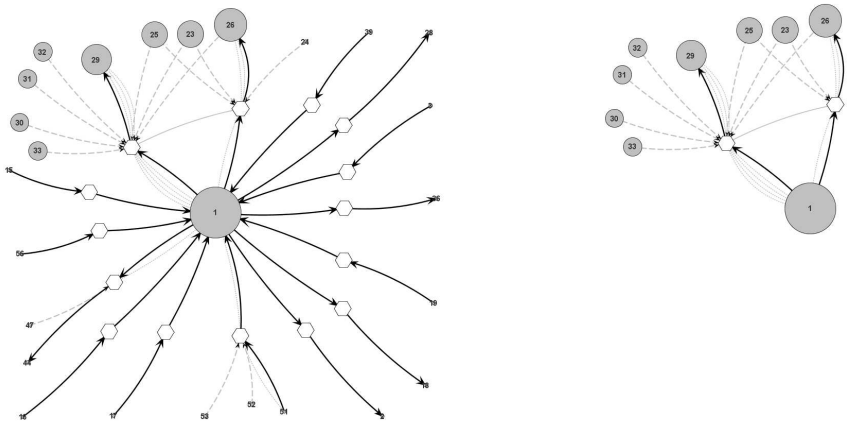
Slika 3.5: Rezultat odkrivanja ključnih entitet v neki sumljivi komponenti (prikazana je mreža nesreč). Velikost vozlišča je sorazmerna s sumljivostjo ustreznega udeleženca. Na levi strani je mreža, ki jo dobimo, če faktorje v enačbah (3.2), (3.3) nastavimo na smiselne vrednosti, na desni pa mreža, pri kateri nastavimo $f_{ent}(\cdot) = 1$ (statične lastnosti entitet zanemarimo).

3.3 Predstavitev ter uporaba znanja

Goljufije ter goljufive skupine je v splošnem nemogoče odkrivati popolnoma avtomatsko. V vsakem primeru je potrebna kasnejša podrobnejša raziskava s strani domenskega analitika, ki pridobi še dodatne podatke o sumljivih entitetah ter, v kolikor spozna, da gre res za goljufijo, sproži ustrezne nadaljnje postopke. Tega sistem ne more storiti, saj pri tem ne sme priti do napake, res pa je tudi, da je dodatne podatke težko pridobiti avtomatsko. To lahko navadno stori le človek.

V sistemu tako rezultate na koncu prikažemo analitiku, za kar seveda zopet, predvsem zaradi jasnosti predstavitve, uporabimo mreže. Pri tem lahko uporabimo različne vrste mrež, bolje rečeno pogledov. Poleg mrež voznikov, sopotnikov in nesreč ponudimo tudi poglede, ki vključujejo vozila in policiste. V pogledih prikažemo poleg povezanosti med entitetami tudi njihove statične lastnosti, kar seveda olajšuje raziskavo. Primeri različnih pogledov so vidni na slikah 3.5, 3.6 in 3.7. Velikost vozlišč, ki ustrezajo udeležencem, je sorazmerna njihovi sumljivosti, pri čemer narišemo zgolj tista, katerih sumljivost je nad povprečjem oziroma nad neko mejo, ki jo določi analitik. Ta lahko tako izpostavi le delež najbolj sumljivih oseb oziroma mejo prilagodi času, ki ga ima na razpolago za samo raziskavo (slika 3.6).

Analitik pri raziskavi sledi naslednjemu postopku. V komponenti najprej razišče najsumljivejšega udeleženca ter z njim povezane nesreče. Nesreče razišče glede na padajočo sumljivost. Če ugotovi, da gre res za goljufijo, ustrezno ukrepa ter raziskavo



