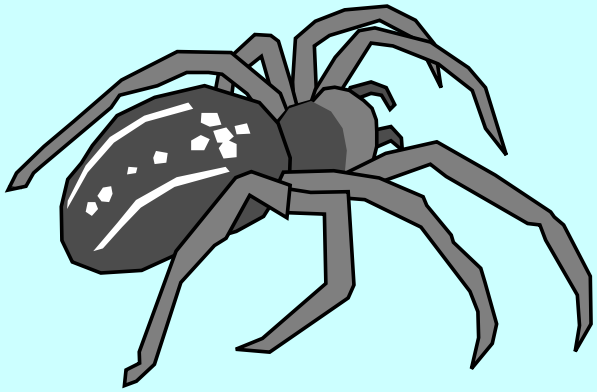
Drawing Genealogies

Andrej Mrvar
Faculty of Social Sciences

Vladimir Batagelj
Faculty of Mathematics and Physics
University of Ljubljana, Slovenia



1. Sources of Genealogies

- Research data
- Genealogies of families and/or teritorial units, e.g. Mormons genealogy:
  `http://www.familytreemaker.com/00000116.html`
  `http://www.genforum.com/mormon/`
- Special Genealogies
  - Students and their PHD thesis advisors: Theoretical Computer Science Genealogy
    `http://sigact.acm.org/genealogy/`
  - gods (antique)

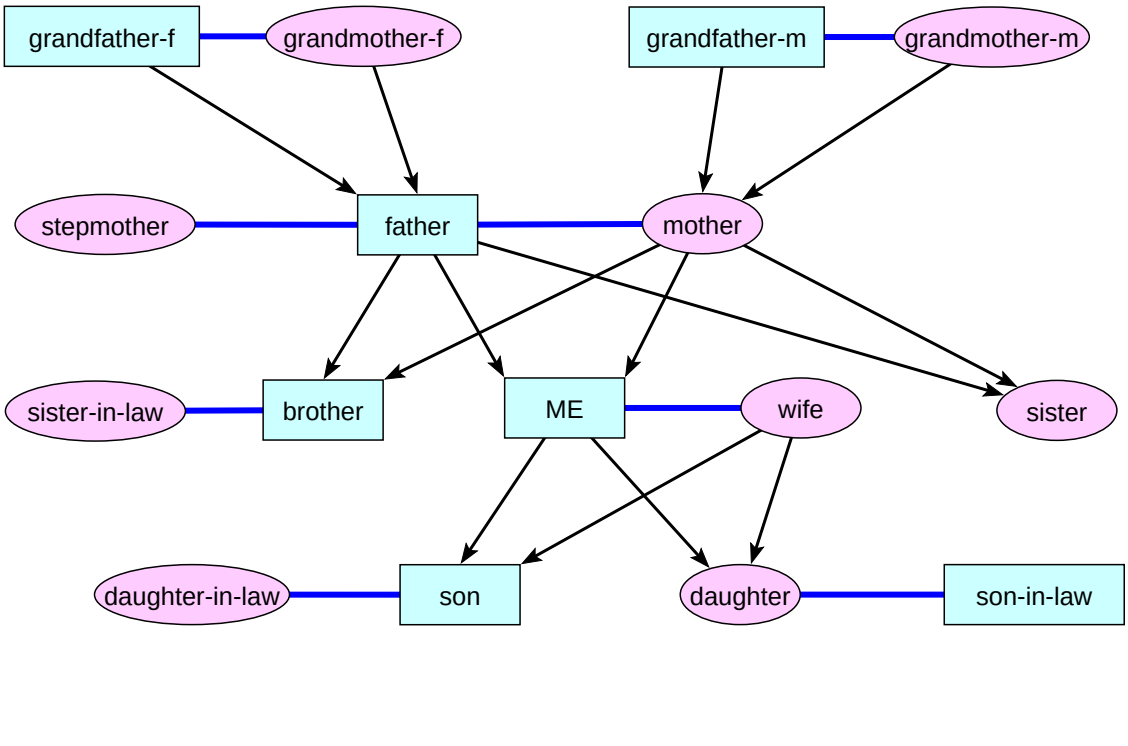
There exist many programs for data entry (`GIM`, `Brother's Keeper`, `Family Tree Maker`,...), but only few analyses can be done using that programs. We use `Pajek` for analyses and visualization of genealogies.

2. GEDCOM Format

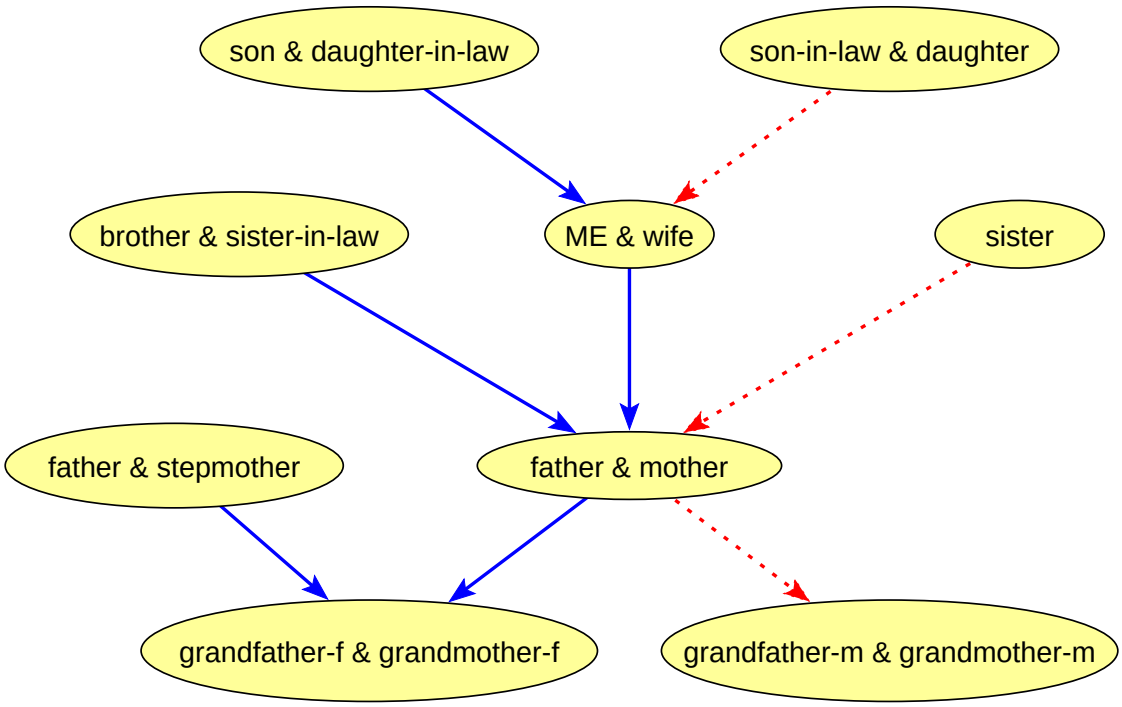
```
0 HEAD
1 SOUR BROSKEEP
2 VERS 5.2 WINDOWS
1 DATE 19 APR 1996
1 CHAR IBMPC
1 FILE F:\BK5\PRES\PRES.GED
...
0 @I1@ INDI
1 NAME William Jefferson /CLINTON/
1 SEX M
1 OCCU US President No. 42
1 BIRT
2 DATE 19 AUG 1946
2 PLAC Hope, Hempstead Co., AR
1 REFN Clinton-1
1 NOTE Born as William Jefferson Blythe IV.
2 CONT Last name was changed to Clinton.
2 CONT on 12 June 1962.
1 FAMS @F1@
1 FAMC @F2@
```

```
...
...
0 @F1@ FAM
1 HUSB @I1@
1 WIFE @I2@
1 CHIL @I66@
1 MARR
2 DATE 11 OCT 1975
0 @F2@ FAM
1 HUSB @I3@
1 WIFE @I4@
1 CHIL @I1@
1 MARR
2 DATE  3 SEP 1943
2 PLAC Texarkana, Miller Co., AR
0 TRLR
```

