Metodološki zvezki, Vol. 1, No. 2, 2004, 407-418

Relinking Marriages in Genealogies

Andrej Mrvar[1] and Vladimir Batagelj[2]

Abstract

Genealogies can be represented as graphs in different ways: as *Ore graphs*, as *p-graphs*, or as *bipartite p-graphs*. p-graphs are usually more suitable for analyses. Some approaches to analysis of large genealogies implemented in program Pajek are presented and illustrated with analysis of some large genealogies.

1 Sources of genealogies

People collect genealogical data for several different reasons/purposes:

- Research on different cultures in history, sociology and anthropology (White et al., 1999), where kinship is taken as a fundamental social relation.
- Genealogies of families and/or territorial units, e.g.,
  - Mormons genealogy (MyFamily.com, 2004)
  - genealogy of Škofja Loka district (Hawlina, 2004)
  - genealogy of American presidents (Tompsett, 1993)
- Special genealogies
  - Students and their PHD thesis advisors:
    Theoretical Computer Science Genealogy (Johnson and Parberry, 1993)
  - gods (antique). See Hawlina (2004).

There also exist many programs for genealogical data entry and maintenance (`GIM, Brother's Keeper, Family Tree Maker,...`), but only few analyses can be done using the programs. We use `Pajek` for analyses and visualization of genealogies.

[1] Faculty of Social Sciences, University of Ljubljana, Slovenia; andrej.mrvar@uni-lj.si
[2] Faculty of Mathematics and Physics, Univ. of Ljubljana, Slovenia; vladimir.batagelj@uni-lj.si

2 GEDCOM standard

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. The following lines are extracted from the GEDCOM file of European Royal families.

```
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
```

From data represented in the described way we can generate several graphs as explained in next chapters.

3 Representation of genealogies using networks

Genealogies can be represented as networks in different ways: as *Ore-graph*, as *p-graph*, and as *bipartite p-graph*.

3.1 Ore-graph

In an Ore graph of genealogy every person is represented by a vertex, marriages are represented with edges and relation *is a parent of* is represented as arcs pointing from each of the parents to their children. See Figure 1.

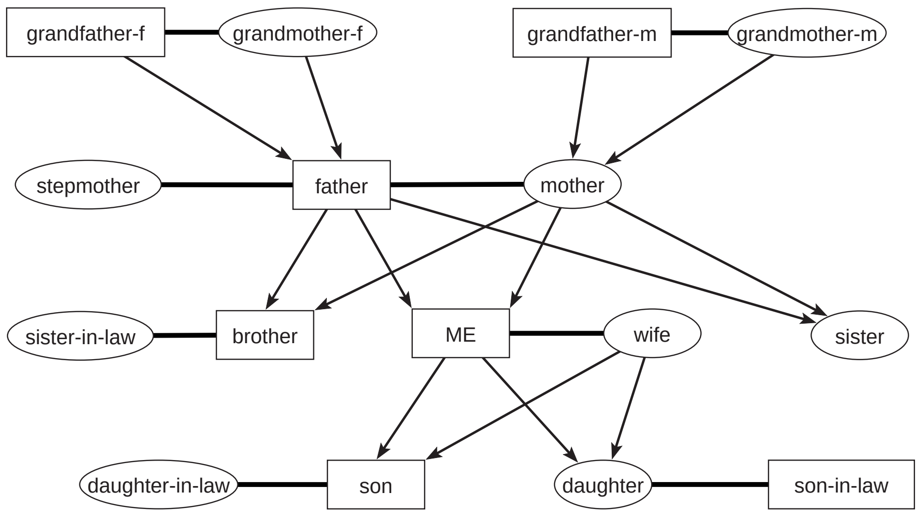

Figure 1: Ore graph.

