Photo: Stefan Ernst,***Gartenkreuzspinne / Araneus diadematus***

Analysis of Genealogies with Pajek

Andrej Mrvar
Vladimir Batagelj

University of Ljubljana
Slovenia

Sunbelt XXIV

Ljubljana, Portorož May 12 - 16, 2004

Sources of genealogies

People collect genealogical data for several different reasons/purposes:

- Research of different cultures in sociology, athropology and history – kinship as fundamental social relation
- Genealogies of families and/or teritorial units, e.g.,
  - Mormons genealogy: `http://www.familytreemaker.com/`
  - genealogy of Škofja Loka district: `http://genealogy.ijp.si`
  - genealogy of American presidents: `ftp://www.dcs.hull.ac.uk/public/genealogy/`
- Special genealogies
  - Students and their PhD thesis advisors:
    - Theoretical Computer Science Genealogy: `http://sigact.acm.org/genealogy/`
    - Mathematics

GEDCOM Format

GEDCOM is standard for storing genealogical data, which is used to interchange and combine data from different programs. 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,England
1 CHR
2 DATE 15 DEC 1948
2 PLAC Buckingham,Palace,Music Room,England
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,Norfolk,England
1 CHR
2 PLAC Sandringham,Church,Norfolk,England
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 Hosp.,Paddington,London,England
1 CHR
2 DATE  4 AUG 1982
2 PLAC Music Room,Buckingham,Palace,England
1 FAMC @F16@
...
```

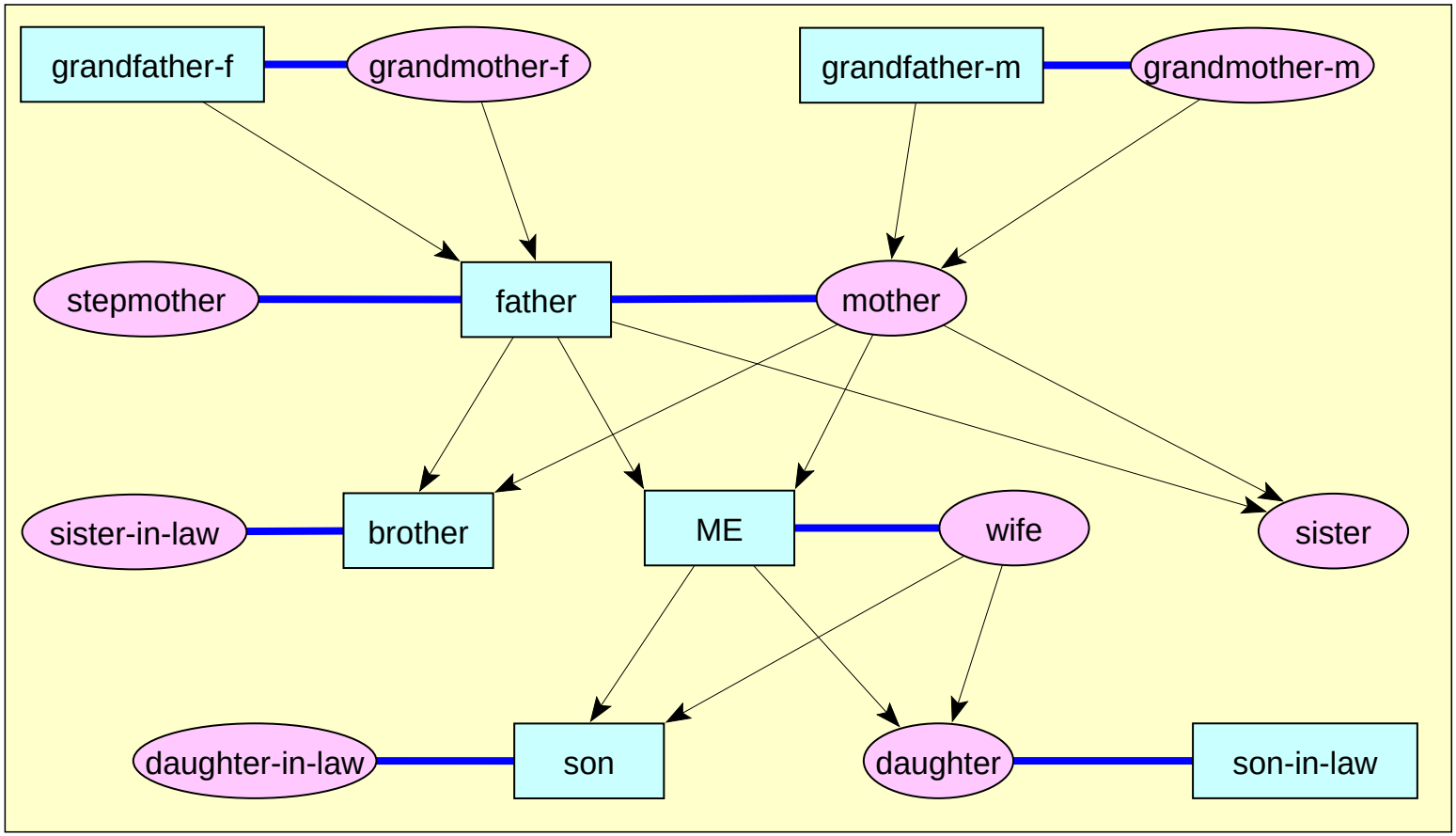
```
...
0 @I116@ INDI
1 NAME Henry Charles Albert/Windsor/
1 TITL Prince
1 SEX M
1 BIRT
2 DATE 15 SEP 1984
2 PLAC St. Mary’s Hosp.,Paddington,London,England
1 FAMC @F16@
...
...
