

**Univerza v Ljubljani
podiplomski študij statistike**

**Analiza omrežij
Omrežja vsepovsod**

**Vladimir Batagej
Anuška Ferligoj**

Univerza v Ljubljani

Ljubljana, 3. november 2003

Kazalo

1	Kako do omrežja?	1
2	Uporaba že zbranih omrežij	2
5	Predelava že zbranih omrežij	5
8	Slovarska omrežja	8
9	Omrežja sklicevanj	9
10	Omrežja sodelovanj	10
12	Najbližji sosedi	12
13	Transformacije	13
14	Internetska omrežja	14
15	Rabutanje	15
16	Slučajna omrežja	16

Kako do omrežja?

Pri zbiranju podatkov o omrežju $N = (V, L, F_V, F_L)$ se moramo odločiti, kaj je množica enot (točk) – *meje omrežja*, kdaj sta točki povezani – *polnost omrežja* in katere lastnosti točk/povezav bomo upoštevali.

Ta vprašanja so še posebej pereča pri merjenju družbenih omrežij (vprašalniki, pogovori, opazovanja, arhivski zapiski, poskusi, ...). Nekateri 'enote' nočeje sodelovati. Nekateri postopki merjenja, na primer, omejujejo število sosedov ...

Pri velikih množicah enot si ne moremo privoščiti polnega opisa. Omrežje izmerimo samo za izbrane enote (in njihove sosede). Tako dobljena omrežja imenujemo *osebna omrežja*.

Merjenju omrežij bo namenjeno posebno predavanje.

Uporaba že zbranih omrežij

Veliko podatkov o omrežjih pa je že zbranih, pogosto so tudi v računalniški obliki.

Pajek omogoča branje več drugih oblik opisa omrežij: UCINETov datoteke DL, grafi projekta Vega, kemični opisi molekul MDLMOL, MAC in BS ter rodovniki v obliki GEDCOM.

Davis.DAT, C84N24.VGR, MDL, 1CRN.BS, DNA.BS, ADF073.MAC, Bouchard.GED.

Rodovniki

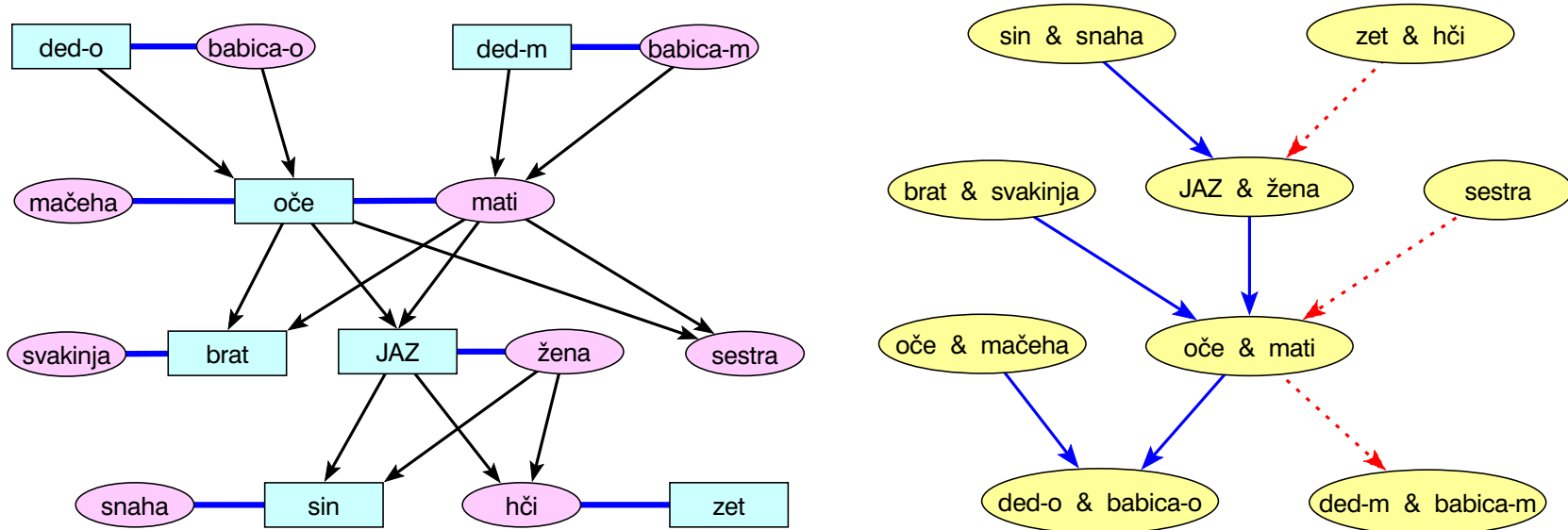
Za opis rodovnikov se najpogosteje uporablja oblika zapisa GEDCOM (*GEDCOM standard 5.5*).