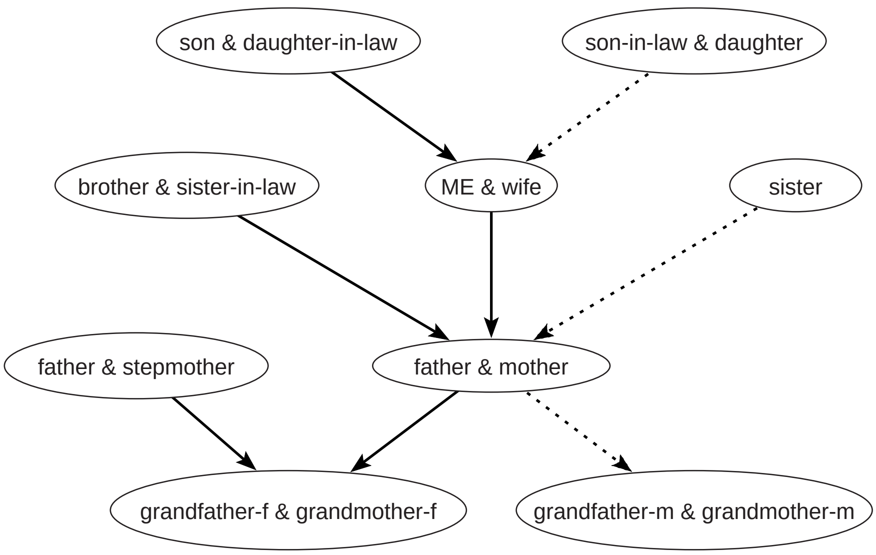

Figure 2: p-graph.

3.2 p-graph

In a p-graph vertices represent individuals or couples. In case that person is not married yet (s)he is represented by a vertex, otherwise the person is represented with the partner in a common vertex. There are only arcs in p-graphs – they point from children to their parents (Figure 2). The solid arcs represent the relation *is a son of* and the dotted arcs represent relation *is a daughter of*.

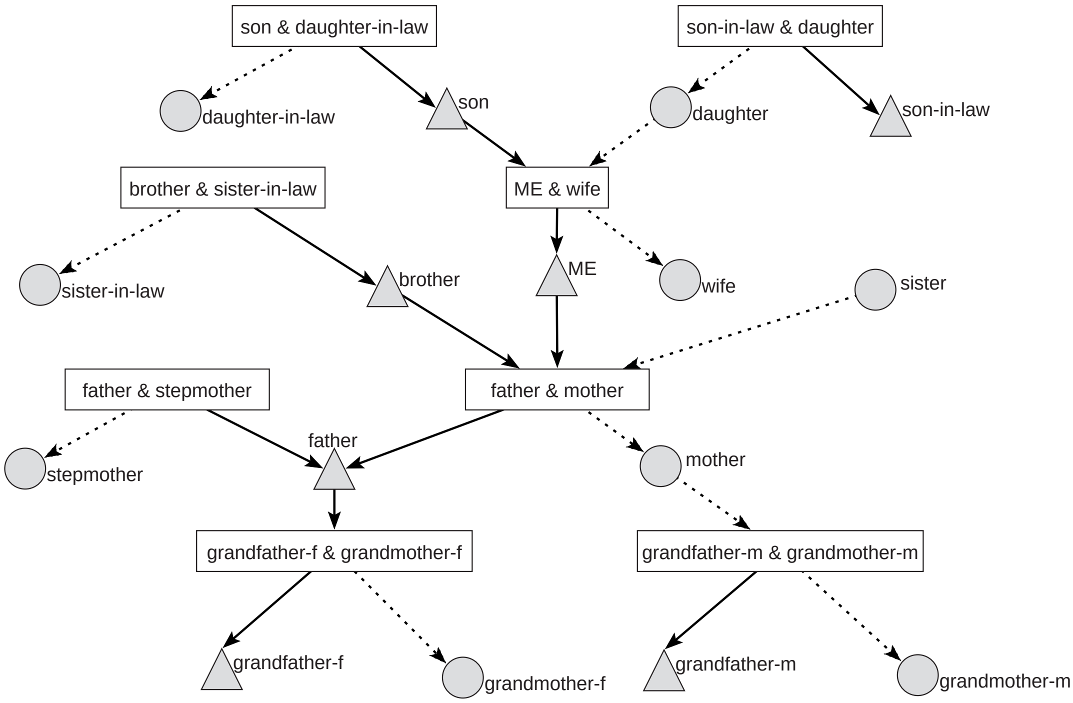

Figure 3: Bipartite p-graph.

