



© Stefan Ernst www.Naturfoto-Online.de

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

Analysis of Genealogies with Pajek

Andrej Mrvar

University of Ljubljana
Slovenia

Sources of genealogies

People collect genealogical data for several different reasons/purposes:

- Research of different cultures in sociology, anthropology and history – kinship as fundamental social relation
- Genealogies of families and/or territorial units, e.g.,
 - genealogy of Ragusan (Dubrovnik) nobel families:
 - 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
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

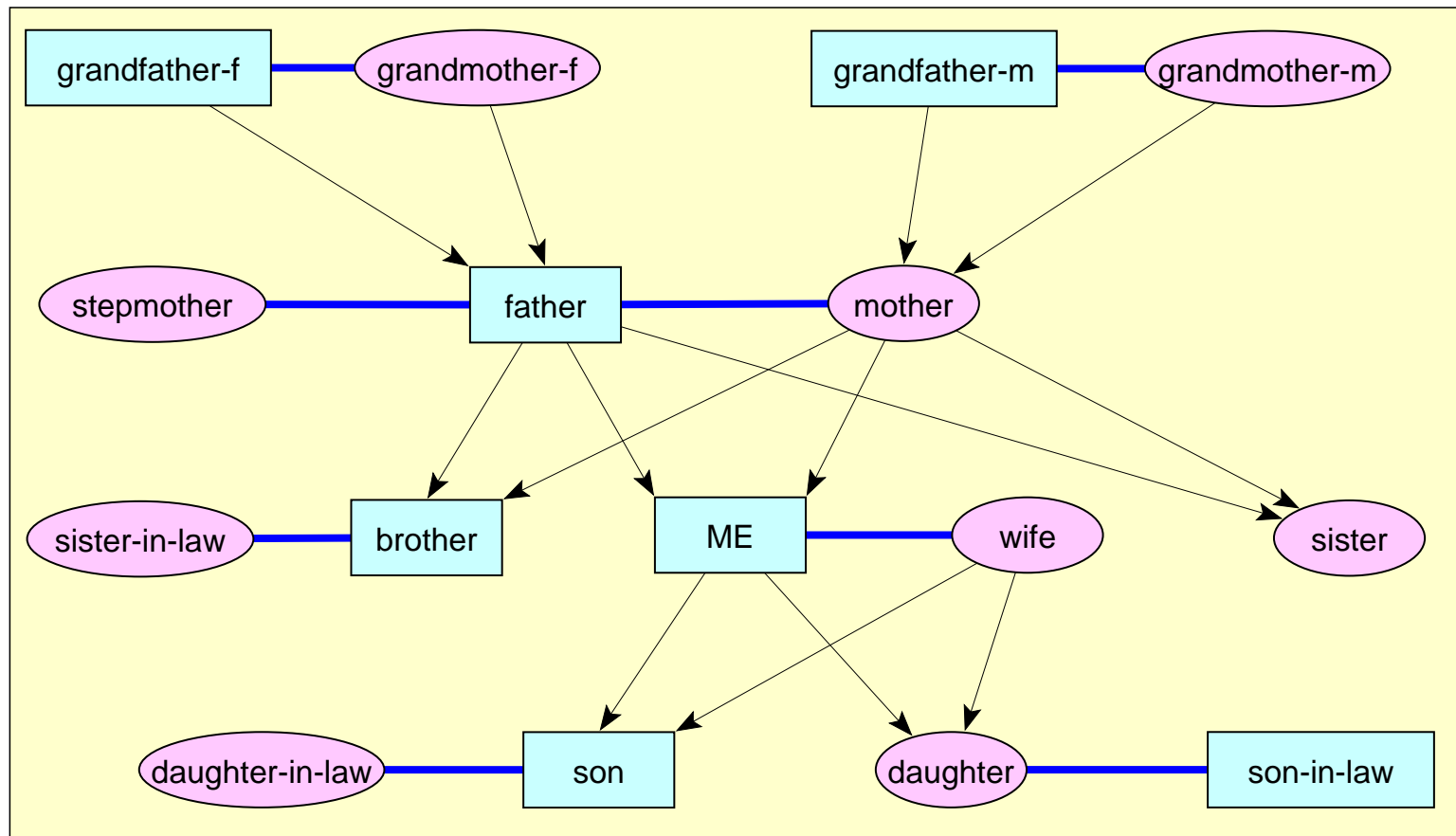
```

Representation of genealogies using networks

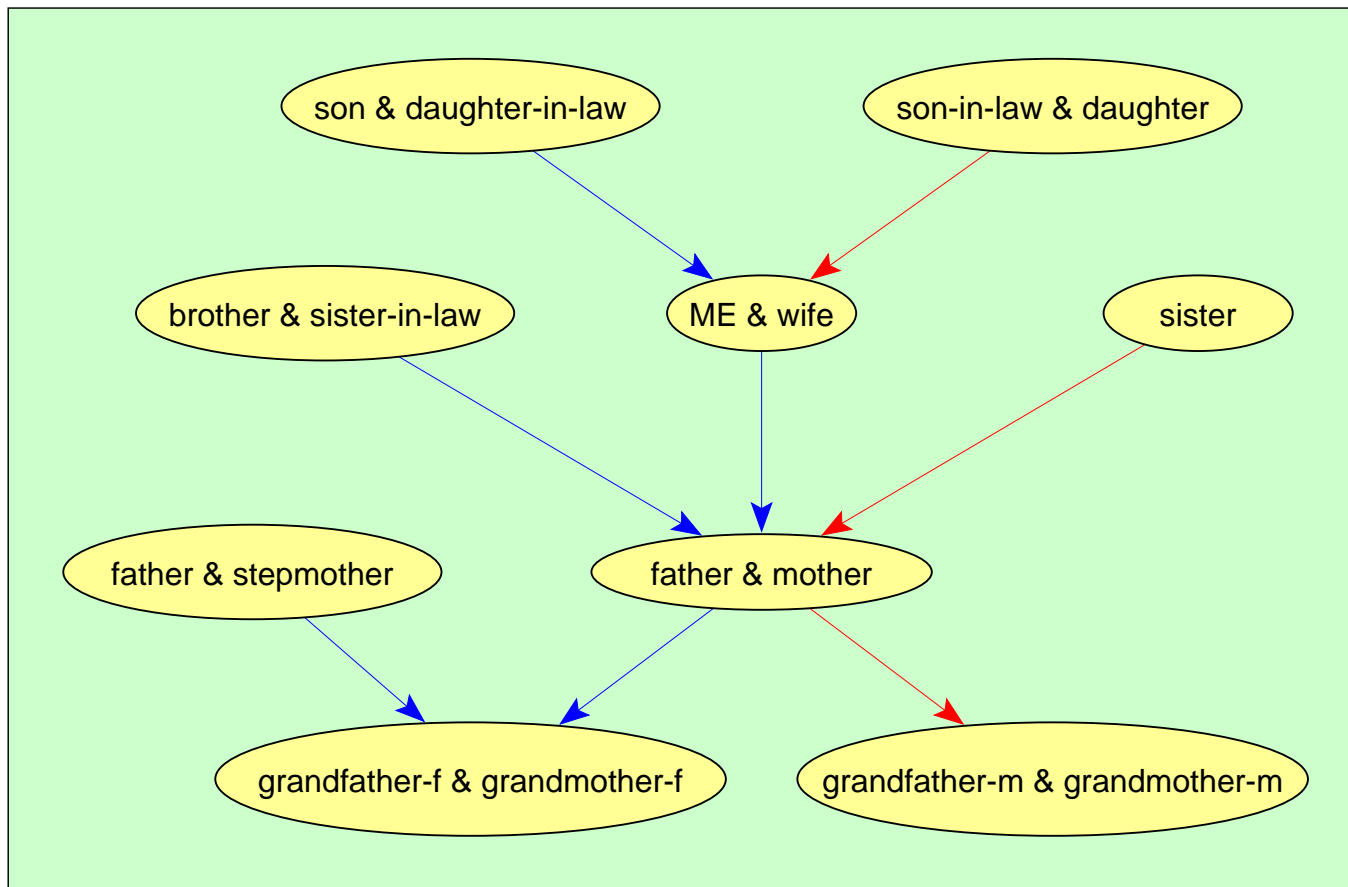
Genealogies can be represented as networks in different ways:

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

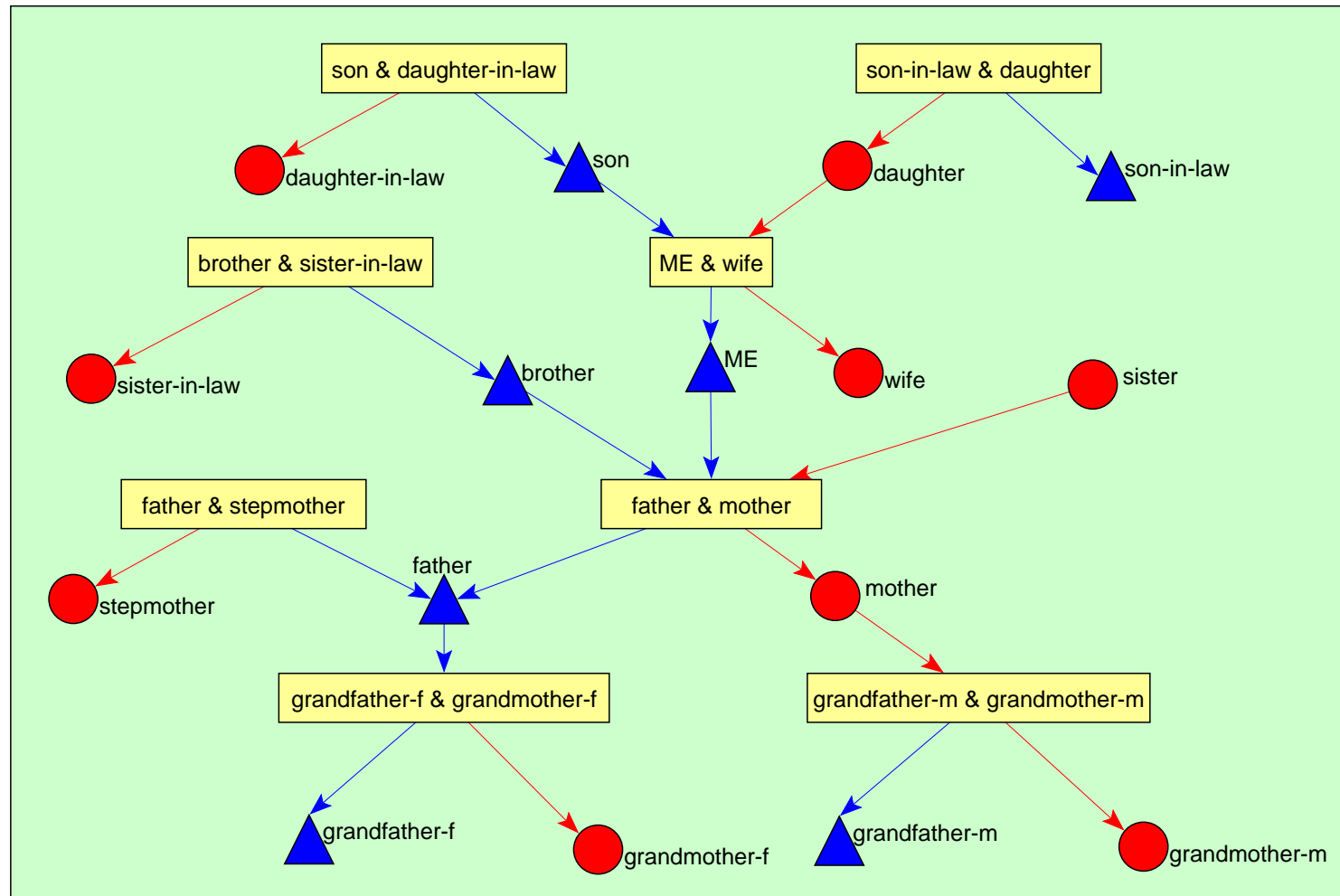
Ore-graph: In Ore-graph every person is represented by a vertex, marriages are represented with edges and relation *is a parent of* as arcs pointing from each of the parents to their children.



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.

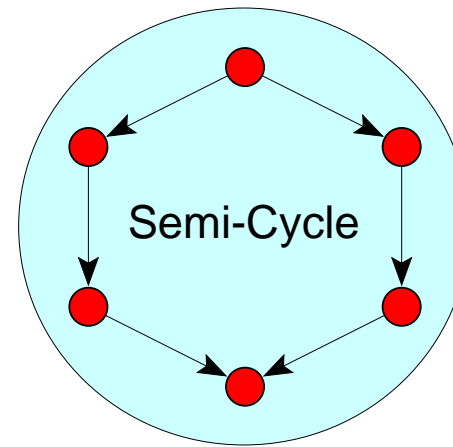
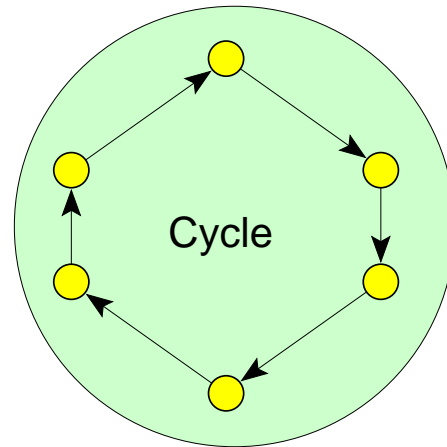


Bipartite p-graph: has two types of vertices – vertices representing couples (rectangles) and vertices representing individuals (circles for women and triangles for men). Arcs again point from children to their parents.



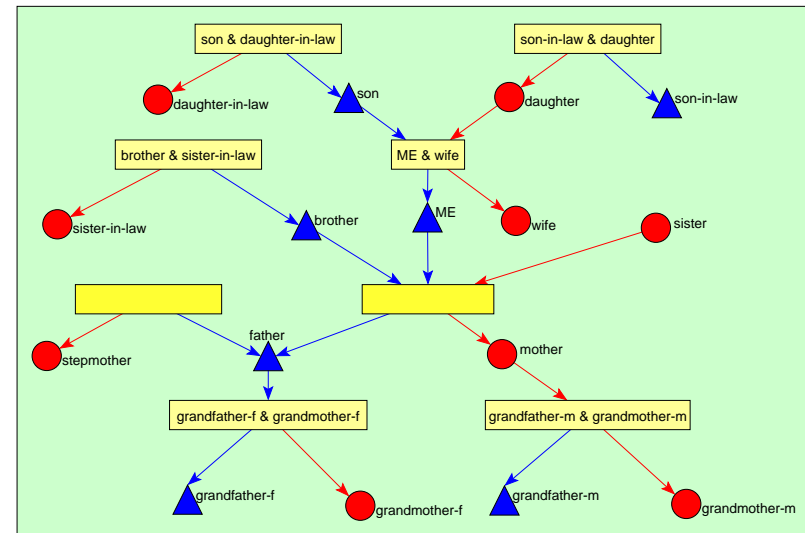
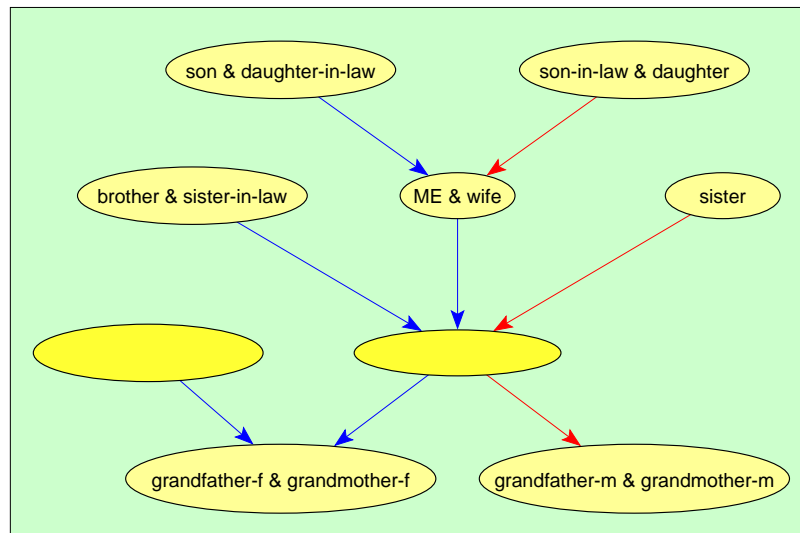
Advantages of p-graphs