0 @F16@ FAM
1 HUSB @I58@
1 WIFE @I65@
1 CHIL @I115@
1 CHIL @I116@
1 DIV N
1 MARR
2 DATE 29 JUL 1981
2 PLAC St. Paul’s,Cathedral,London,England
```

Representation of genealogies using networks

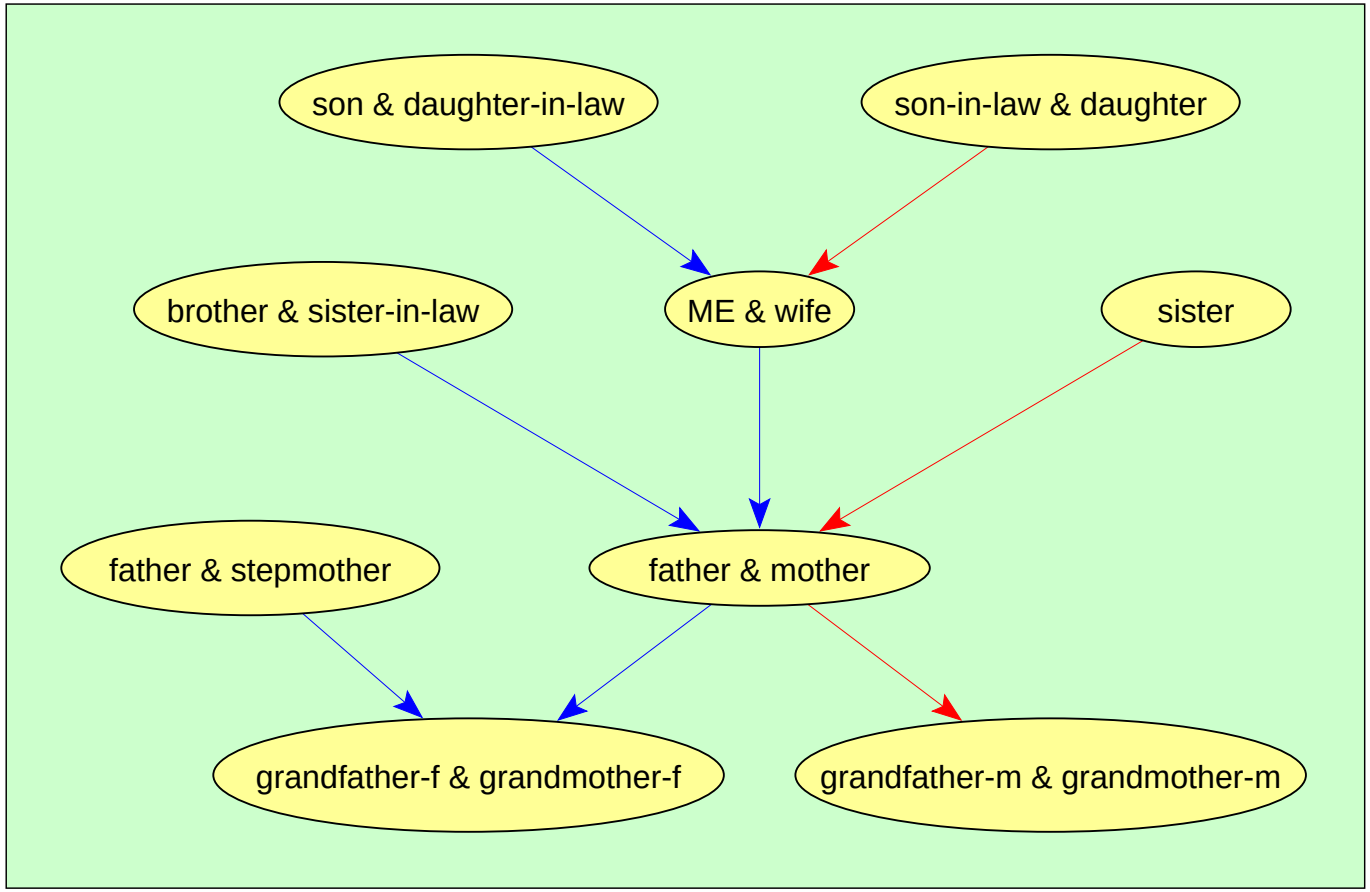
Genealogies can be represented as networks in different ways:

- as Ore-graph,
- as p-graph,
- as bipartite p-graph.

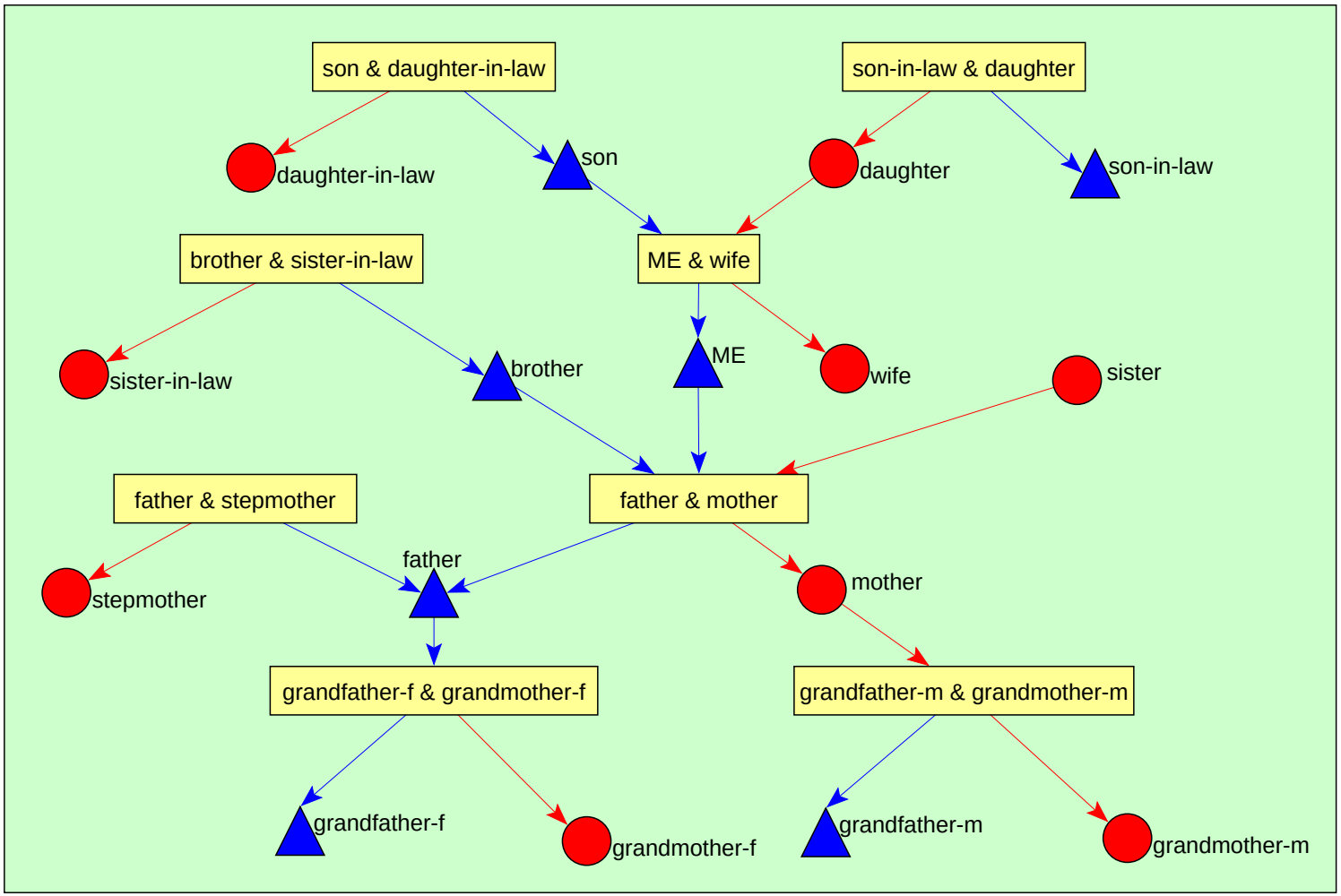
Ore-graph: In Ore-graph every person is represented by a vertex, marriages are represented with edges and relation *is a parent of* as arcs pointing from each of the parents to their children.



p-graph: In p-graph vertices represent individuals or couples. In the case that person is not married yet (s)he is represented by a vertex, otherwise person is represented with the partner in a common vertex. There are only arcs in p-graphs – they point from children to their parents.



Bipartite p-graph: has two types of vertices – vertices representing couples (rectangles) and vertices representing individuals (circles for women and triangles for men). Arcs again point from children to their parents.