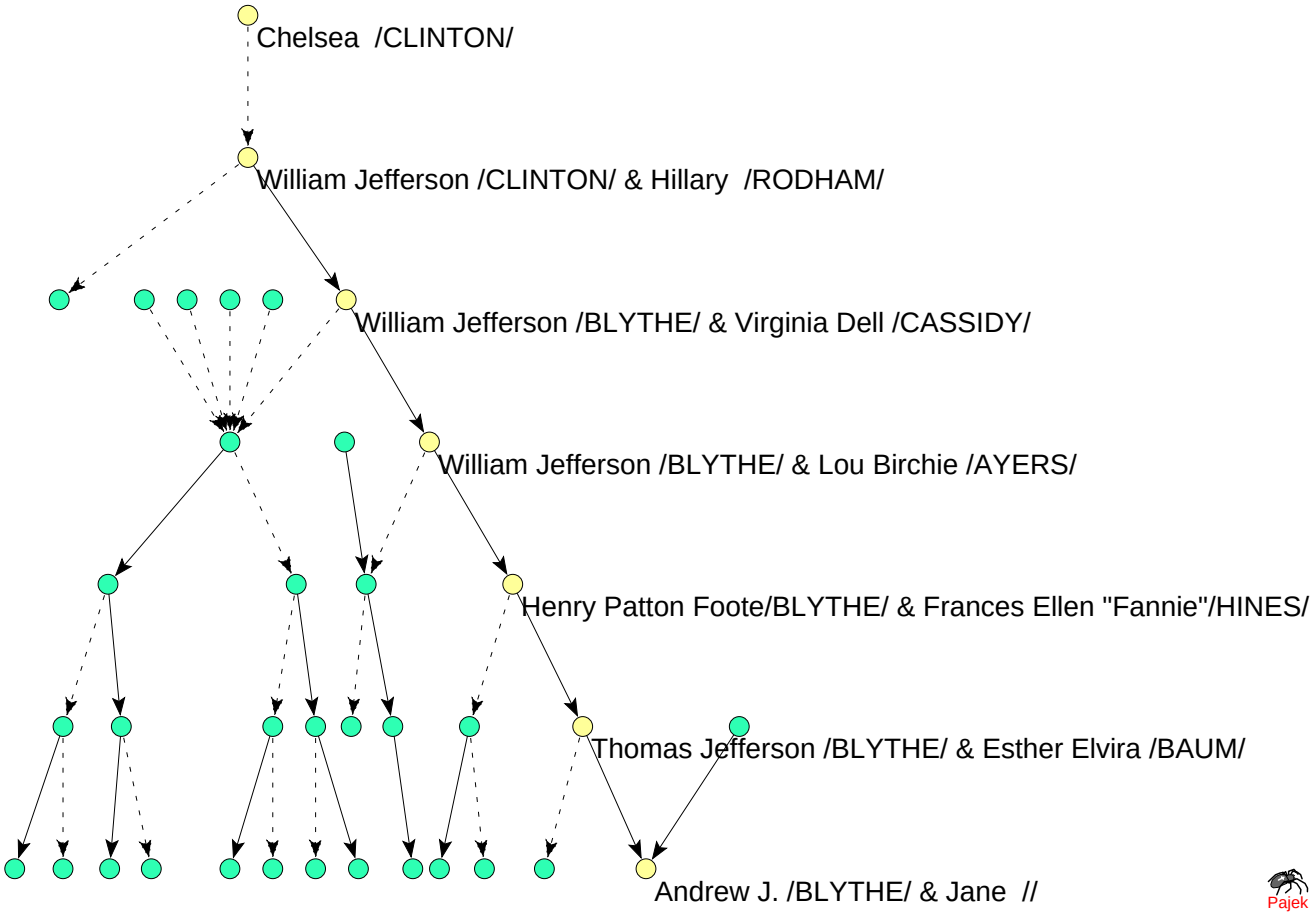
3. Ore-Graph


grandfather-f
grandmother-f
grandfather-m
grandmother-m
stepmother
father
mother
sister-in-law
brother
ME
wife
sister
daughter-in-law
son
daughter
son-in-law
Pajek

4. p-graph – D. R. White



5. Genealogy of Bill Clinton



6. Sparse networks

Regular genealogy – every person has at most two parents.

Genealogies are *sparse* networks.

For directed part of *regular Ore genealogy*:

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

A – set of directed lines, V – set of vertices, $d_{in}(v)$ – input degree of vertex v, $d_{in}(v) \leq 2$. Most of the persons are married only once, some are not married.