3.3 Bipartite p-graph

A bipartite p-graph has two kinds of vertices – vertices representing couples (rectangles) and vertices representing individuals (circles for women and triangles for men) – therefore each married person is involved in two kinds of vertices (or even more if he/she is involved in multiple marriages). Arcs again point from children to their parents (see Figure 3).

3.4 Comparison of different presentations

p-graphs and bipartite p-graphs have many advantages (see White et al., 1999):

- 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.
  - 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 (see Figures 2 and 3). This property enables us, for example, to find marriages between half-brothers and half-sisters.

4 Genealogies are sparse networks

We will call genealogy *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. In this section some approximations and bounds on the number of lines in different kinds of regular genealogies are given.

For the directed part of an *regular Ore genealogy* the approximation of the number of arcs A is:

$$|A| = \sum_{v \in V} d_{in}(v) \leq 2|V|$$

where V is set of vertices, and $d_{in}(v)$ input degree of vertex v, $d_{in}(v) \leq 2$. Most of the persons are married only once, some are not married. For the undirected part of an *Ore* genealogy the number of edges (E) is

$$|E| \leq \frac{1}{2}|V|$$

Therefore

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

p-graphs are almost trees – deviations from trees are caused by relinking marriages. Let us denote the number of vertices of *p-graph* with $|V_p|$ and the number of multiple marriages with n_{mult}. Then, since $|E|$ equals to the number of couples,

$$|V_p| = |V| - |E| + n_{mult}$$

and therefore

$$|V| \geq |V_p| \geq |V| - |E| \geq \frac{1}{2}|V|$$

The number of arcs in p-graph is

$$|A_p| = \sum_{v \in V_p} d_{out}(v) \leq 2|V_p|$$

where $d_{out}(v)$ is output degree of vertex v.

For the number of vertices V_b in a bipartite p-graph, we have

$$|V_b| = |V| + |E|$$

Since $|E| \leq \frac{1}{2}|V|$ we get

$$|V| \leq |V_b| \leq \frac{3}{2}|V|$$

Table 1: Number of vertices and number of lines in Ore graphs and p-graphs for some large networks.

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

For the number of arcs A_b we have

$$|A_b| = |A_p| - n_{mult} + 2|E| \leq 2(|V_p| + |E|) - n_{mult} = 2|V| + n_{mult}$$

To check the results we take several large genealogies and look at the corresponding Ore and p-graphs. A comparison of Ore and p-graph is given in Table 1. In the table the following notation is used:

- *Ore genealogy*: $|V|$ – number of vertices; $|E|$ – number of edges; $|A|$ – number of arcs; $|L| = |E| + |A|$ – total number of lines.
- *p-graph*: $|V_i|$ – number of individuals; n_{mult} – number of multiple marriages; $|V_p| = |V_i| + |E|$ – total number of vertices; $|A_p|$ – number of arcs.

p-graphs are usually used also for visual representation of genealogies. Since they are acyclic graphs the vertices can be assigned to levels (see Figure 4 and Figure 5).

5 Relinking index

The relinking index is a measure of relinking by marriages among persons belonging to the same families. A special case of relinking is a blood-marriage in which the man and woman from the couple have a common ancestor.

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, $M \geq 1$).
If a p-graph is a forest (consists of trees), then $m = n - k$, or $k + m - n = 0$.

