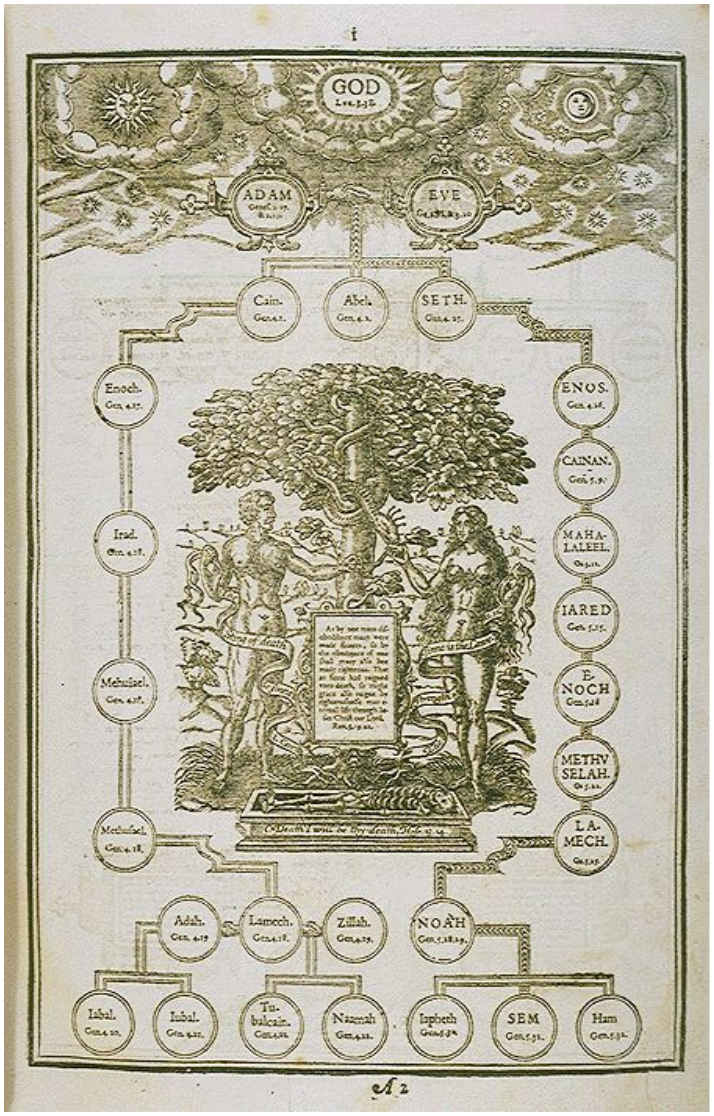


Analysing Large Genealogical Networks with Pajek

Vladimir Batagelj
Andrej Mrvar

University of Ljubljana
Slovenia

Sources et ressources pour les sciences sociales

EHESS, Paris, 9-11. december 2004

Contents

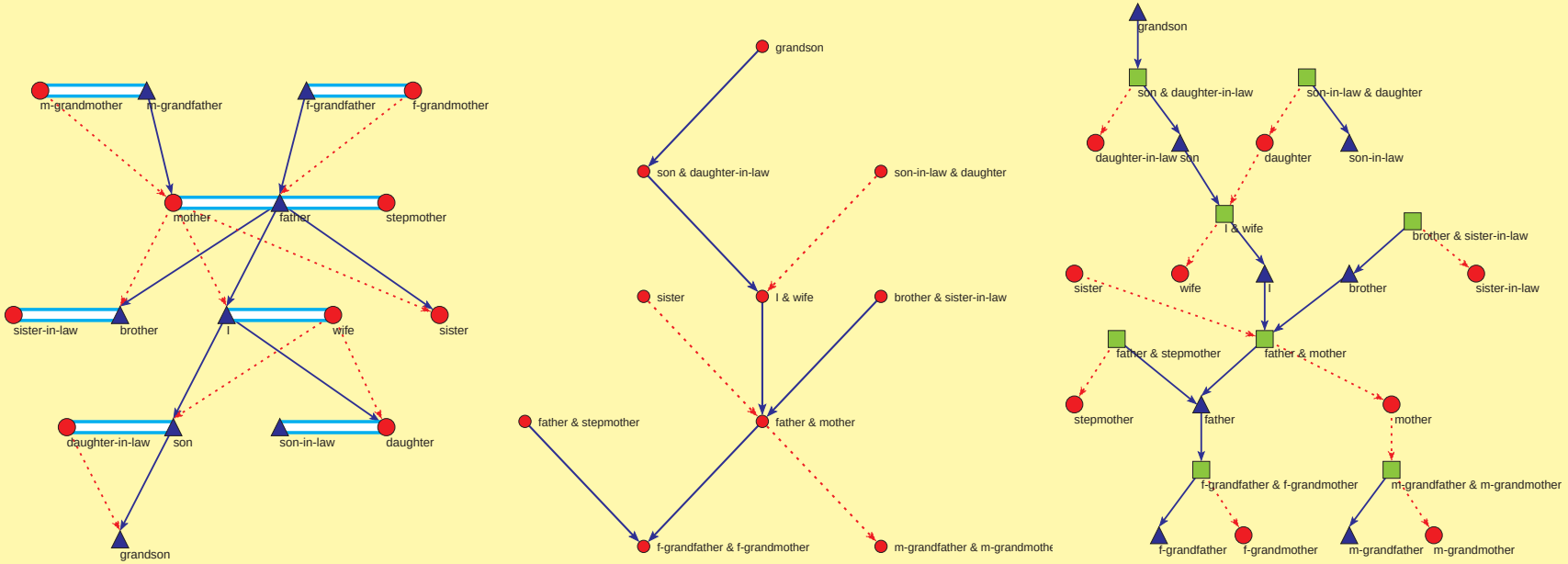
GEDCOM

GEDCOM is a standard for storing genealogical data, which is used to interchange and combine data from different programs, which were used for entering the data.

```
0 HEAD
1 FILE ROYALS.GED
...
0 @I58@ INDI
1 NAME Charles Philip Arthur/Windsor/
1 TITL Prince
1 SEX M
1 BIRT
2 DATE 14 NOV 1948
2 PLAC Buckingham Palace, London
1 CHR
2 DATE 15 DEC 1948
2 PLAC Buckingham Palace, Music Room
1 FAMS @F16@
1 FAMC @F14@
...
...
0 @I65@ INDI
1 NAME Diana Frances /Spencer/
1 TITL Lady
1 SEX F
1 BIRT
2 DATE 1 JUL 1961
2 PLAC Park House, Sandringham
1 CHR
2 PLAC Sandringham, Church
1 FAMS @F16@
1 FAMC @F78@
...
...
```

```
0 @I115@ INDI
1 NAME William Arthur Philip/Windsor/
1 TITL Prince
1 SEX M
1 BIRT
2 DATE 21 JUN 1982
2 PLAC St.Mary's Hospital, Paddington
1 CHR
2 DATE 4 AUG 1982
2 PLAC Music Room, Buckingham Palace
1 FAMC @F16@
...
0 @I116@ INDI
1 NAME Henry Charles Albert/Windsor/
1 TITL Prince
1 SEX M
1 BIRT
2 DATE 15 SEP 1984
2 PLAC St.Mary's Hosp., Paddington
1 FAMC @F16@
...
0 @F16@ FAM
1 HUSB @I58@
1 WIFE @I65@
1 CHIL @I115@
1 CHIL @I116@
1 DIV N
1 MARR
2 DATE 29 JUL 1981
2 PLAC St.Paul's Cathedral, London
```

Network representations of genealogies



Ore graph, p-graph, and bipartite p-graph

Properties of representations

p-graphs and bipartite p-graphs have many advantages:

- there are less vertices and lines in p-graphs than in corresponding Ore graphs;
- p-graphs are directed, acyclic networks;
- every semi-cycle of the p-graph corresponds to a *relinking marriage*. There exist two types of relinking marriages: *blood* marriage: e.g., marriage among brother and sister, and *non-blood* marriage: e.g., two brothers marry two sisters from another family.
- p-graphs are more suitable for analyses.

Bipartite p-graphs have an additional advantage: we can distinguish between a married uncle and a remarriage of a father. This property enables us, for example, to find marriages between half-brothers and half-sisters.

Genealogies are sparse networks

A genealogy is *regular* if every person in it has at most two parents. Genealogies are *sparse* networks – number of lines is of the same order as the number of vertices.

