Network Analysis

Andrej Mrvar
Faculty of Social Sciences
Ljubljana

Postgraduate study - 3rd cycle

Introduction

Let $\mathbf{U} = \{X_1, X_2, \ldots, X_n\}$ be a finite set of units. Connections among units are described using one or more *binary relations* $R_t \subseteq \mathbf{U} \times \mathbf{U}, \quad t = 1, \ldots, r$, which determine *a network* $\mathcal{N} = (\mathbf{U}, R_1, R_2, \ldots, R_r)$.

Example: A relation can represent friendship, negative relation, kinship relation (. . .is a child of. . ., . . .is a daughter of . . ., . . .is married to. . .), citations. . .

In the following we will use mostly one relation R.

$X_i R X_j$ is read as:
unit X_i is in relation R with unit X_j.

Example: if R corresponds to relation 'liking', then $X_i R X_j$, means that person X_i likes person X_j.

A network defined using relation R can be represented in different ways:

- Representation using corresponding **binary matrix**
  $\mathbf{R} = [r_{ij}]_{n \times n}$, where

$$r_{ij} = \begin{cases} 1 & X_i R X_j \\ 0 & \text{otherwise} \end{cases}$$

  Sometimes r_{ij} is a real number, expressing the strength of relation R between units X_i in X_j.

- **list of neighbours**
  A network can be described by specifying the list of all other units with which the unit is in relation.

- description by a **graph** $G = (V, L)$ where V is set of vertices and L set of (directed or undirected) lines. Vertices represent units of a network, lines represent the relation. A graph is usually represented by a picture: vertices are drawn as small circles, directed lines are drawn as arcs and undirected lines as edges connecting the corresponding two vertices.

$X_i R X_j \Rightarrow$ there exists directed line from unit X_i to X_j in corresponding graph. Vertex X_i is called *initial*, vertex X_j is called *terminal* vertex.

A line whose initial and terminal vertices are the same is called *a loop*.

If directed lines between two vertices exist in both directions, they are sometimes replaced by a single undirected line.

Sometimes we do the opposite: an undirected line is replaced by two directed lines in opposite directions.

We will denote the number of vertices in a graph with n, and number of directed lines with m.

	1	2	3	4	5
1	0	1	0	1	1
2	0	0	1	0	0
3	1	0	0	1	0
4	0	0	0	0	1
5	1	0	0	0	0

Arcslist

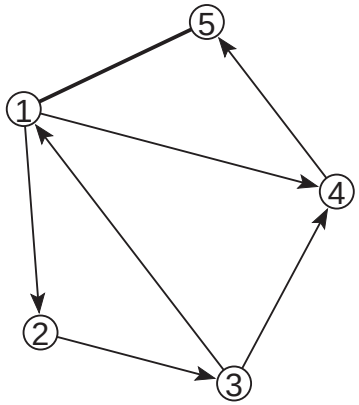
1: 2 4 5

2: 3

3: 1 4

4: 5

5: 1



Types of networks

- *undirected network* – the relation is symmetric – all lines are undirected – edges, $L = E$.
- *directed network* – the relation is not symmetric – all lines are directed – arcs, $L = A$.
- *mixed network* – both arcs and edges exist in a corresponding graph – $L = A \cup E$.
- *two-mode network*
  A two-mode network consists of two sets of units (e. g. people and events), relation connects the two sets, e. g. participation of people in social events.

Small and large networks

Networks with some 100 units and lines are called *small networks*, while networks with some 10000 units and lines are called *large networks*.

Dense and sparse networks

A network is called *sparse* if the number of lines in the corresponding graph is of the same order as the number of vertices ($n \approx km$). Large networks that are sparse can still be efficiently analysed with some algorithms. In real life we often find very large but sparse networks.

In general, the number of lines can be much higher than the number of vertices. Such networks are called *dense*.

If every unit is connected to every other unit the number of lines is n^2 (number of elements in matrix).

If every unit is connected to every other unit except to itself (graph without loops), the number of lines is $n(n-1)$ (number of elements in matrix without diagonal).

According to that the density of a network can be defined:

For networks with loops:

$$Density1 = \frac{m}{n^2}$$

For networks without loops:

$$Density2 = \frac{m}{n(n-1)}$$

If at most one line can exist among any two vertices the density is a real number between 0 and 1.

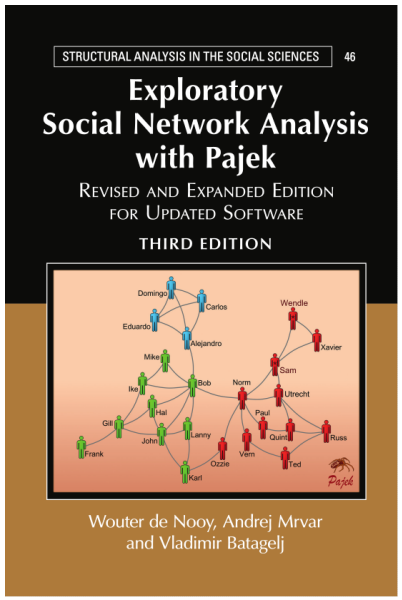
Density of a network is one of the measures by which we can compare different networks.

Example of large networks

- social networks
  - connections among people (friendship);
  - relation among political parties;
  - trade among organizations, countries;
  - genealogies;
  - citation networks;
  - computer networks (local networks, Internet, links among home pages);
  - telephone calls;
- flow charts in computer science;
- Petri nets;
- organic molecule in chemistry;
- connections among words in text;
- transportation networks (airlines, streets, electric networks...).

Pajek

Pajek is a program package for Windows 32 and 64, which enables analyses of *large networks*. Program is freely available at:
http://mrvar.fdv.uni-lj.si/pajek/



de Nooy, Mrvar, Batagelj (2018):
Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software. Third Edition.
Cambridge University Press, New York.

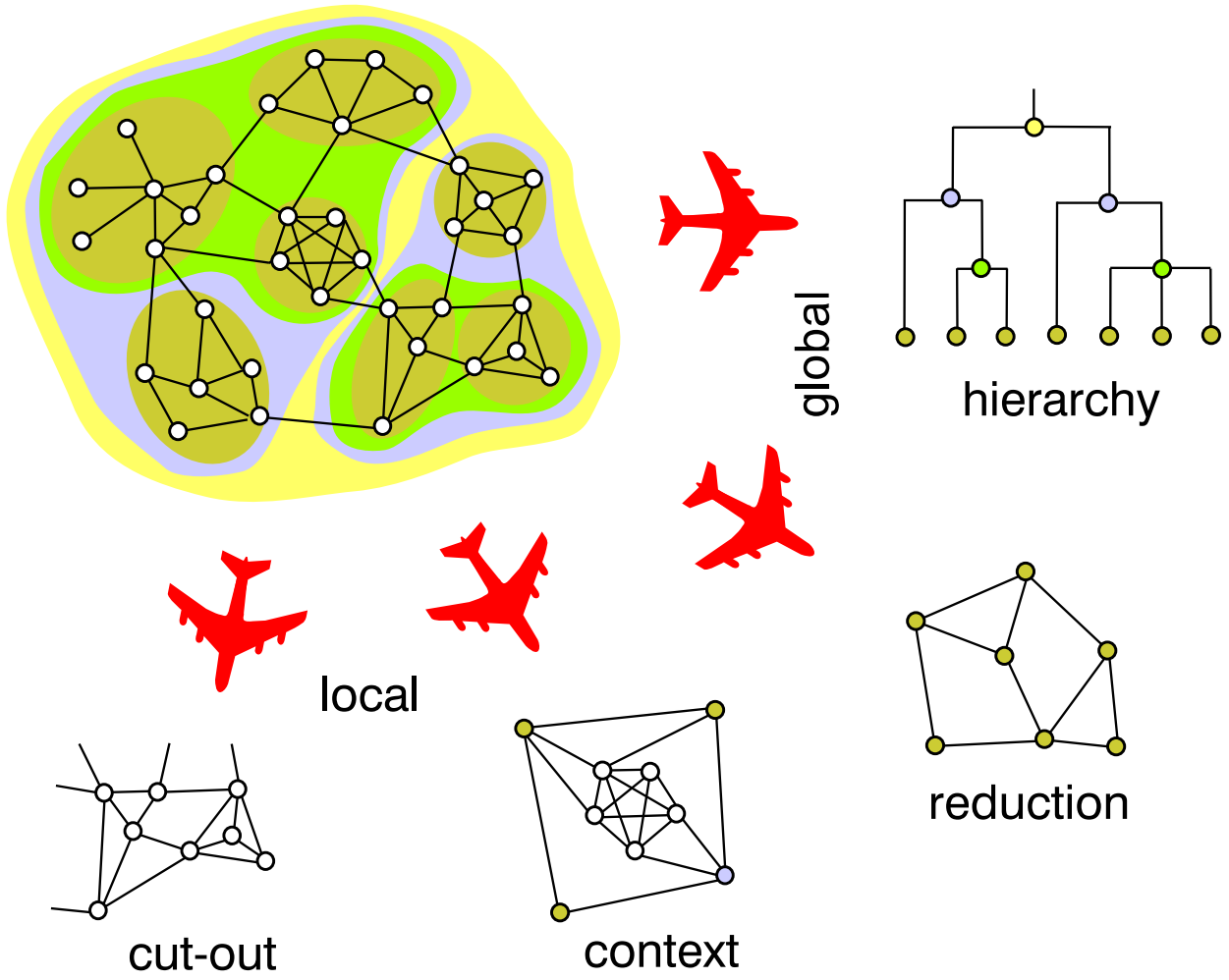
Analyses in Pajek are performed using six data structures:

1. network,
2. partition,
3. cluster,
4. permutation,
5. vector,
6. hierarchy.

Main goals in designing Pajek

The main goals in the design of Pajek are:

- to support abstraction by (recursive) decomposition of a large network into several smaller networks that can be treated further using more sophisticated methods;
- to provide the user with some powerful visualization tools;
- to implement a selection of efficient (*subquadratic*) algorithms for analysis of large networks.


global
hierarchy
local
reduction
cut-out
context

Two-mode networks

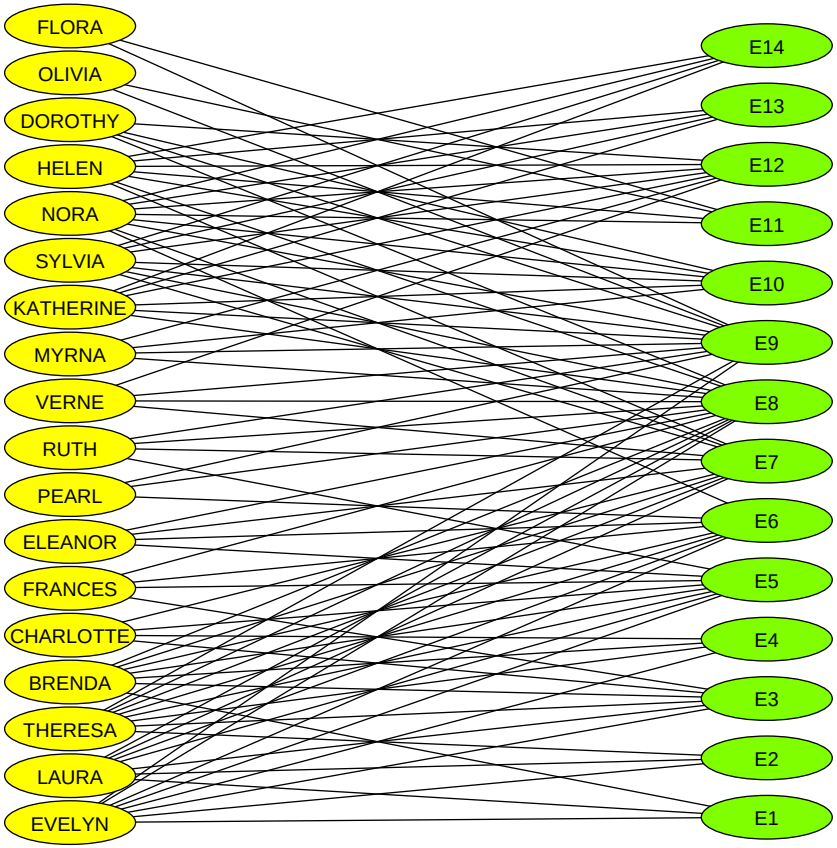
A two-mode network consists of two sets of units (e. g. people and events), relation connects the two sets, e. g. participation of people in social events.