- 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 $RI = 0$.

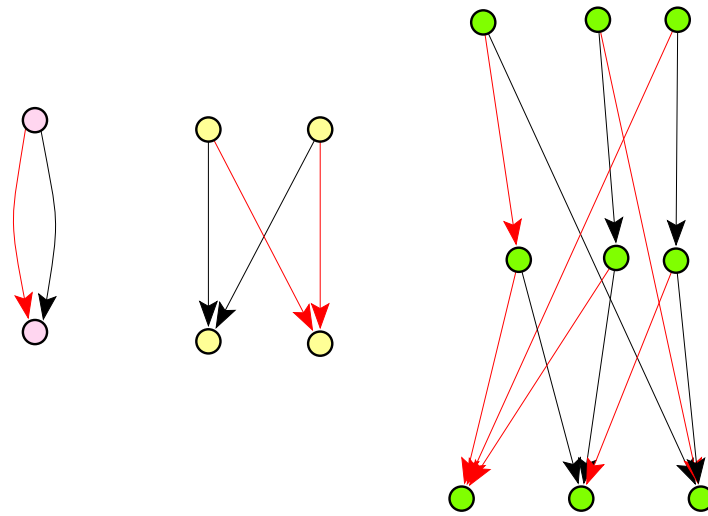
* $0 \leq RI \leq 1$

* 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



Comparing genealogies

For comparison, we took five genealogies:

- Loka.ged – genealogy of Škofja Loka district, 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



Silba, Dubrovnik; Croatia



Aydin Southwest-Anatolia, Turkey



Ragusan nobility - some history

1023: As a result of the segregation of the town's community into patricians and commoners, a new class, the nobles (*nobiles*) is mentioned for the first time.

1332: The Ragusan nobility was finally recognized by statute. After 1332 no new family was accepted until the large earthquake in 1667.

In Ragusa all political power was in the hands of male nobles older than 18 years. They were members of the **Great Council** (*Consilium majus*) which had the *legislative function*. Every year, 11 members of the **Small Council** (*Consilium minus*) were elected. Together with a **duke** it had both *executive and representative functions*. The main power was in the hands of the **Senat** (*Consilium rogatorum*) which had 45 members elected for one year.

This organization prevented any single family unlike the Medici in Florence, from prevailing. Nevertheless the historians agree that the Sorgo family was all the time among the most influential.

A major problem facing the Ragusan noble families was also that by decreases of their numbers and the lack of noble families in the neighbouring areas (which were under Turkish control), they became more and more closely related – the marriages between relatives of only 3rd and 4th removes were frequent.

