

## 11 Genealogies and citations

### 11.1 Introduction

Time is responsible for a special kind of asymmetry in social relations, since it orders events and generations in an irreversible way. Social identity and position is partially founded on common ancestors, whether in a biological sense (birth) or in an intellectual manner: citations by scientists or references to predecessors by artists. This is social cohesion by common descent, which is slightly different from cohesion by direct ties. Social communities and intellectual traditions can be defined by a common set of ancestors, by structural relinking (families which intermarry repeatedly), or by long-lasting co-citation of papers.

Pedigree is also important for the retrospective attribution of prestige to ancestors. For example, in citation analysis the number of descendants (citations) is used to assign importance and influence to precursors. Genealogy is the basic frame of reference here, so we will discuss the analysis of genealogies first.

### 11.2 Example I: Genealogy of the Ragusan nobility

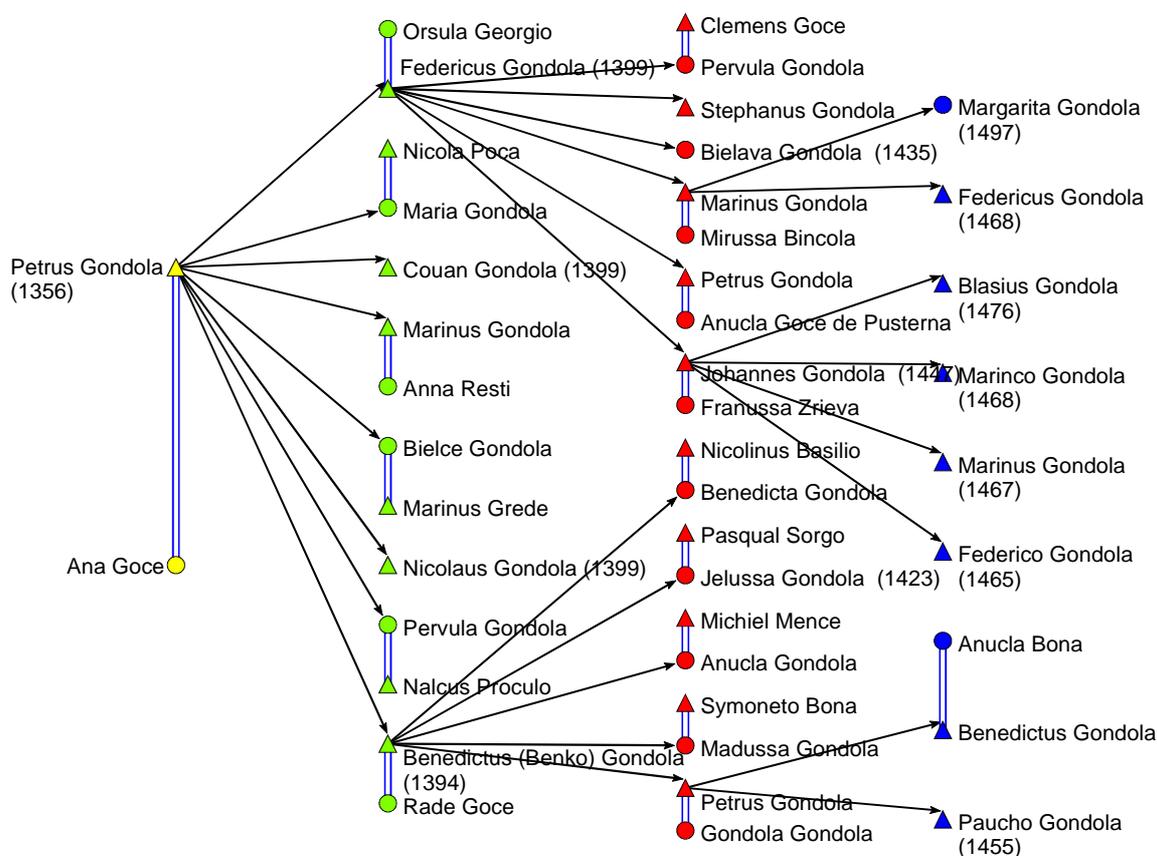
Ragusa, which is now known as Dubrovnik, was settled on the coast of the Adriatic Sea (Europe) in the 7th century. For a time, it was under Byzantine protection, becoming a free commune as early as the 12th century. Napoleon, having destroyed the Venetian Republic in 1797, put an end to the Republic of Ragusa in 1806. It came under Austrian control until the fall of the Austro-Hungarian monarchy in 1918.

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, the Small Council 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.

The Ragusan nobility evolved in the 12th century through the 14th century and was finally established by statute in 1332. After 1332, no new family was accepted until the large earthquake in 1667. A major problem facing the Ragusan noble families was that by decreases of their numbers and the lack of noble families in the neighboring areas, which were under Turkish control, they became more and more closely related – marriages between relatives in the 3rd and 4th remove were frequent. It is interesting to analyze how families of a privileged

social class organized their mutual relations by marriage and how they coped with the limited number of potential spouses for their children.

The file `Ragusan.ged` contains the members of the Ragusan nobility from the 12th to the 16th century, their kinship relations (parent-child), their marriages, and their (known) years of birth, marriage and death. Note that this is not an ordinary network file, since it contains attributes and relations of vertices. The extension `.ged` indicates that it is a GEDCOM-file, which is the standard format for genealogical data as we will explain in the next section. The genealogy is large, it contains 5999 persons. For illustrative purposes, we selected the descendants of one nobleman, Petrus Gondola, in the file `Gondola_Petrus.ged` (336 persons).