There exist several two-mode networks:

- *Membership in institutions* - people, institutions, *is a member*, e.g. directors and commissioners on the boards of corporations.
- *Voting for suggestions* - polititians, suggestions, *votes for*.
- *'Buying articles in the shop'*, where first set consists of consumers, the second of articles, the connection tells which *article was bought by a consumer*.
- *Readers and magazines*.
- *Citation network*, where first set consists of authors,
  the second set consists of articles/papers,
  connection is a relation *author cites a paper*.
- *Co-autorship networks* - authors, papers, *is a (co)author*.

A corresponding graph is called *bipartite graph* – lines connect only vertices from one to vertices from another set – inside sets there are no connections.

Example: participation of women in social events



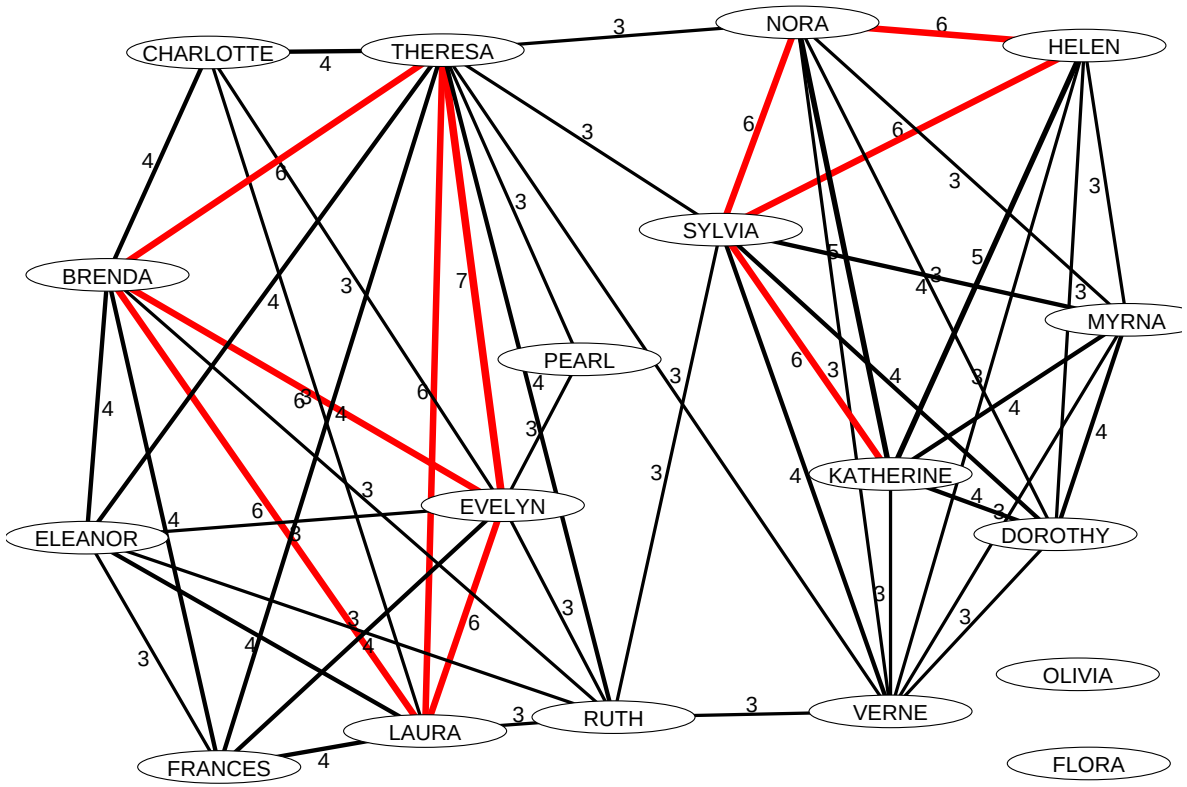
Davis.net

Transforming two-mode networks to ordinary valued networks

Two-mode network can be transformed to 'ordinary' network, where units are only units from first or only units from the second set.

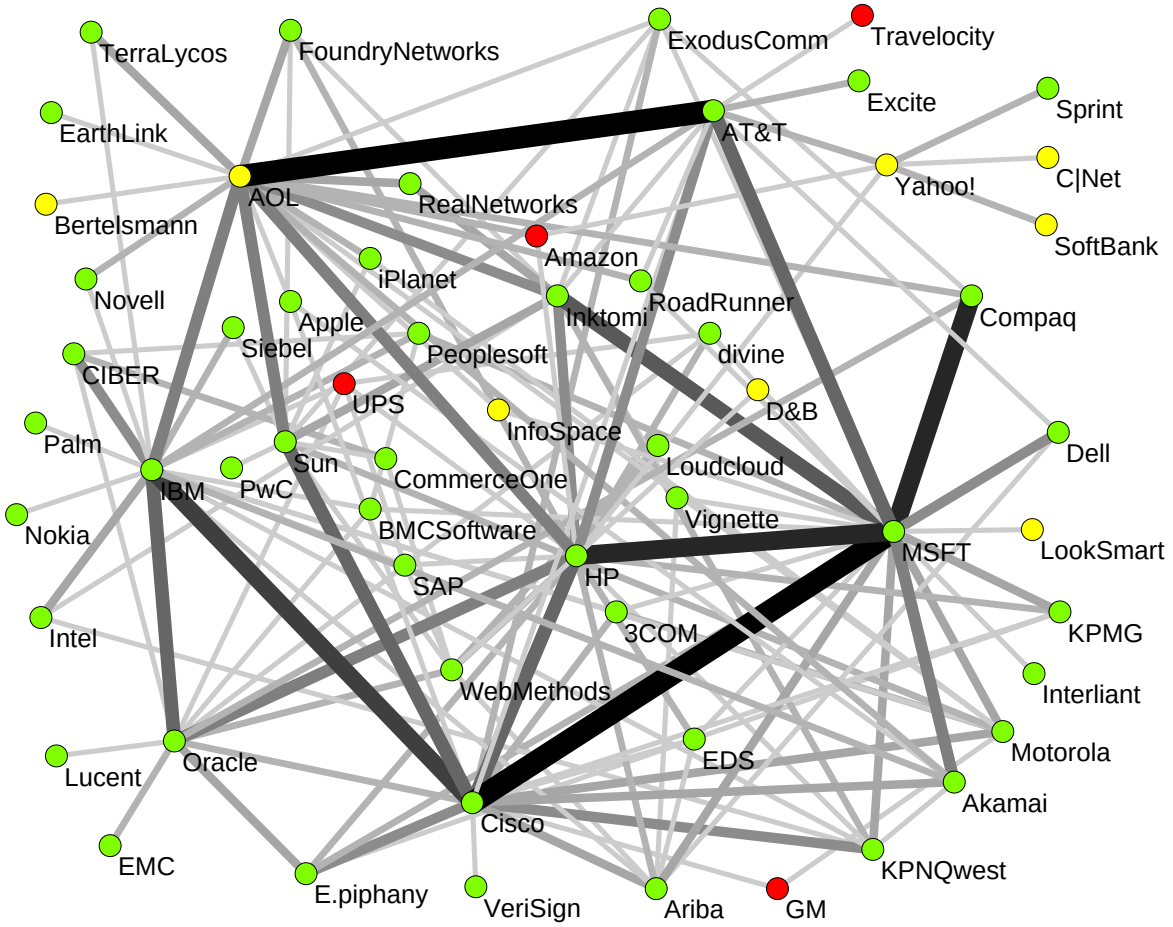
The previous two-mode network (women-events) can be transformed to ordinary network, where units are women. Two women are in relation (in the corresponding graph there exists an undirected line) if they took part in at least one common event. The line value tells the number of events where both of them took place. Loop values represent total number of events for each woman.

But if we transform network to ordinary network where units are events, the two events are in relation (in the corresponding graph there exists an undirected line) if there exists at least one woman who took part in both events. The line value between two events tells the number of women who took place in both events. Loop values represent total number of women present at each event.


CHARLOTTE
THERESA
NORA
HELEN
SYLVIA
BRENDA
MYRNA
PEARL
KATHERINE
EVELYN
ELEANOR
DOROTHY
OLIVIA
RUTH
VERNE
LAURA
FRANCES
FLORA

Example: Internet Industry Partnerships

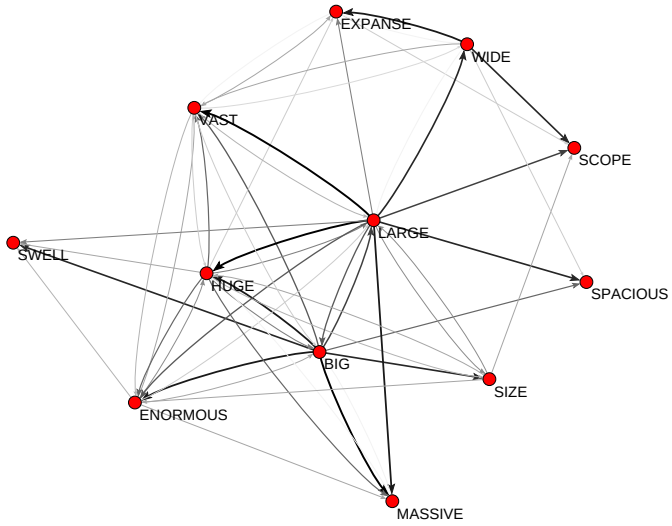
Most important lines

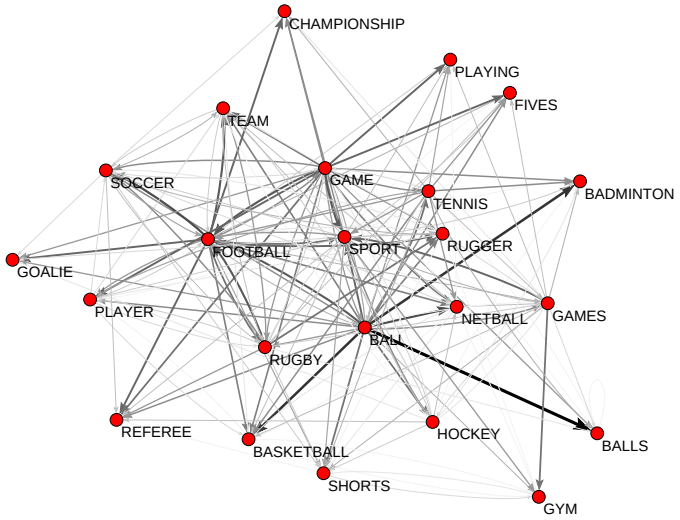


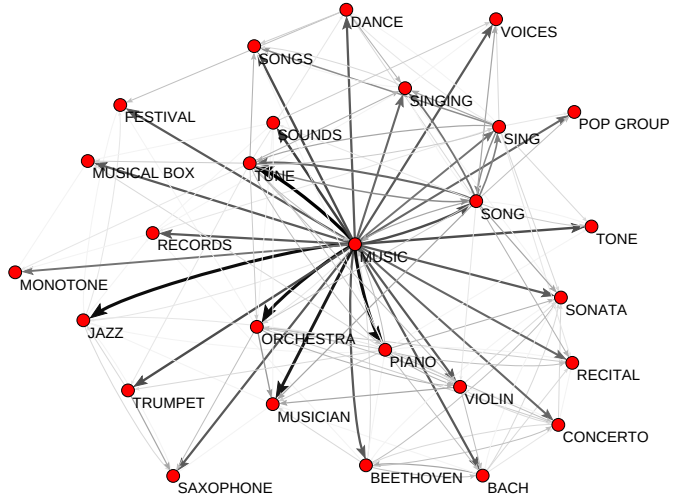
Example: The Edinburgh Associative Thesaurus

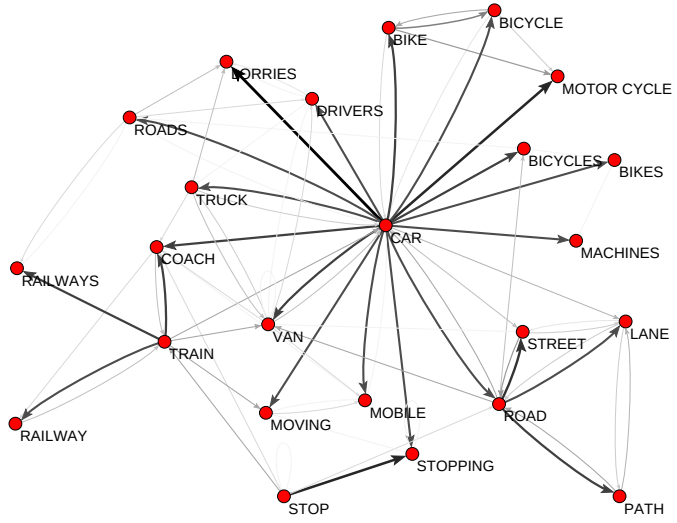
- The Edinburgh Associative Thesaurus (EAT) is a set of word association norms showing the counts of word association as collected from students.
- It is a big network – 23219 vertices (words) and 325624 directed lines.