For undirected part of *Ore* genealogy

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

E – set of undirected lines. So

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

P-graphs are almost trees – deviations from trees are caused by marriages among relatives. $|V_p|$ – number of vertices of *p-graph*, n_{mult} – number of multiple marriages

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

$|V| - 2|E| + n_{mult}$ – approximation of number of non-married persons, $|E|$ – approximation of number of couples.

$$|V_p| \geq |V| - |E|$$

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

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

Advantages of p-graphs:

- Less vertices and lines in `p-graphs`;
- `p-graphs` are directed, acyclic;
- `p-graphs` are more suitable for analyses.

7. Drawing genealogies

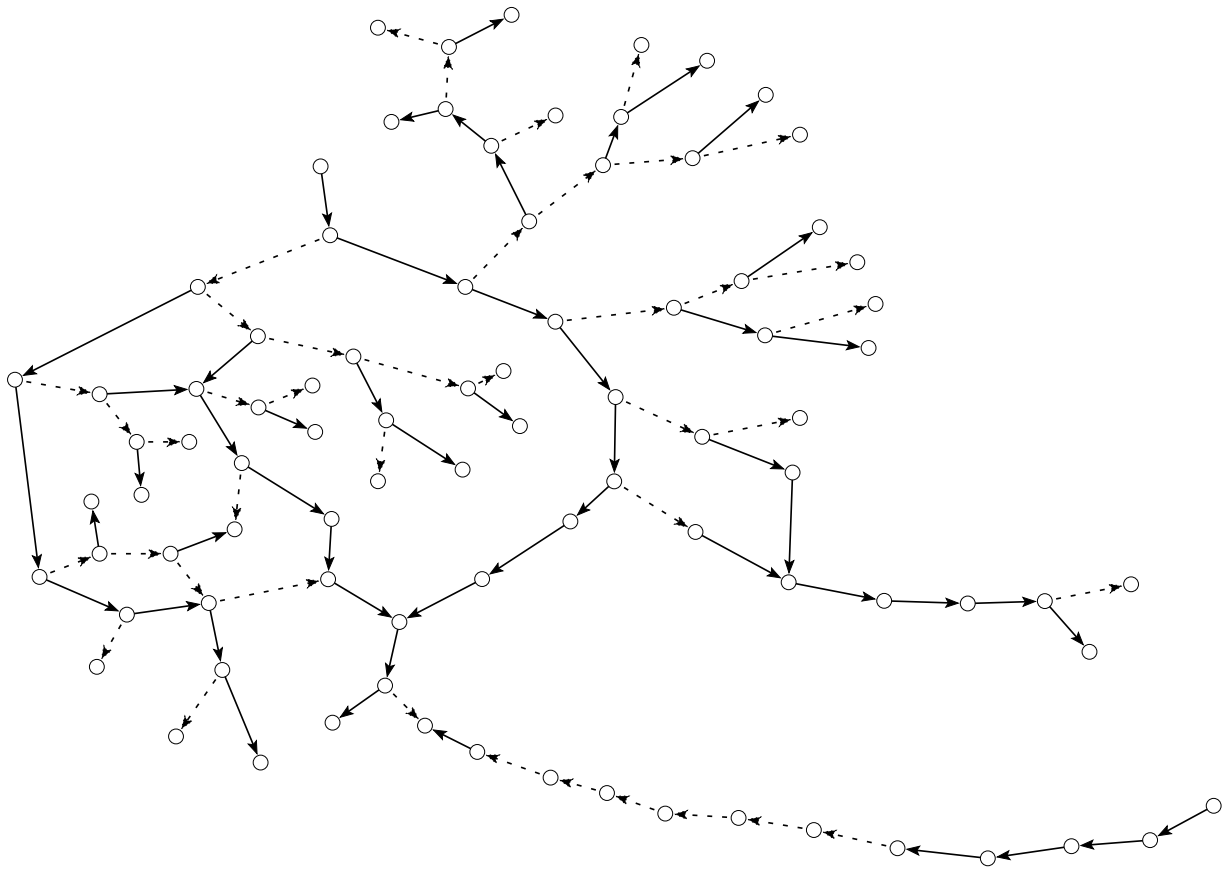
Standard automatic layout algorithms (spring embedders) can be used, but generations (time) cannot be easily seen in obtained pictures. Our goal is: Layout of a genealogy in layers, such that there exist no connections among vertices within layers.

- Determine layers / generations – y coordinates.
  - delete all tree subgraphs;
  - determine all first and last vertices;
  - determine all longest paths among first and last vertices;
  - determine the levels of vertices along the longest paths;
  - normalize the levels;
  - extend levels to the entire genealogy.

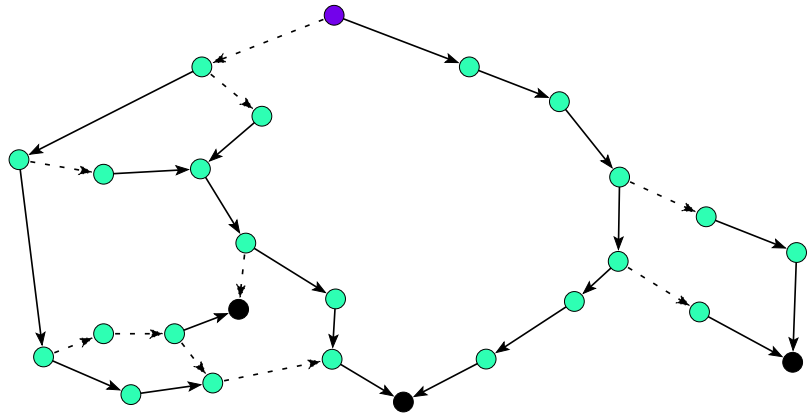
- Determine positions of vertices in layers.
  - put vertices at random positions in their layers;
  - for each vertex set its x-coordinate as the average of x coordinates of all its neighbours; repeat this until the coordinates stabilize;
  - displace vertices that are too close to a selected minimum distance.
- Optimize the total length of lines using relocation algorithm.

It is easy to generalize the algorithm for drawing genealogies in space (3D) – layers are planes on differents heights – z coordinate.