**Figure 1** - Four generations of descendants to Petrus Gondola (known years of birth between brackets).

### 11.3 Family trees

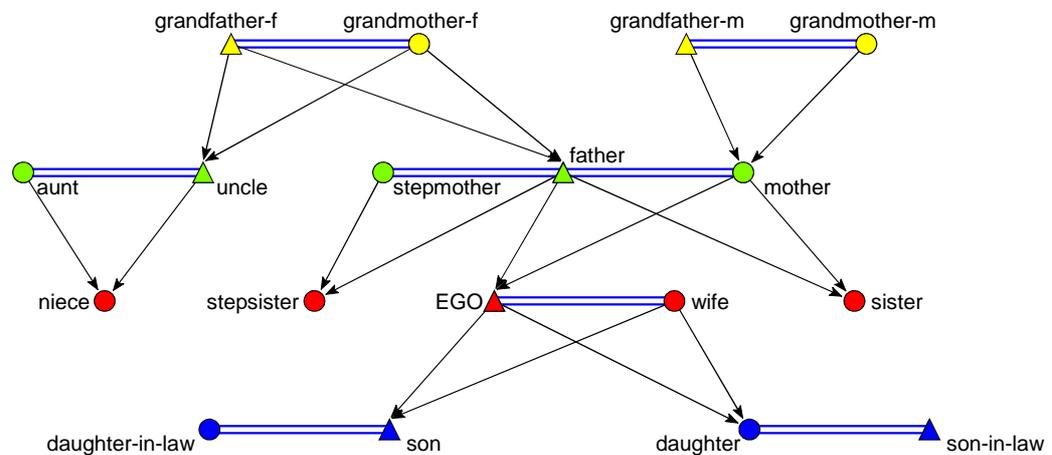
Across the world, many people are assembling their family trees. They visit archives to collect information about their ancestors in registers of births, deaths, and marriages. Since family names are the usual entries in these registers and family names are the father's surname in most Western societies, they reconstruct a patrilineal genealogy, in which father-child relations connect generations rather

than mother-child relations. In addition to father-child relations, marriages are included in the family tree.

Figure 1 shows a part of the Gondola family tree, which includes three generations of descendants to Petrus Gondola, who was born in 1356. Note that children born to a Gondola father are included because they receive the Gondola surname. Children of a Gondola mother are not included because their surname assigns them to another family in this historiography of a family name. An exception would be a Gondola mother who married a Gondola father but this does not occur among the descendants in Figure 1.

In principle, genealogies contain persons as units and two types of relations among persons: birth and marriage. A person may belong to two nuclear families: a family in which it is a child and a family in which it is a parent. The former family is called the **family of child or orientation** and the latter is **family of spouse or procreation**. Petrus Gondola's family of procreation, for example, contains his wife and eight children and it is identical to the family of orientation of each of his children. A husband and wife have the same family of procreation, but they have different families of orientation unless they are brother and sister.

The standard data format for genealogies (GEDCOM) uses the double coding according to family of orientation and family of procreation. In addition, it has facilities to store all sorts of information about the persons and events, e.g., about their marriage, so we advise to use this data format for the collection and storage of genealogical data. On the internet, excellent free software and several databases of genealogical data are available (see Section *Further Reading*).



**Figure 2** - Ore graph.

In a representation of a genealogy as a network, family codes are translated to arcs between parents and children. In the most common sociogram of kinship ties, which is known as the **Ore graph** (Figure 2), men are represented by triangles, women by ellipses, marriages by (double) lines, and parent-child relations by arcs. Note that the arcs point from parent to child following the flow of time.

In contrast to the family tree, fathers *and* mothers are connected to their children in an Ore graph. This greatly simplifies the calculation of kinship relations because the length and the direction of the shortest semipath between

two individuals defines their kinship relation, e.g., my grandparents are the vertices two steps ‘up’ from me in the Ore graph. They are relatives in the second remove because two births are included in this path. In a patrilineal family tree, relatives from my mother’s side, e.g., her parents and brother, are not included so it is impossible to establish my kinship relation with them. In the Ore graph, it is possible to distinguish between blood relations and marriage relations, so we may calculate the remove in a strict sense, that is, ignoring marital relations, or in a loose sense, including them and considering them relations with zero distance.

In the standard display of a kinship network, marriages and siblings are drawn at the same layer and layers are either top-down (Figure 2) or they are ordered from left to right (Figure 1). A layer contains a **genealogical generation**: grandparents versus parents, uncles and aunts versus children, nieces, and nephews. Such are the generations that we experience during our lives. From a social point of view, however, we define generations as birth cohorts, e.g., the generation of 1945-1960. In contemporary Western societies, **social generations** contain people who were born in a period of approximately fifteen years. Genealogical generations overlap with social generations to a limited extent. For four or more generations, genealogical generations may group people of very different ages as a result of early marriage and childbearing in one branch of the family and late marriage in another branch. The ages of the great-grandchildren of Petrus Gondola, for instance, range from 1455 (Paucho) to 1497 (Margarita). Biologically, the former could have been the latter’s grandfather.

The Ore graph is a very useful instrument for finding an individual’s ancestors (pedigree) and descendants both from the father’s side and the mother’s side. In addition, it is easy to count siblings and to trace the closest common ancestor of two individuals. This allows us to assign people to **descent groups**, which are groups of people who have a common ancestor among the people who are alive at a particular moment.

### Application

*Options>Read/Write  
>GEDCOM - Pgraph*

Genealogical data in GEDCOM format can be read directly by Pajek. To obtain the Ore graph, make sure that the option *GEDCOM - Pgraph* in the

*Options>Read/Write  
>Ore: 1-Male, 2-Female  
links*

*Options>Read/Write* submenu is **not** selected before you open the GEDCOM file. Then, open a GEDCOM file in the usual way with the *File>Network>Read* command, but select the option *Gedcom files (\*.ged)* in the *File Type* drop list of the Read dialog screen. When you check the option *Ore: 1-Male, 2-Female links*, father-child relations have line value one and mother-child relations have value two. This is particularly useful if you want to extract patrilineal relations from the Ore graph.

*Info>Vector*

Reading the GEDCOM file, Pajek translates family numbers to parent-child relations and it creates a partition and three vectors. The partition identifies vertices which are brothers and sisters, that is, children born to the same father and mother. Stepbrothers and stepsisters from a parent’s remarriage are grouped separately. The vectors contain the years of birth, marriage, and death of the

people in the network. Unknown dates are represented by vector value 999998. You may inspect the dates with the *Info>Vector* procedure in the usual way.

*Operations>Extract*  
*>to Gedcom*

A GEDCOM file contains several relations and attributes, including dates, so we advise to read data directly from these files. When you want to save an isolated branch from a genealogy in GEDCOM format, you can use the *to Gedcom* command in the *Operations>Extract* submenu. This command saves one or more classes of vertices, which you must define in a (weak) components partition first, from an Ore graph as a new GEDCOM file. Note that this command only creates a valid GEDCOM file if the subnetwork is isolated from the part of the genealogy which is not saved.