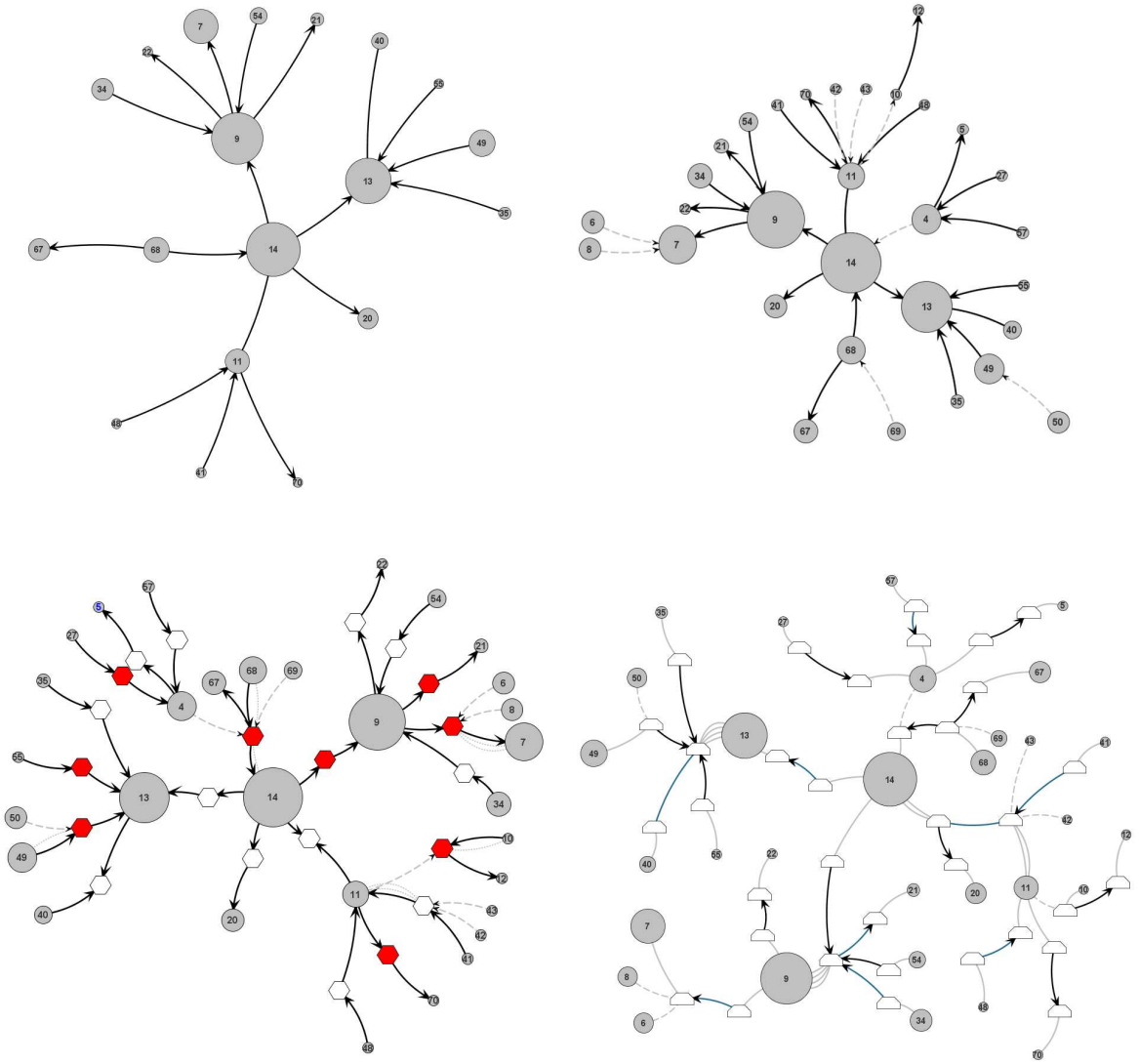
Slika 3.6: Pri predstavitvi rezultatov prikažemo le udeležence z nadpovprečno visoko sumljivostjo (na sliki mreža nesreč). Na desni strani pogled, ki prikazuje zgolj nesreče, ki povezujejo prej omenjene udeležence.

nadaljuje pri naslednji najbolj sumljivi entiteti. V kolikor pa ugotovi, da določene entitete niso goljufive, ponovi tretji del sistema (razdelek 3.2.3), pri čemer sumljivost omenjenih entitet nastavi na nič in jih v postopku ne spreminja. Sistem tako izračuna novo sumljivost entitet, kjer upošteva še dodatno znanje, pridobljeno z raziskavo. Podobno lahko analitik v primeru, ko odkrije neko goljufijo, goljufivim entitetam nastavi sumljivost na neko neničelno vrednost ter postopa tako kot prej. Vendar v tem primeru ni jasno, kakšna naj bo ta vrednost.