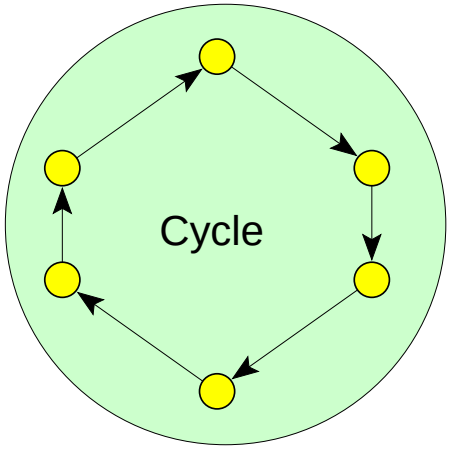


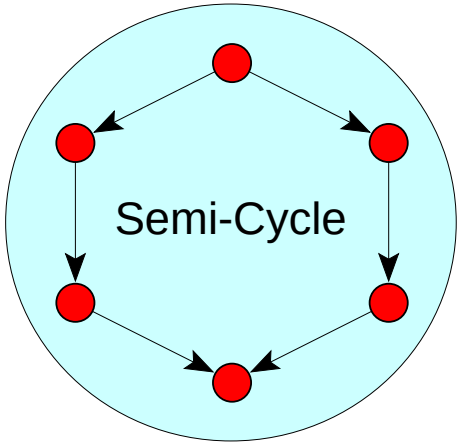
Genealogies are sparse networks

	Ore-graph				p-graph			
data	$\|V\|$	$\|E\|$	$\|A\|$	$\frac{\|L\|}{\|V\|}$	$\|V_{ip}\|$	$\|V_{cp}\|$	$\|A_p\|$	$\frac{\|A_p\|}{\|V_p\|}$
Bruno	15512	4841	18664	1.52	6000	5289	10053	0.89
Combo	20350	7248	26199	1.64	6931	7945	14845	1.00
Dodderer	16761	5650	22425	1.68	6029	5652	11765	1.01
Drame	29606	8256	41814	1.69	13254	8939	21862	0.99
Little	25968	8778	34640	1.67	9212	8850	18233	1.01
President	2145	978	2223	1.49	218	1042	1222	0.97
Tillotsn	42559	12796	54043	1.57	15177	15959	31234	1.00
Loka	47956	14154	68052	1.71	19189	16039	36192	1.03
Silba	6427	2217	9627	1.84	2001	2479	5281	1.18
Ragusa	5999	2002	9315	1.88	2066	2310	5336	1.22
Tur	1269	407	1987	1.89	0	956	1114	1.17
Royal	3010	1138	3724	1.62	719	1422	2259	1.06

Advantages of p-graphs

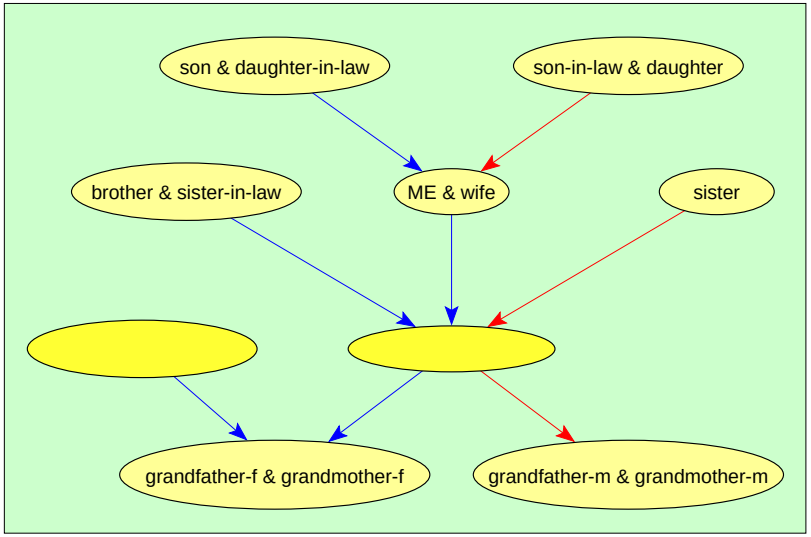
- there are less vertices and lines in p-graphs;
- p-graphs are directed, acyclic networks;

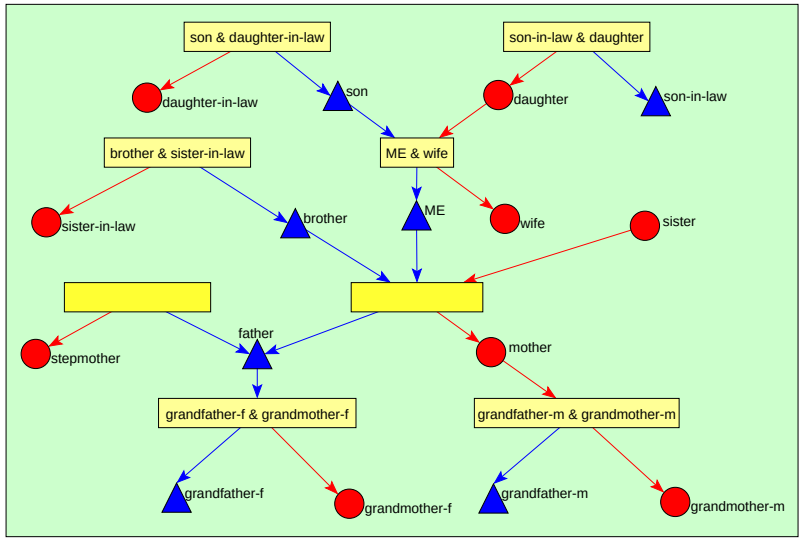




- every semi-cycle corresponds to a *relinking marriage*. There exist two types of relinking marriages:
  - blood marriage: e.g., marriage among brother and sister.
  - non-blood marriage: e.g., two brothers marry two sisters from another family.

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





Relinking index

Relinking index is a measure of relinking by marriages among persons belonging to the same families. Special case of relinking is a blood-marriage.