Selected themes from EAT

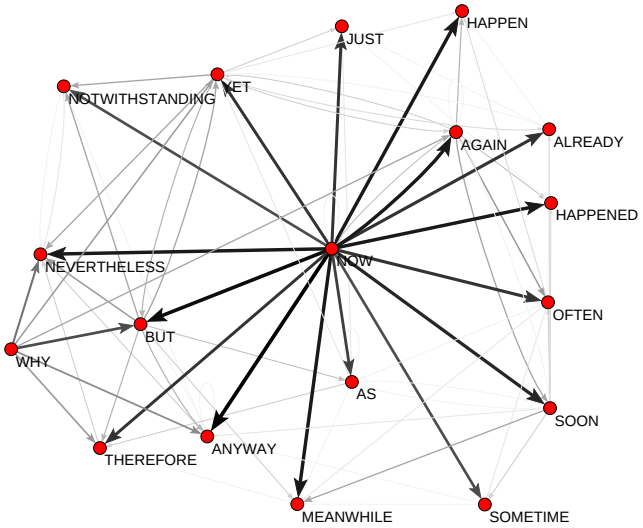


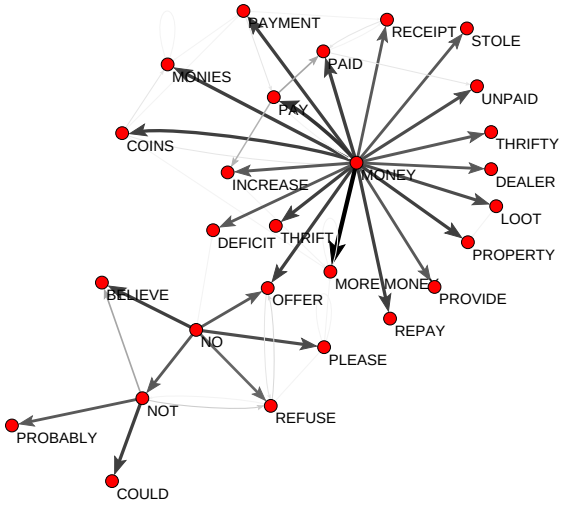


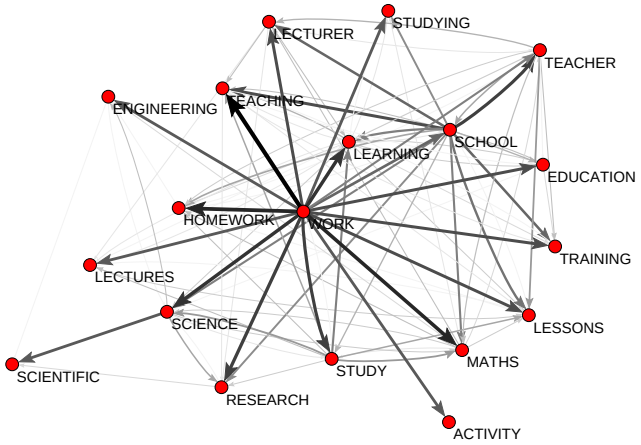


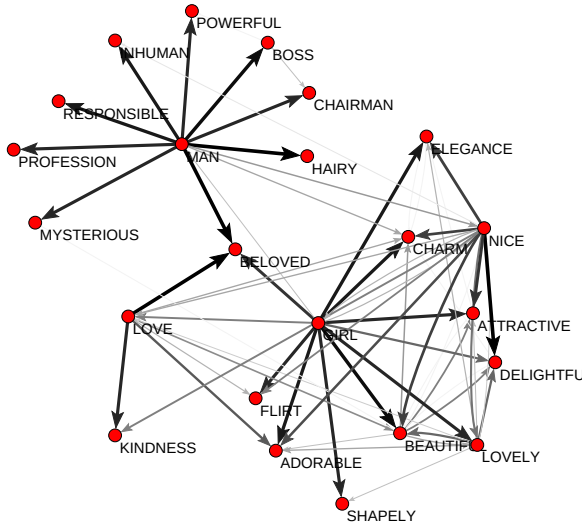


Selected themes from EAT









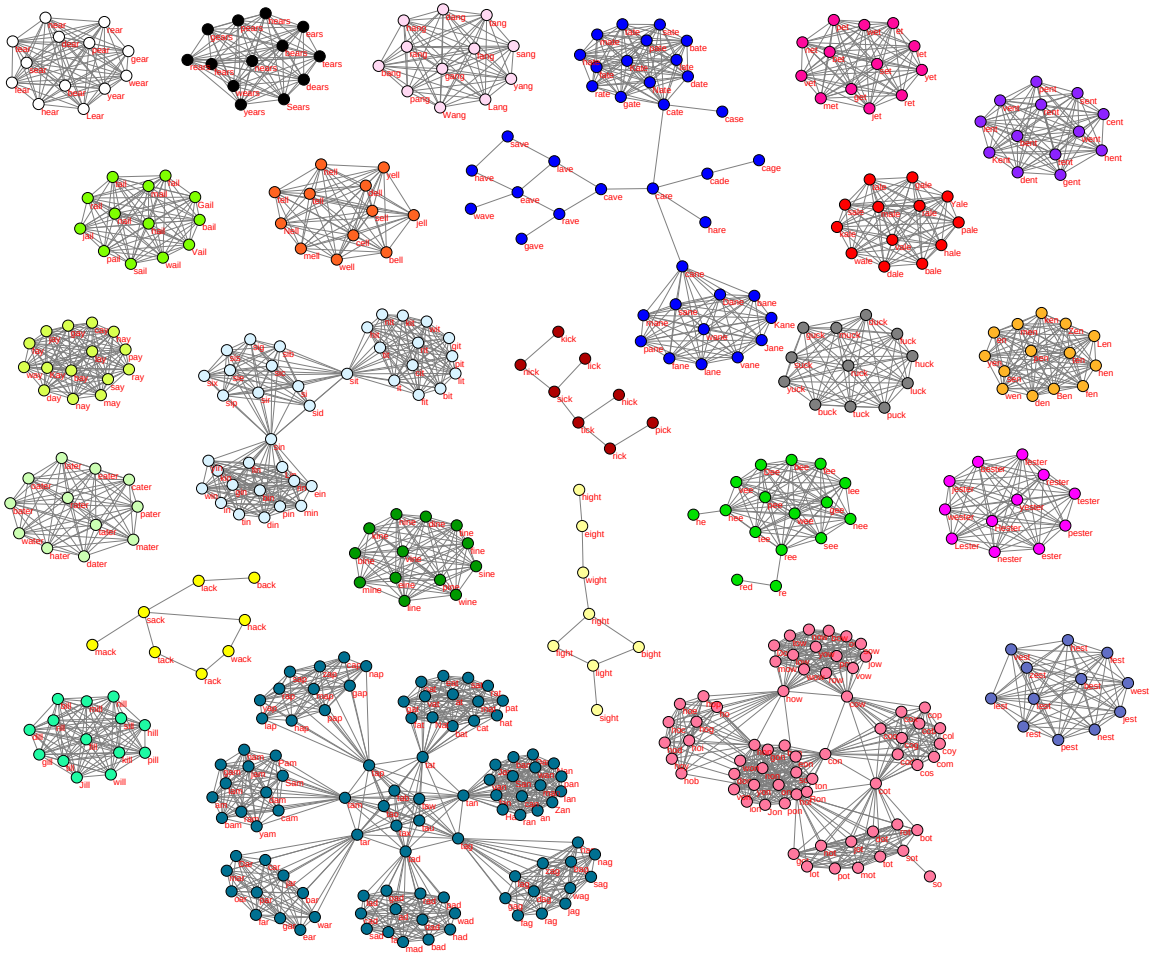
Example: The Knuth's English Dictionary

A large network can be generated from words of dictionary. Two words are connected using an undirected line if we can reach one from the other by

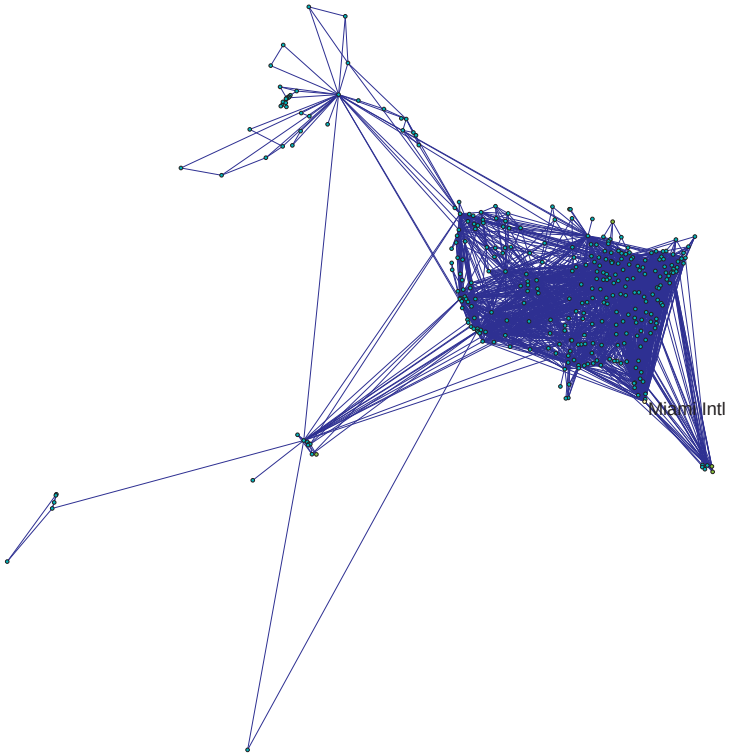
- changing a single character (e. g., work – word)
- adding / removing a single character (e. g., ever – fever).

Knuth's dictionary was used. There exist 52,652 words having 2 to 8 characters. The obtained network has 92,307 edges. The network is sparse: density is 0.0000666.

Selected clusters in Knuth's Dictionary



Airline connections among 332 US airports (332 vertices, 2116 lines)



Paths in a graph

In a *directed graph*:

- A sequence of vertices $(v_1, v_2, \ldots, v_k)$ is called ***a walk*** if

$$(v_i, v_{i+1}) \in A, i = 1 \ldots k-1$$

(consequent vertices must be connected with arcs).

- A sequence of vertices $(v_1, v_2, \ldots, v_k)$ is called ***a chain*** if

$$(v_i, v_{i+1}) \in A \text{ or } (v_{i+1}, v_i) \in A, i = 1 \ldots k-1$$

(consequent vertices must be connected, direction of lines is not important).