1189: Kulin Ban allowed the people of Dubrovnik the liberty of trading in Bosnia without the payment of taxes. In return they gave him whatever they thought appropriate. In the agreement the *Croatian name, Dubrovnik, appeared for the first time.*

1348: The black death struck down 110 members of the Great Council and 7,000 townsfolk. The Plague re-occurred in 1357, 1366, 1371, 1374 and 1391. To protect themselves, the people of Dubrovnik brought into effect **quarantine** for all ships.

1667: A terrible earthquake and fire demolished Dubrovnik. Over 4,000 citizens were left under the ruins. Between 2,000 and 3,000 people survived. Dubrovnik recovered thanks to the trade. Due to the fact that a number of aristocrats were killed and their family names died with them, the Great Council accepted amongst its aristocracy 10 of the town's families. In the year 1673 another five families were accepted amongst the aristocracy.

1763: The conflict between the old and new nobility (the latter were those co-opted into the ranks after the earthquake) resulted in reforms of election rights for the state administration in favour of the newcomers.

Collecting data on Ragusan genealogy

Irmgard Mahnken (1960): Das Ragusanische Patriziat des XIV. Jahrhunderts – Ph.D. thesis.

Data collected from *private documents* and *notes of Great Council (Consilium Maius)*.

She presented data as trees (more than 100 hand written pages).