8. Complete genealogy

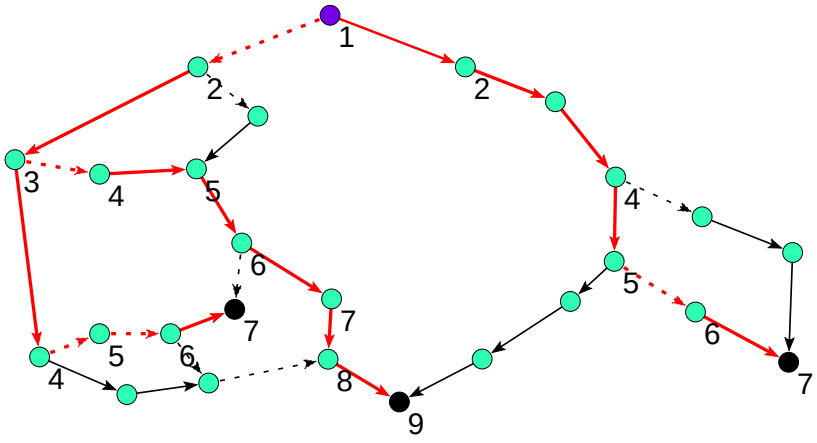


9. Non-tree part



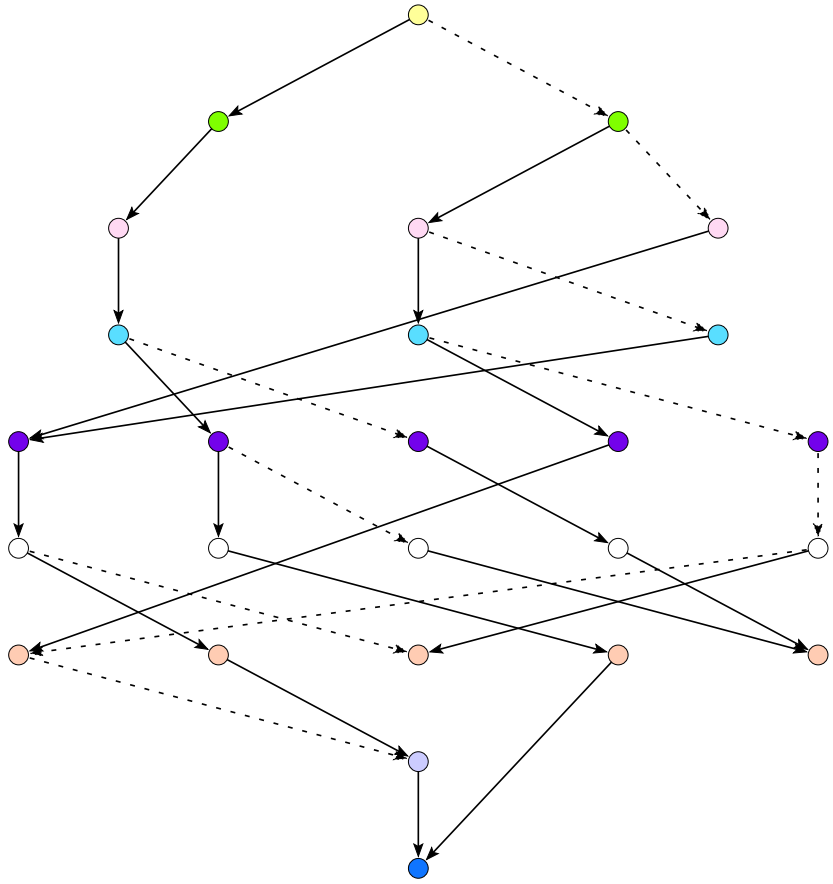
Pajek

10. Longest paths

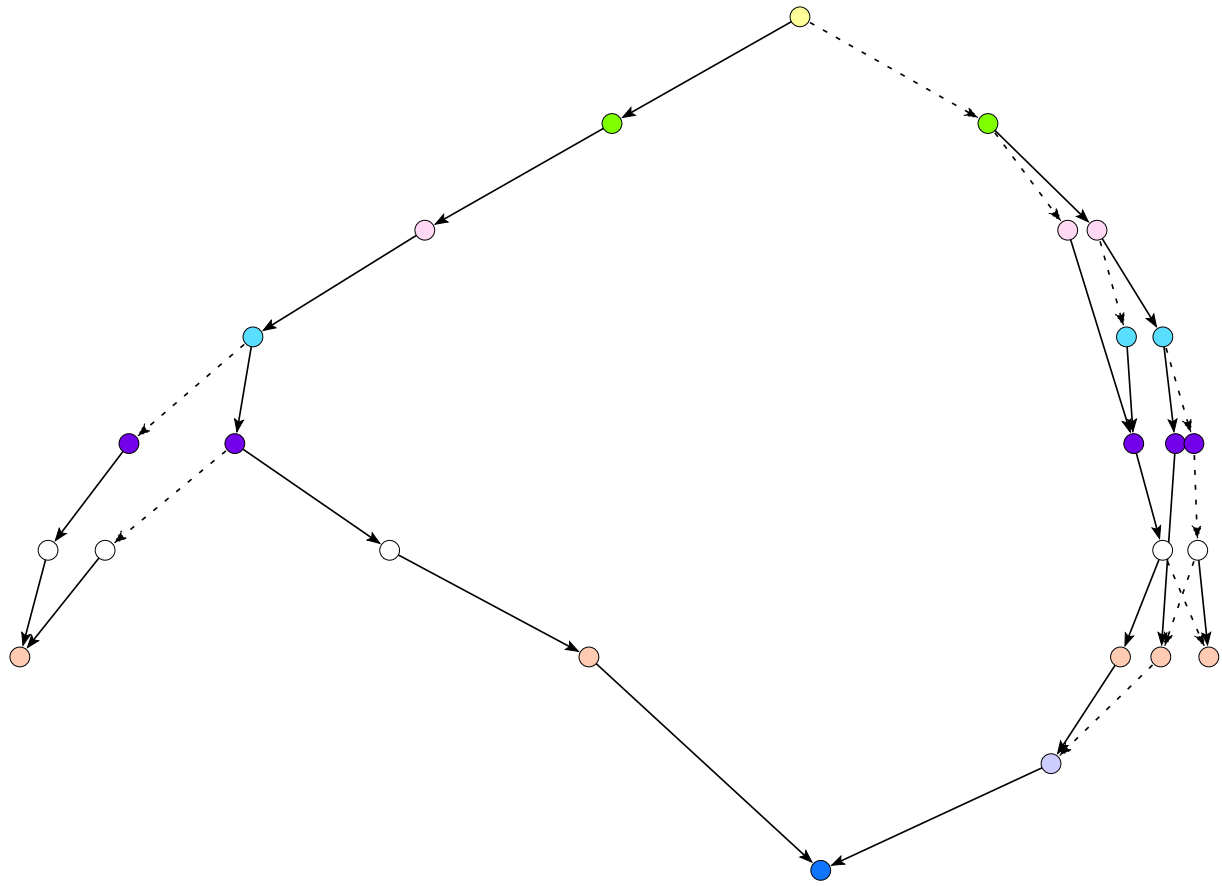


Pajek

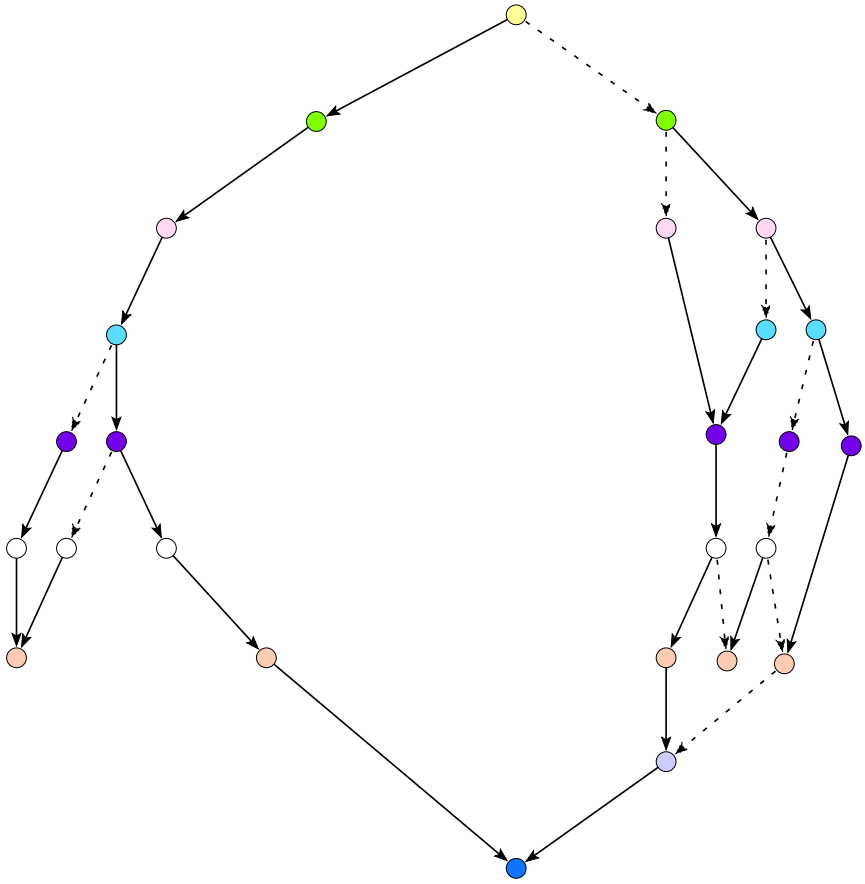
11. Genealogy drawn in layers



12. x coordinates averaged

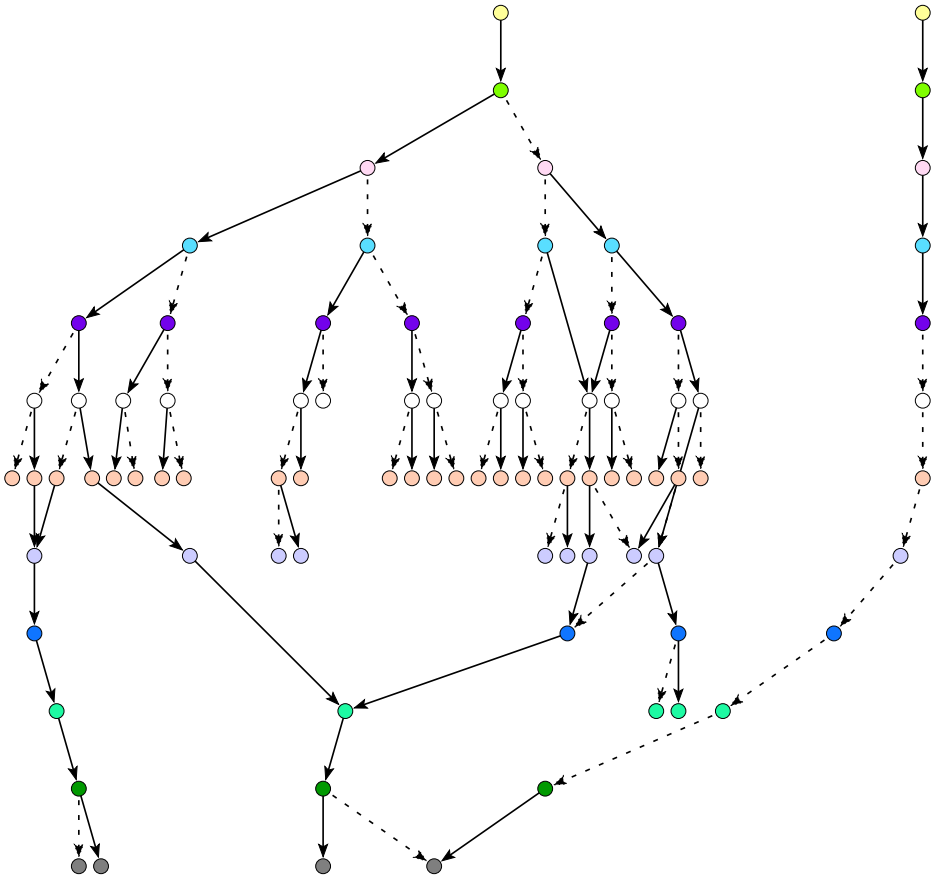


13. Optimized layers



Pajek

14. Complete genealogy – final layout



15. Analyses of genealogies

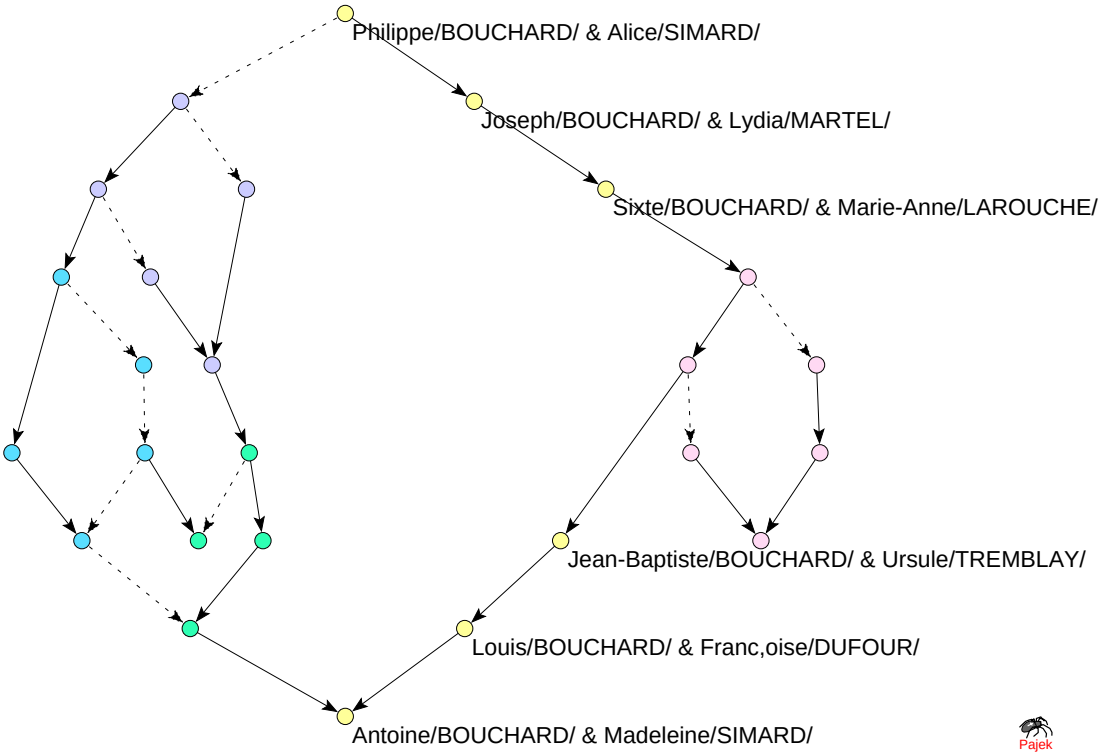
- Searching for the shortest kinship paths among persons.