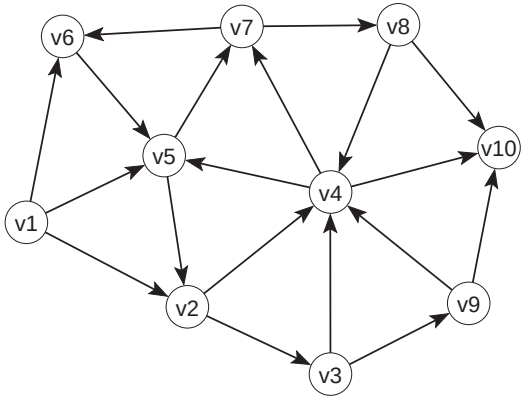
- A walk is ***elementary*** if all its vertices, except maybe initial and terminal, are different. We will call an elementary walk *a path*.
- A walk is ***simple*** if all its lines are different.
- If $v_1 = v_k$, the walk is called ***a closed walk*** or ***circuit*** (the walk starts and ends in the same vertex).
- ***A cycle*** is a closed walk of at least three vertices in which all lines are distinct

and all vertices except the initial and terminal are distinct.

- The chain starting and ending in the same vertex is called ***a closed chain***.
- If $v_1 = v_k$ and $k = 1$, the circuit is called ***a loop*** (a connection of a vertex to itself).
- ***The length of a path*** is $k - 1$ (the number of lines traveled).
- There can exist several paths between two vertices. The most interesting is ***the shortest path*** (all shortest paths).

  If a relation means *telling news* the shortest path will tell the shortest route of information traveling between two persons. – e. g., if all shortest paths from people to a selected person are short the selected person is well or quickly informed.

  In the case of a relation *was in contact to*, the shortest path tells the most probable way of passing an infection. In this case shortest paths are not welcome.
- The length of the longest shortest path in a graph is called its ***diameter***.



$v1 - v6 - v4 - v8$ is not even a chain

$v1 - v6 - v7 - v8$ is a chain, but not a walk

$v8 - v4 - v5 - v2 - v4 - v10$ is a simple, not elementary walk

$v8 - v4 - v5 - v2 - v3 - v9$ is an elementary walk (path)

$v4 - v10 - v8 - v4$ is a closed chain

$v6 - v5 - v7 - v6$ is a cycle

$v1 - v2 - v3 - v9 - v10$ is a path between $v1$ and $v10$, but not the shortest path (length 4).

The shortest path is: $v1 - v2 - v4 - v10$ (length 3).

The diameter is 5: $v7 - v6 - v5 - v2 - v3 - v9$

Some interesting results

Networks with large number of vertices usually have short shortest paths among vertices. For example:

- The average length of shortest path of the WWW, with over 800 million vertices, is around 19. Albert, R., Jeong, H., and Barabasi, A.-L. (1999): Diameter of the World-Wide Web. *Nature*, **401**, 130-131.
- Social networks (whom do you know) with over six billion individuals are believed to have a average length of shortest path around six. Milgram, S. (1967): The small-world problem. *Psychol. Today*, **2**, 60-67.

  Small world experiment: A psychologist Stanley Milgram made the following experiment with letters: The letter should reach a target person. The persons involved in experiment were asked to send the letter with these instructions to the target person (if they personally know him/her) or (if they do not know him/her personally) to their friend who was more likely to know the target. Letters were sent from Omaha (Nebraska) to target person in Boston (Massachusetts). The average length of the successful paths was 6.

Components

Three types of components will be defined: *strong*, *weak* and *biconnected*.

A subset of vertices in a network is called ***a strongly connected component*** if (taking directions of lines into account) from every vertex of the subset we can reach every other vertex belonging to the same subset.

If direction of lines is not important (where we consider the network to be undirected), such a subset is called ***a weakly connected component***.

Example: Let the vertices of network correspond to buildings in the city, and lines to streets that connect the buildings. Some streets are ordinary (undirected lines), some are one way only (directed lines).

All buildings that can be reached from one to another using the car, belong to the same strongly connected component (we must take one way streets into account).

All buildings that can be reached from one to another by walking, belong to the same weakly connected component (one way streets are no restriction).

Lets take the relation *whom would you invite to the party*. For all persons belonging to the same strongly connected component:

- everybody will (at least indirectly) invite everybody else from the same strongly connected component
- everybody will be (at least indirectly) invited by everybody else from the same strongly connected component

If the network is undirected, a strongly connected component os the same as a weakly connected one, so they are simply called connected components.

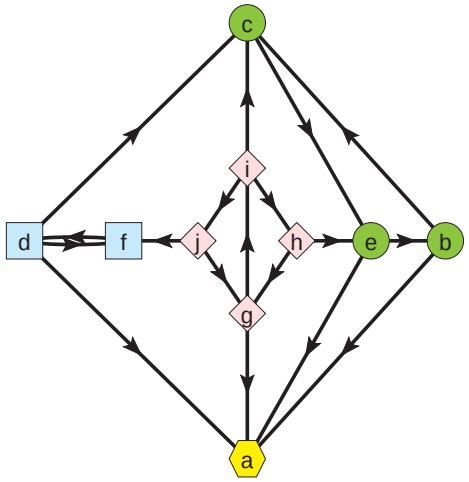
Connection between ***walks*** in a network and ***components***

Between any two vertices from the same strongly connected component there always exists a *walk*. We also say that the two vertices are strongly connected.

Between any two vertices from the same weakly connected component there always exists a *chain*. We also say that the two vertices are weakly connected.

A network is called ***strongly/weakly connected*** if every pair of vertices is strongly/weakly connected.

Example



The network is weakly connected.

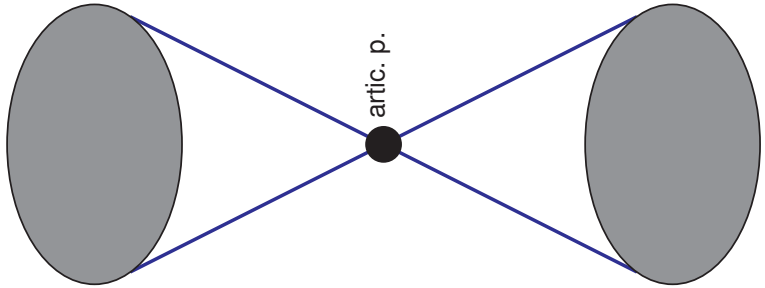
There exist 4 strongly connected components:

1. (a), 2. (b, c, e), 3. (d, f), 4. (g, h, i, j)

Vertices that belong to the same strongly connected components are drawn using the same shape.

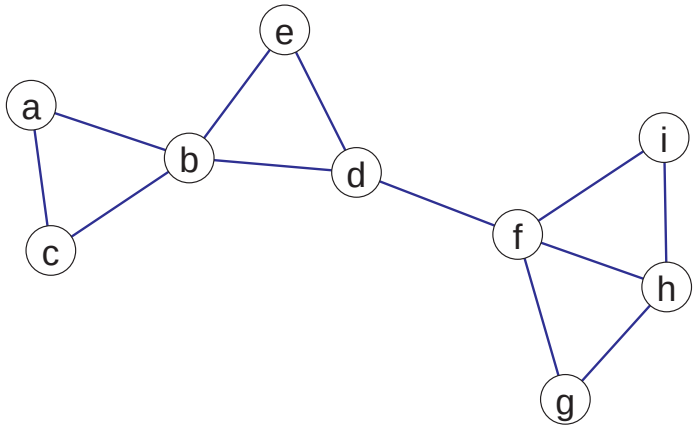
Biconnected components

Lets take the undirected connected network. Vertex a of the network is ***an articulation point*** of the network if there exist two other, different vertices v and w, so that every chain between the two vertices includes also vertex a. Simply saying: vertex a is articulation point, if removing the vertex from the network causes the network to become disconnected.



A network is called *biconnected* if for every triple of vertices a, v and w there exists a chain between v and w which does not include vertex a.

Example



Te network in the picture (aho1.net) consists of 4 biconnected components:

- (a, b, c)
- (b, d, e)
- (d, f)
- (f, g, h, i)

The articulation points are b, d and f.

Example: Airlines connections network

Cores

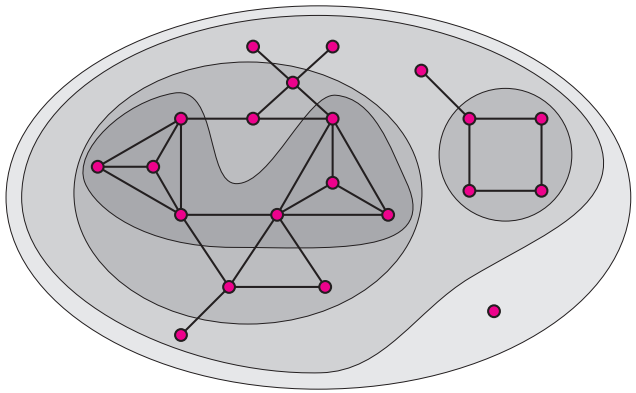
A subset of vertices is called a k**-core** if every vertex from the subset is connected to at least k vertices from the same subset.

Formally:
Let $N = (V, L)$, $L \subseteq V \times V$ be a network.

A maximal subgraph $H = (W, \cap W \times W)$ induced by the set W is a *k-core* iff $\forall v \in W : \deg_H(v) \geq k$.

Special example:
A subset of vertices is called ***a clique*** if every vertex from the subset is connected to every other vertex from the subset.

0, 1, 2 and 3 core:



Properties of cores:

- Cores are nested (see figure):

$$i < j \Longrightarrow H_j \subseteq H_i$$

- They are not necessarily connected subnetworks (see figure).
- Cores can be generalized to networks with values on lines.

Example: Airlines connections network

Community detection methods

Communities - dense clusters for which there are more lines inside than among clusters (values of lines are taken into account too).

In Pajek two community detection methods are available: *Louvain method* and *VOS Clustering*.

When applying Louvain method we search for partition into clusters with the highest value of *modularity (Q)*. Modularity is defined in the following way:

$$Q = \frac{1}{2m} \sum_s (e_s - r * \frac{K_s^2}{2m})$$

- m – total number of lines in network,
- s – cluster (community),
- $e_s = \sum_{ij \in s} A_{ij}$ – 2 times the number of lines in community s

- $K_s = \sum_{i \in s} k_i$ – sum of degrees in community s
- r – *resolution parameter*, default value 1 means modularity as originally defined

Similar method is *VOS Clustering*, where *VOS quality function* is taken into account instead of modularity.

By changing *resolution parameter* - r we can get larger or smaller communities. By default resolution parameter is set to 1. Setting r larger than 1 means searching for larger number of smaller communities. Setting r smaller than 1 means searching for smaller number of larger communities.

Examples: football.net, import.net.

Some more examples

Measures of centrality and prestige

One of the most important uses of network analysis is identification of 'most central' units in a network.

Measures of centrality and prestige can be defined in two different ways:

- for each unit respectively – **unit centrality** (one number for each unit)
- for the whole network – **network centralisation** (only one number for the whole network).

Unit centrality and prestige

When we talk about centrality we must distinguish between undirected and directed networks:

- the term **centrality measures** is used for **undirected** networks;
  Example: A city is *central*, if a lot of roads are passing through it.
- the term **prestige** is used for **directed** networks. In this case we can define two different types of prestige:
  - one for outgoing arcs (**measures of influence**),
  - one for incoming arcs (**measures of support**).

