Pajek – program za analizo obsežnih omrežij Uporaba v rodoslovju

Andrej Mrvar
Fakulteta za družbene vede
Univerza v Ljubljani

Vladimir Batagelj
FMF, Oddelek za matematiko
Univerza v Ljubljani

1 Uvod

Pri svojem delu se večkrat srečamo z obsežnimi omrežji, kjer gre število točk/povezav v tisoče. Do teh omrežij lahko pogosto pridemo avtomatično iz drugih virov. Naštejmo nekaj primerov: rodovniki, diagrami poteka programskih sistemov, organske molekule, računalniška omrežja, transportna omrežja, vodovodna in električna omrežja, referenčna omrežja, družboslovna omrežja, itd. Obvladovanje tako velikih omrežij predstavlja tako časovno, kot tudi prostorsko zahteven problem. Trenutno še ne obstaja veliko programov, ki bi omogočali delo z njimi, zato sva se novembra 1996 odločila, da začneva z razvojem programskega paketa za analizo velikih omrežij `Pajek`. Program je prosto dostopen na naslovu:

`http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

2 Uporaba programa v rodoslovju

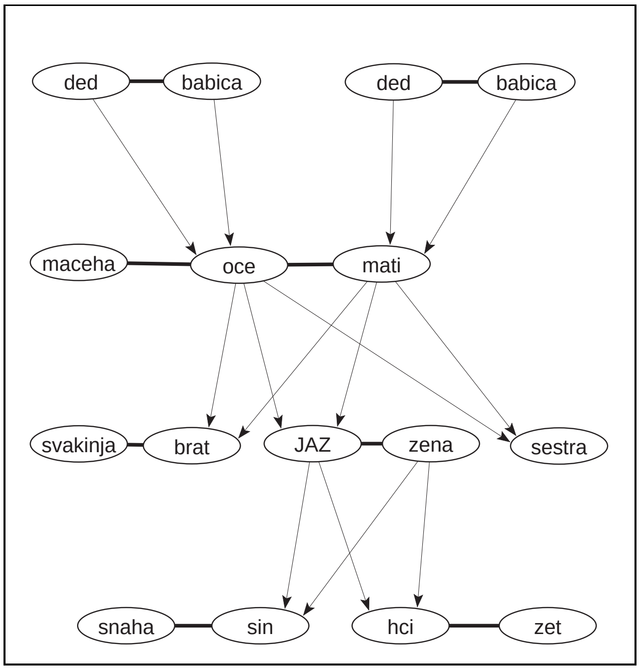
Za izmenjavo rodoslovnih podatkov se uporablja oblika zapisa `GEDCOM`. To obliko prepozna tudi program `Pajek` in jo predela v graf. Pravzaprav lahko rodoslovne podatke predelamo v dve vrsti grafov – *rodovnikov*:

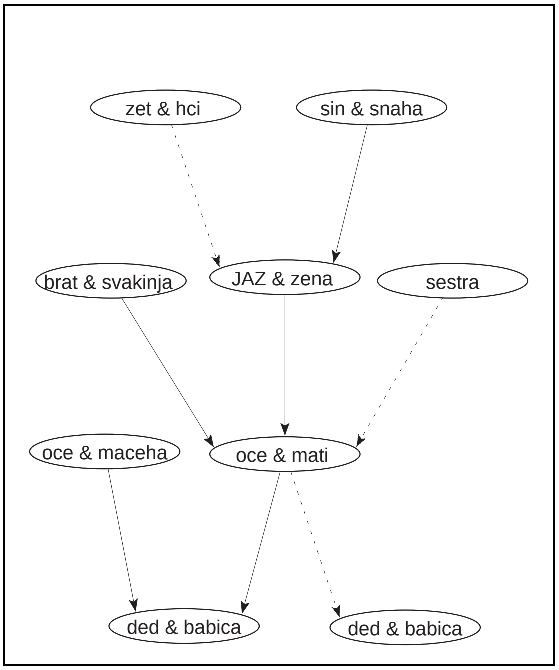
V *navadni* obliki je vsaka oseba predstavljena s točko grafa, poroke so označene z neusmerjeno povezavo med točkama, usmerjene povezave pa vodijo od staršev do otrok. Glej primer na levi strani slike 1.