Pri predstavitvi rezultatov z mrežami velja omeniti tudi pomembnost postavitve vozlišč z namenom, da je slika čim bolj jasna. Slednje je seveda ključno za samo predstavitev znanja. Razviti so bili različni algoritmi, ki poskušajo vozlišča postaviti tako, da se čim manj povezav med seboj seka, pri čemer so te približno enake dolžine. Dobre rezultate dajo na primer metode, kjer povezave predstavljajo vzmeti, vozlišča pa električne naboje [13, 17]. Metoda nariše mrežo tako, da sistem, ki ga opisujejo vzmeti in naboji, doseže ravnovesje (glej sliko 3.6 ter 3.7 spodaj desno). Obstaja pa še vrsta drugih metod, ki temeljijo na spektralni analizi, simetrijah v mreži, hierarhiji vozlišč ter tudi na nevronskih mrežah [23].

Ker so komponente mreže nesreč razmeroma majhne (razdelek 3.2.1) ter skoraj drevesa[4], se v tem primeru izkaže že nekoliko bolj preprosta metoda. V mreži zgolj poiščemo vozlišče z največjo centralnostjo, ki ga proglasimo za koren mreže. Le-to nato narišemo kot da bi bila drevo[5], pri čemer postavljamo vozlišča na koncentrične kroge okrog korena (glej mreži na sliki 3.5).

[4]Odstraniti bi bilo potrebno zgolj manjše število povezav, da bi mreža postala drevo (brez ciklov).
[5]Dejansko poiščemo v mreži minimalno vpeto drevo ter narišemo tega.



Slika 3.7: Primeri različnih pogledov za isti nabor podatkov. Mreža voznikov zgoraj levo, mreža sopotnikov zgoraj desno, levo spodaj mreža nesreč ter desno spodaj pogled, ki vsebuje tudi vozila – nepravilni šestkotniki (pomen povezav je podoben kot pri mreži nesreč). Mreža nesreč prikazuje tudi, pri katerih nesrečah je prišlo do sumljivih poškodb (rdeče), pogled z vozili pa, katere nesreče so se zgodile v poznih večernih urah (temno modre povezave).

Poglavje 4

Eksperimentalni rezultati

Za namene naloge je bilo pridobljenih 40 policijskih zapisnikov o nesrečah v času osmih let. V te nesreče je bilo vključenih 71 oseb, od tega 47 voznikov, obravnavalo jih je 48 različnih policistov. Nesreče so se zgodile na 35 različnih lokacijah, v njih pa je bilo udeleženih 68 različnih vozil. Podatki so bili seveda neoznačeni, poudarimo pa, da ti niso bili izbrani naključno. Relacijski podatki v splošnem ne dopuščajo naključnega vzorčenja, saj na ta način uničimo povezave med entitetami. Načrtno so bile zato izbrane razmeroma povezane nesreče, in kot bomo videli v nadaljevanju, tudi goljufive.

Na podlagi analize ter v sodelovanju z domenskimi eksperti smo vse udeležence označili. Pri tem bi bilo seveda bolj pravilno označiti goljufive nesreče ter upoštevati tudi ceno pri vsaki goljufiji (izgubo za zavarovalnico), a je bil to pretežak zalogaj. Izkazalo se je, da gre za razmeroma goljufive nesreče in udeležence. Kar 24 od 71 udeležencev je bilo označenih za goljufe. Na to pa je nakazovalo tudi že samo razmerje med številom voznikov in nesreč, ki znaša $\frac{47}{40} = 1.175$. V popolnoma neodvisnih nesrečah je to razmerje enako 2.