*Net>Partitions>Depth*  
*>Genealogical*

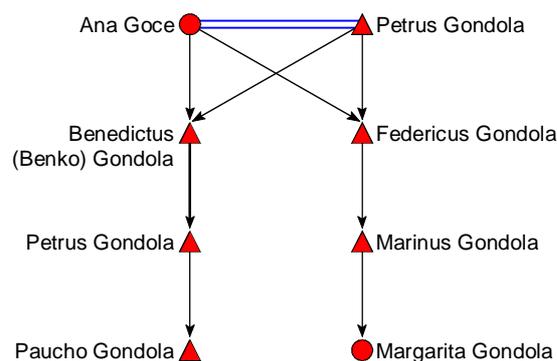
The genealogical generations of the Ore graph can be obtained with the command *Genealogical* from the *Net>Partitions>Depth* submenu. An acyclic depth partition is not possible because the marriage edges are cyclic: a husband is married to his wife and a wife is married to her husband at the same time. Draw the network in layers according to the genealogical depth partition and optimize it in the usual way. To focus on the distinct branches in the genealogy rather than the vertices, use the *Averaging x coordinate* procedure from the *Layers* menu. Usually, the *Forward* option works fine.

*Layers>Averaging x*  
*coordinate*

*Net>Transform>Remove*  
*>all edges*

The length of the geodesic (shortest path) in a symmetrized Ore graph is the remove or degree of a family relation. First, decide whether you want to include marital relations in the calculation. If not, remove the edges from the network (*Net>Transform>Remove>all edges*). Then, symmetrize the Ore graph and use the *Paths between vertices>All Shortest* command to obtain the geodesics between two individuals in the network. When asked, do not ignore (forget) the values of the lines, because a marriage link should not contribute to the length of the semipath, hence to the remove of the relation. The length of the shortest paths, which is the distance between the vertices, is printed in the Report screen. Among the descendants of Petrus Gondola (Figure 1), for instance, Paucho Gondola (1455) is a relative of Margarita Gondola (1497) in the sixth remove.

*Paths between vertices*  
*>All Shortest*



**Figure 3** - Shortest paths between Paucho and Margarita Gondola.

Pajek creates a new network of the geodesics it has found and a partition which identifies the vertices on the geodesics in the original network provided that you requested this in one of the dialog boxes. If we extract these vertices from the original directed network, we obtain Figure 3. It is easy to see that Petrus Gondola and his wife Anna Goce are the closest common ancestors of Paucho

and Margarita. In Figure 1, we can easily check this visually, but we need the shortest paths procedure in large networks such as the genealogy of the entire Ragusan nobility, which are too complicated to analyze by eye-balling methods. Note, however, that the computer may need quite some time to find longer paths in large genealogies.

*Net>k-Neighbours*

The ancestors (pedigree) or descendants of a person are easily found with the *k-Neighbours* procedure in the Ore graph. Ancestors are connected by paths towards an individual, so they are input neighbors of the individual. Descendants are reachable from the individual: they are output neighbors in the Ore graph. You may restrict the selection of ancestors to a limited number of generations in the *Maximal distance* dialog box of the *k-Neighbours* procedure. Note that the number of generations that you select is one more than the largest distance that you specify because the selected person, who also represents a generation, is placed in class zero.

In research of kinship relations, it is interesting to focus on the people who are alive at a particular moment. It is, for example, interesting to know which people are connected by kinship ties through living people because living family members may pass on information and they may organize events at which the family meets. The people who are alive at a particular moment can be identified by their dates of birth and death: select all individuals who were born but did not pass away before a particular moment. Note that this procedure requires full information about the date of birth and death of the persons in the genealogical network.

*Partitions>Intersection*

First, translate the vectors with years of birth or death into partitions by truncating them. Then binarize each partition such that all people born between year one (assuming we have no people born before the start of our era) and the chosen year are in class one of the binarized birth partition. In the death partition, class one must contain all people who died in this year or later (use Pajek's missing value code 999998 or 999999 as an upper limit). Now, you can obtain the intersection of both partitions with the *Intersection* command in the *Partitions* menu, provided that you select the binarized partitions as *First* and *Second Partition* in this menu. The *Intersection* of two binary partitions assigns vertices which are selected (class one) in both partitions to the first class of a new partition. With the intersection partition, you can extract the people who are alive at the chosen moment from the Ore graph. In the extracted network, weak components are descent groups or clusters of descent groups connected by marriages.

#### 11.4 Social research on genealogies

Kinship is a fundamental social relation, which is extensively studied by anthropologists and historians. In contrast to people who assemble their private family trees, social scientists are primarily interested in the genealogies of entire communities, such as the nobility of Ragusa.

These genealogies, which are usually very large, enable the study of overall patterns of kinship relations which, for instance, reflect cultural norms for marriage: who are allowed to marry? Property is handed over from one generation to the next along family lines, so marriages may serve to protect or enlarge the wealth of a family; family ties parallel economic exchange. Demographic data on birth, marriage and death reflect economic and ecological conditions, e.g., a famine or deadly disease causes high mortality rates.

The number of marriages and the age of the marital couple, the size of sibling groups, nuclear families, or extended families are determined and compared across different societies or different periods. Differences are related to external conditions and internal systems of norms or rules.

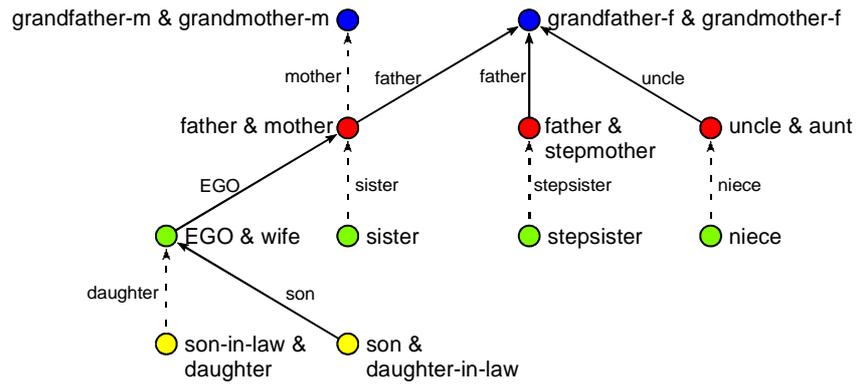
Table 1 compares the number of children of Ragusan noblemen across two periods: men born in 1200-1250 and 1300-1350. Unfortunately, many birth dates are unknown, so we added the parents' children and the children's in-laws from the kinship network assuming that they will belong to the same generation. In the Ore graph, the simple outdegree of a vertex specifies the number of children of a person. Table 1 summarizes the output degree frequencies. In the first half of the 14th century, a large proportion of the noblemen had no children in comparison to the previous century. Perhaps, less men got married because no new families were admitted to the nobility as of 1332. On the other hand, some men may have died young as a consequence of the black death epidemic which struck the town in 1348.