V obliki *p-graph* pa so točke grafa lahko posamezniki ali pari. Več o tej obliki najdemo na predstavitveni strani dr. D. R. Whitea z Univerze v Kaliforniji (Irvine):

`http://eclectic.ss.uci.edu/~drwhite/pgraph/p-graphs.html`

V primeru, da neka oseba še ni poročena, je ta oseba v grafu predstavljena s svojo točko; če pa je poročena, je v grafu predstavljena skupaj s svojim zakoncem s skupno točko. V tej obliki imamo torej samo usmerjene povezave, ki vodijo od otrok do staršev. Ker





Slika 1: Predstavitev rodovnika v navadni obliki (levo) in v obliki p-graph (desno).

so točke lahko tudi pari, moramo posebej označiti ali se povezava nanaša na moškega ali na žensko. Povezave, ki se nanašajo na moške (kažejo na moževe starše) so označene z neprekinjeno črto, povezave, ki se nanašajo na ženske (kažejo na ženine starše) pa s prekinjeno (pikčasto) črto. V primeru, da je neka oseba večkrat poročena, se ta oseba pojavi v toliko točkah, kolikor je porok. Glej primer na desni strani slike 1.

3 Analiza rodovnikov s Pajkom

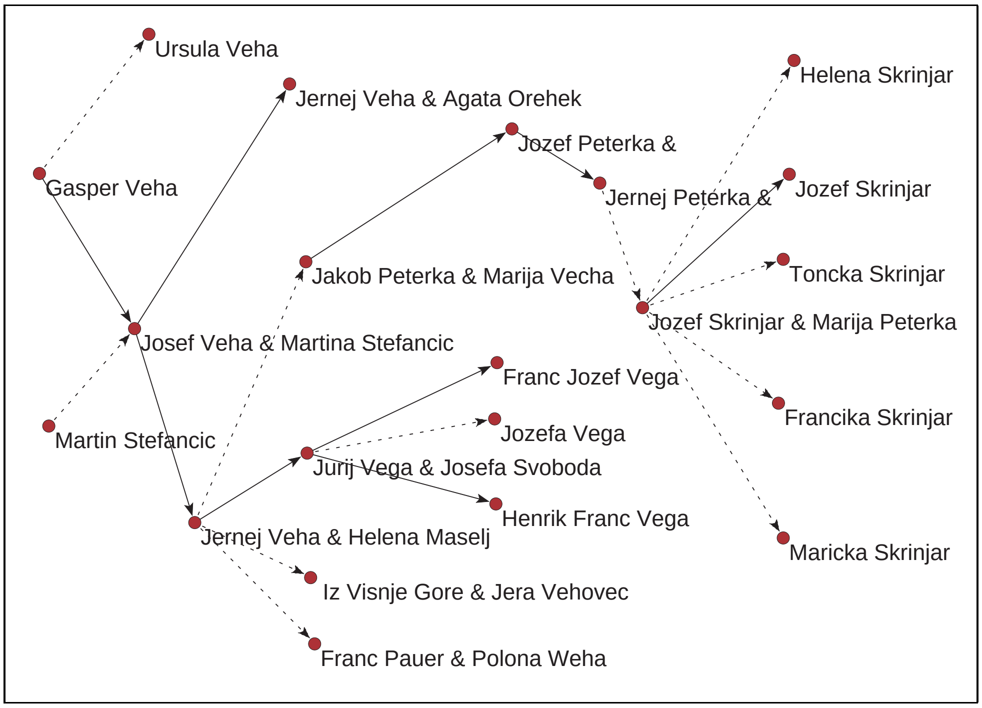
V nadaljevanju so opisani osnovni koraki, ki jih moramo opraviti pri uporabi programa `Pajek`, da dobimo željene rezultate.

3.1 Branje datoteke GEDCOM

Datoteko z rodoslovnimi podatki (v obliki `GEDCOM`) preberemo z ukazom `File/ Network/ Read` (poševne črte pomenijo zaporedne izbire) ali pritisnemo ikono za branje, ki se nahaja v glavnem oknu desno od besede `Network` (če se je dotaknemo z miško, se izpiše `Read Network`). Datoteka se prebere v obliki p-graph. Če želimo datoteko prebrati v navadni obliki, moramo pred branjem spremeniti nastavitev s klikom na `Options/ ReadWrite/ GEDCOM - Pgraph`.

3.2 Kako obrnemo smeri puščic

Kot smo opisali v prejšnjem podpoglavju, kažejo v obliki p-graph puščice od otrok proti staršem. Za razumevanje je morda bolje obrniti smeri puščic, kar naredimo z ukazom:



Slika 2: Rodovnik Jurija Vege (oblika p-graph).

`Net/ Transform/ Transpose`. V nadaljevanju je v vseh primerih kjer je rodovnik prebran v obliki p-graph uporabljena oblika z obrnjenimi puščicami (puščice kažejo od staršev proti otrokom).