Veliko rodovnikov (datoteke * .GED) najdemo na spletu – na primer *Roper's GEDCOMs* ali *Isle-of-Man GEDCOMs*.

Za pripravo in vzdrževanje rodovnikov je na voljo več programov: prosti *GIM* in tržni *Brothers Keeper* (obstaja tudi slovenska različica – *SRD*).

Orejevi in parni grafi

Običajna predstavitev rodovnika z grafom, *Orejev* graf, ima za točke posameznike in združuje dve relaciji: *sta poročena* (modra neusmerjena) in *ima otroka* (črna usmerjena). V *parnem grafu* so točke poročeni pari ali neporočeni posamezniki, in relaciji *je sin* (polna modra) in *je hči* (črtkana rdeča).



Več o parnih grafih *D. White*.

Predelava že zbranih omrežij

Ragusa – rodovniki v doktoratu:

Mahnken, Irmgard. 1960. Dubrovački patricijat u XIV veku. Beograd, Naučno delo.

The Internet Movie Database.

Omrežja molekul

V zbirki **Brookhaven Protein Data Bank** lahko najdemo veliko velikih organskih molekul (npr. Simian / 1AZ5.pdb) opisanih v obliki PDB.

Molekulo si lahko ogledamo s programom Rasmol (*RasMol*, *program*, *RasWin*) ali *Protein Explorer*.

Molekulo v obliki PDB lahko predelamo v obliko BS, ki jo pozna **Pa jek**, s programom *BabelWin* + *Babel16*.

Odnosi med državami

Paul Hensel's International Relations Data Site,

International Conflict and Cooperation Data,

Correlates of War,

Kansas Event Data System *KEDS*,

KEDSi na Pajkovih datotekah,

Prekodirni programi v R-ju.

Slovarska omrežja

Online Dictionary of Library and Information Science *ODLIS*, *Odlis.net*,
2909 / 18419

Free On-line Dictionary of Computing *FOLDOC*, *Foldoc2b.net*, 133356 /
120238

Wordnet, *ConceptNet*, *OpenCyc*.

Artlex.

The Edinburgh Associative Thesaurus (*EAT*), *EAT*,

Članek.