**Table 1** - Size of sibling groups\* in 1200-1250 and 1300-1350.

Size of sibling group	1200 - 1250		1300 - 1350	
0 (no children)	10	9.1%	298	42.1%
1	23	20.9%	99	14.0%
2	20	18.2%	73	10.3%
3	17	15.5%	69	9.7%
4	11	10.0%	52	7.3%
5	10	9.1%	35	4.9%
6 - 10	19	17.3%	79	11.2%
11 - 21	-	-	3	0.4%
Total (# sibling groups)	110	100%	708	100%

\* number of children from one father.

This type of research may use network analysis but it can also be done by database counts, for instance, calculations on a GEDCOM genealogy database. A second type of research, however, is inherently relational and must use network analysis as a tool. It focuses on structural relinking between families and the economic, social, and cultural reasons or rules for structural relinking. **Structural relinking** refers to the phenomenon that families intermarry more than once in the course of time. Intermarriage or **endogamy** is an indicator of social cohesion within a genealogy. If families are linked by more kinship ties, they are more likely to act as a clan: sharing cultural norms, entertaining tight relations, and restricting ties to families outside the clan.

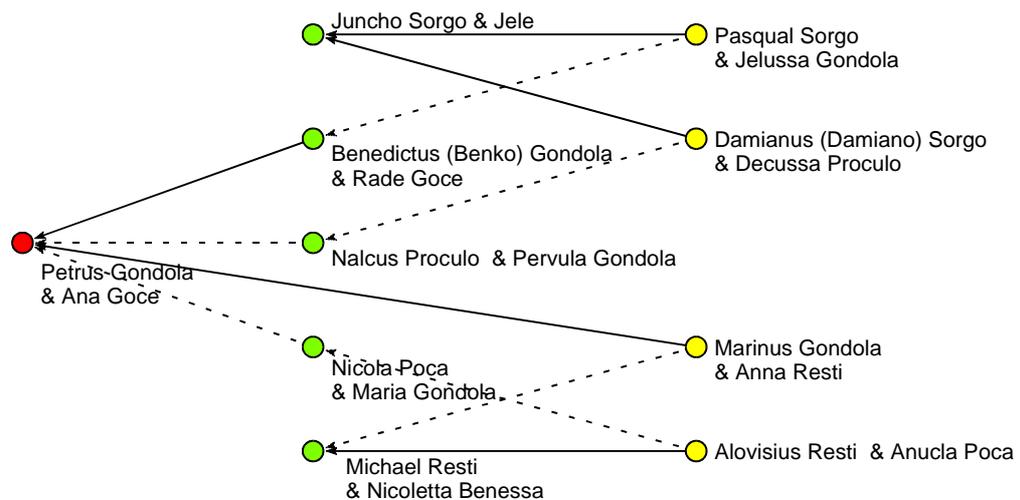


**Figure 4** - P-graph.

A **blood-marriage** is a special kind of structural relinking, namely the marriage of people with a close common ancestor, e.g., a marriage between brother and sister or between a granddaughter and a grandson. The occurrence of this type of relinking tells us which types of intermarriages are culturally allowed and which are not.

Structural relinking is best investigated in a special kind of genealogical network: the parentage graph or P-graph. In the **P-graph**, couples and unmarried individuals are the vertices and arcs point from children to parents. The type of arc shows whether the descendant is male (full arc) or female (dotted arc). In Figure 4, for instance, my son and his wife are connected by a full arc to me and my spouse; my daughter and her husband are connected by a dotted arc.

The P-graph has several advantages. It contains fewer vertices but the path distance in a symmetrized P-graph still shows the remove of a relation, although it is not possible to exclude marital relations from the calculation. The main advantage of the P-graph, however, is the fact that it is acyclic. There are no edges between married people, so every semicycle and bi-component indicates relinking, which is either a blood-marriage or another type of relinking.

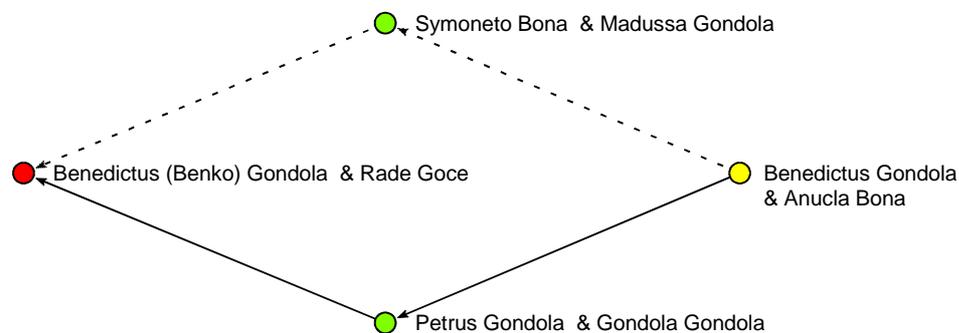


**Figure 5** - Relinking between different families.

Non-blood relinking often serves economic goals, namely to keep the wealth and power within selected families. Figure 5 shows non-blood marriages between

children and grandchildren of Petrus Gondola: two granddaughters marry brothers from the Sorgo family (Pasqual and Damianus), which is acknowledged to be the most influential family among the Ragusan nobility. Furthermore, a son and a granddaughter marry into the family of Michael Resti, which causes a generation jump. It is impossible to draw this network with all siblings and married couples in one layer because Marinus Gondola is the brother-in-law of Alovsius and his uncle at the same time.