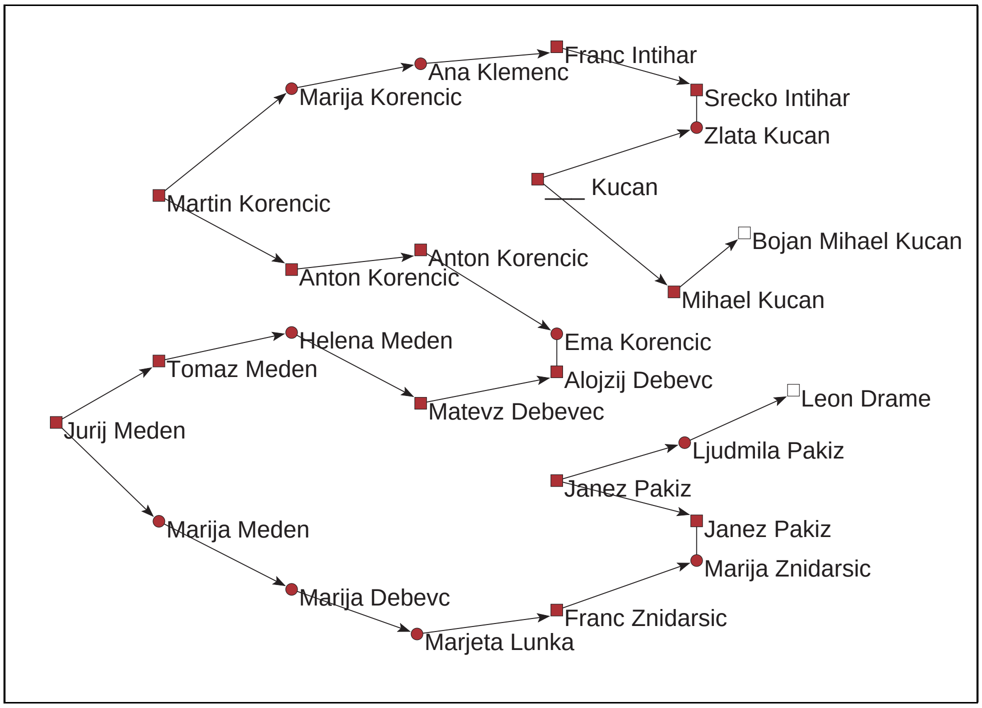
3.3 Risanje rodovnika

Za risanje rodovnikov se je smiselno odločiti, če graf ni prevelik (največ 200 do 300 točk), sicer je pred tem smiselno opraviti še kakšno analizo in na osnovi te analize izločiti in narisati samo izbrani podgraf.

Graf lahko narišemo z uporabo energijskega risanja (`Draw/ Energy/ Kamada-Kawai` ali (`Draw/ Energy/ 2D Fruchterman-Reingold`), ali pa z uporabo risanja prirejenega za risanje rodovnikov. To vrsto risanja poženemo z uporabo ukaza makro: `Macro/ Play` in nato poiščemo kje se nahaja datoteka `Layers.mcr` (ponavadi na podpodročju `Genea`). Dobljeno sliko lahko še izboljšamo z uporabo ukaza `Draw/ Draw-Partition/ Layers/ Optimize layers in xy`, ki nam skuša dobljeno razvrstitev po nivojih narisati s čim krajšimi povezavami (izbiramo lahko še med različnimi natančnostmi v podizbiri `Draw/ Draw-Partition/ Layers/ Resolution`. Na sliki 2 je na ta način narisan rodovnik Jurija Vege (v obliki p-graph z obrnjenimi puščicami).

3.4 Iskanje najkrajše sorodstvene zveze med dvema osebama

Tudi za iskanje in risanje najkrajše sorodstvene zveze med dvema osebama je pripravljena datoteka z ukazom makro – `path.mcr`. Poženemo jo na isti način kot prejšnje, le da



Slika 3: Najkrajša pot (navadna oblika).