V nadaljevanju podamo rezultate testiranja nad omenjenim naborom podatkov, vendar pa je potrebno te jemati nekoliko z rezervo. Omenimo glavne razloge za to:

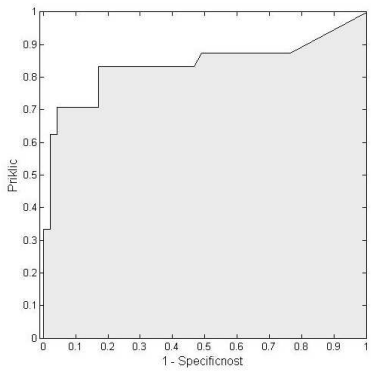
- kljub temu da sistem podatkov ne uporablja za učenje, smo jih uporabljali pri sami zasnovi sistema. Tako se pričakuje, da bo uspešnost metode nad tem naborom visoka;
- podatke je težko pravilno označiti, saj nikoli ne moremo biti popolnoma prepričani, ali gre v določenem primeru res za goljufijo;
- vzorec je premajhen, da bi iz njega lahko sklepali o uspešnosti sistema v praksi. To velja predvsem za ocenjevanje uspešnosti drugega dela sistema;
- vzorec ni reprezentativen, saj gre za razmeroma goljufive podatke – kar 34 % udeležencev v nesrečah je bilo spoznanih za goljufe. Sicer bi pričakovali, da bo ta odstotek veliko manjši, seveda pa bi bil reprezentativen vzorec podobne velikosti za namene testiranja popolnoma neuporaben.

Pri predstavitvi z mrežami (prvi del sistema) podatki razpadejo na štiri povezane komponente (mreža nesreč). Sistem nato v drugem delu dve od teh komponent zavrže, drugi dve pa označi kot sumljivi. S tem zavrže 14 udeležencev, od tega 3 goljufe. Ostalim udeležencem nato ocenimo stopnjo sumljivosti s pomočjo predstavljene metode, zavrženim udeležencem pa jo nastavimo na 0. Povprečna sumljivost udeleženca je enaka $\overline{x} = 0.30$ s standardno deviacijo $s = 0.33$. Kot sumljive sedaj izpostavimo vse udeležence, ki imajo nadpovprečno visoko sumljivost (večjo od $\overline{x}$). Tako nizka meja je primerna zgolj zato, ker gre za razmeroma goljufiv nabor podatkov. Pričakujemo, da bi sicer v praksi navadno izpostavili udeležence s sumljivostjo nad na primer $\overline{x} + 2\,s$.

Metoda doseže klasifikacijsko točnost $CA = 83.10\,\%$ (*classification accuracy*), priklic je enak 83.33 % (*recall*), specifičnost pa 82.98 % (*specificity*). Podrobnejši rezultat klasifikacije je viden v tabeli 4.1. Pravilno identificiramo 20 goljufov, poleg teh pa še dodatnih 8, ki niso goljufi. Štirih goljufov ne identificiramo, pri čemer 3 od teh izločimo že v drugem delu sistema. Goljufi spadajo v dve povezani komponenti ter tako tvorijo dve goljufivi skupini.

	Goljuf	Ni goljuf
Klasificiran kot goljuf	20	8
Klasificiran kot negoljuf	4	39

Tabela 4.1: Rezultati odkrivanja goljufov na testnem naboru podatkov.



Slika 4.1: ROC krivulja za rezultate testiranja.

Tako natančnost kot priklic sta zelo visoka, vendar iz rezultata ne moremo sklepati o uspešnosti metode v splošnem. Ker je naš namen zgolj rangirati udeležence glede na sumljivost, je navadno boljša mera uspešnosti mera *AUC* (*area under curve*). Metoda doseže $AUC = 83.87\,\%$, na sliki 4.1 pa je prikazana še ROC krivulja (*receiver operating characteristic*). Podobno kot prej je tudi v tem primeru rezultat zelo dober.

Rezultati so bili prikazani tudi analitikom slovenske avtomobilske zavarovalnice, ki so bili z njimi zelo zadovoljni, tako z odkrivanjem kot tudi s samim prikazom rezultatov.