Relinking within a family (blood-marriage) did also occur. A grandson of Benko Gondola, who is a son of Petrus Gondola, married a granddaughter, who was a relative in the fourth degree (see Figure 6). Blood marriages between closer relatives - a son who married a daughter, a child who married a grandchild - do not occur among the Ragusan nobility. Apparently, these marriages were not allowed.



**Figure 6** - Relinking within one family.

The amount of relinking in a P-graph is measured by the relinking index. In order to understand this index, we must introduce the concept of a tree in graph theory: a connected graph which does not contain semicycles. A tree has several interesting properties but for our purposes the fact that it does not contain cycles and semicycles is most important.

A **tree** is a connected graph which does not contain semicycles.

In a P-graph, every semicycle indicates structural relinking because the people or couples on the semicycle are linked by (at least) two chains of family ties, e.g., common grandparents on the father's side and on the mother's side. As a consequence, a P-graph which is a tree or a set of distinct trees (a **forest**) has no relinking and its relinking index is zero. Given the number of people and the assumption that a marriage links exactly one man and one woman, the maximum amount of relinking within the P-graph of a genealogy can be computed, so the actual number of relinking can be expressed as a proportion of this maximum. This is the **relinking index**, which is one in a genealogy with maximum relinking and it is zero in a genealogy without relinking.

We advise to calculate the relinking index on bi-components within the P-graph rather than on the entire P-graph. Genealogies have no natural borders; kinship ties extend beyond the boundaries of the data collected by the researcher, but boundary setting is important to the result of the relinking index. The largest

bi-component within a genealogy is a sensible boundary because it demarcates families which are integrated into a system by at least one instance of relinking. In general, structural relinking may be used to bound the field of study, which means that you limit your analyses to the families within the largest bi-component of a genealogy.

Let us calculate the amount of structural relinking among the Ragusan nobility in the period 1200-1350, in which new families were admitted to the nobility, and 1350-1500 when the nobility was chartered and no new families were admitted. Because of lacking birth dates, we add the parents' children and children's in-laws to the couples in which at least one spouse is known to be born in the selected period. Between 1200 and 1350, a small number of the couples (128 out of 1383 vertices or 9.3 percent) were connected by two or more family ties, so the relinking index is low for the network in this period (0.02). Within this bi-component, the relinking index is higher (0.25), so there is a small core of families, the Sorgo family among them, which are tightly related by intermarriages. In the period 1350-1500, the bi-component is larger, containing 476 couples (23.8 percent) and featuring many members of the Goce, Bodacia, and Sorgo families. The relinking index of the entire network is 0.20 and the proportion of relinking is 0.69 within the bi-component. Both values are much larger than in the period before 1350, which shows increased endogamy among the Ragusan nobility.

In the P-graph, each person is represented by one arc except in the case of a remarriage. Since each marriage is a separate vertex, e.g., my father and mother or my father and stepmother in Figure 4, men and women who remarry are represented by two or more arcs. In the P-graph, it is impossible to distinguish between a married uncle and a remarriage of a father or between stepsisters and nieces. This problem is solved in the bipartite P-graph, which has vertices for individuals and vertices for married couples. The bipartite P-graph, however, has the drawback of containing considerably more vertices and lines than the P-graph and path distance does not correspond to the remove of a kinship relation. We will not use bipartite P-graphs in this book.

### Application

*Options>Read/Write*  
*>GEDCOM-Pgraph,*  
*Bipartite graph,*  
*Pgraph+labels*

The format of a genealogy which is read from a GEDCOM data file depends on the options checked in the *Options>Read/Write* menu. As we noted before, Pajek transforms a GEDCOM data file into an Ore-graph if the option *GEDCOM-Pgraph* is **not** checked and a regular P-graph is created if this option is checked but the option *Bipartite Pgraph* is not. If the option *Pgraph+labels* is also checked, the name of a person is used as the label of an arc. Pajek does not create a brothers and sisters partition in conjunction with a Pgraph. It stores the years of birth of men and women in separate vectors because a couple has two birth dates. This also applies to the years of death.

*Info>Partition*

The Ore graph is most suited for finding brothers and sisters and count the size of sibling groups in a genealogical network. Pajek automatically creates a

brothers/sisters partition, which identifies children of the same parental couple. Each class is a sibling group, so the number of vertices within a brothers and sisters class represents the size of a sibling group. Unfortunately, it is not easy to obtain a frequency distribution of the size of sibling groups from this partition because the *Info>Partition* command lists each sibling group (class) separately.

It is possible to obtain a frequency distribution of the size of sibling groups which have the same father or the same mother. In the Ore graph, the outdegree of a vertex is equal to its number of children provided that marriage lines are disregarded. Ideally, every child has a father and a mother in the genealogical network, so we may count the number of children for each father or for each mother. In the case of a single marriage, the father and mother have the same number of children but these numbers may differ in the case of remarriages. In the example (Figure 2), my father remarried: he has three children (my stepsister, sister, and me) whereas my mother has only two children (my sister and me). We must look at the outdegree of fathers or mothers, not to both at the same time.

*Net>Partitions>Degree  
>Output*

*Net>Partitions  
>Vertex Shapes*

*Partitions> Extract Second  
from First*

This is achieved in the following way. First, remove the marriage lines (*Net>Transform>Remove>all edges*) from the Ore graph. Now, the outdegree of a vertex is equal to an actor's number of children. Then, create an outdegree partition with the *Net>Partition>Degree>Output* command and select it as the first partition in the *Partitions* menu. Next, create a partition on vertex shape (*Net>Partitions>Vertex Shapes*). Recall that men are represented by triangles and women by ellipses in the re graph. In the vertex shape partition, one class contains the men and another contains the women. Draw the network with this partition in order to find out which class represents the men or the women. Finally, select the partition according to vertex shape as the second partition in the *Partitions* menu and execute the command *Extract Second from First*. Choose the vertex shapes class which contains the gender that you want to select and Pajek will create a new partition containing the outdegree of the selected vertices. The *Info>Partition* command will produce the desired frequency tabulation. Maybe, you want to correct this tabulation for people who cannot have children (yet) in your genealogy.