Genealogy was entered to electronic form using **GIM** by P. Dremelj (1999).

Available data: names, surnames, date of birth, date of marriage, date of death.

In some occasions additional data were given: profession, children born out of wedlock (*filius*), persons entering the monastery...

Several missing data: dates corresponding to women, data about first generations...

Other problems

Surnames are written sometimes in Latin, sometimes in Croatian language, sometimes in combination:

Menčetić – Mence,

Sorkočević – Sorgo,

Djurdjević – Georgio,

Gundulić – Gondola,

Gučetić – Gozze (Goce),

Bunić – Bona,

Crijević – Crieva (Zrieva),

Lukarević – Luccari (Lucaro),

Bobaljević – Babalio,

Budačić – Bodazza (Bodaca),

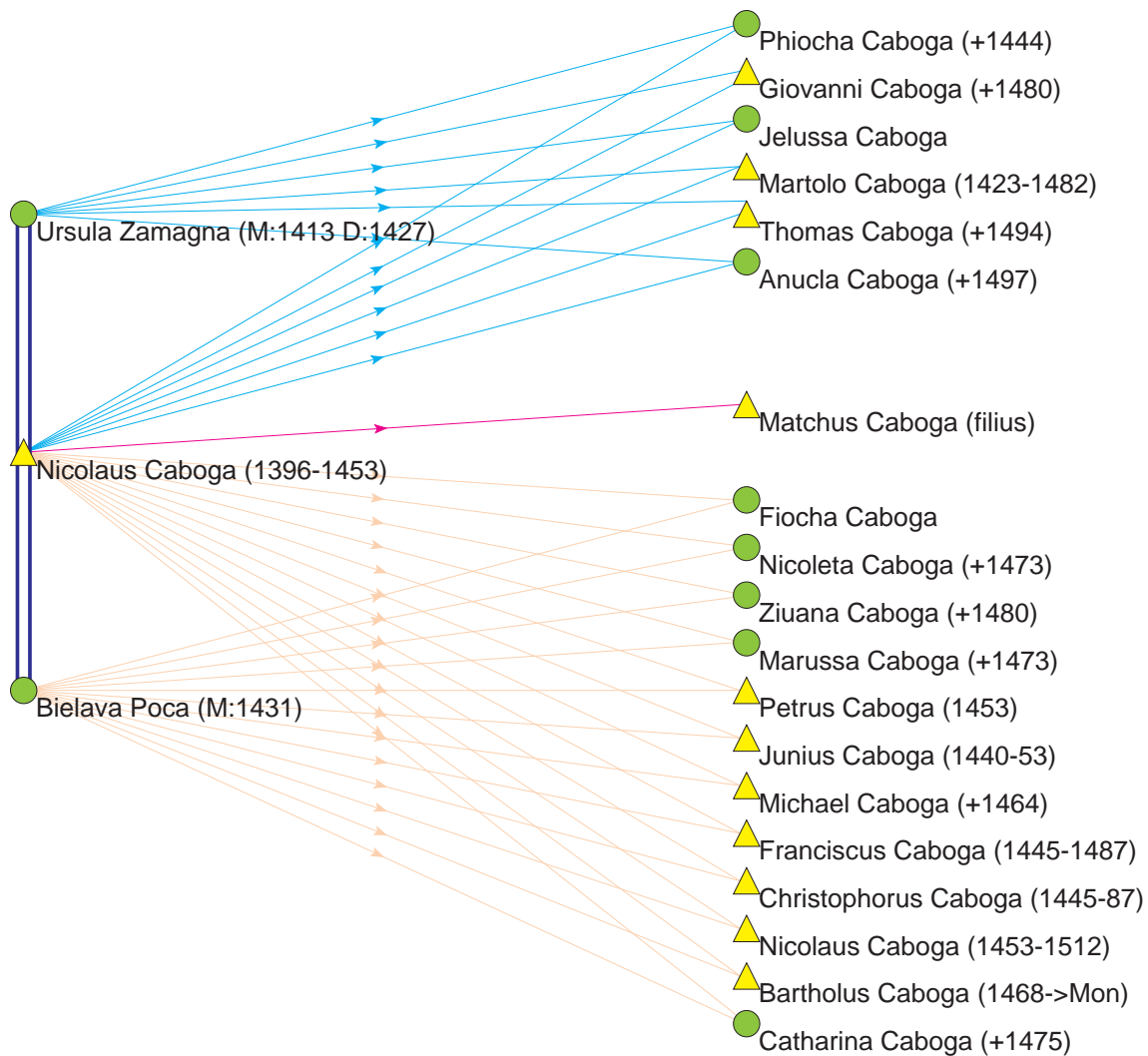
Pucić – Poca...

some individuals changed their surnames or took another ...

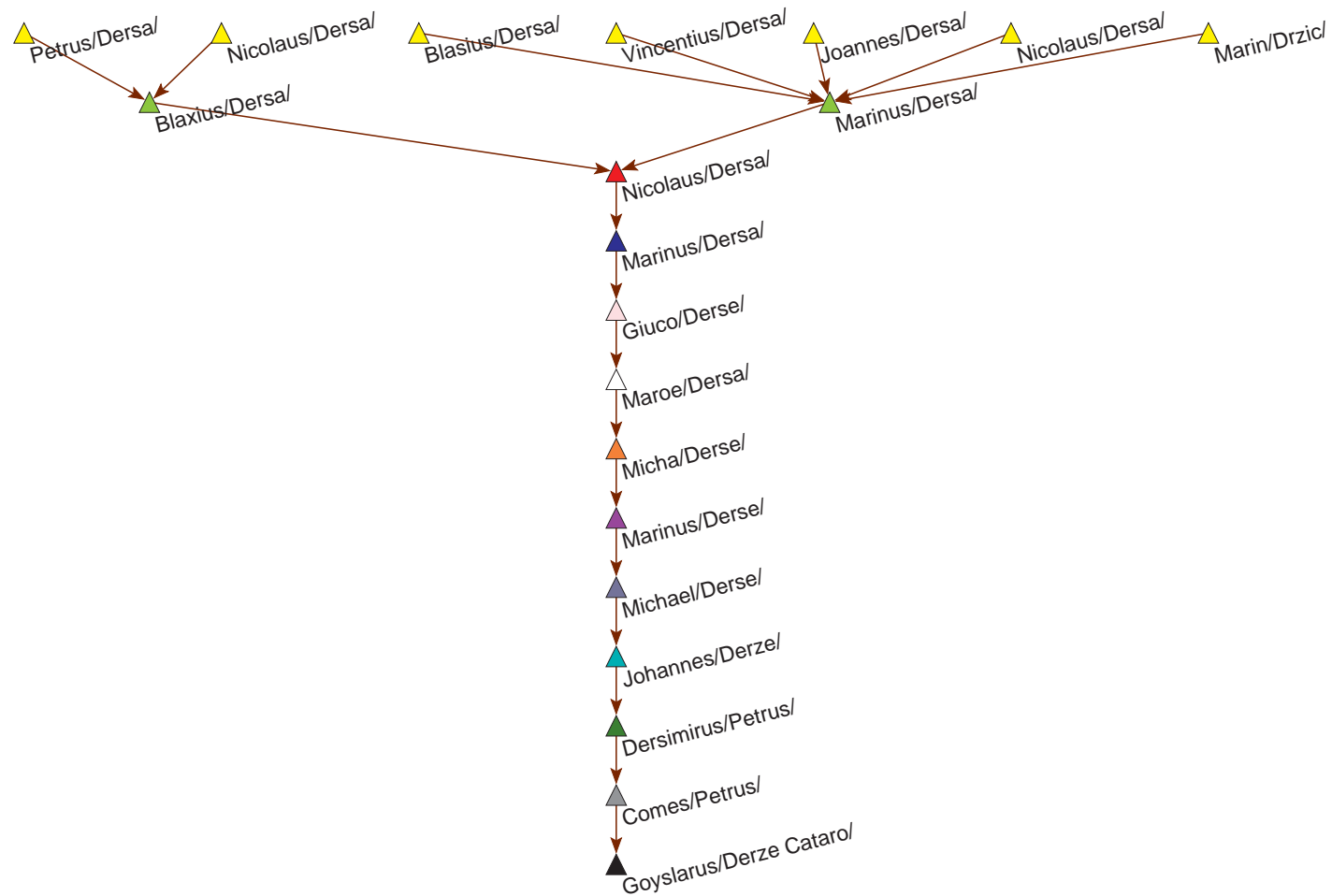
Some numbers

- 5999 persons from 254 families. Largest families:
 - Goce (8.64%)
 - Mence (6.21%)
 - Sorgo (5.41%)
 - Bona (5.28%)
 - Georgio (4.65%)
 - Gondola (4.13%)
 - Zrieva (3.15%)
- only one person (unknown) isolated,
all others in single weakly connected component
- Ore-graph
 - 5999 vertices
 - 2002 undirected (marriages) and 9315 directed (a child of) lines

Ore graph – The highest number of children



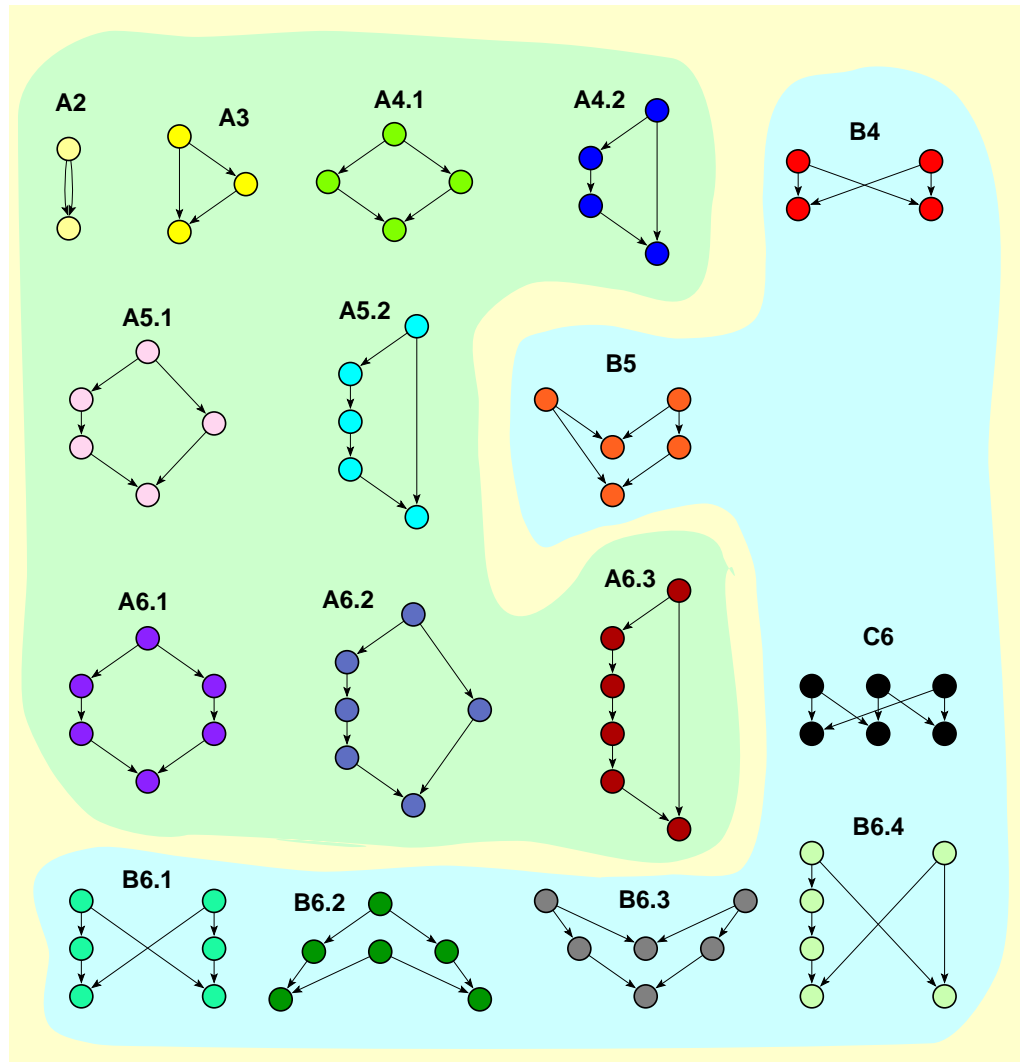
Ore graph – The longest patrilineage



Macro/Play/LongestPatrilineage.mcr

Layers/Optimize layers in x direction/Complete

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



p-graph – Relinking marriages

Read genealogy as p-graph