In a *regular* genealogy, $m \leq 2(n - M) = 2n - 2M$. Thus:

$$0 \leq k + m - n \leq k + n - 2M$$

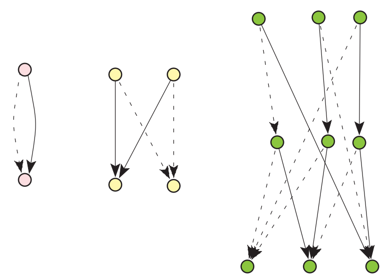

Figure 4: Patterns with relinking index 1 (p-graph).

or

$$0 \leq \frac{k+m-n}{k+n-2M} \leq 1$$

This is called *the relinking index* (RI):

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

If we take a connected genealogy (selected weakly connected component) we get

$$RI = \frac{m-n+1}{n-2M+1}$$

For a trivial graph (having only one vertex) we define $RI = 0$. See also White et al., 1999.

RI has some interesting properties:

- $0 \leq RI \leq 1$
- If a network is a forest/tree, then $RI = 0$ (no relinking).
- For a cycle $h = \frac{m}{2} = \frac{n}{2}$, $RI = \frac{1}{2h-1}$ (the higher depth the weaker relinking). For a cycle of depth 3 (6 vertices) $RI = \frac{1}{5}$.
- There exist genealogies having $RI = 1$ (the highest relinking). Figure 4 shows such situations.
  - marriage between brother and sister ($n = 2, m = 2, k = 1, M = 1$),
  - two brothers married to two sisters from another family ($n = 4, m = 4, k = 1, M = 2$),
  - more complicated situations ($n = 9, m = 12, k = 1, M = 3$).

  Arbitrary large genealogies with $R = 1$ exist.

Usually we compute the relinking index over the biconnected subgraph (or its largest component) of a given genealogy.