- Determining all predecessors and successors of a selected person.
- Extracting neighbourhood of a selected person.
- Searching for interesting patterns in a genealogy – marriages among relatives, children having many parents, persons married several times...
- Statistics: average number of children, maximun number of children,...

16. Polititians of Quebec

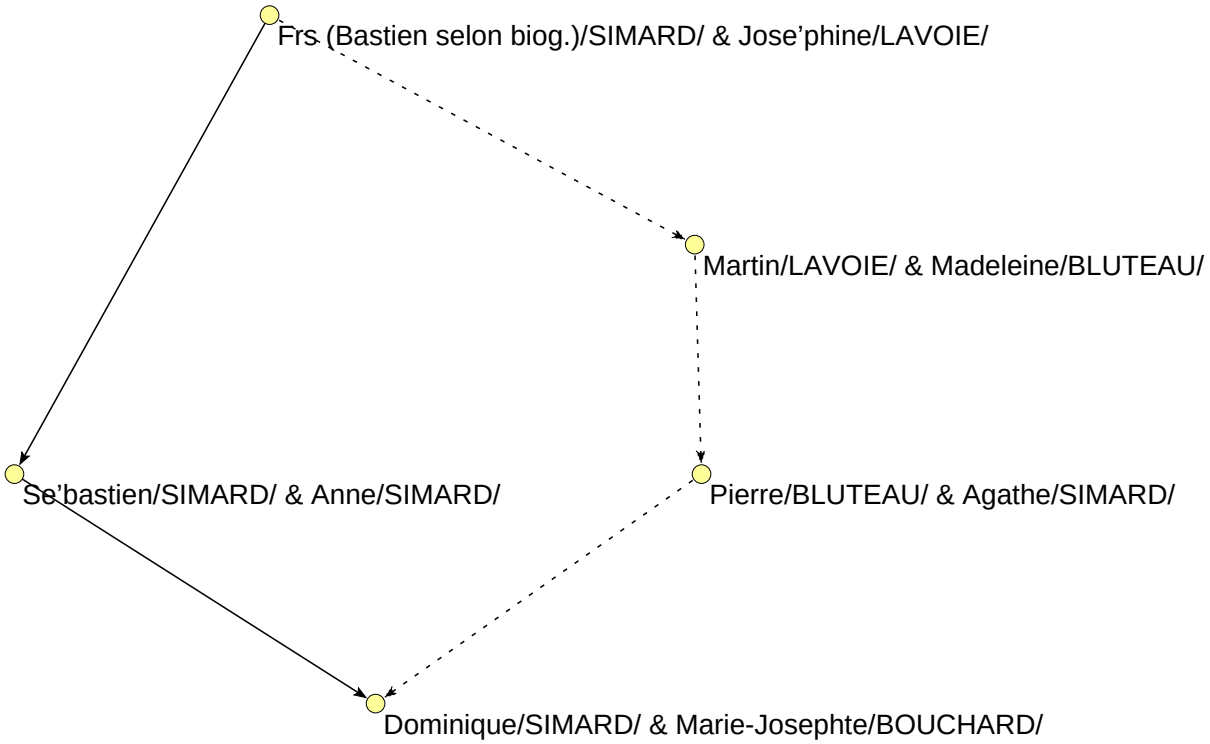
`ftp://ftp.cam.org/users/beaur/bouchard.ged`

`http://www.cam.org/~beaur/gen/politiciens.html`