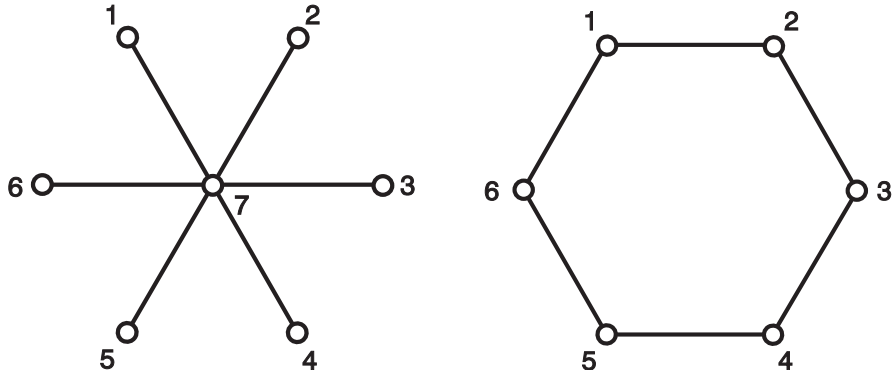
  Examples:
  An actor has high *influence*, if he gives commands to several other actors.
  An actor has high *support*, if a lot of people vote for him.

Unit centrality measures

Selected unit is central

- if it has high **degree**,
- if it is easily accessible (**close to**) all other units,
- if it lies on several geodesics (shortest paths) **between** other units.

Star and cycle



According to all criteria mentioned, unit 7 in the star is the most central, while all units in the cycle are equally central.

Degree Centrality

The simplest measure – unit is central in a network, if it is active enough in the sense that is has a lot of links to other units (unit 7 in the star). In the case of cycle all units are equally central.

Degree centrality is defined by a degree of unit x

$$c_D(x) = \text{degree of unit } x$$

Such measures are called **absolute measures of centrality**. Absolute measures cannot be used to compare centralities of networks with different number of units. Therefore such measures are normalised to get measure in interval from 0 to 1, where 0 means the smallest possible and value 1 the highest possible centrality. Measures normalised in this way are called **relative measures of centrality**.

Relative degree centrality is

$$C_D(x) = \frac{c_D(x)}{\text{highest degree}} = \frac{c_D(x)}{n-1}$$

if n is number of units in a network, the highest possible degree (in a network without loops) is $n-1$.

The same centrality measure can be used as a measure of *prestige for directed networks*. In this case there are two possibilities

- to choose (*influence* – out degree: number of arcs going out)
- to be chosen (*support* – in degree: number of arcs coming into).

Closeness Centrality

Sabidussi (1966) suggested the measure of centrality according to closeness of unit x:

$$c_C(x) = \frac{1}{\sum_{y \in U} d(x,y)}$$

where $d(x,y)$ is the graph theoretic distance (length of shortest path) between units x and y, U is set of all units.

If network is not strongly connected, we take only reachable vertices into account, but we weight the result with number of reachable vertices.

The most central units according to closeness centrality can quickly interact to all others because they are close to all others.

This measure is preferable to degree centrality, because it does not take into account only direct connections among units but also indirect connections.

Betweenness Centrality

In the case of communication networks the distance from other units is not the only important property of a unit. More important is which units lie on the shortest paths among pairs of other units. Such units have control over the flow of information in the network.

Idea of betweenness centrality measures: unit is central, if it lies on several shortest paths among other pairs of units.

Freeman (1977) defined the centrality measure of unit x according to betweenness in the following way:

$$c_B(x) = \sum_{y<z} \frac{\text{\# of shortest paths between } y \text{ and } z \text{ through unit } x}{\text{\# of shortest paths between } y \text{ and } z}$$

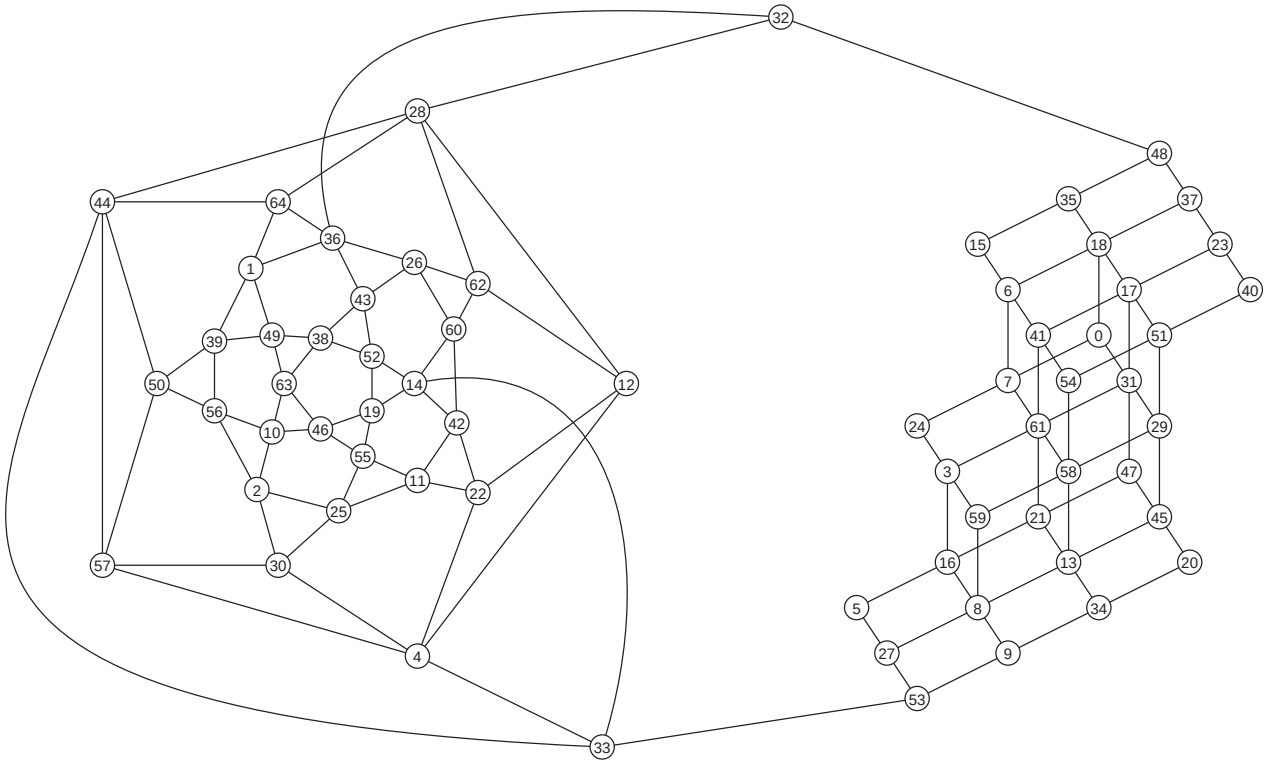
Suppose that communication in a network always passes through shortest available paths: Betweenness centrality of unit x is the sum of probabilities across all possible pairs of units, that the shortest path between y and z will pass through unit x.

Choosing suitable centrality measure

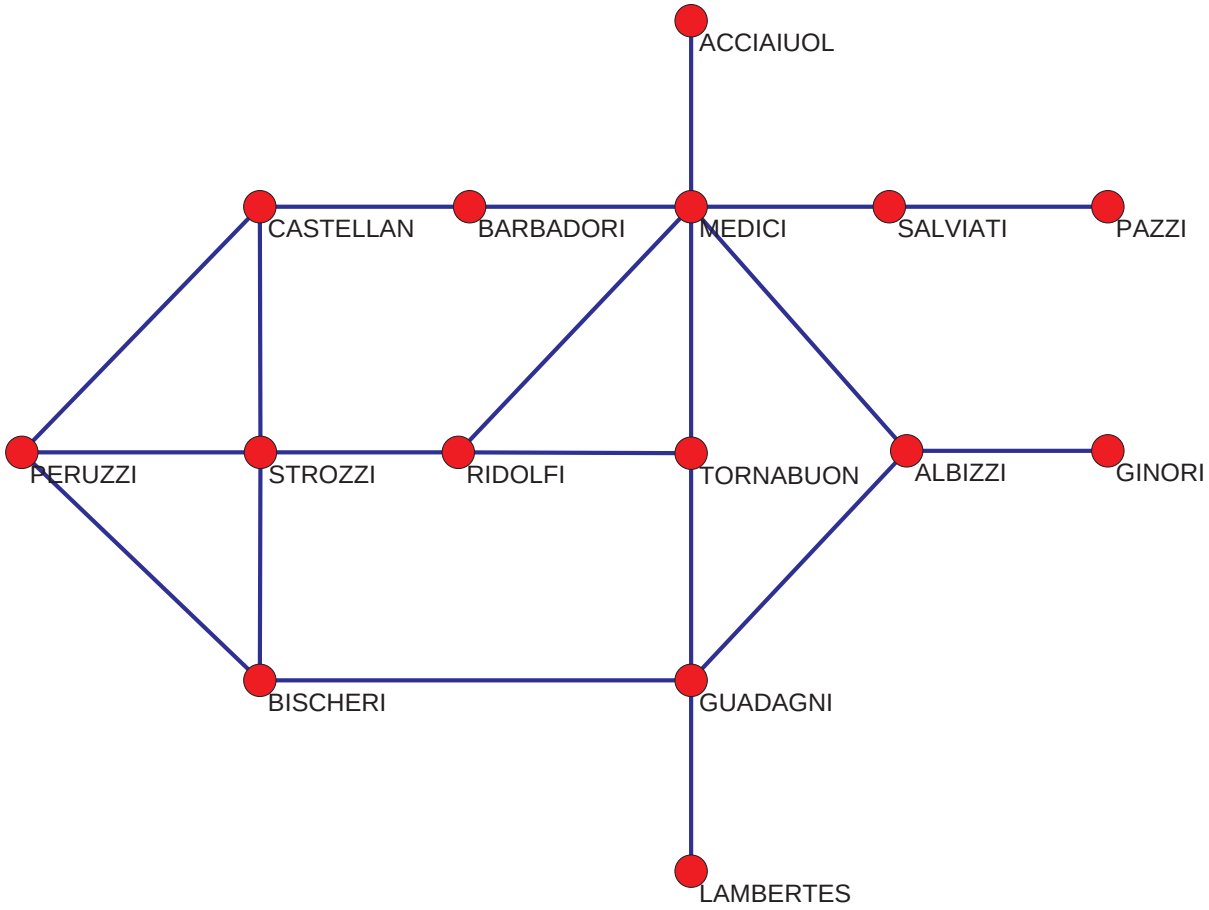
Different centrality measures can give quite different results. Therefore we must be very careful which centrality measure to choose for a given network:

It can happen, that some units have low degrees, but high betweenness centrality.

Example: units 33, 53, 32 in 48 are the most central according to betweenness centrality, although several other vertices with higher degree centrality exist (unit 61 has degree 6, some units have degree 5). The four units mentioned have degree only 3 or 4.