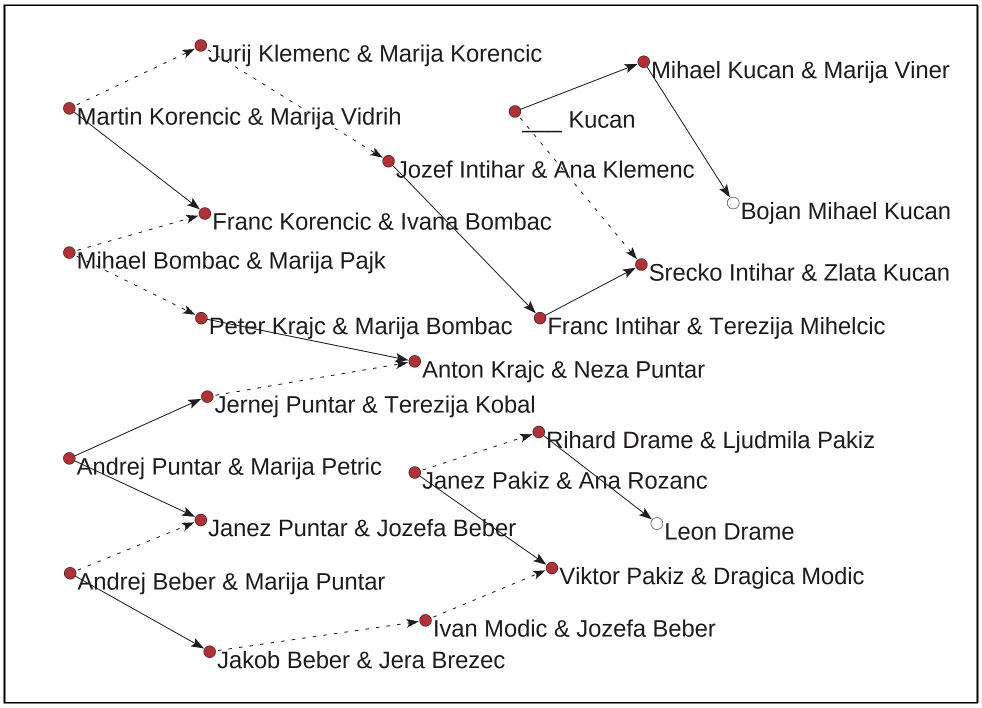
moramo še določiti obe osebi (odtipkamo lahko številko ali vsaj začetek imena).

Za primer smo uporabili podatke z datoteke `Drame.ged` (rodovnik Leona Drameta), v kateri se nahajajo podatki o 29606 osebah. Datoteka je sestavljena iz 703 ločenih komponent, od katerih je ena zelo velika (z 19854 osebami), ostale pa so manjše. Število komponent ugotovimo z ukazom `Net/ Components/ Weak` ter dobljeni rezultat pogledamo z `Info/ Partition`. Če si zapomnimo številko neke komponente, jo lahko kasneje izločimo iz grafa z ukazom `Operations/ Extract from Network/ Partition`, kjer dvakrat odtipkamo številko željene komponente.

Slika 3 prikazuje najkrajšo pot med Leonom Drametom in Bojanom Mihaelom Kučanom, če uporabimo navaden rodovnik, slika 4 pa najkrajšo pot, če uporabimo rodovnik v obliki p-graph (v tem primeru imamo v grafu namesto 29606 samo 22193 točk). Opazimo lahko, da se poti malo razlikujeta, ker smo v primeru navadne oblike šteli za korak tudi poroko (neusmerjeno povezavo).

3.5 Iskanje okolice izbrane osebe

Če imamo opravka z velikim grafom, se je smiselno omejiti na neko okolico izbrane osebe. Podgraf, ki ga sestavljajo osebe, ki so od izbrane osebe oddaljene največ za tri generacije dobimo tako, da poženemo makro `3bef3lat.mcr` in dvakrat odtipkamo željeno osebo.



Slika 4: Najkrajša pot (p-graph).

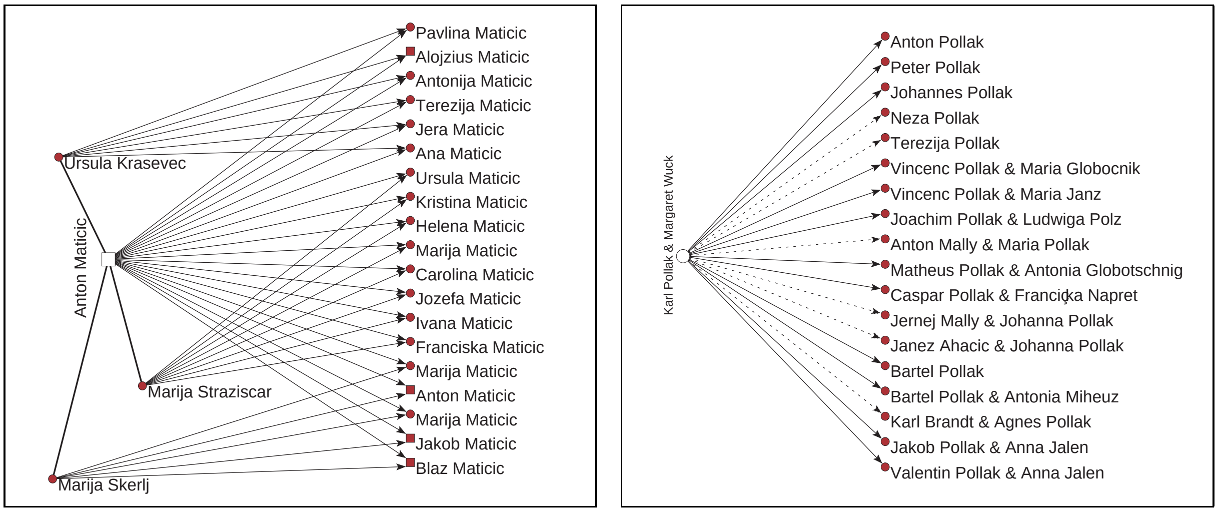
3.6 Število neposrednih potomcev izbrane osebe