- 173 records about inidividuals;
- 86 families and 1 individual
- one bicomponent having 26 vertices



17. Polititians – 5 cycle



18. TCS Genealogy

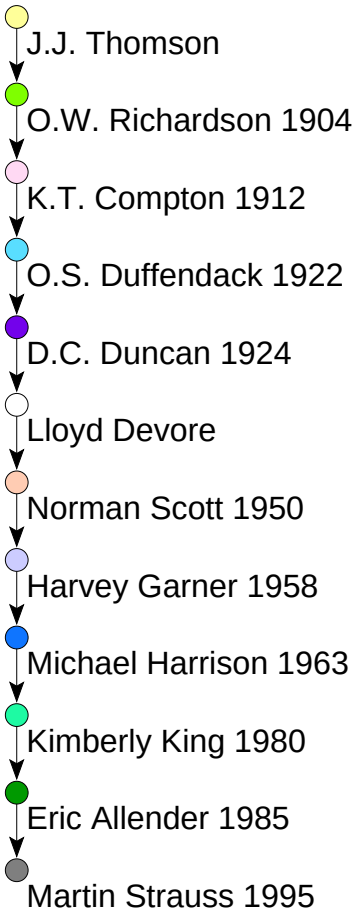
The Genealogy of Theoretical Computer Science is available as a text file on Internet.
`http://sigact.acm.org/genealogy/`
The following fields are available:

- the student's name,
- the name of the student's thesis adviser,
- an acronym for the university granting the doctoral degree, and
- the year the degree was granted.