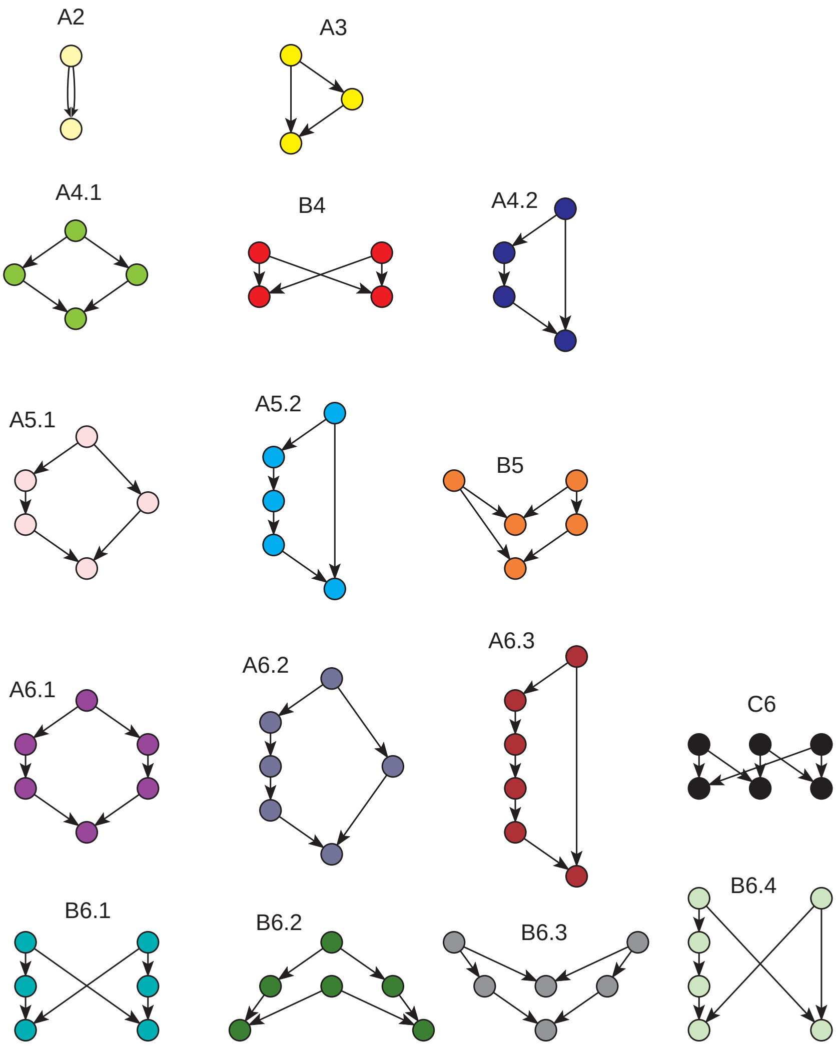

Figure 5: Relinking marriages (p-graphs with 2 to 6 vertices).

6 Relinking patterns in p-graphs

In Figure 5 all possible relinking marriages in p-graphs containing from 2 up to 6 vertices are presented (subtypes and variants as to sex are not included). Patterns are labeled in the following way:

- first character: A – pattern with a single first vertex (vertex without incoming arcs), B – pattern with two, and C – pattern with three first vertices.
- second character: number of vertices in pattern (2, 3, 4, 5, or 6).
- last character: identifier (if the two first characters are identical).

It is easy to see that patterns denoted by A are exactly the blood marriages. Also, in every pattern the number of first vertices equals to the number of last vertices.

Table 2: Comparison of genealogies according to distribution of patterns.

pattern	Loka	Silba	Ragusa	Tur	Royal	$\sum$
A2	1	0	0	0	0	1
A3	1	0	0	0	3	4
A4.1	12	5	3	65	21	106
B4	54	25	21	40	7	147
A4.2	0	0	0	0	0	0
A5.1	9	7	4	15	13	48
A5.2	0	0	0	0	0	0
B5	19	11	47	19	8	104
A6.1	28	28	2	69	13	140
A6.2	0	2	0	0	1	3
A6.3	0	0	0	0	0	0
C6	10	12	19	15	5	61
B6.1	0	1	2	0	0	3
B6.2	27	39	63	53	12	194
B6.3	47	30	82	46	13	218
B6.4	0	0	5	3	0	8
blood-marriages	51	42	9	149	51	302
relinking-marriages	157	118	239	176	45	735
no. of individuals	47956	6427	5999	1269	3010	
vertices in p-graph	35228	4480	4376	956	2141	
no. of couples	14154	2217	2002	407	1138	
no. of bicon. comp.	29	4	2	3	5	
largest bicon. comp.	4095	1340	1446	250	435	
RI (largest bicon. comp.)	0.55	0.78	0.74	0.75	0.37	

6.1 Comparing genealogies

Using frequency distributions for different patterns we can compare different genealogies As examples we take five genealogies:

- `Loka.ged` – genealogy in Škofja Loka district (western part of Slovenia). Data collected by P. Hawlina.
- `Silba.ged` – genealogy of island Silba, Croatia. Data collected by P. Hawlina. Here we expect high relinking because of special geographical position (isolation).
- `Ragusa.ged` – genealogy of Ragusan noble families between 12 and 16 century (Mahnken, 1960; Dremelj et al., 2002). High relinking is expected because of very restricted marriage rules: member of a noble family is supposed to marry another member of a noble family.
- `Tur.ged` – genealogy of Turkish nomads (White et al., 1999). A relinking marriage is a signal of commitment to stay within the nomad group.
- `Royal.ged` – genealogy of European royal families.

Frequency distributions are given in Table 2. We can make the following observations:

- Probability of generation jump for more than one generation is very low (patterns A4.2, A5.2 and A6.3 do not appear in any genealogy, pattern A6.2 appears twice in Silba genealogy and once in Royal, pattern B6.4 appears five times in Ragusa and thee times in Tur).
- In Tur there are a lot of marriages of types A4.1 and A6.1 (marriages among grandchildren and grand grandchildren).
- For all genealogies number of relinking 'non-blood' marriages (e.g. patterns B4, B5, C6, B6.1, B6.2, B6.3 and B6.4) is much higher than number of blood marriages. That is especially true for Ragusa where for 'critical' marriages a special permission of pope was needed.