```
Options/ReadWrite/GEDCOM-Pgraph [checked]
```

```
File/Network/Read/Ragusa.ged (Second)
```

Read project file with 16 fragments defined:

```
File/Pajek Project File/Read/frag16.paj (First)
```

Select first fragment as first network and genealogy as second, then

```
Nets/Fragment(First in Second)/Find
```

Repeat the command on another 15 fragments and on the same network:

```
Macro/Repeat Last Command/Fix Second Network
```

















```
Macro/Repeat Last Command [15]
```

Select one of the results (fragments found) and then:

```
Macro/Play Layers1.mcr
```

```
Layout/Tile Components
```


Frequencies of fragments

	pattern	Loka	Silba	Ragusa	Tur	Royal	Σ
	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 families 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

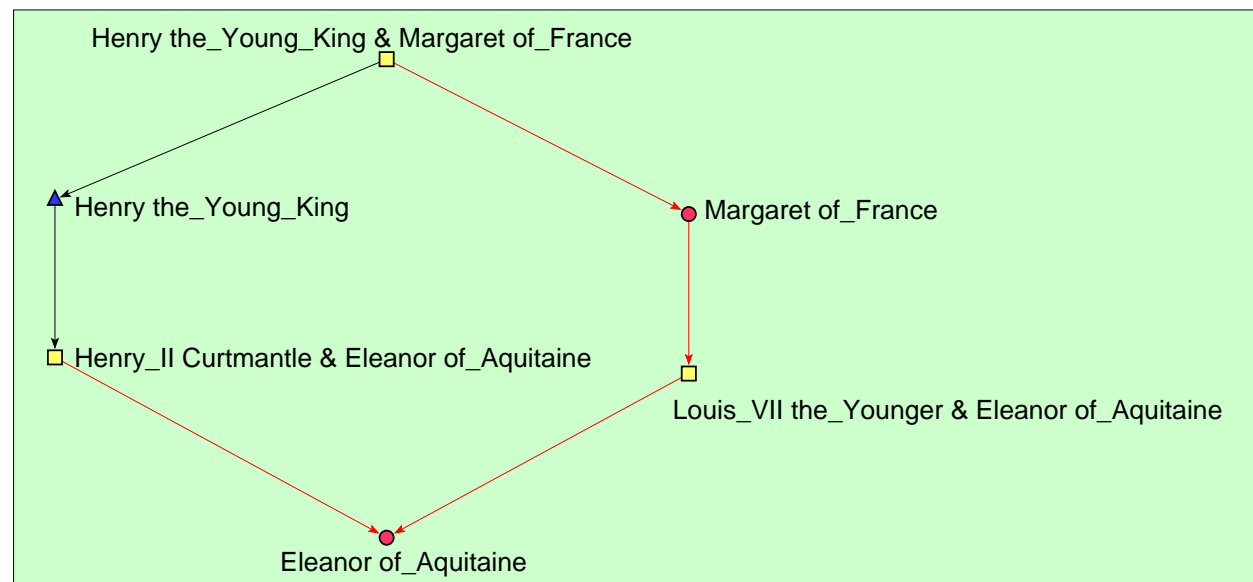
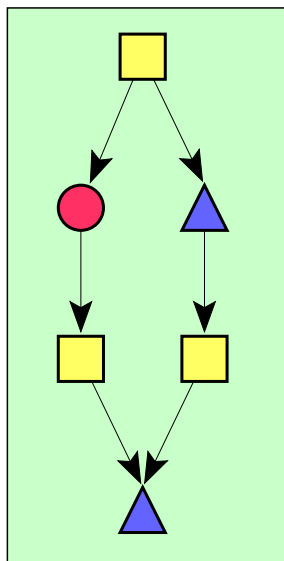
Number of individuals in genealogy Tur is much lower than in others, Silba and Ragusa are approximately of the same size, while Loka is much larger genealogy, what we must also take into account.

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
	Σ	14.70	72.17	123.88	798.53	84.36

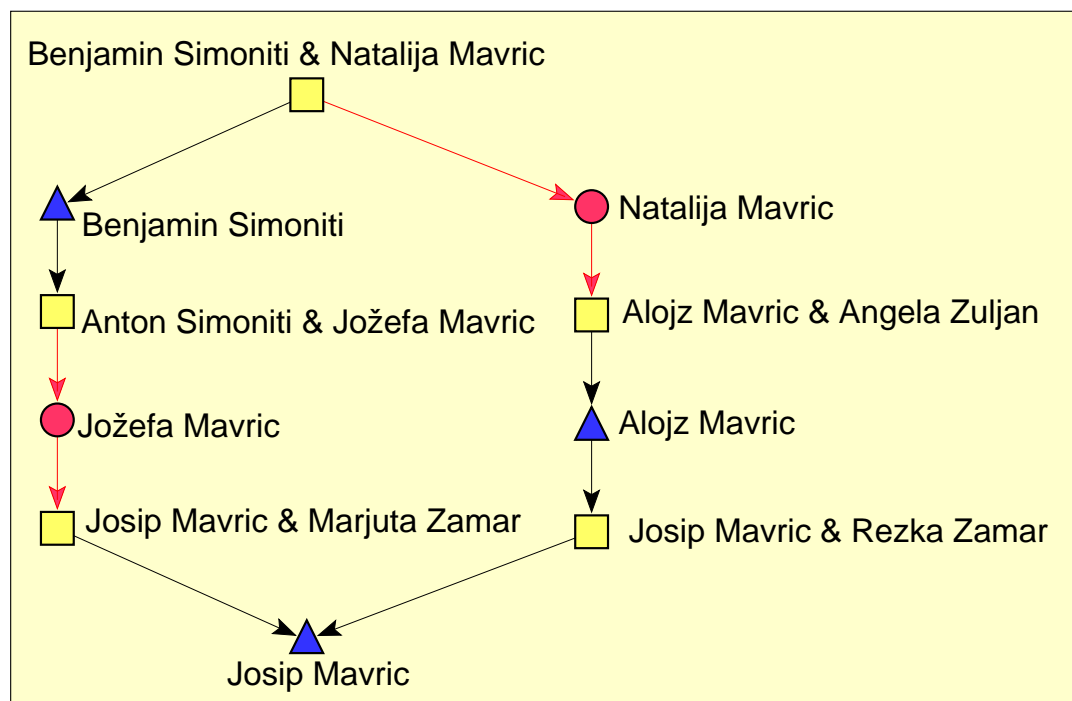
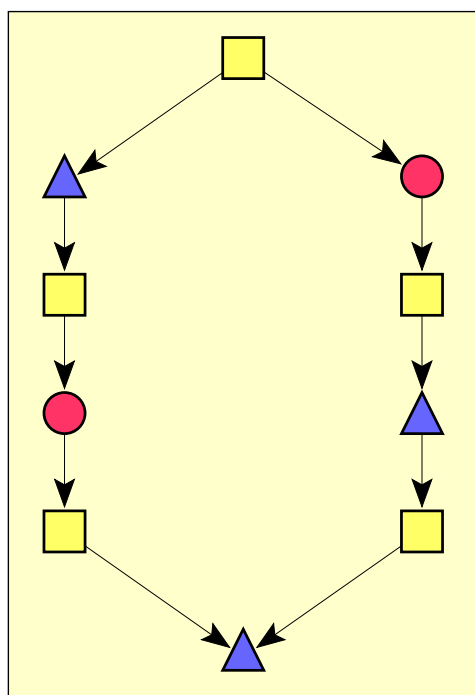
Bipartite p-graphs: Marriage between half-brother and half-sister

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.

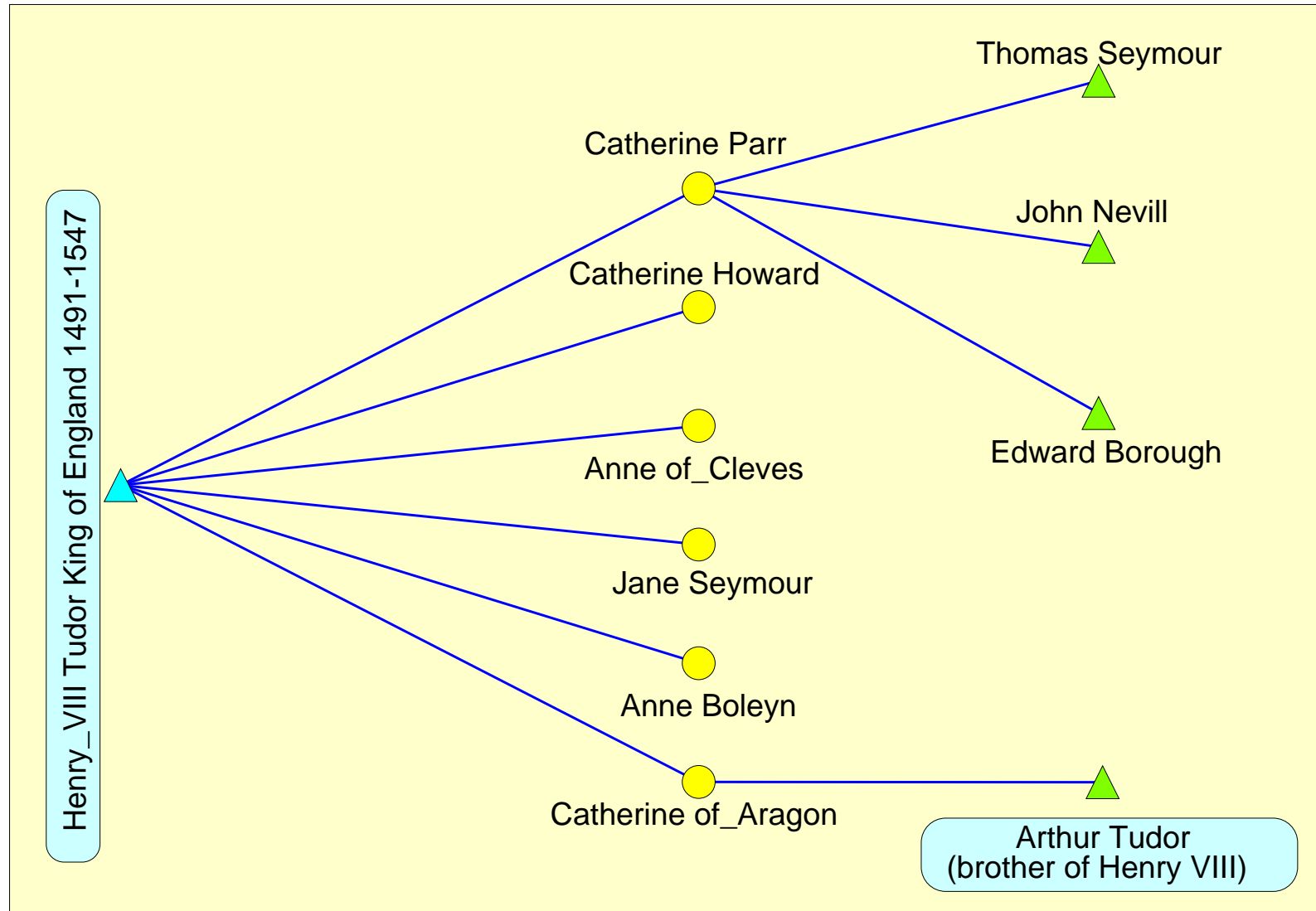


Other analyses

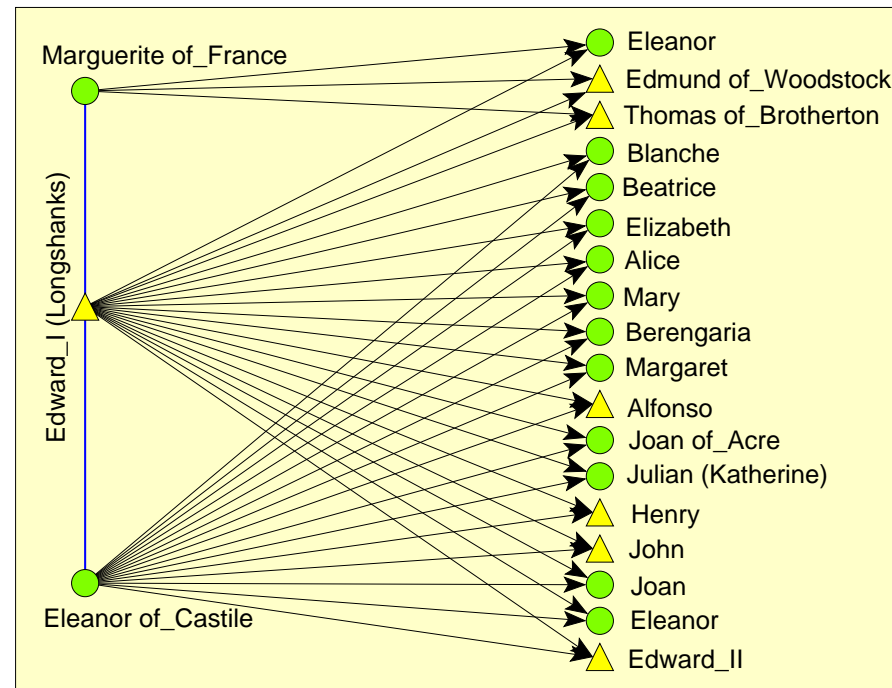