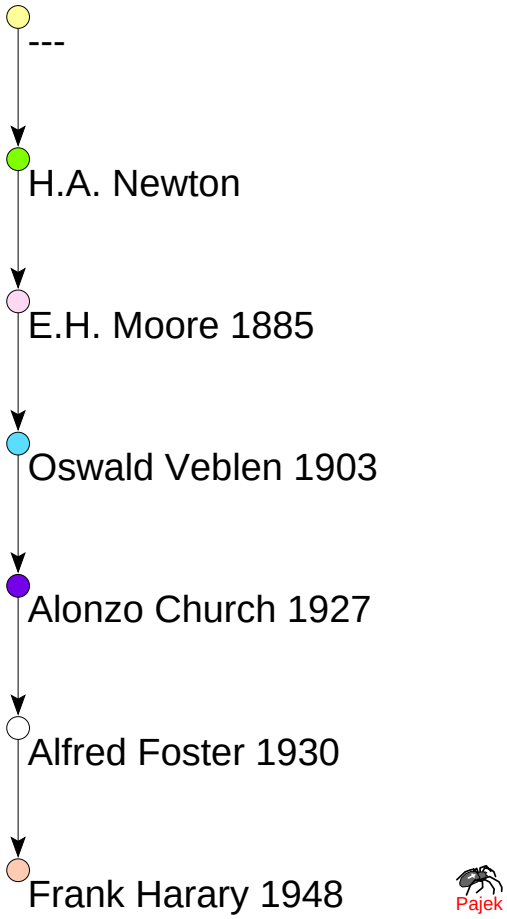
From the data a network can be constructed.

- 1882 vertices and 1740 directed lines;
- 168 weakly connected components, one large (1025 vertices);