Poglavje 5

Sklepne ugotovitve

V nalogi predstavimo nov sistem za odkrivanje goljufij v avtomobilskem zavarovalništvu. Ta se osredotoča na odkrivanje goljufivih skupin posameznikov, ki so s stališča zavarovalnic najbolj zanimive. Sistem, za razliko od nekaterih drugih rešitev, pri predstavitvi podatkov uporablja mreže, ki so verjetno najnaravnejša predstavitev, poleg tega pa omogočijo formulacijo kompleksnih relacij med entitetami, kar je ključno pri odkrivanju takih goljufij. Izjemno pomembno za sam problem je tudi, da mreže omogočijo jasno predstavitev končnih rezultatov domenskemu ekspertu, ki opravi nadaljnjo raziskavo. Sistem ga pri tem vodi ter upošteva tudi novo pridobljeno znanje. Sicer v sistemu ne potrebujemo označenega začetnega nabora podatkov, je razmeroma enostaven za implementacijo ter dopušča vključitev poljubnega dodatnega znanja o domeni.

Z nastavljanjem parametrov sistema je tega moč v veliki meri prilagoditi potrebam oziroma zahtevam uporabnika. S parametri drugega dela sistema lahko vplivamo na sam priklic, podobno velja za mejo sumljivosti, ki jo nastavimo na koncu. Tudi faktorje tretjega dela lahko posebej prilagodimo vrsti goljufij, ki jih želimo odkrivati. V te je tako mogoče vnesti ogromno količino domenskega znanja.

Sistem bi bilo moč na mnogih področjih še izboljšati, predvsem v kolikor bi uspeli pridobiti večji označen nabor podatkov. V drugem delu sistema tako ne bi potrebovali naključnih mrež za ocenjevanje realnega sveta oziroma še bolje, za odkrivanje sumljivih komponent bi uporabili kar eno od metod strojnega učenja. Slednje bi izboljšalo odkrivanje ter tudi pohitrilo sam sistem.[1] Podobno bi se lahko v tretjem delu naučili, kakšna je verjetnost, da je entiteta z določenimi lastnostmi goljufiva. Namesto faktorjev, ki jih sicer določi domenski ekspert, bi tedaj lahko uporabili kar verjetnosti, ki jih vrne metoda strojnega učenja. Odkrivanje goljufij bi bilo moč izboljšali tudi v primeru, če bi uspeli pridobiti več podatkov o posamezni entiteti oziroma podatke o novih entitetah. Tako bi v obravnavo lahko vključili tudi slednje. Pri velikem številu entitet bi se verjetno izkazalo, da bi bilo smiselno uporabiti hipergrafe oziroma hipermreže. V

[1] Izboljšava bi bila pri večji količini podatkov verjetno kar nujna, saj si v tem primeru ne moremo privoščiti konstruiranja naključnih mrež.

nasprotnem primeru pogosto ni jasno, kako entitete smiselno povezati med seboj tako, da v mreži ne ustvarimo umetnih ciklov.

Pomanjkanje (označenih) podatkov je bila verjetno ena glavnih težav v nalogi. Kljub temu je bil sistem preizkušen na realnem naboru, kjer je dosegel odlične rezultate. Ker je bil vzorec nekoliko manjši in nereprezentativen, ne moremo sklepati o uspešnosti sistema v splošnem, vendar pa smo z doseženim zelo zadovoljni. Podobno je bilo čutiti tudi s strani domenskih analitikov, katerim so bili prikazani rezultati testiranja. Tako lahko zaključimo, da so bili cilji naloge v veliki meri doseženi.