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

If we take a connected genealogy we get

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

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

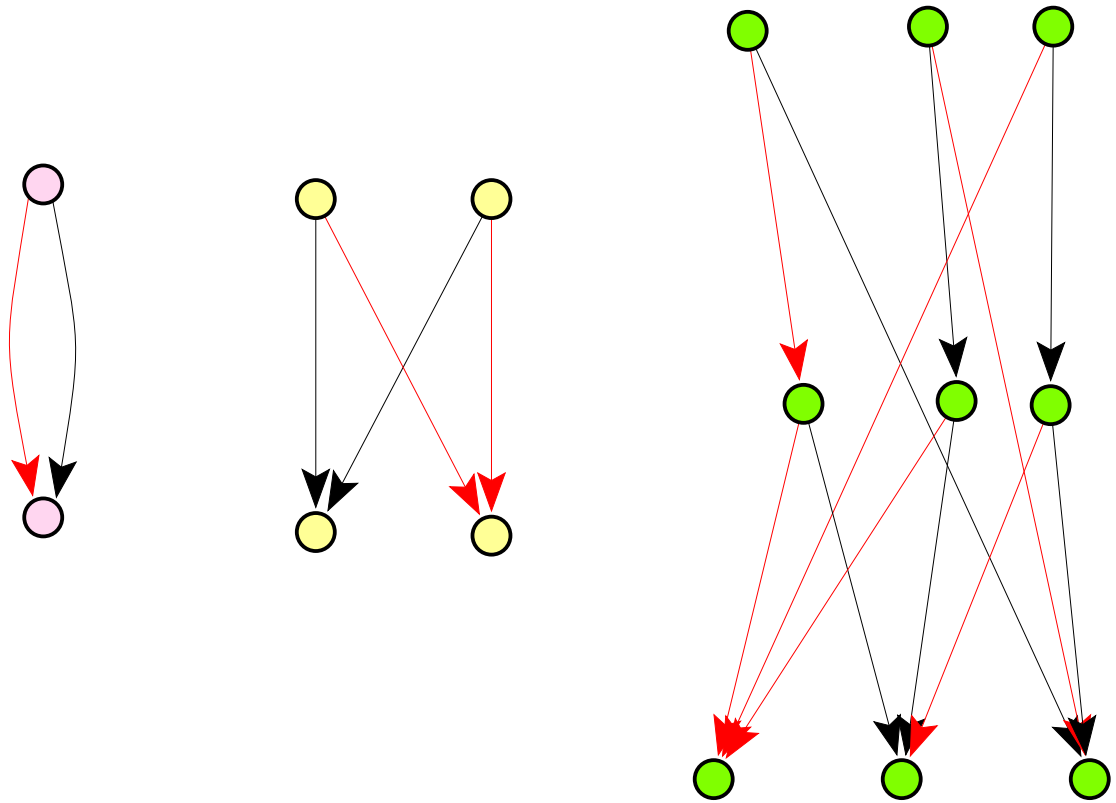
* $0 \leq RI \leq 1$

* If network is a forest/tree, then $RI = 0$ (no relinking).

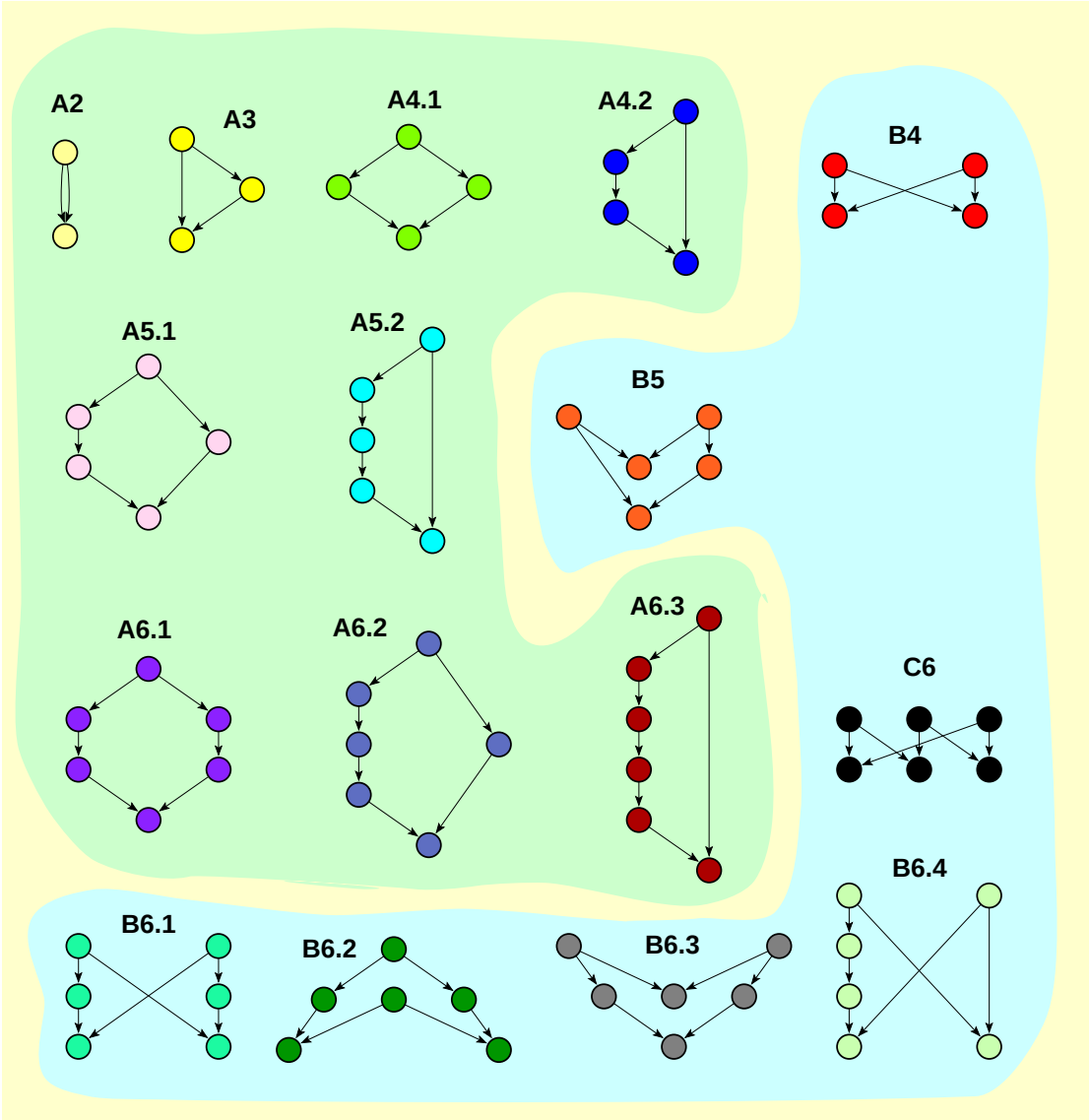
* There exist genealogies having $RI = 1$ (the highest relinking).

* Relinking is usually computed for the largest biconnected component.

Patterns with Relinking Index = 1



Relinking marriages (p-graphs with 2 to 6 vertices)



Comparing genealogies

Using frequency distributions for different patterns we can compare different genealogies. As an example we took five genealogies:

- `Loka.ged` – genealogy of Škofja Loka dictrict, Slovenia (P. Hawlina).
- `Silba.ged` – genealogy of the island Silba, Croatia (P. Hawlina).
  Special geographical position.
- `Ragusa.ged` – marriages among Ragusan (Dubrovnik) noble families between 12 and 16 century. Data collected by I. Mahnken (1960); entered to electronic form by P. Dremelj (1999).
  Very restricted marriage rules.
- `Tur.ged` – genealogy of Turkish nomads, Yörük. Data collected by Ulla C. Johansen and D.R. White (2001)
  A relinking marriage is a signal of commitment to stay within the nomad group.
- `Royal.ged` – genealogy of European royal families.