People collecting data about their families are interested in several other 'standard' analyses:

- 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 → errors made in data entry (network consistency check).

The largest number of marriages...



The largest number of children...

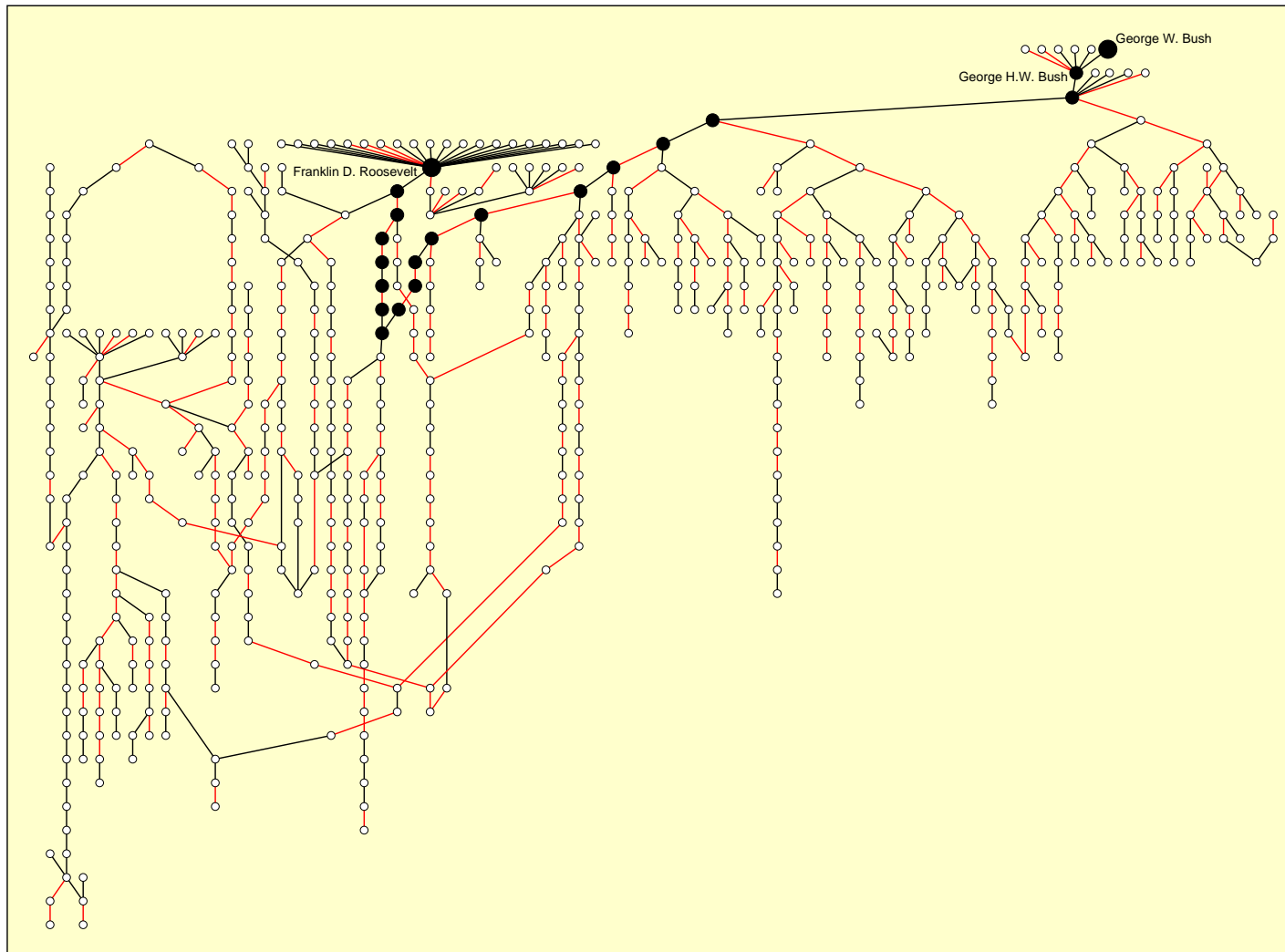


English king Edward I (1237-1307) and his wife Eleanor (1241-1290) had 16 children who were born between 1255 and 1284 (in the picture a daughter without given name is missing). The youngest son (Edward) was the first among sons who survived his childhood. Eleanora had to try sixteen times to fulfill her most important duty as a queen: to give a birth to a men successor who later became a king. 10 out of 16 children died before age 10, only 3 of them lived longer than 40

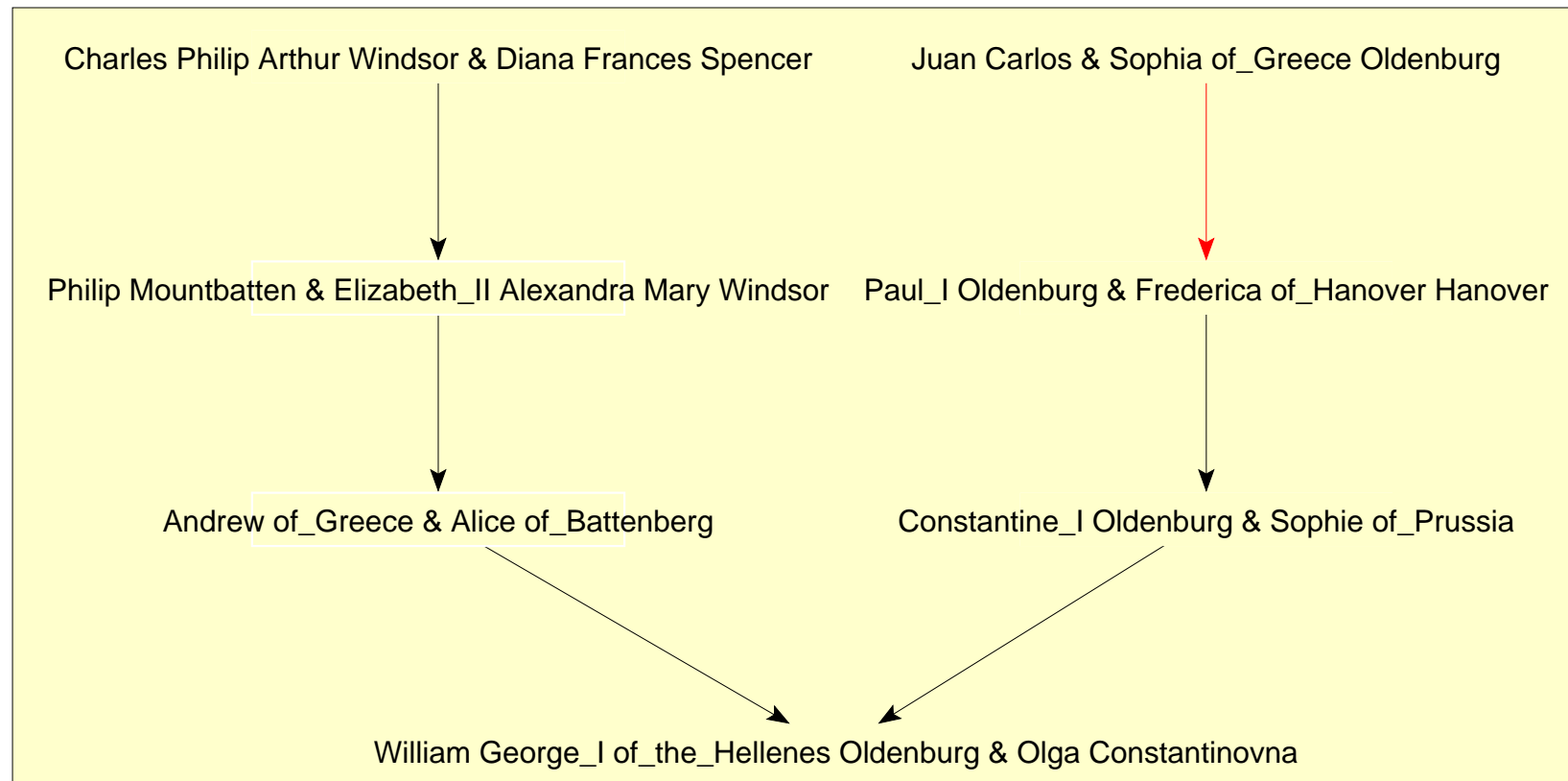
years.



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



Basic Kin Types

Anthropologists typically use a basic vocabulary of kin types to represent genealogical relationships. One common version of the vocabulary for basic relationships:

KinType	EnglishType
P	Parent
F	Father
M	Mother
C	Child
D	Daughter
S	Son
G	Sibling
Z	Sister
B	Brother
E	Spouse
H	Husband
W	Wife

Calculating kinship relations

Pajek generates three relations when reading genealogy as Ore graph:

F: *_ is a father of _*

M: *_ is a mother of _*

E: *_ is a spouse of _*