For a *regular Ore genealogy* we have (V – vertices, A – arcs, E – edges):

$$|A| \leq 2|V|, \quad |E| \leq \frac{1}{2}|V|, \quad |L| = |A| + |E| \leq \frac{5}{2}|V|$$

p-graphs are almost trees – deviations from trees are caused by relinking marriages (V_p, A_p – vertices and arcs of p-graph):

$$|V_p| = |V| - |E| + n_{mult}, \quad |V| \geq |V_p| \geq \frac{1}{2}|V|, \quad |A_p| \leq 2|V_p|$$

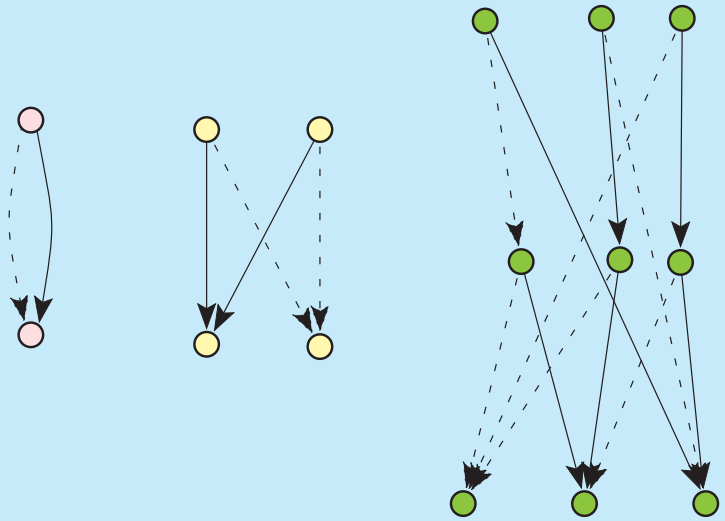
and for a bipartite p-graph, we have

$$|V| \leq |V_b| \leq \frac{3}{2}|V|, \quad |A_b| \leq 2|V| + n_{mult}$$

Number of vertices and lines in Ore and p-graphs

data	$\|V\|$	$\|E\|$	$\|A\|$	$\frac{\|L\|}{\|V\|}$	$\|V_i\|$	n_{mult}	$\|V_p\|$	$\|A_p\|$	$\frac{\|A_p\|}{\|V_p\|}$
Drame	29606	8256	41814	1.69	13937	843	22193	21862	0.99
Hawlina	7405	2406	9908	1.66	2808	215	5214	5306	1.02
Marcus	702	215	919	1.62	292	20	507	496	0.98
Mazol	2532	856	3347	1.66	894	74	1750	1794	1.03
President	2145	978	2223	1.49	282	93	1260	1222	0.97
Royale	17774	7382	25822	1.87	4441	1431	11823	15063	1.27
Loka	47956	14154	68052	1.71	21074	1426	35228	36192	1.03
Silba	6427	2217	9627	1.84	2263	270	4480	5281	1.18
Ragusa	5999	2002	9315	1.89	2347	379	4376	5336	1.22
Tur	1269	407	1987	1.89	549	94	956	1114	1.17
Royal92	3010	1138	3724	1.62	1003	269	2141	2259	1.06
Little	25968	8778	34640	1.67	8412				1.01
Mumma	34224	11334	45565	1.66	11556				1.00
Tilltson	42559	12796	54043	1.57	16967				1.00

Relinking index



Let n denotes number of vertices in p-graph, m number of arcs, k number of weakly connected components, and M number of maximal vertices (vertices having output degree 0, $M \geq 1$).