Škofja Loka


Gornja Radgona
Lenart
Maribor
Murska Sobota
Lendava
Ljutomer
Slovenj Gradec
Ptuj
Ormož
Jesenice
Radovljica
Velenje
Slovenska Bistrica
Slovenske Konjice
Kranj
Žalec
Tolmin
Kamnik
Celje
Šmarje pri Jelšah
Škofja Loka
Šentjur
Domžale
Trbovlje
Ljubljana
Litija
Idrija
Sevnica
Nova Gorica
Krško
Grosuplje
Brežice
Ajdovščina
Vrhnika
Trebnje
Cerknica
Novo mesto
Postojna
Sežana
Kočevje
Koper
Črnomelj
Piran
Ilirska Bistrica

Silba, Dubrovnik; Croatia

Aydin Southwest-Anatolia, Turkey

Frequencies 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	65	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	54	12	194
B6.3	47	30	82	46	13	218
B6.4	0	0	5	3	0	8
No. indi.	47956	6427	5999	1269	3010	
Largest bic.	4095	1340	1446	250	435	
RI	0.55	0.78	0.74	0.75	0.37	

Observations

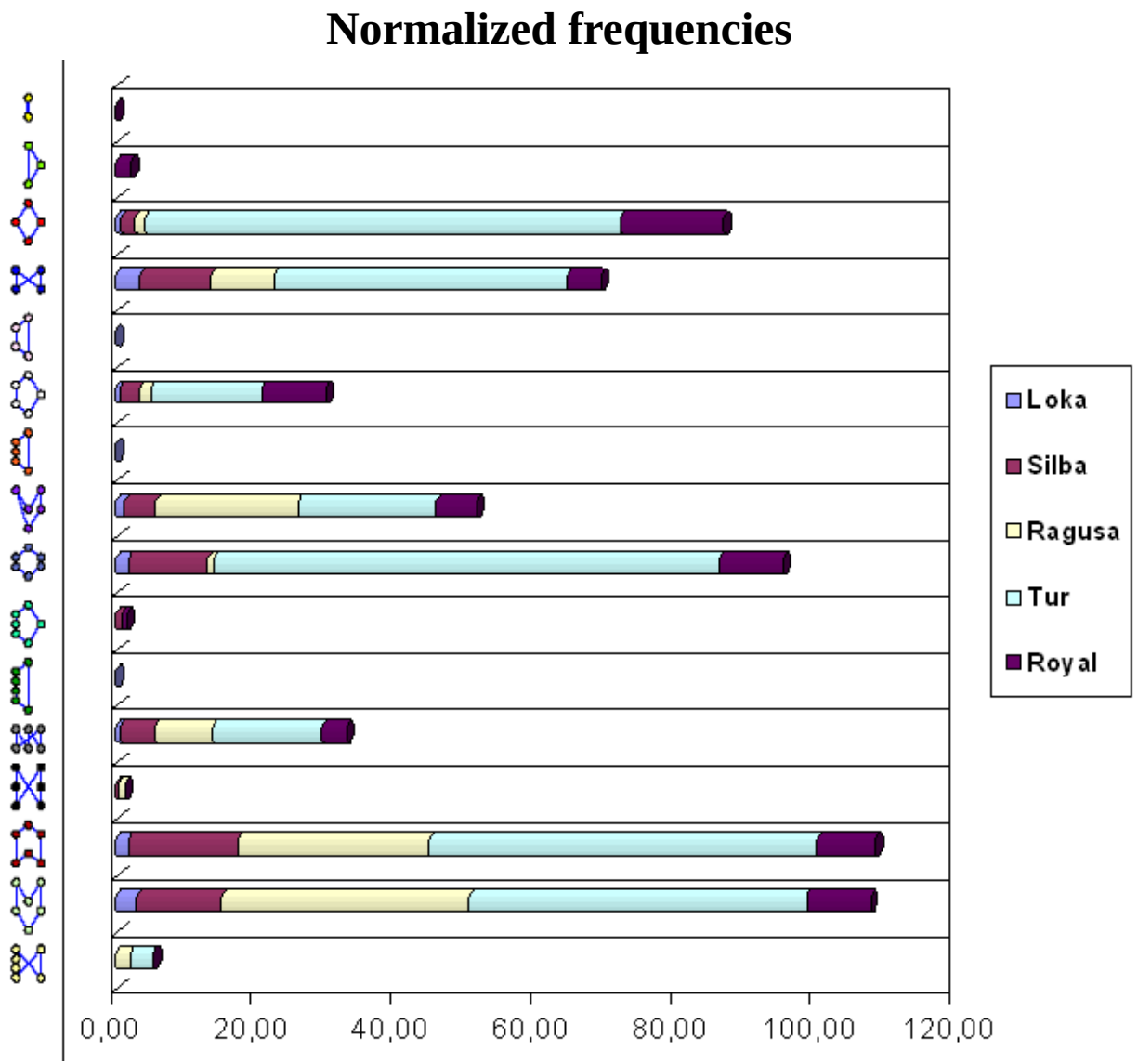
- Generation jumps for more than one generation are very unlikely.
- There are many marriages B6.3 (two grandchildren married into the same family) and B6.2 (two familes were relinked by a marriage between children and again in the next generation by a marriage between grandchildren)
- In Tur there are many marriages of types A4.1 and A6.1.
- For all genealogies number of relinking 'non-blood' marriages is much higher than number of blood marriages (this is especially true for Ragusa, exception is Royal). There were economic reasons for non-blood relinking marriages: to keep the wealth and power within selected families.

type of marriage	Loka	Silba	Ragusa	Tur	Royal
blood-marriages	51	42	9	149	51
relinking-marriages	157	118	239	176	45

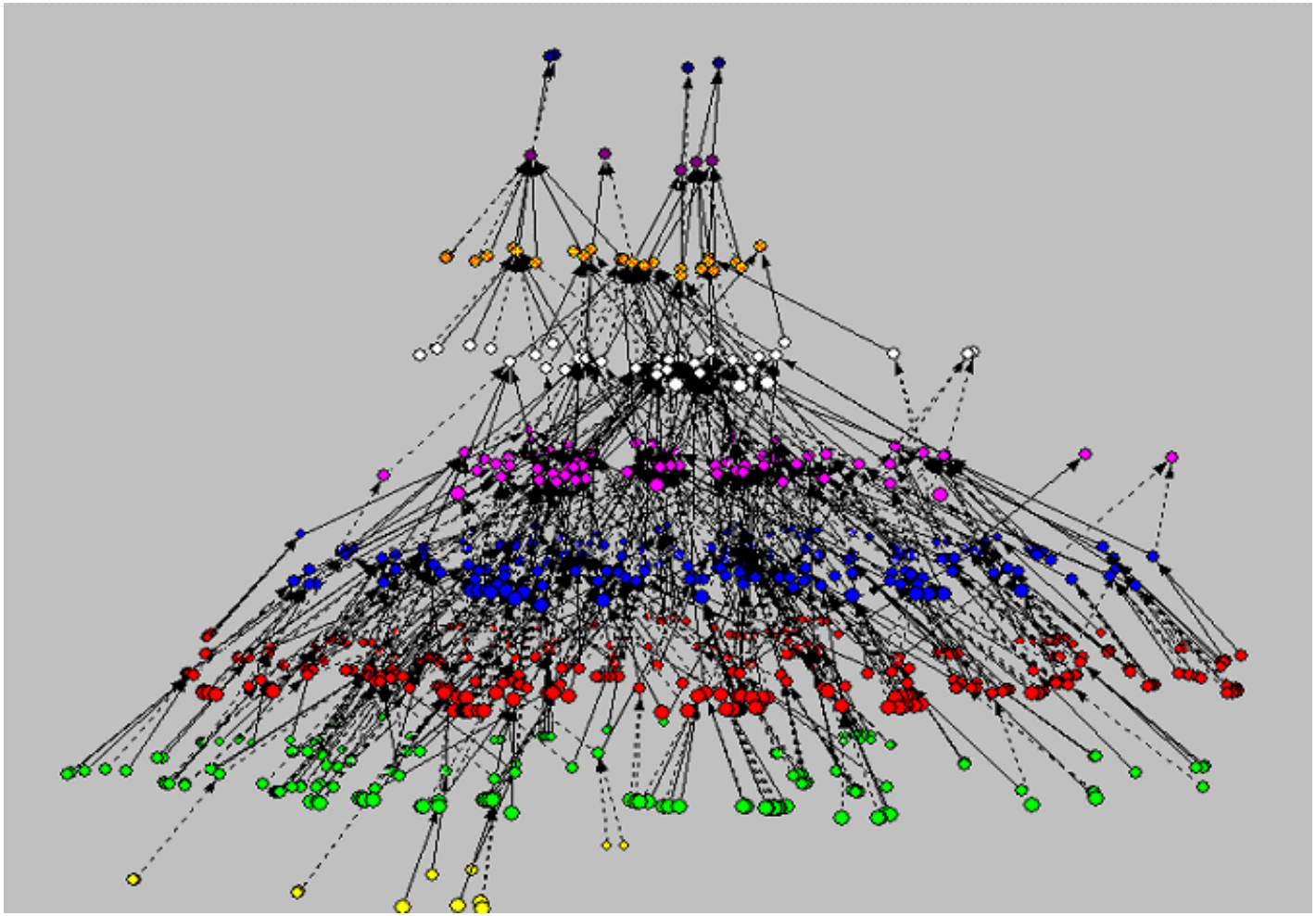
Number of individuals in genealogy Tur is much lower than in others, Silba and Ragusa are approximately of the same size, while Loka is much larger genealogy, what we must also take into account.