Additionally we must generate two binary diagonal matrices, to distinguish between male and female:

L: *_ male _* / 1-male, 0-female

J: *_ female _* / 1-female, 0-male

Other *basic* relations can be obtained by running given macros:

– <i>is a parent of</i> –	P	$=$	$F \cup M$
– <i>is a child of</i> –	C	$=$	P^{-1}
– <i>is a son of</i> –	S	$=$	$L * C$
– <i>is a daughter of</i> –	D	$=$	$J * C$
– <i>is a husband of</i> –	H	$=$	$L * E$
– <i>is a wife of</i> –	W	$=$	$J * E$
– <i>is a sibling of</i> –	G	$=$	$((F^{-1} * F) \cap (M^{-1} * M)) \setminus I$
– <i>is a brother of</i> –	B	$=$	$L * G$
– <i>is a sister of</i> –	Z	$=$	$J * G$

Several *derived* relations can be computed, e.g.:

– <i>is an uncle of</i> –	U	$=$	$B * P$
– <i>is an aunt of</i> –	A	$=$	$Z * P$
– <i>is a semi-sibling of</i> –	G_e	$=$	$(P^{-1} * P) \setminus I$
– <i>is a grandparent of</i> –	GP	$=$	P^2
– <i>is a grandfather of</i> –	GF	$=$	$F * P = L * GP$
– <i>is a niece of</i> –	N_i	$=$	$D * G$

Sizes of kinship relations in genealogies

Kin Type	Turks	Ragusa	Loka	Silba	Royal
P-Parent	1987	9315	68052	9627	3724
F-Father	1022	4956	34330	4998	2010
M-Mother	965	4359	33722	4629	1714
C-Child	1987	9315	68052	9627	3724
D-Daughter	857	3577	32647	4518	1589
S-Son	1130	5738	35405	5109	2135
G-Sibling	2485	8782	69347	7803	2858
Z-Sister	2256	6949	66874	7314	2634
B-Brother	2714	10615	71820	8292	3082
E-Spouse	407	2002	14154	2217	1138
H-Husband	407	2002	14154	2217	1138
W-Wife	407	2002	14154	2217	1138
U-Uncle	3816	16665	81695	11372	3453