*Partition>Binarize*

A birth cohort can be identified with the vector containing years of birth. Translate the vector into a partition by truncation, and binarize the partition according to the period you want to select. When you want to select the people who were born between 1350 and 1500, specify 1350 as the lowest class and 1500 as the highest class in the dialog boxes which appear on execution of the *Partition>Binarize* command.

*Partitions>Add Partitions*

In the case of a P-graph, the situation is more complicated because both spouses have a year of birth. You must create separate binary partitions for the men and the women born in the selected period. In the P-graph of the Ragusan nobility (*Ragusan.ged*), for instance, 1488 men and 396 women are known to be born between 1350 and 1500. If we combine the two binarized birth partitions by means of the *Partitions>Add Partitions* command, we find 1732 bachelors or couples with one spouse (known to be) born in the period (class one) and 76

couples of man and wife who were both born in the selected period (class two); the birth dates of 2568 couples and bachelors are either unknown or completely outside the selected period (class zero). If we are satisfied with at least one spouse to be born in the selected period, we binarize the resulting partition selecting classes one and two: 1808 couples and bachelors.

*Macro>Play*

In the Ragusan nobility genealogy, many birth dates are missing. Assuming that all children of the same parents and all parents and in-laws of children belong approximately to the same birth cohort, we add them to the people of whom we know that they were born in the required period. We need these indirect neighbors to preserve the structure of the genealogical network. The procedure is stored in the macro `expand_generation.mcr`, which can be executed with the *Macro>Play* command. A genealogical network (Ore graph or P-graph) must be selected in the Network drop list and the binary partition identifying the selected birth cohort (see previous paragraph) must be selected in the Partition drop list. The macro creates a new partition with the extended birth cohort in class number one: 2004 bachelors and couples in our example.

*Partitions>Extract  
Second from First*

The macro can be executed several times to increase the number of selected vertices but generation jumps may extend the range of birth dates enormously.

*Info>Partition*

We advise to apply the macro only once and check the range of known birth years among the selected vertices afterwards. To this end, extract the vertices selected in the partition from the year of birth vector(s): make sure the expanded birth cohort partition is selected in the Partition drop list and a year of birth vector in the Vector drop list and execute the *Vector>Extract Subvector* command (select class 1 only). You may inspect the extracted years with the *Info>Vector* command, which reports the lowest and highest values: there should not be years which fall widely outside the selected period. In the case of a P-graph, you must check the birth dates of men and women separately. With the men, the known birth dates range from 1280, which is 70 years before the selected period, to 1500. The women were born between 1298 and 1498. Even in its first step, the expansion macro lengthens the range of birth dates considerably.

*Info>Network>Indices*

The relinking index is calculated by the *Info>Network>Indices* command and it is printed in the Report screen. Note that the index is valid only for P-graphs. On request, Pajek will compute it for any network, but then its value is meaningless. In the P-graph with the extended birth cohort of 1350-1500, which can be extracted with the partition created in the previous paragraph, the relinking index is 0.20.

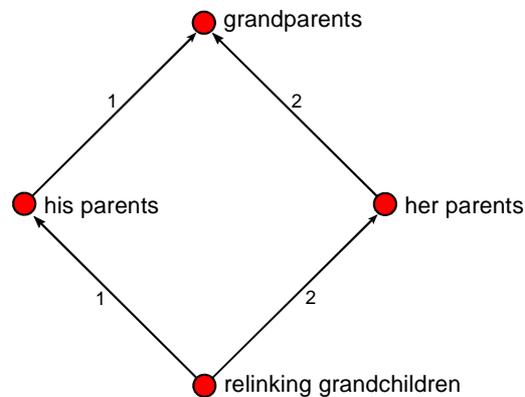
*Net>Components  
>Bi-Components*

If you want to calculate the relinking index for the largest bi-component in this P-graph, you have to identify the bi-components and extract the largest bi-component first. The *Net>Components>Bi-Components* command, introduced in Chapter 7, identifies the bi-components. You may neglect very small bi-components by setting the minimum size of a bi-component to three or more. As you have learned in previous chapters, bi-components are stored as a hierarchy, so inspect the hierarchy (*File>Hierarchy>Edit*) to find the sequential number and size of the largest bi-component. Extract this bi-component from the network in

*Hierarchy>Make Cluster*

*Operations  
>Extract from Network  
>Cluster*

the following way: translate the required class of the hierarchy into a cluster with the *Hierarchy>Make Cluster* command, specifying the sequential number of the bi-component in the hierarchy, and execute the *Extract from Network>Cluster* command from the *Operations* menu. Finally, calculate the relinking index with the *Info>Network>Indices* command. For the extended 1350-1500 birth cohort in the Ragusan nobility P-graph, the relinking index is 0.69.



**Figure 7** - Fragment of relinking grandchildren.

*Nets>Fragment (1 in 2)* Particular types of relinking can be found with the *Fragments* commands in the *Nets* menu, which we also used to trace complete subnetworks (Chapter 3). Create a network which represents the relinking structure that you want to find, e.g., a marriage between two grandchildren of the same grandparents (see Figure 7), with the *Net>Random Network* command and manual editing in the Draw screen. Select it as the first network in the *Nets* menu and select the P-graph as the second network, then find the fragments with the *Nets>Fragment (1 in 2)>Find* command. If you want to find a fragment with a particular pattern of male and female lines, make sure that the lines have the right values in the fragment (1 for male and 2 for female) and select the *Check values of lines* option in the *Nets>Fragment (1 in 2)>Options* menu.