Dodatek A

Seznam uporabljenih kratic in simbolov

	Opis
NP	nedeterministično polinomski
AUC	ploščina pod krivuljo (*area under curve*)
CA	klasifikacijska točnost
N	mreža
K	povezana komponenta mreže
V	množica vozlišč
E	množica povezav
n	število vozlišč
m	število povezav
v_i	i-to vozlišče
e_i	i-ta povezava
k	stopnja vozlišča
k_{max}	največja stopnja vozlišča v mreži
l	povprečna razdalja med parom vozlišč
C	koeficient razvrščanja
$B(\cdot)$	vmesnost vozlišča oziroma povezave
e	baza naravnega logaritma

Slike

Literatura

[1] L. A. N. Amaral, A. Scala, M. Barthélémy, H. E. Stanley, "Classes of small-world networks", *Proceedings of the National Academy of Sciences of the United States of America*, št. 97, zv. 21, str. 11149–11152, 2000.

[2] M. Artís, M. Ayuso, M. Guillén, "Detection of automobile insurance fraud with discrete choice models and misclassified claims", *The Journal of Risk and Insurance*, št. 63, zv. 3, str. 325–340, 2002.

[3] M. Bajec, S. Vidmar, Š. Furlan, "Odkrivanje goljufij v zdravstvu", *Bilten : ekonomika, organizacija, informatika v zdravstvu*, št. 23, str. 5, 2007.

[4] A. L. Barabasi, R. Albert, "Emergence of scaling in random networks", *Science*, št. 286, str. 509–512, 1999.

[5] R. J. Bolton, D. J. Hand, "Statistical fraud detection: A review", *Statistical Science*, št. 17, str. 235–249, 2002.

[6] R. L. Breiger, S. A. Boorman, P. Arabie, "An algorithm for clustering relations data with applications to social network analysis and comparison with multidimensional scaling", *Journal of Mathematical Psychology*, št. 12, str. 328–383, 1975.

[7] S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, št. 30, str. 107–117, 1998.

[8] P. L. Brockett, R. A. Derrig, L. L. Golden, A. Levine, M. Alpert, "Fraud classification using principal component analysis of RIDITs", *The Journal of Risk and Insurance*, št. 69, zv. 3, str. 341–371, 2002.

[9] D. Eppstein, J. Wang, "A steady state model for graph power laws", v zborniku *Second International Workshop on Web Dynamics*, Honolulu, Hawaii, 2002, str. 15–24.

[10] P. Erdös, A. Rényi, "On random graphs", *Publicationes Mathematicae Debrecen*, št. 6, str. 290–297, 1959.

[11] A. Firat, S. Chatterjee, M. Yilmaz, "Genetic clustering of social networks using random walks", *Computational Statistics and Data Analysis*, št. 51, zv. 12, str. 6285–6294, 2007.

[12] L. Freeman, "A set of measures of centrality based upon betweenness", *Sociometry*, št. 40, str. 35–41, 1977.

[13] T. M. J. Fruchterman, E. M. Reingold, "Graph Drawing by Force-Directed Placement", *Software: Practice and Experience*, št. 21, zv. 11, str. 1129–1164, 1991.

[14] Š. Furlan, S. Vidmar, M. Bajec, "Sistem za odkrivanje goljufij v sistemih zdravstvenega zavarovanja", v zborniku *Dnevi slovenske informatike*, Portorož, Slovenija, 2008, str. 10.

[15] M. Girvan, M. E. J. Newman, "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences of the United States of America*, št. 99, zv. 12, str. 7821–7826, 2002.

[16] G. Jeh, J. Widom, "SimRank: a measure of structural-context similarity", v zborniku *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, New York, 2002, str. 538–543.

[17] T. Kamada, S. Kawai, "An algorithm for drawing general undirected graphs", *Information Processing Letters*, št. 31, zv. 1, str. 7–15, 1989.

[18] D. Kempe, J. Kleinberg, E. Tardos, "Maximizing the spread of influence through a social network", v zborniku *Proceedings of the 9th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*,, Washington, D.C., 2003, str. 137–146.

[19] J. Kleinberg, "Authoritative sources in a hyperlinked environment", v zborniku *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, California, 1998, str. 668–677.

[20] B. Kondža, "Kaskaderstvo uspeva tudi v Sloveniji, ne le v Hollywoodu", *Delo, Finančni tednik*, str. 36, 9. 6. 2008.