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

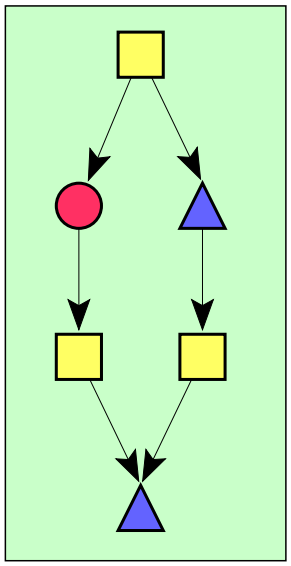

Normalized frequencies
0,00
20,00
40,00
60,00
80,00
100,00
120,00
Loka
Silba
Ragusa
Tur
Royal

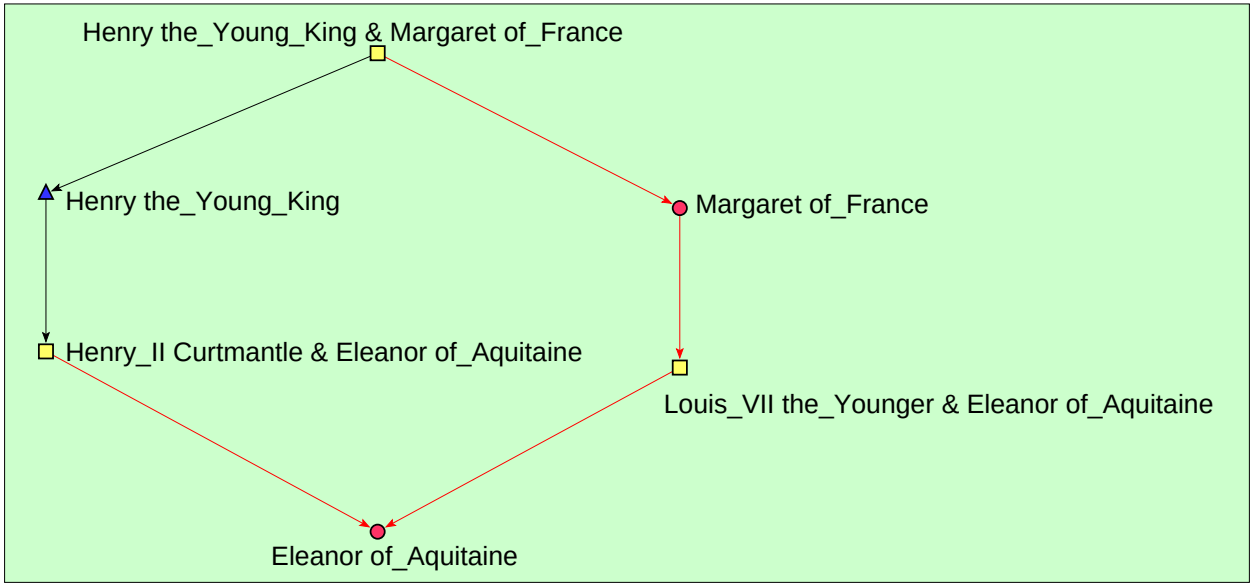
Relinking of Tur genealogy



Bipartite p-graphs: Marriage between half-brother and half-sister

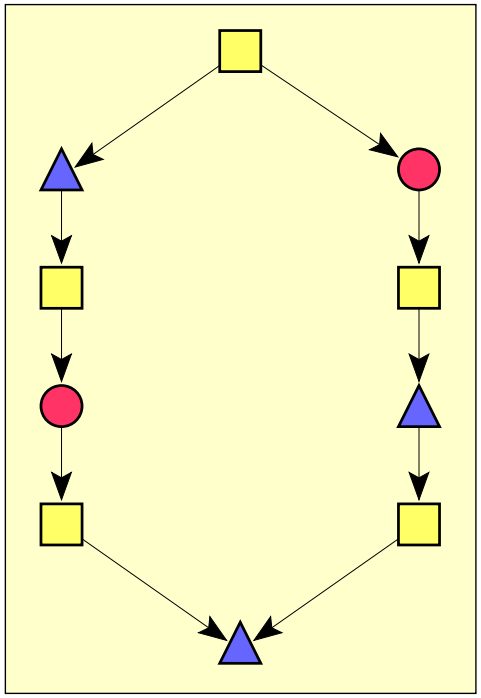
Using p-graphs we cannot distinguish persons married several times. In this case we must use bipartite p-graphs. Using bipartite p-graphs we can find marriages between half-brothers and half-sisters. In our five genealogies we found only one such example in Royal.ged.