### 11.5 Example II: Citations among papers on network centrality

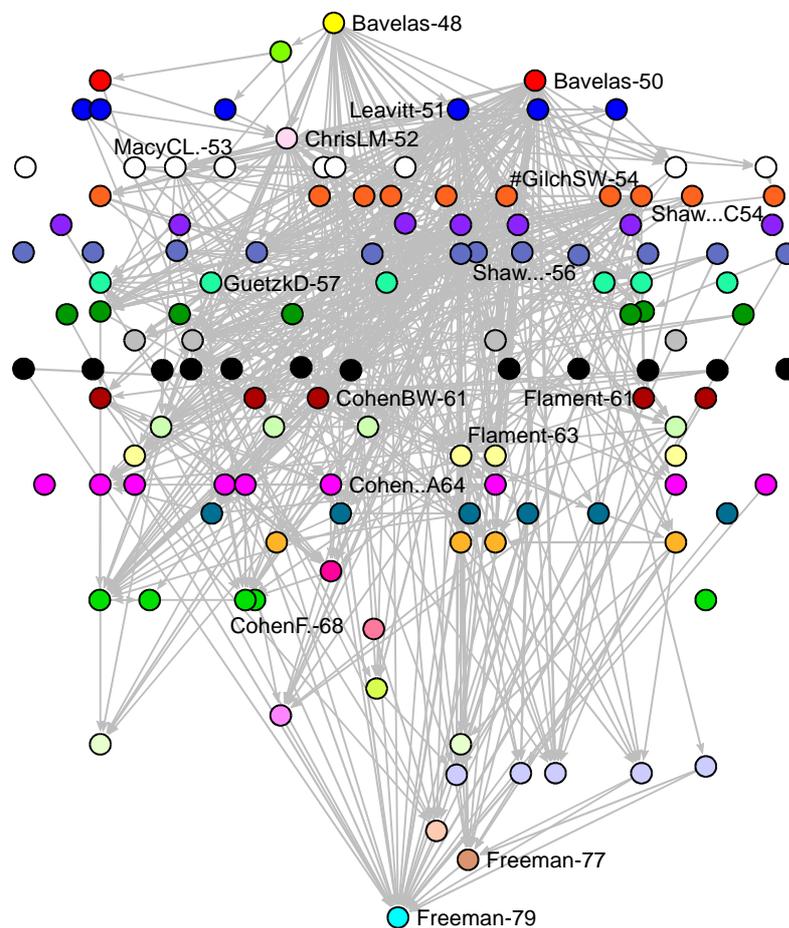
In several social domains, genealogical terminology is used as a metaphor for non-biological affinity. Artists who were trained by the same ‘master’ or who are influenced by the same predecessors are considered to belong to the same ‘family’ or tradition. A work of art has a ‘pedigree’: a list of former owners. In a similar way, scientists are classified according to their intellectual pedigree: the theories and theorists which they use as a frame of reference in their work.

In science, citations make explicit this frame of reference, so they are a valuable source of data for the study of scientific development and scientific communities in scientometrics, history, and sociology of science. They reveal the impact of papers and their authors on later scientific work and they signal scientific communities or specialties which share knowledge.

In this chapter, we will analyze the citations among papers which discuss the topic of network centrality. In 1979, Linton Freeman published a paper which

defined several kinds of centrality. His typology has become the standard for network analysis, so we used it in Chapter 6 of this book. Freeman, however, was not the first to publish on centrality in networks. His paper is part of a discussion which dates back to the 1940s. The network depicted in Figure 8 ([centrality\\_literature.net](http://centrality_literature.net)) shows the papers which discuss network centrality and their cross-references until 1979. Arcs represent citations; they point from the cited paper to the citing paper.

In principle, papers can only cite papers which appeared earlier, so the network is acyclic. Arcs never point back to older papers just like parents cannot be younger than their children. However, there are usually some exceptions in a citation network: papers which cite one another, e.g., papers appearing at about the same time and written by one author. We eliminate these exceptions by removing arcs which are going against time or by shrinking the papers by an author which are connected by cyclic citations. In the centrality literature network, we used the latter approach (e.g., two publications by Gilch in Figure 8).



**Figure 8** - Centrality literature network.

There are important differences between a genealogical network and a citation network. A citation network contains one relation, whereas a genealogical network contains two: parenthood and marriage. In addition, a paper may cite all previous papers notwithstanding their distance in time. In a genealogical network, children have two (biological) parents and parenthood relations always link two

successive generations. The concept of a generation is not very useful in the context of a citation network, so we order the papers by publication date. In Figure 8, layers and vertex colors represent the year of publication (partition `centrality_literature_year.clu`), which is also indicated by the last two digits in the label of a vertex.

### 11.6 Citations

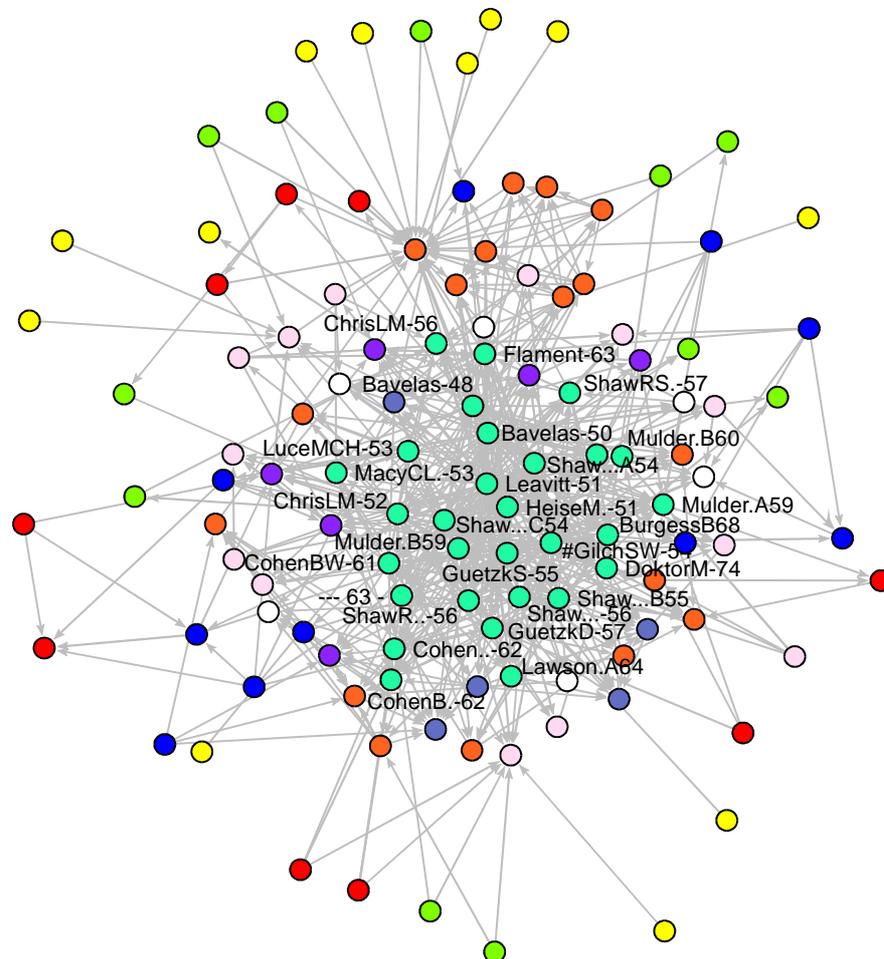
Nowadays, citations are being used to assess the scientific importance of papers, authors, and journals. In general, an item receiving more citations is deemed more important. Databases of citations, e.g., the *Science Citation Index* and the *Social Science Citation Index* compiled by the Institute for Scientific Information (ISI®) list the citations in a large number of journals. Simple calculations yield indices of scientific standing, e.g., the **impact factor** of a journal (the average number of citations to papers in this journal) and the **immediacy index** (the average number of citations of the papers in a journal during the year of its publication). In each year, journals are ranked by their scores on these indices. Compared over longer periods, these indices show differences between scientific disciplines. In the liberal arts, for instance, it is rare for authors to cite recent publications, whereas this is very common in the natural sciences.