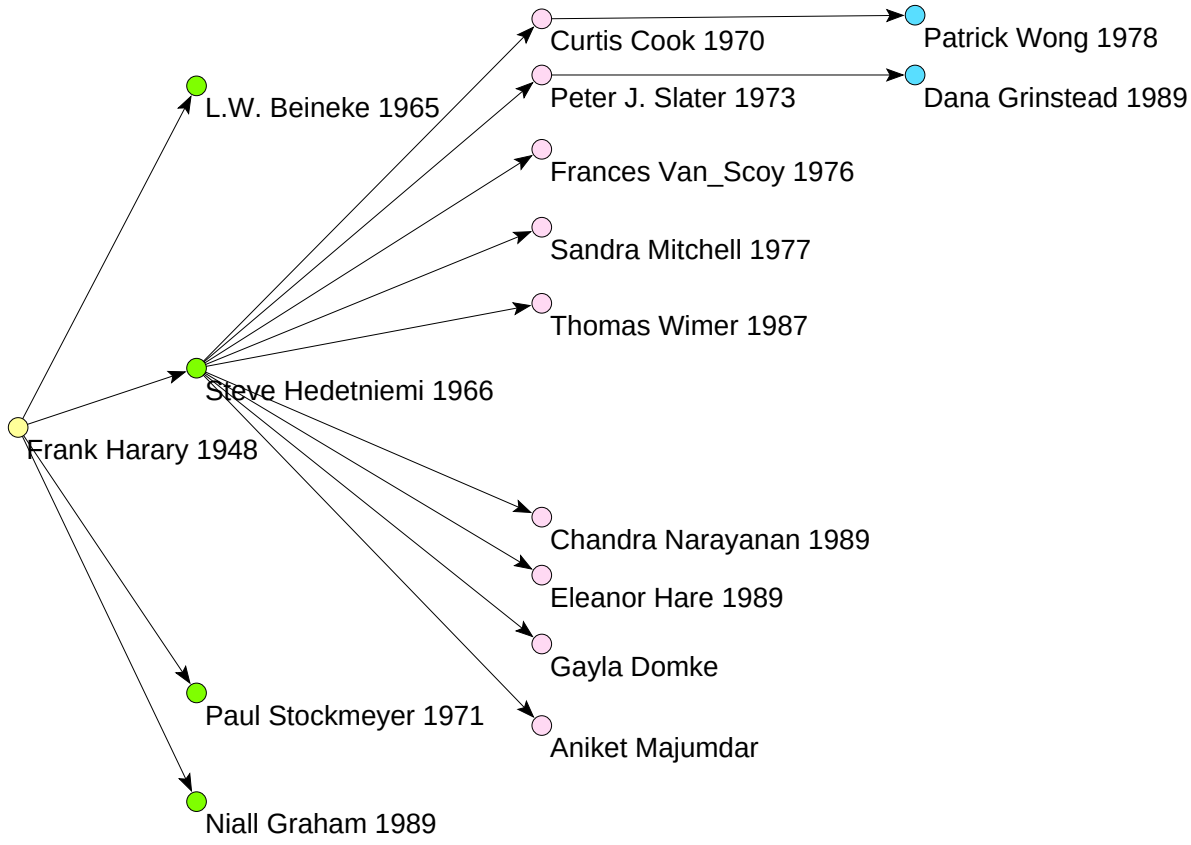
19. Longest Shortest Path



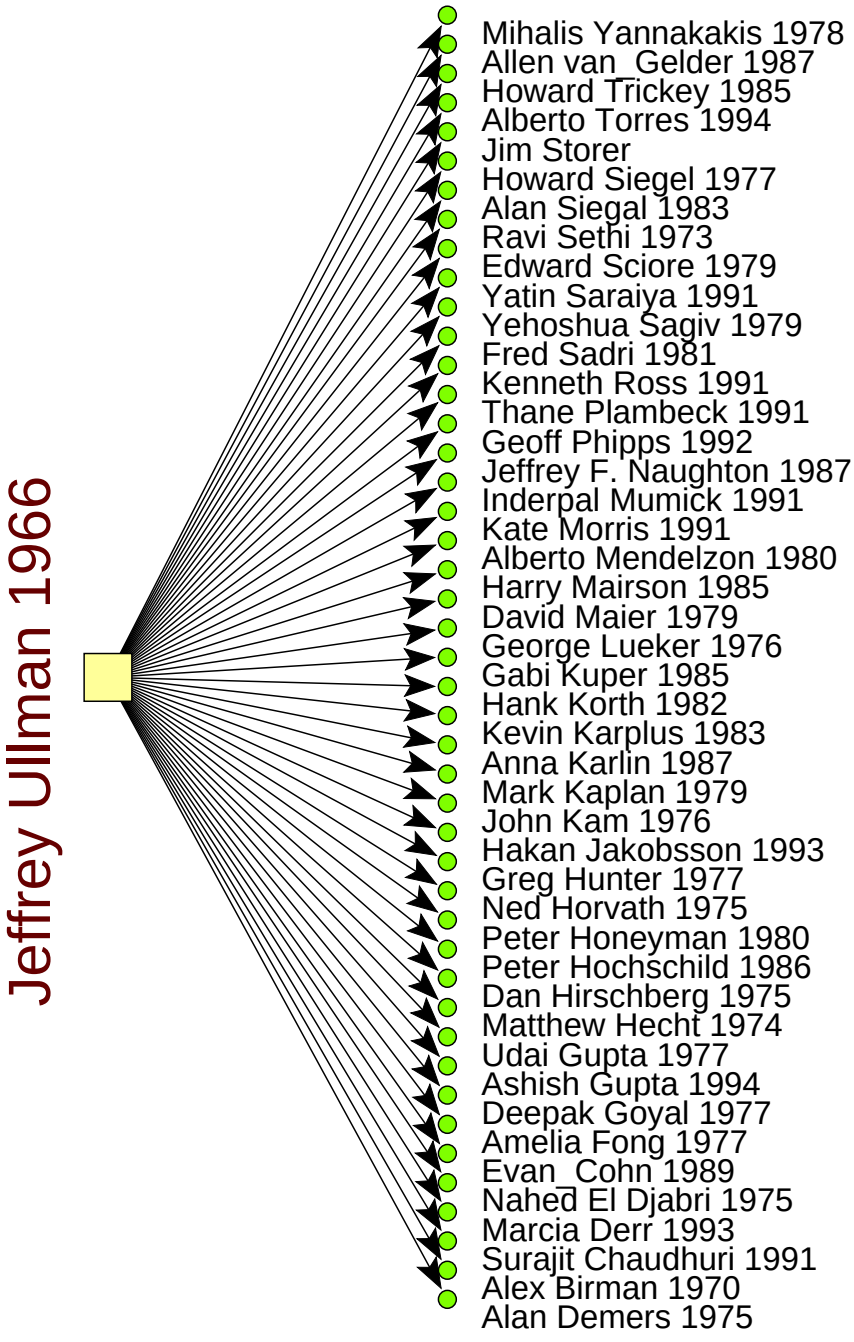
20. All Predecessors of Harary



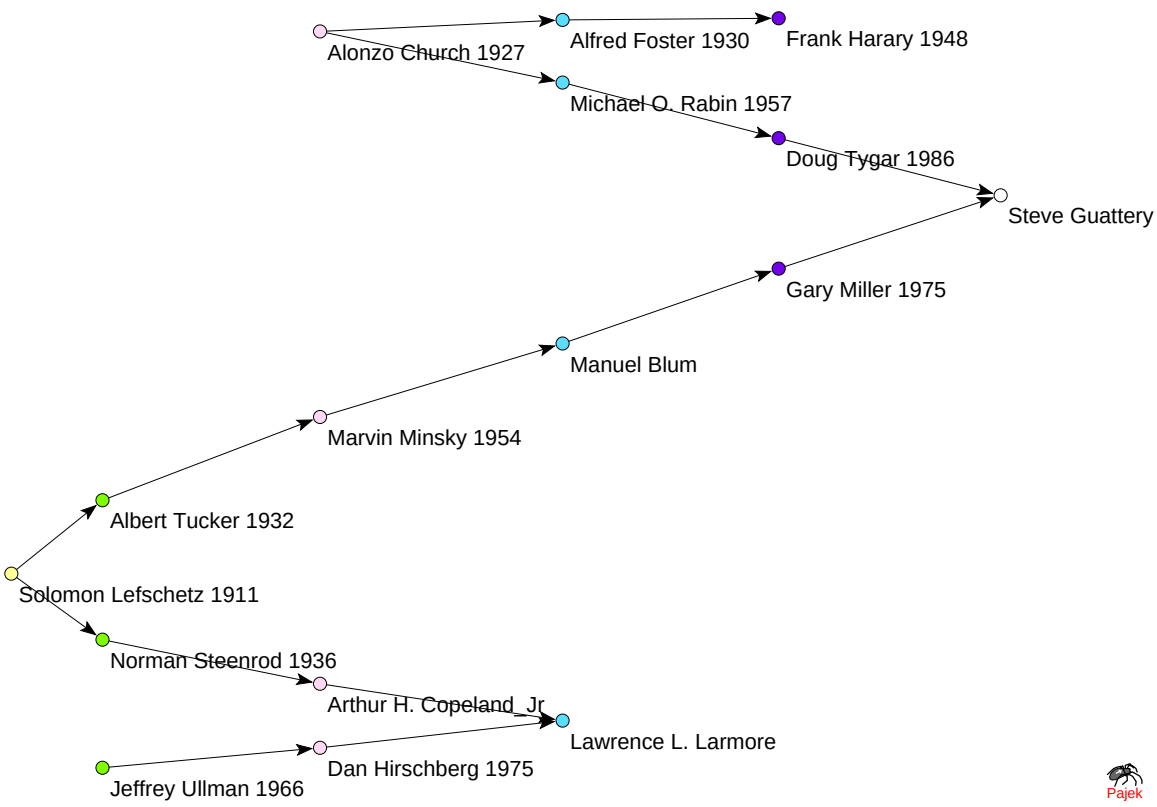
21. All Successors of Harary



22. Direct successors of Ullman



23. Shortest path from Harary to Ullman



24. Index of connectedness

n – number of vertices,

m – number of lines,

k – number of weakly connected components,

M – number of maximal vertices (vertices having output degree 0, $M \geq 1$).

If G is a forest (consists of trees), then
$m = n - k$, or $k + m - n = 0$.

Genealogy is *regular* if everyone has at most 2 parents.

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

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

or

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

Index of connectedness:

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

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

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

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

P has some interesting properties:

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

25. Genealogies having $P = 1$