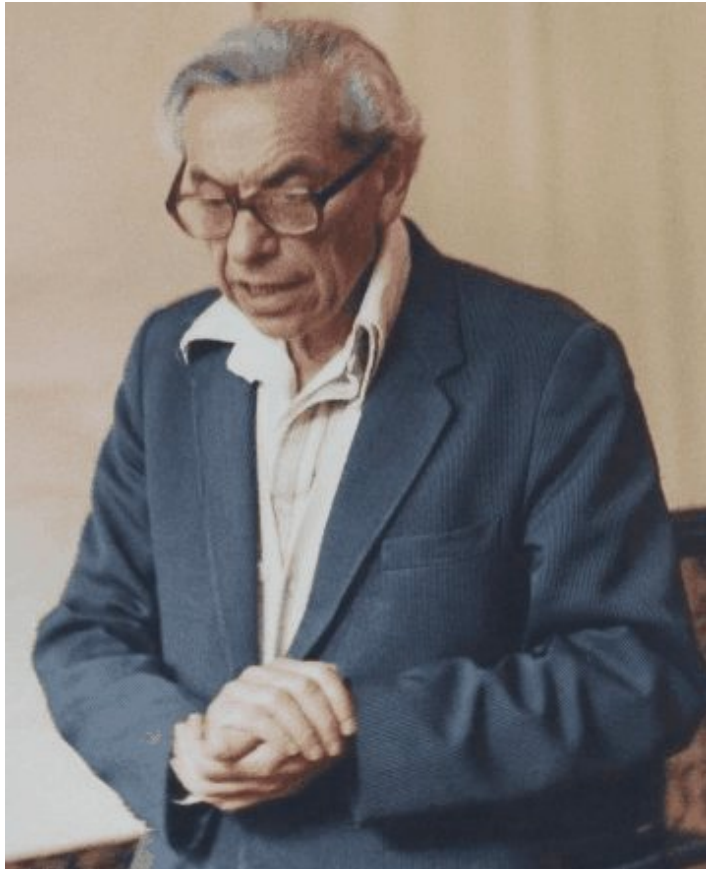
Omrežja sklicevanj



E. Garfield: *HistCite* / *Pajek*,
članki.

Primer velikega omrežja sklicevanj je
omrežje *US Patents* / *Nber*,
 $n = 3774768$, $m = 16522438$ (1 zanka).

Omrežja sodelovanj



V omrežjih sodelovanj so enote osebe ali ustanove. Enoti sta povezani, če sta sodelovali pri skupnem delu. Utež povezave je število skupnih del.

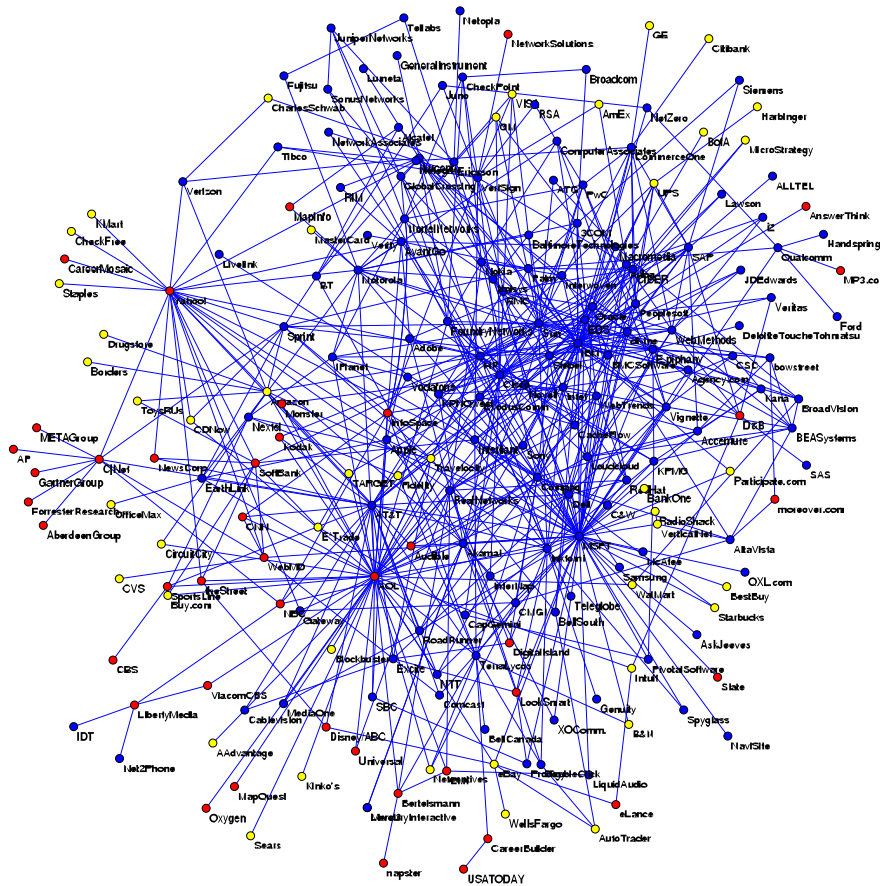
Najbolj poznano omrežje sodelovanja je *The Erdős Number Project*, *Erdos.net*.