[21] Y. Kou, C. T. Lu, S. Sirwongwattana, Y. P. Huang, "Survey of fraud detection techniques", v zborniku *Proceedings of IEEE International Conference on Networking, Sensing and Control*, Taipei, Taiwan, 2004, str. 749–754.

[22] C. F. Mann, D. W. Matula, E. V. Olinick, "The use of sparsest cuts to reveal the hierarchical community structure of social networks", *Social Networks*, št. 30, 2008.

[23] B. Meyer, "Self-organizing graphs - A neural network perspective of graph layout", *Lecture Notes in Computer Science*, št. 1547, str. 246–262, 1998.

[24] S. Milgram, "The small world problem", *Psychology Today*, št. 2, str. 60–67, 1967.

[25] J. Neville, D. Jensen, "Iterative classification in relational data", v zborniku *Proceedings of the Workshop on Statistical Relational Learning, 17th National Conference on Artificial Intelligence*, Austin, Texas, 2000, str. 42–49.

[26] M. E. J. Newman, "Detecting community structure in networks," *The European Physical Journal B*, št. 38, str. 321–330, 2004.

[27] M. E. J. Newman, "Mathematics of networks", v *The New Palgrave Encyclopedia of Economics, 2nd edition*, England: Palgrave Macmillan, 2008.

[28] M. E. J. Newman, "The structure and function of complex networks", *SIAM Review*, št. 45, str. 167–256, 2003.

[29] M. E. J. Newman, "The structure of scientific collaboration networks", *Proceedings of the National Academy of Sciences of the United States of America*, št. 98, zv. 2, str. 404–409, 2001.

[30] C. C. Noble, D. J. Cook, "Graph-based anomaly detection", v zborniku *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, D.C., 2003, str. 631–636.

[31] J. M. Perez, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, J. I. Martin, "Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance," v zborniku *Third International Conference on Advances in Pattern Recognition*, Bath, United Kingdom, 2005, str. 381–389.

[32] C. Phua, V. Lee, K. Smith, R. Gayler, "A comprehensive survey of data mining-based fraud detection research", *Artificial Intelligence Review*, 2005.

[33] R. Solomonoff, A. Rapoport, "Connectivity of random nets", *The Bulletin of mathematical biophysics*, št. 13, str. 107–117, 1951.

[34] S. H. Strogatz, "Exploring complex networks", *Nature*, št. 410, str. 268–276, 2001.

[35] J. Sun, H. Qu, D. Chakrabarti, C. Faloutsos, "Relevance Search and Anomaly Detection in Bipartite Graphs", *ACM SIGKDD Explorations Newsletter*, št. 7, zv. 2, str. 48–55, 2005.

[36] S. Viaene, G. Dedene, R. A. Derrig, "Auto claim fraud detection using Bayesian learning neural networks", *Expert Systems with Applications*, št. 29, str. 653–666, 2005.

[37] S. Viaene, R. A. Derrig, B. Baesens, G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection", *The Journal of Risk and Insurance*, št. 69, zv. 3, str. 373–421, 2002.

[38] H. I. Weisberg, R. A. Derrig, "Quantitative methods for detecting fraudulent automobile bodily injury claims", *Risques*, št. 35, str. 75–101, 1998.

[39] Home Page of Neo Martinez. Dostopno na: http://userwww.sfsu.edu/~webhead

[40] Operation Bumper Cars: Collision Fraud Schemes in Texas. Dostopno na: http://www.autobodynews.com/southwest-news/operation-bumper-cars-collision-fraud-schemes-in-texas.html

[41] Seven Bridges of Königsberg. Dostopno na: http://en.wikipedia.org/wiki/Seven_Bridges_of_Königsberg

Zahvala

Zahvalil bi se mentorju izr. prof. dr. Marku Bajcu ter somentorju doc. dr. Matjažu Kukarju za pomoč in usmerjanje med izdelavo diplomskega dela.

Posebna zahvala pa gre tudi Štefanu Furlanu in Juretu Leskovcu, ki sta s svojimi nasveti in idejami močno pripomogla k uspešnemu zaključku.