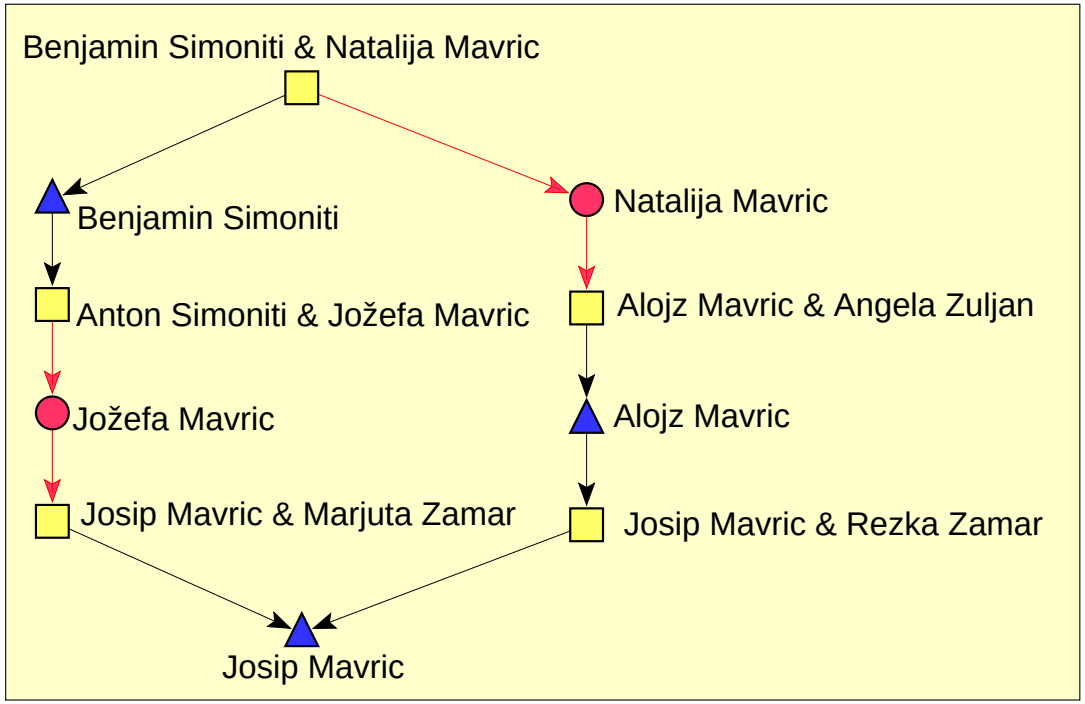




Bipartite p-graphs: Marriage among half-cousins

There also do not exist many marriages between half-cousins. We found one such marriage in Loka genealogy and four in Turkish genealogy.



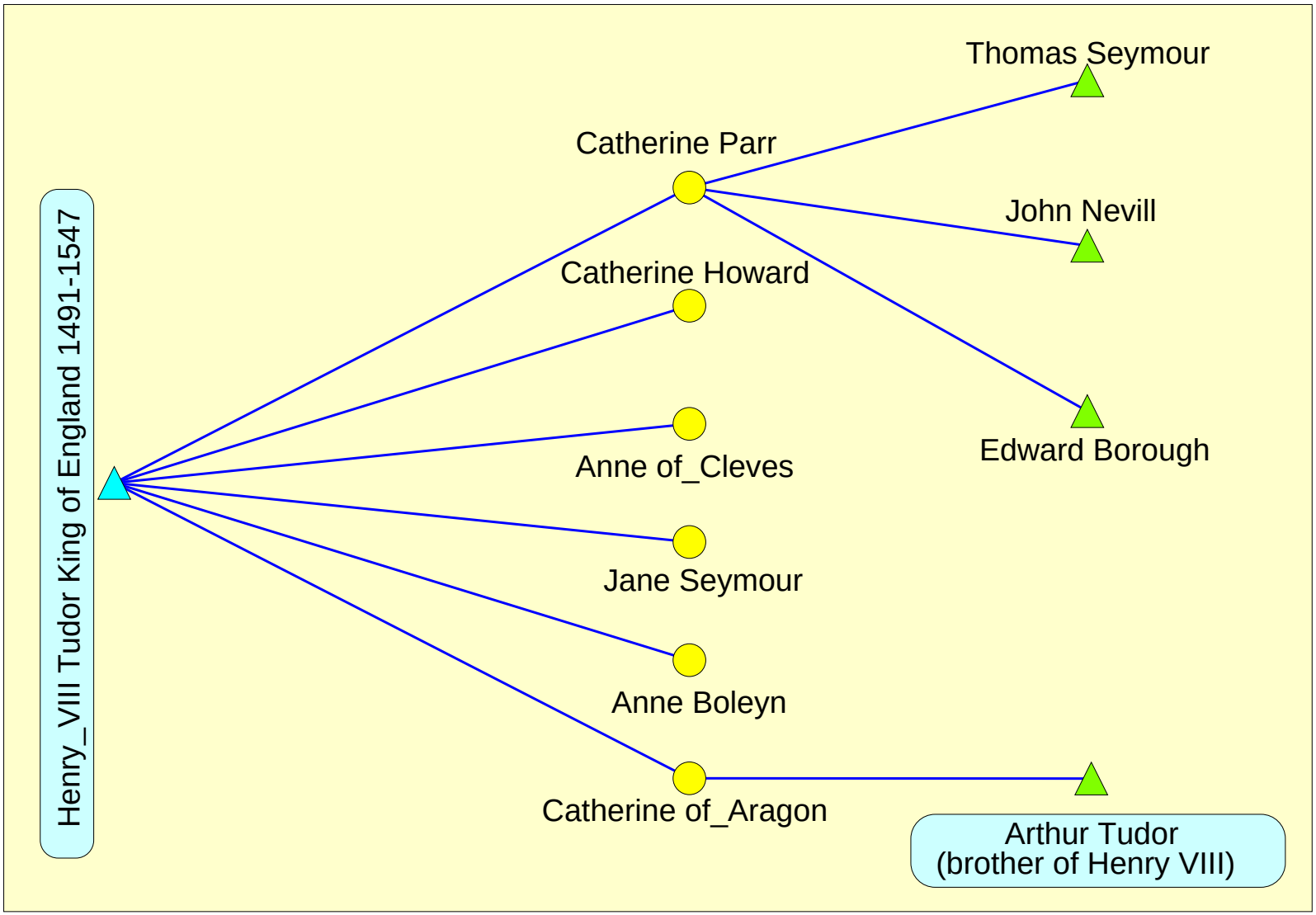


Other analyses

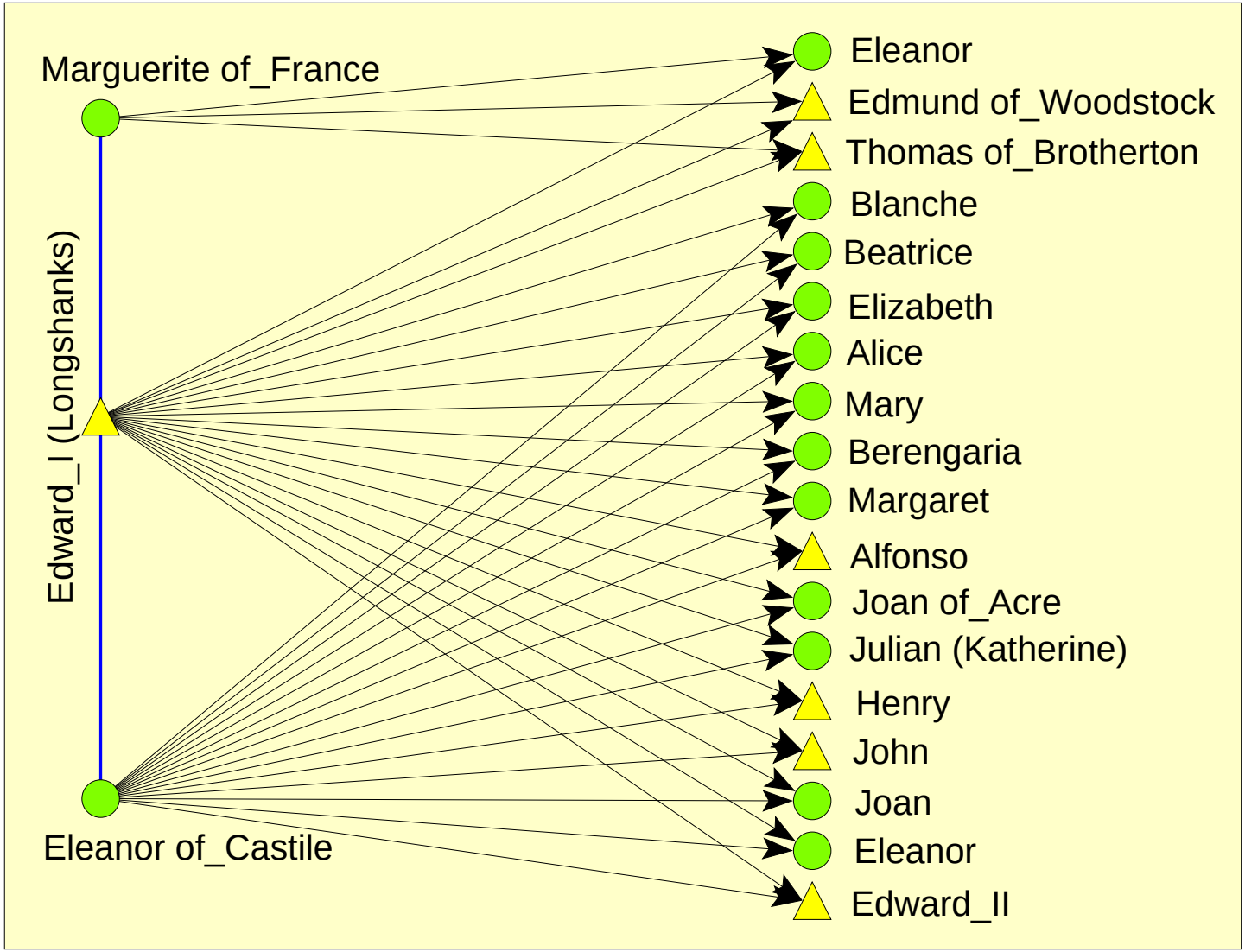
People collecting data about their families are interested in several other 'standard' analyses:

- changes in relinking patterns over time;
- special situations: persons married several times, persons having the highest number of children;
- checking whether the two persons are relatives and searching for the shortest genealogical path between them;