  There were also economic reasons for non-blood relinking marriages: to keep the wealth and power within selected families.

The number of individuals in genealogy Tur is much lower than in others, Silba and Ragusa are approximately of the same size, while Loka is much larger genealogy, what we must also take into account.

We take this into account in Table 3 with normalized frequencies for number of couples in the p-graph x 1000. It can be easily noticed that most of the relinking marriages happened in the genealogy of Turkish nomads; the second is Ragusa while relinking marriages in other genealogies are much less frequent.

Table 3: Frequencies normalized with number of couples in p-graph × 1000.

pattern	Loka	Silba	Ragusa	Tur	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

Using p-graphs, we cannot distinguish persons married several times. In this case we must use bipartite p-graphs. Using bipartite p-graphs we can find marriages between half-brothers and half-sisters (as pattern shown on the left side of Figure 6). In the five genealogies we found only one such example in Royal.ged (right side of Figure 6).

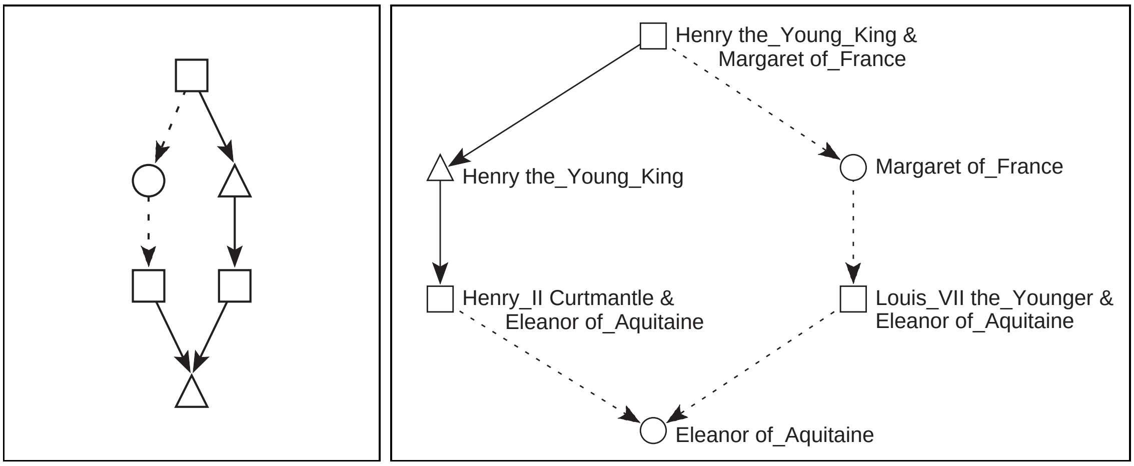

Figure 6: Bipartite p-graphs: Marriage between half-brother and half-sister (left) and example of such marriage (right).

There exist marriages between half-cousins (Figure 7, left). We found one such marriage in the Loka genealogy (right side of Figure 7) and four in the Turkish genealogy.