A-Aunt	3477	10644	80995	10564	2973
Ge-Semi-sibling	2926	10763	76746	8972	3372
# Individuals	1269	5999	47956	6427	3010

Relative sizes of kinship relations in genealogies

Kin Type	Turks	Ragusa	Loka	Silba	Royal
P-Parent	1.000	1.000	1.000	1.000	1.000
F-Father	0.514	0.532	0.504	0.519	0.540
M-Mother	0.486	0.468	0.496	0.481	0.460
C-Child	1.000	1.000	1.000	1.000	1.000
D-Daughter	0.431	0.384	0.480	0.469	0.427
S-Son	0.569	0.616	0.520	0.531	0.573
G-Sibling	1.250	0.943	1.019	0.811	0.767
Z-Sister	1.135	0.746	0.983	0.760	0.707
B-Brother	1.366	1.140	1.055	0.861	0.828
E-Spouse	0.205	0.215	0.208	0.230	0.306
H-Husband	0.205	0.215	0.208	0.230	0.306
W-Wife	0.205	0.215	0.208	0.230	0.306
U-Uncle	1.920	1.789	1.200	1.181	0.927
A-Aunt	1.750	1.143	1.190	1.097	0.798
Ge-Semi-sibling	1.473	1.155	1.128	0.932	0.905

Ore graph – Relation Uncle

Options/ReadWrite/Ore: Different relations for male and female links [checked]

Options/ReadWrite/GEDCOM-Pgraph [unchecked]

File/Network/Read/Ragusa.ged

Select Gender partition in the First Partition box

Macro/Play/AddAllRelations.mcr

Info/Network/Multiple Relations

Net/Transform/Multiple Relations/Extract Relations [13]

Find person who is in position of an uncle the highest number of times:

Net/Partitions/Degree/Output

Info/Partition --> Federicus/Goce/

Visualize the person and his neighbourhood:

Net/k-Neighbours/Output --> Federicus/Goce/

Operations/Extract from Network/Partition [0 1]

Layout/Circular/Using Partition