- searching for all predecessors/successors of selected person and searching for person with the largest number of known predecessors or successors;
- the largest difference in age between husband and wife,
  the oldest/youngest person at the time of marriage,
  the oldest/youngest person at the time of child's birth;
- searching for the longest patrilineage and matrilineage;
- special situations $\longrightarrow$ errors made in data entry (network consistency check).

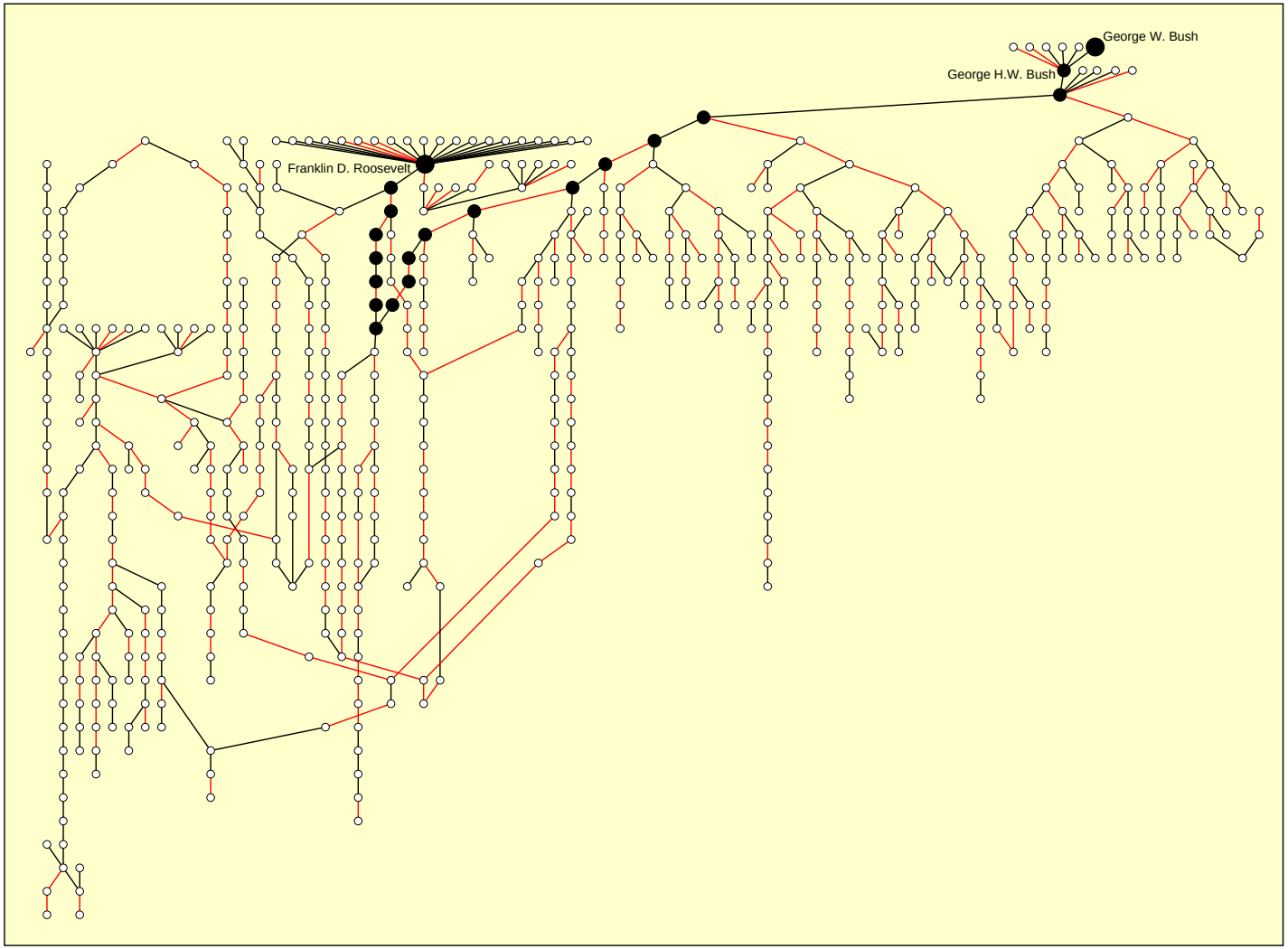
The largest number of marriages…



The largest number of children…



The largest connected component in the genealogy of American presidents



The shortest genealogical path between *Charles Philip Arthur Windsor* (Prince of UK) and *Juan Carlos* (King of Spain) in Royal.ged