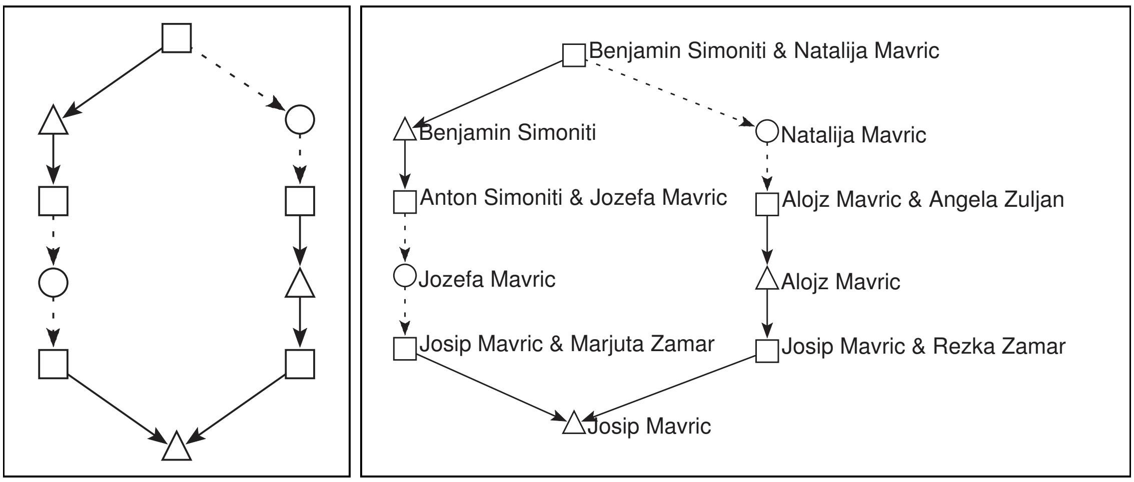

Figure 7: Bipartite p-graphs: Marriage among half-cousins (left), and example of such marriage (right).

References

[1] Batagelj, V. (1996): Ragusan families marriage networks. *Developments in Data Analysis, Metodološki zvezki*, **12**, Ljubljana: FDV, 217-228.

[2] Dremelj, P., Mrvar, A., and Batagelj, V. (1999): Rodovnik dubrovniških plemiških družin med 12. in 16. stoletjem. *Drevesa*. Bilten slovenskega rodoslovnega društva. **1-4**, 4-11.

[3] Dremelj, P., Mrvar, A., and Batagelj, V. (2002): Analiza rodoslova dubrovačkog vlasteoskog kruga pomoču programa Pajek. *Anali Zavoda povij. znan. Hrvat. akad. znan. umjet. Dubr.* **40**, 105-126.

[4] Family History Department (1996): Standard GEDCOM. http://homepages.rootsweb.com/~pmcbride/gedcom/55gctoc.htm

[5] Hawlina, P. (2004): Slovenian Genealogical Society. http://genealogy.ijp.si/slovrd/rd.htm

[6] Johnson, D.S. and Parberry, I. (1993): Theoretical Computer Science Genealogy. http://sigact.acm.org/genealogy/

[7] Krivošič, S. (1990): *Stanovništvo Dubrovnika i Demografske Promjene u Prošlosti*. Dubrovnik: Zavod za povijesne znanosti JAZU u Dubrovniku.

[8] Mahnken, I. (1960): *Dubrovački Patricijat u XIV Veku*. Beograd: Naučno delo.

[9] Mrvar, A. and Batagelj, V. (1997): Pajek – program za analizo obsežnih omrežij. Uporaba v rodoslovju. *Drevesa*. Bilten slovenskega rodoslovnega društva. **12**, december 1997, 4-6.

[10] MyFamily.com (2004): Mormons genealogy. http://www.familytreemaker.com/00000116.html

[11] de Nooy, W., Mrvar, A., and Batagelj, V. (2004): *Exploratory Social Network Analysis with Pajek*. New York: Cambridge University Press. (Forthcoming)

[12] Ore, O. (1963): *Graphs and Their Uses*. New York: Random House.

[13] Tompsett, B. (2004): American presidents GEDCOM file. http://www3.dcs.hull.ac.uk/public/genealogy/presidents/gedx.html

[14] White, D.R., Batagelj, V., and Mrvar, A. (1999): Analyzing Large Kinship and Marriage Networks with Pgraph and Pajek. *Social Science Computer Review – SSCORE*, **17**, 245-274.

[15] White, D.R. and Jorion, P. (1992): Representing and Computing Kinship: A New Approach. *Current Anthropology*, **33**, 454-462.

[16] White, D.R. and Jorion, P. (1996): Kinship Networks and Discrete Structure Theory: Applications and Implications. *Social Networks*, **18**, 267-314.