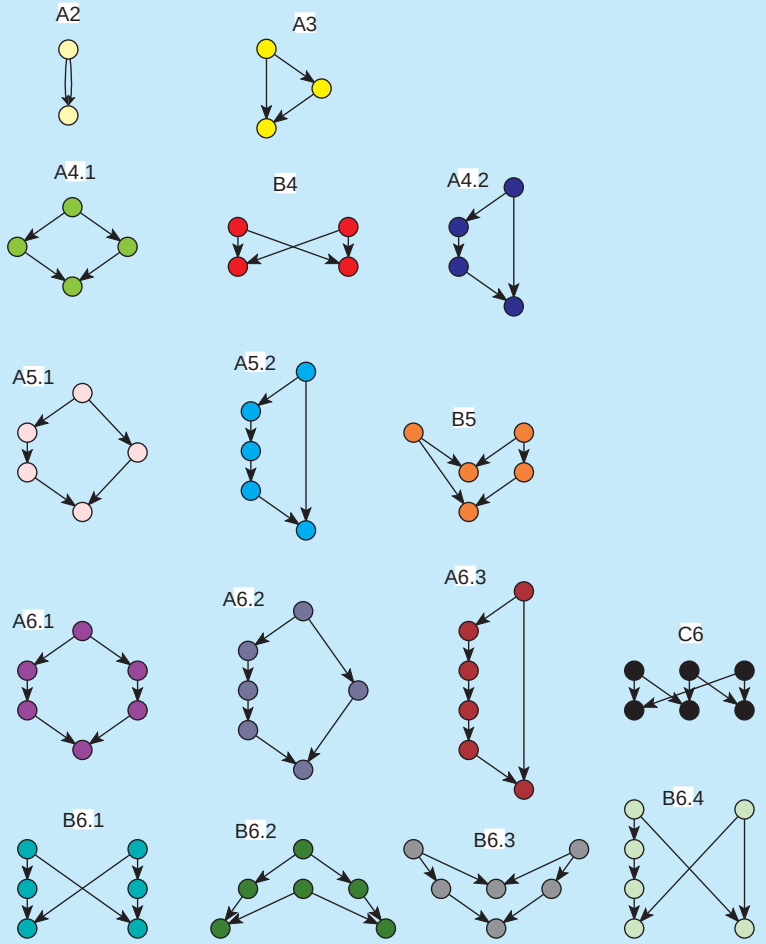
The *relinking index* is defined as:

$$RI = \frac{k + m - n}{k + n - 2M}$$

For a trivial graph (having only one vertex) we define $RI = 0$.

It holds $0 \leq RI \leq 1$. $RI = 0$ iff network is a forest.

Relinking patterns in p-graphs



All possible relinking marriages in p-graphs with 2 to 6 vertices. Patterns are labeled as follows:

- first character – number of first vertices: A – single, B – two, C – three.
- second character: number of vertices in pattern (2, 3, 4, 5, or 6).
- last character: identifier (if the two first characters are identical).

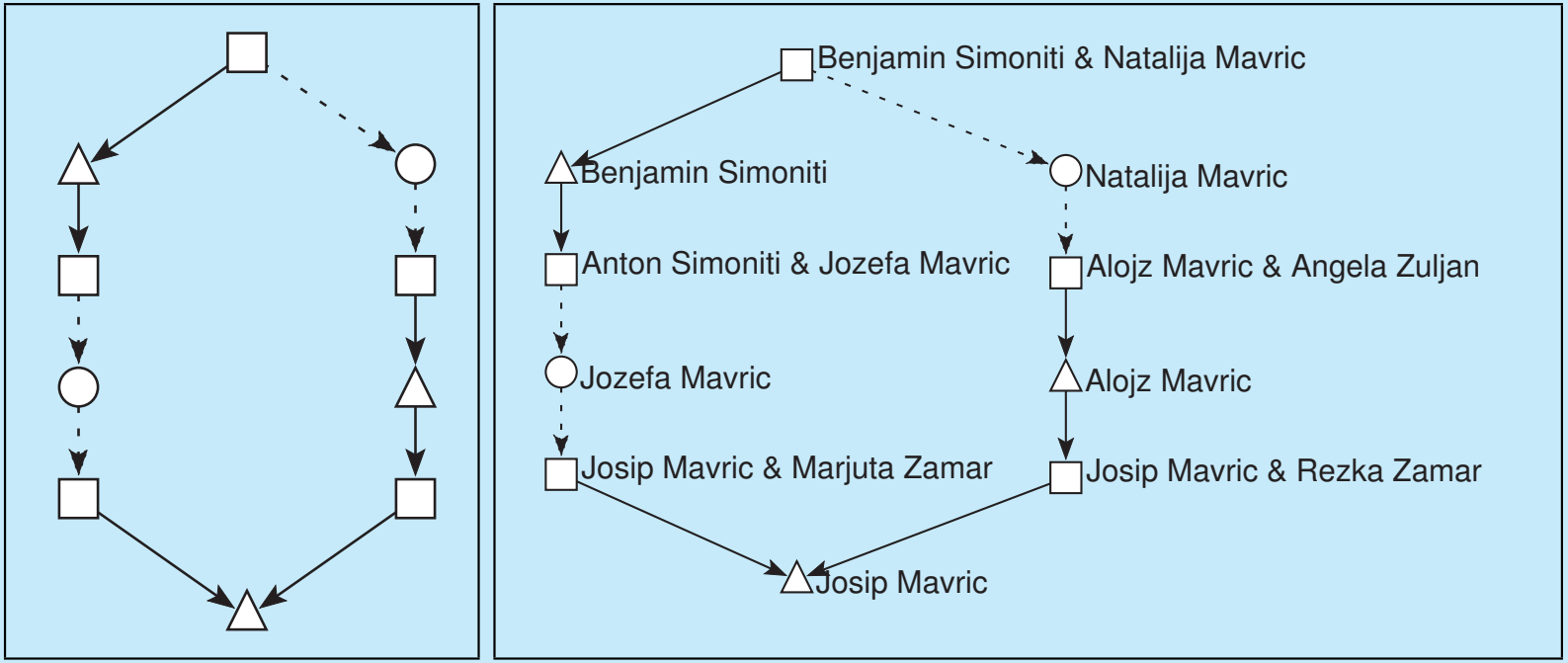
Patterns denoted by A are exactly the blood marriages. In every pattern the number of first vertices equals to the number of last vertices.