Bogat vir podatkov za izgradnjo omrežij sodelovanj so bibliografije v Bib_TE_Xu *Nelson H. F. Beebe's Bibliographies Page*.

Taka je na primer B. Jones-ova bibliografija računalniške geometrije *Computational geometry database* (2002), *FTP*, *Geom.net*.

Pri pripravi omrežja sodelovanj iz izvornih podatkov si lahko pomagamo z ustreznimi programi. Nato pa sledi mukotržno čiščenje - ugotavljanje enot.

Krebsova Internetska podjetja



Omrežje sestavljajo izbrana
Internetska podjetja v obdobju
1998 do 2001.

$$n = 219, m = 631.$$

rdeča – vsebina,
modra – podpora,
rumena – trgovina.

Podjetji sta povezani, če sta
objavili skupni posel ali sode-
lovanje.

Spletni naslov: <http://www.orgnet.com/netindustry.html>.

Recode, InfoRapid.

Najbližji sosedi

Recimo, da imamo na množici enot V dano mero različnosti $d(u, v)$. Glede na d lahko vpeljemo dve vrsti omrežij:

Najbližji sosedi: $N(k) = (V, A, d)$

$(u, v) \in A \Leftrightarrow v$ je med k najbližjimi sosedi točke u

Omrežje okolic: $N(r) = (V, E, d)$

$(u : v) \in E \Leftrightarrow d(u, v) \leq r$

Ta omrežja so povezava z običajno analizo podatkov. Še vedno odprto vprašanje: učinkoviti postopki za določanje teh omrežij.

Multivariatne pajčevine – določitev omrežij sosedov v R-ju.

Transformacije

Besedni graf – točke so besede; besedi sta povezani, če lahko eno dobimo iz druge s spremembo ene črke. *DIC28, Članek.*

Grafi iger – točke so stanja, povezave pa dovoljeni prehodi med njimi.

Omrežja iz besedil – besedi sta povezani, če se v besedilu pojavita dovolj blizu skupaj. Utež povezave je število takih ponovitev. Primer *CRA.*

Internetska omrežja

Internet Mapping Project.

Sosednost na spletu (Najdi.si), Grobelnik.

E-mail in druge storitve.

Dnevnik dosegov strežnika.

Cybergeography, CAIDA.

Rabutanje

Za pridobivanje izbranih podatkov iz (večih) spletnih strani lahko napišemo posebne programe *web wrappers*. Ti iz posamezne strani izluščijo iskane podatke in jih shranijo – pogosto v obliki XML.

Primeri v R-ju: *Naslovi patentov*, *Knjige z Amazon*.

Ker je pisanje teh programov za običajnega uporabnika prezapleteno, je bilo razvitih več *orodij*, ki jih ustvarijo iz uporabnikovih opisov/zahtev (*članek / seznam / LAPIS*).

Med prostimi orodji sta zanimiva še XWRAP (*opis / stran*) in TSIMMIS (*opis / stran*).

Med tržnimi orodji trenutno prevladuje *lixto*.

Še nekaj naslovov *1*, *2*, *3*.

Slučajna omrežja

Omrežja lahko tudi sami ustvarimo z nekim slučajnim postopkom. **Ozadja** teh postopkov bomo spoznali kasneje. Vgrajeni so v Pajka (Net / Random network), lahko pa jih tudi sami zapišemo z razmeroma kratkimi **postopki** v R-ju.

Na voljo je tudi program **GeneoRnd** za ustvarjanje slučajnih rodovnikov.