32
28
48
44
64
36
35
37
1
26
15
18
23
62
6
17
40
43
41
0
51
39
49
38
60
52
7
54
31
50
63
14
12
24
61
29
56
19
42
10
46
3
58
47
55
59
21
45
2
11
22
25
16
13
20
57
30
5
8
34
27
9
4
53
33

Example: Padgett's Florentine Families



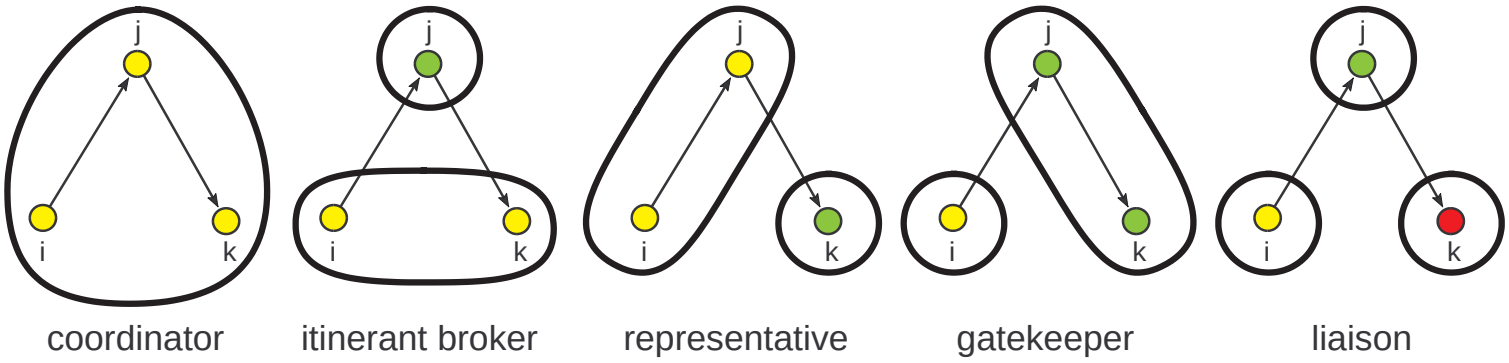
Relative centrality measures

No.	Family	C_D	C_C	C_B
1.	Acciaiuoli	0.071	0.368	0.000
2.	Albizzi	0.214	0.483	0.212
3.	Barbadori	0.143	0.438	0.093
4.	Bischeri	0.214	0.400	0.104
5.	Castellani	0.214	0.389	0.055
6.	Ginori	0.071	0.333	0.000
7.	Guadagni	0.286	0.467	0.255
8.	Lamberteschi	0.071	0.326	0.000
9.	Medici	*0.429*	*0.560*	*0.522*
10.	Pazzi	0.071	0.286	0.000
11.	Peruzzi	0.214	0.368	0.022
12.	Ridolfi	0.214	0.500	0.114
13.	Salviati	0.143	0.389	0.143
14.	Strozzi	0.286	0.438	0.103
15.	Tornabuoni	0.214	0.483	0.092

Brokerage roles

In a network in which groups are defined (e.g. men and women) each unit can be involved in the following brokerage roles:

- coordinator,
- itinerant broker,
- representative,
- gatekeeper,
- liaison.



Genealogies as large networks

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@
...
...
```

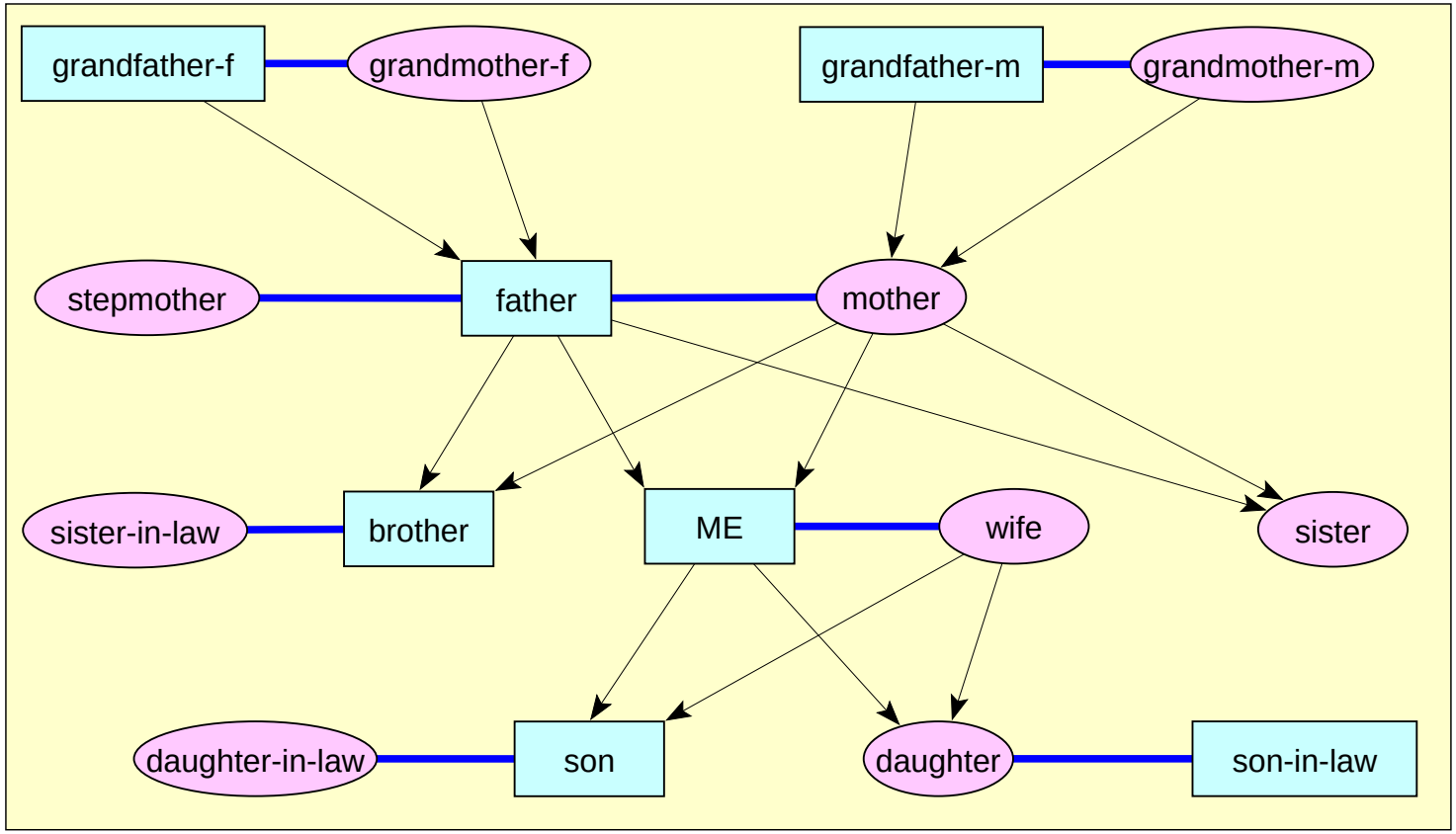
```
0 @I115@ INDI
1 NAME William Arthur Philip/Windsor/
1 TITL Prince
1 SEX M
1 BIRT
2 DATE 21 JUN 1982
2 PLAC St.Mary's Hospital, Paddington
1 CHR
2 DATE 4 AUG 1982
2 PLAC Music Room, Buckingham Palace
1 FAMC @F16@
...
0 @I116@ INDI
1 NAME Henry Charles Albert/Windsor/
1 TITL Prince
1 SEX M
1 BIRT
2 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

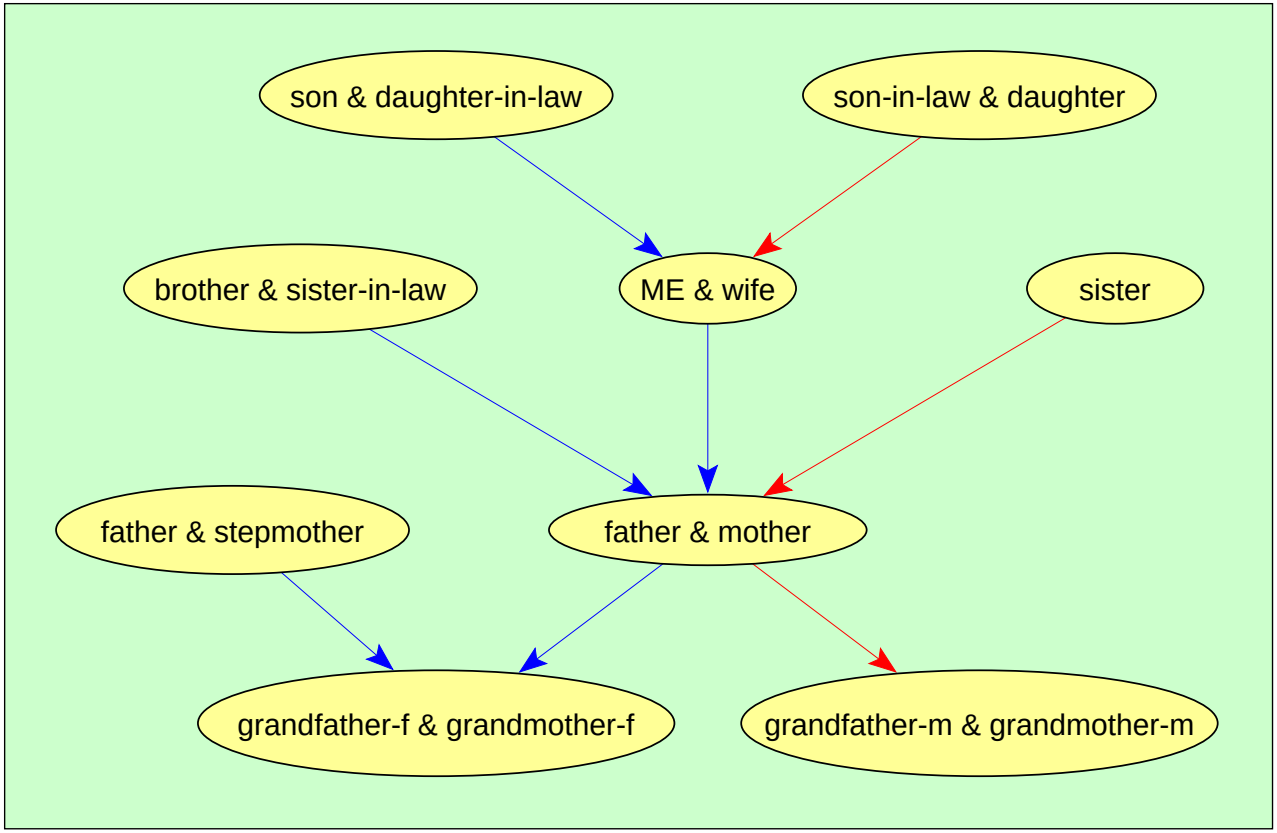
Genealogies can be represented as networks in different ways:

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

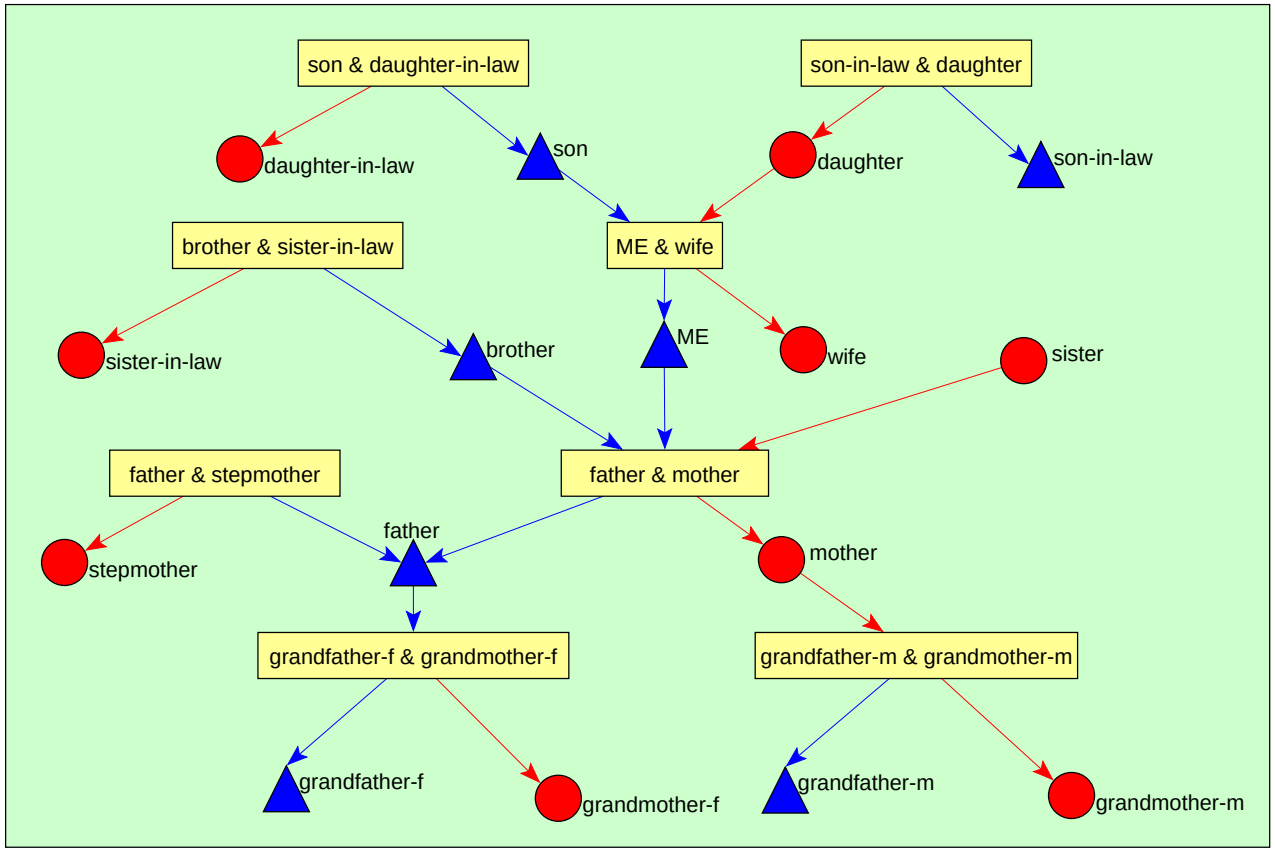
Ore-graph: In Ore-graph every person is represented by a vertex, marriages are represented with edges and relation *is a parent of* as arcs pointing from each of the parents to their children.