Frequencies normalized with number of couples in p-graph $\times$ 1000.

pattern	Loka	Silba	Ragusa	Turcs	Royal
A2	0.07	0.00	0.00	0.00	0.00
A3	0.07	0.00	0.00	0.00	2.64
A4.1	0.85	2.26	1.50	**159.71**	18.45
B4	3.82	11.28	10.49	**98.28**	6.15
A4.2	0.00	0.00	0.00	0.00	0.00
A5.1	0.64	3.16	2.00	36.86	11.42
A5.2	0.00	0.00	0.00	0.00	0.00
B5	1.34	4.96	23.48	46.68	7.03
A6.1	1.98	12.63	1.00	**169.53**	11.42
A6.2	0.00	0.90	0.00	0.00	0.88
A6.3	0.00	0.00	0.00	0.00	0.00
C6	0.71	5.41	9.49	36.86	4.39
B6.1	0.00	0.45	1.00	0.00	0.00
B6.2	1.91	17.59	31.47	**130.22**	10.54
B6.3	3.32	13.53	40.96	**113.02**	11.42
B6.4	0.00	0.00	2.50	7.37	0.00
Sum	14.70	72.17	123.88	798.53	84.36

Most of the relinking marriages happened in the genealogy of Turkish nomads; the second is Ragusa while in other genealogies they are much less frequent.

Bipartite p-graphs: Marriage among half-cousins



Additional options in Pajek

Drawing: drawing of acyclic networks – macro **`layers`**; fixing movement in y-direction in manual drawing; grid; vector values (year) as y-axis.

Operations on sequences of networks (implicit loops).

GeneoRnd - random genealogies generator

For generating random genealogies a special program GeneoRnd was written (running in command mode - DOS). **GeneoRnd** produces required number of random genealogies exported as Pajek project file (Ore graph or/and p-graph format) or/and GEDCOM files. Besides networks, GeneoRnd adds on ***.paj** file also the vector of birth year for each person and the gendre partition (1-Female/2-Male). The values of lines are the years when they were created (marriage, birth). On the file **report.lst** it writes the trace of generation process. A 'random' layout for each network is provided on the ***.paj** file: random x-coordinate, y-coordinate proportional to year (birth, marriage for couples in p-graphs). Additional information - circle/triangle for sex, solid/dotted for son/dauther is also added.

. . . GeneoRnd

Program GeneoRnd can be controled with the following parameters: format (Ore-graph, p-graph, GEDCOM), number of genealogies, number of vertices, number of initial vertices , polygamy allowed, sex F probability, death F probability, death M probability, divorce probability, marriage probability, reproduction probability, F reproduction start, F reproduction end, F life span, M reproduction start, M reproduction end, and M life span. Program GeneoRnd saves the values of parameters on the file **`gen.par`**.

Links

Programs Pajek and GeneoRnd are freely available at

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

http://vlado.fmf.uni-lj.si/pub/networks/pajek/howto/geneoRnd.htm

See also

A. Mrvar, V. Batagelj: *Relinking Marriages in Genealogies.*
Metodološki zvezki, Vol. 1, No. 2, 2004, 407-418.

These slides are available at

http://vlado.fmf.uni-lj.si/pub/networks/doc/seminar/Geneo04.pdf