Z ukazom `Net/ Partition/ Degree/ Output` ugotovimo, koliko povezav vodi od posameznih točk v grafu. Informacijo si lahko v skrčeni obliki pogledamo z ukazom `Info/ Partition` ter odtipkamo neko številko, npr. 10 – želimo videti prvih 10 oseb z največjim številom otrok. Številko ali ime osebe, ki ima največ otrok si zapomnimo, in poženemo `Net/ k-Neighbours/ Output` (izračunamo oddaljenosti od te osebe), ter z ukazom `Operations/ Extract from Network/ Partition` (vnesemo 0 in 1) izločimo le osebo samo (ki je na oddaljenosti 0) in njene potomce (na oddaljenosti 1).

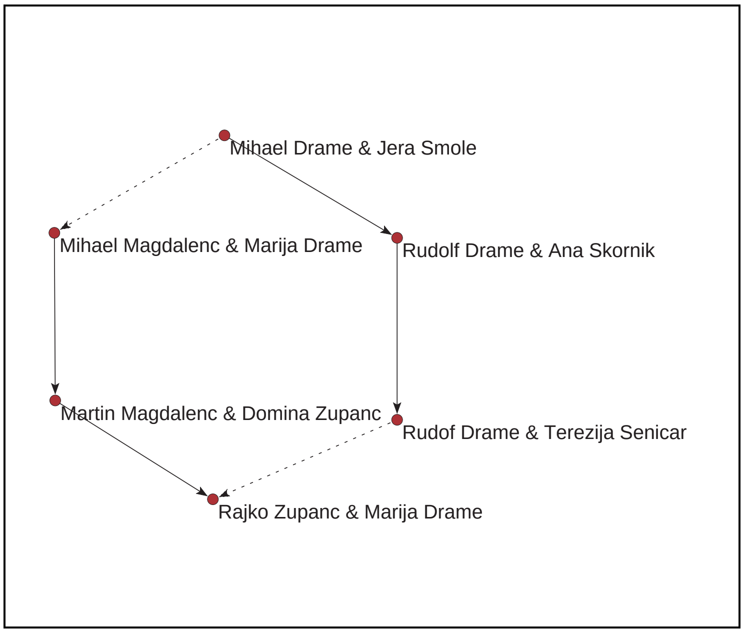
Vzemimo spet podatke `Drame.ged`. Če preberemo datoteko v navadi obliki, dobimo kot rezultat osebo z največ otroki (skupno), kar je prikazano na levi strani slike 5; če pa datoteko preberemo v obliki p-graph, dobimo zakonski par z največ otroki (slika 5, desno).

3.7 Iskanje zanimivih vzorcev

S programom `Pajek` lahko odkrijemo tudi zanimive vzorce v rodovnikih (poroke med sorodniki). Rodovnik najprej preberemo v obliki p-graph. Nad njim izračunamo dvopovezane komponente z ukazom `Net/ Components/ Bi-Components`, pogledamo informacijo o komponentah `Info/ Hierarchy` ter odtipkamo neko številko, npr. 10 – želimo videti 10 največjih dvopovezanih komponent. Če se v rezultatu pojavijo samo komponente velikosti 2, je vse v redu. Večje komponente pa so bolj zanimive. Številko izbrane komponente si zapomnimo in jo izločimo z ukazom `Hierarchy/ Extract Cluster` (dobimo množico oseb v tej komponenti), nato pa to množico izločimo iz grafa z ukazom `Operations/`



Slika 5: Posameznik z največ otroki (navadna oblika, levo) in zakonski par z največ otroki (p-graph, desno).



Slika 6: Poroka med sorodniki.

`Extract from Network/ Cluster`, ter jo po želji narišemo kot smo že opisali. Slika 6 prikazuje enega od zanimivih vzorcev v datoteki `Drame.ged`.