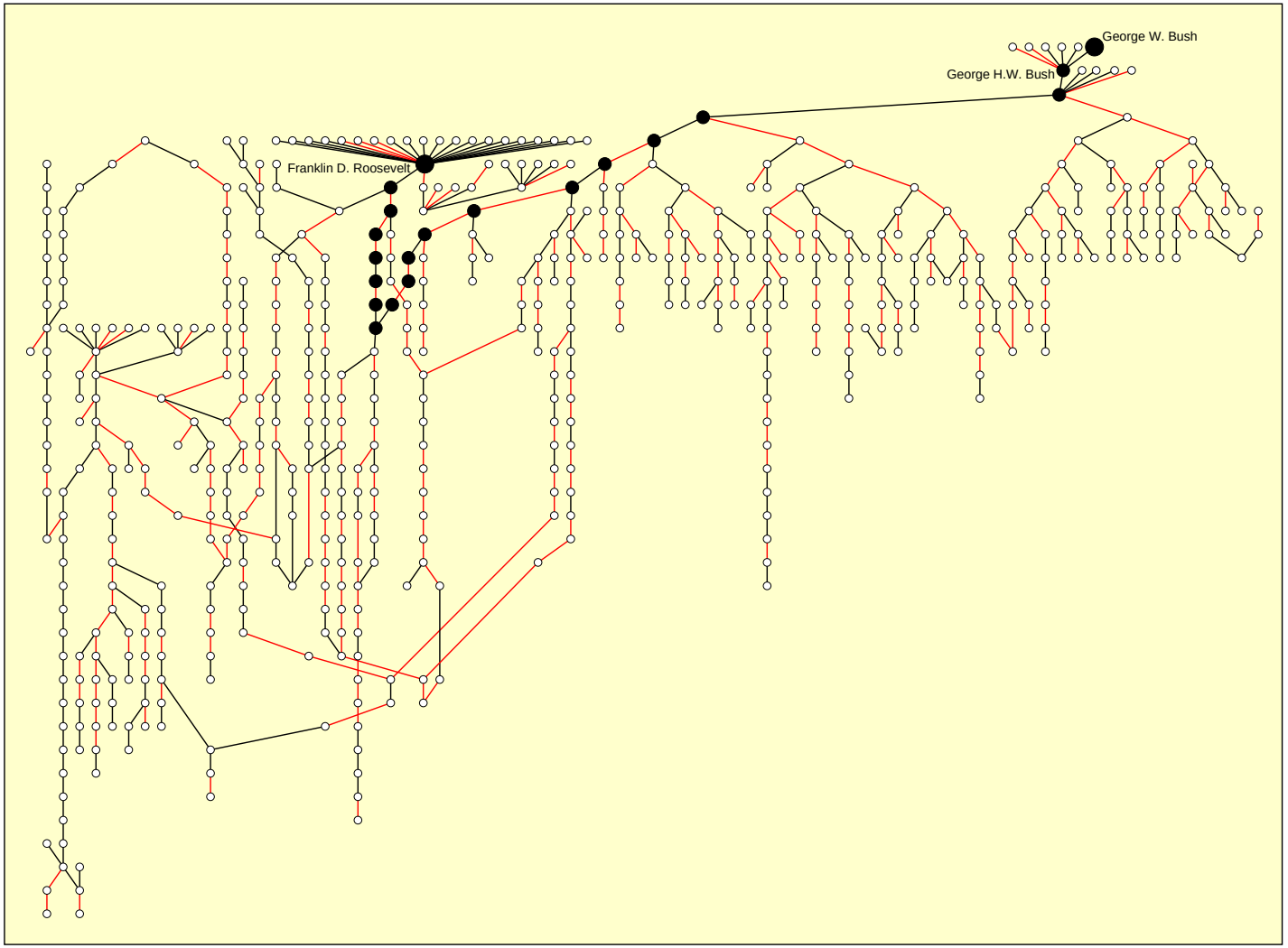
p-graph: In p-graph vertices represent individuals or couples. In the case that person is not married yet (s)he is represented by a vertex, otherwise person is represented with the partner in a common vertex. There are only arcs in p-graphs – they point from children to their parents.



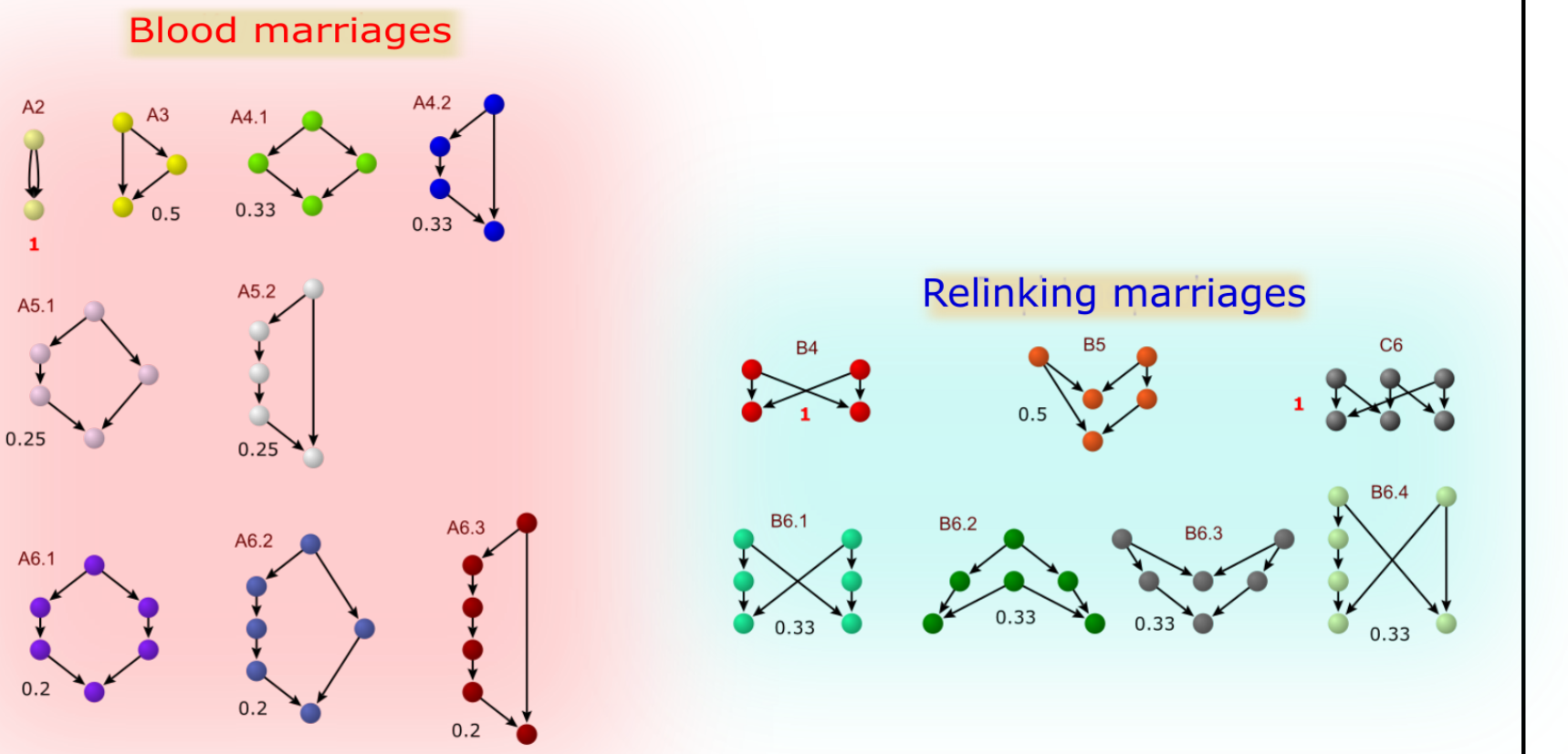
Bipartite p-graph: has two types of vertices – vertices representing couples (rectangles) and vertices representing individuals (circles for women and triangles for men). Arcs again point from children to their parents.



Largest component of genealogy of American presidents

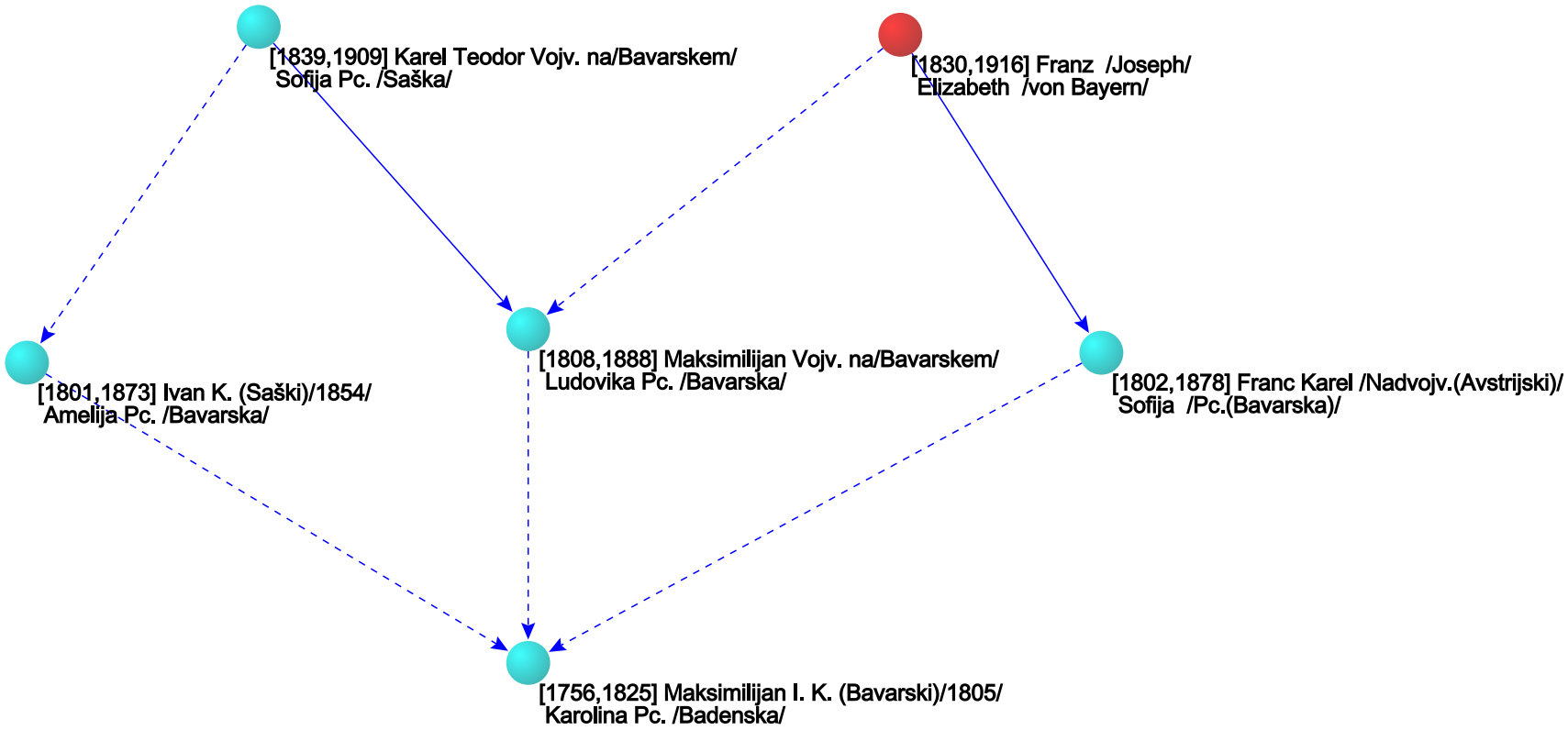


Relinking marriages (p-graphs with 2 up to 6 vertices)

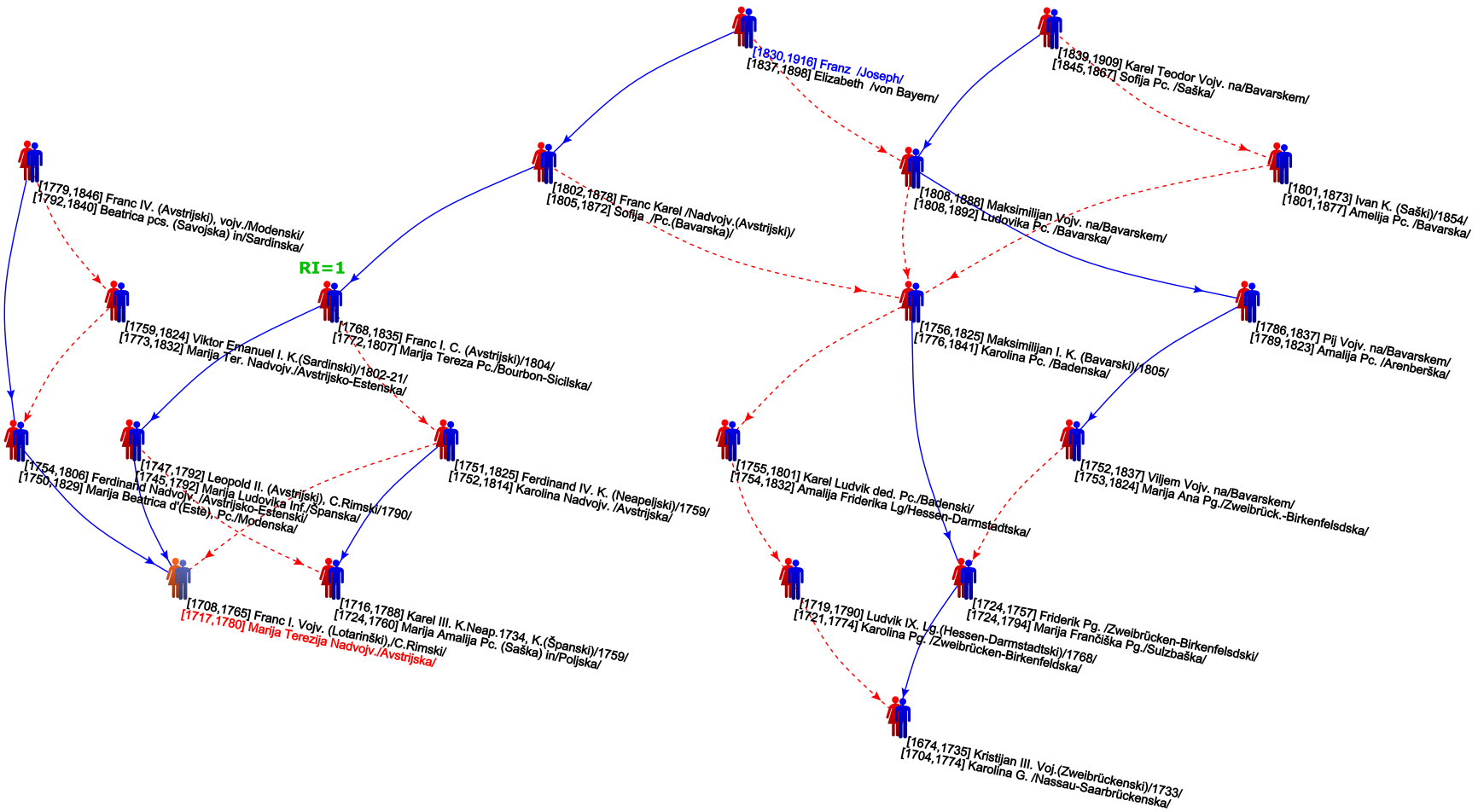


Relinking marrages in genealogy of European noble families

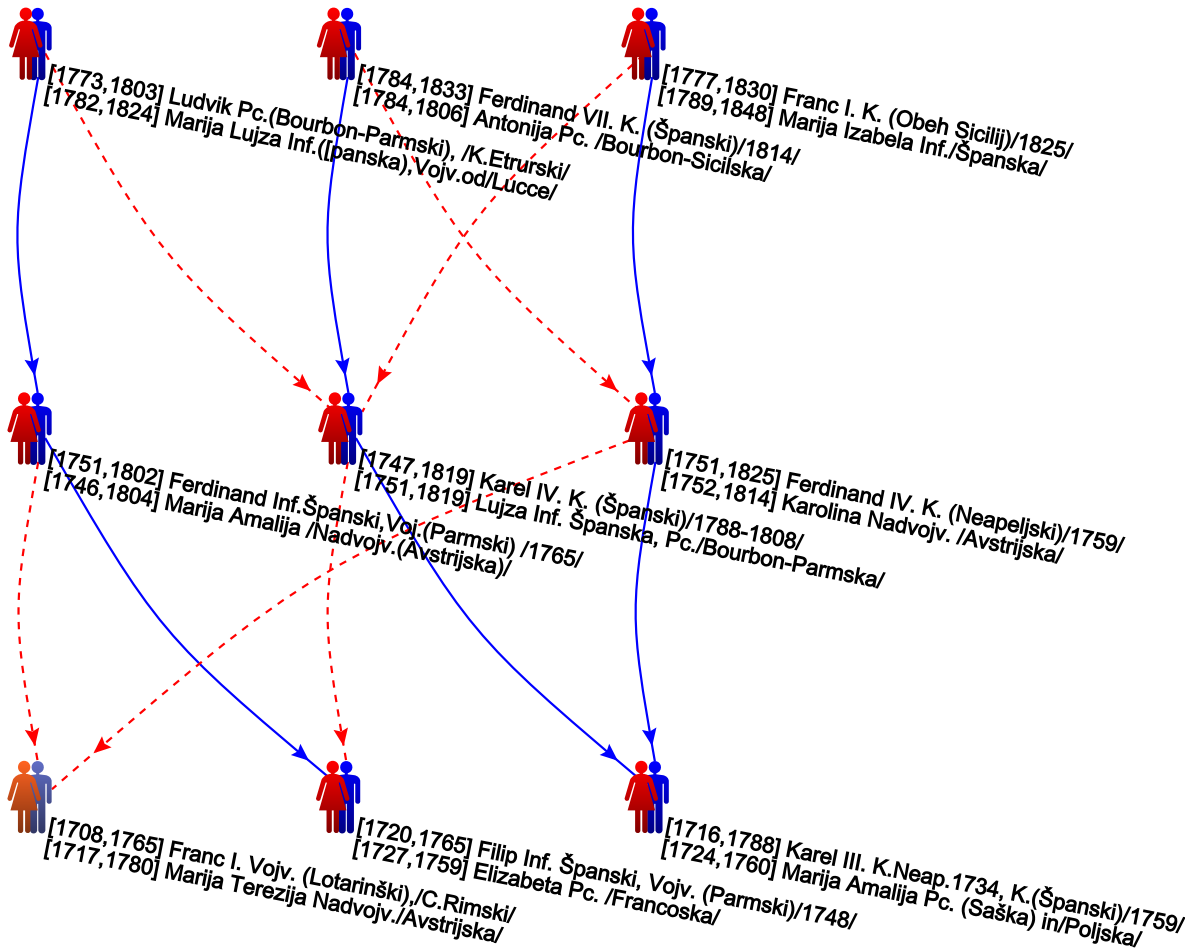
Genealogies of noble families are much more relinked than the 'usual' one. In 2016 we celebrated 100 years since Franc Jožef died. He was married to his cousin 'Sisi'.



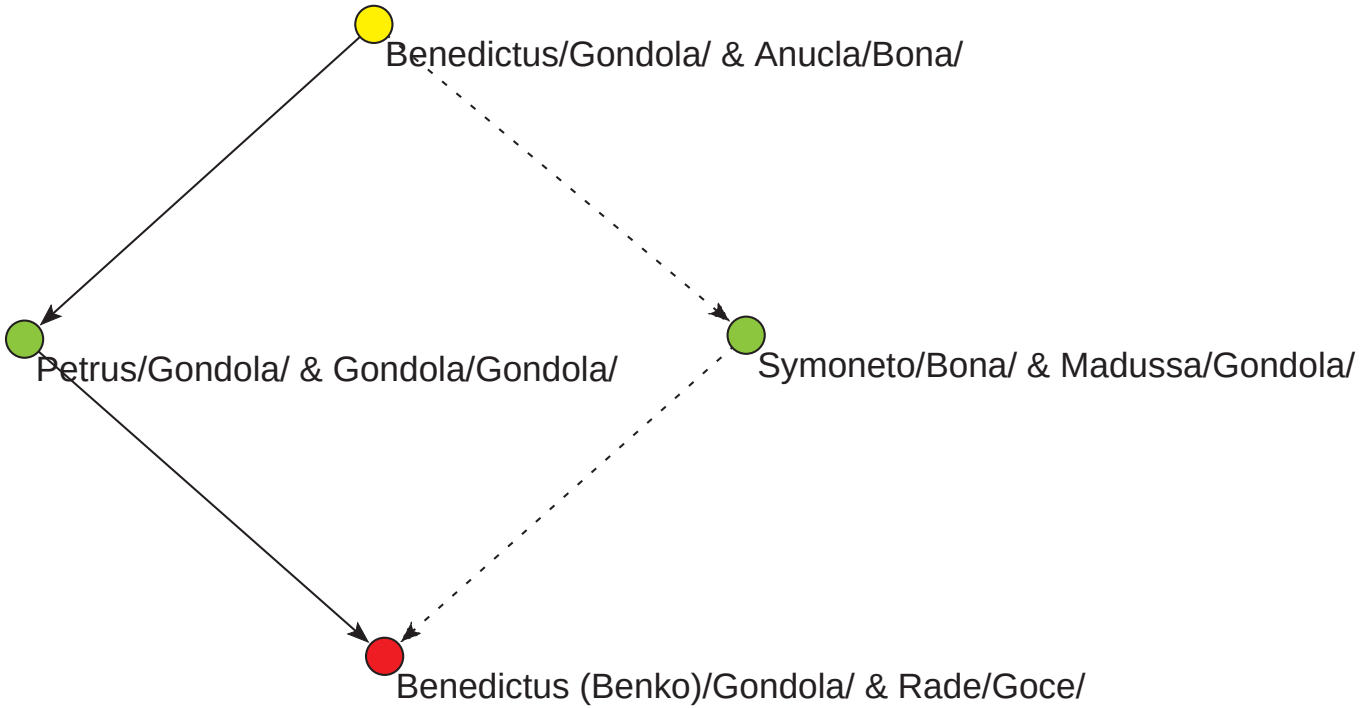
...More relinking marraiges



...More relinking marraiges



Blood marriage grandson-granddaughter in Genealogy of Ragusan noble families



Non-blood relinking marriages