Citation analysis is not exclusively interested in the assessment of scientific standing. It also focuses on the identification of specialties, the evolution of research traditions, and changing paradigms. Researchers operating within a particular subject area or scientific specialty tend to cite each other and common precursors. Citation analysis reveals such cohesive subgroups and it studies their institutional or paradigmatic background. Scientific knowledge is assumed to increment over time: previous knowledge is used and expanded in new research projects. Papers which introduce important new insights will be cited until new results modify or contradict them. Citation analysis, therefore, may spot the papers which influence the research for some time and link them into a research tradition which is the backbone of a specialty. Scientific revolutions, that is, sudden paradigmatic changes resulting from new insights, are reflected by abrupt changes in the citation network.

Network analysis is the preferred technique to extract specialties and research traditions from citations. Basically, specialties are cohesive subgroups in the citation network, so they can be detected with the usual techniques. Weak components identify isolated scientific communities which are not aware of each other or who see no substantial overlap between their research domains. Within a weak component, a bi-component identifies sections where different ‘lines’ of citations emanating from a common source text meet again. This is similar to the concept of relinking in genealogical research.

In most citation networks, however, these criteria are not strong enough because almost all papers are linked into one bi-component. *k*-cores (Chapter 3) offer a more penetrating view. The centrality literature network, for example,

contains one large weak component and 11 isolates. There is one large bi-component and twelve vertices are connected by one citation. The network contains a 10-core of 29 papers which is the central ‘summit’ of this network (Figure 9). Each of the papers in this core is connected to at least ten other papers by citations but we do not know which papers are cited often and which cite a lot.



**Figure 9** -  $k$ -cores in the centrality literature network (without isolates).

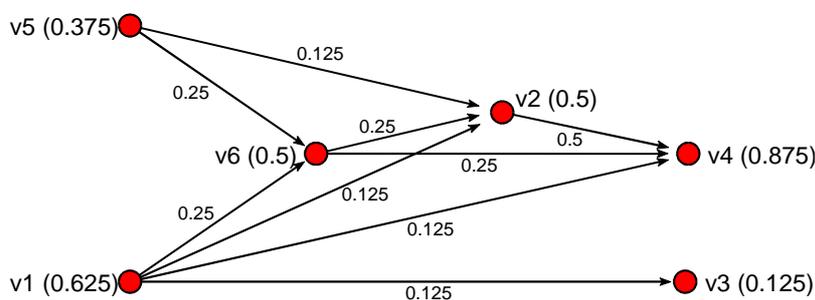
The cohesion concept does not take time into account. It does not reflect the incremental development of knowledge, nor does it identify the papers which were vital to this development. Therefore, a special technique for citation analysis was developed which explicitly focuses on the flow of time. It was proposed by N. Hummon and colleagues and it is called **main path analysis**.

Let us think of a citation network as a system of channels which transport scientific knowledge or information. A paper which integrates information from several previous papers and adds substantial new knowledge, will receive many citations and it will make citations to previous papers more or less redundant. As a consequence, it is an important junction of ‘channels’ and a great deal of knowledge flows through it. If knowledge flows through citations, a citation which is needed in paths between many papers is more crucial than a citation which is hardly needed to link papers. The most important citations constitute one or more main paths, which are the backbones of a research tradition.

Main path analysis calculates the extent to which a particular citation or paper is needed to link papers, which is called the traversal count or traversal weight of a citation or paper. First, the procedure counts all paths from each source - a paper which is not citing within the data set - to each sink - a paper which is not cited within the data set - and it counts the number of paths which include a particular citation. Next, it divides the number of paths which use a citation by the total number of paths between source and sink vertices in the network. This proportion is the traversal weight of a citation. In a similar way, you can obtain the traversal weight of each paper.

In an acyclic network, a **source vertex** is a vertex with zero indegree.  
 In an acyclic network, a **sink vertex** is a vertex with zero outdegree.  
 The **traversal weight** of an arc or vertex is the proportion of all paths between source and sink vertices which contain this arc or vertex.

Figure 10, for example, shows a citation network of six papers ordered in time from left to right. There are two sources ( $v_1$  and  $v_5$ ) and two sinks ( $v_3$  and  $v_4$ ). One path connects source  $v_1$  and sink  $v_3$  but there is no path from  $v_5$  to  $v_3$ . Four paths reach  $v_4$  from  $v_1$  and three paths from  $v_5$ . In sum, there are eight paths from sources to sinks. The citation of paper  $v_1$  by paper  $v_3$  is included in one of the eight paths, so its traversal weight is 0.125. The citation of  $v_2$  in paper  $v_4$  is contained in exactly half of all paths. The traversal weights of the vertices, which are reported between brackets, are calculated in a similar way.



**Figure 10** - Traversal weights in a citation network.

Now that we have defined and calculated the traversal weights of citations, we may extract the paths or components with the highest traversal counts on the lines, the main paths or main path components, which are hypothesized to identify the main stream of a literature. We can analyze their evolution over time and search for patterns which reflect the integration, fragmentation, or specialization of a scientific community.

In a citation network, a **main path** is the path from a source vertex to a sink vertex with the highest traversal weights on its arcs. Several methods have been proposed to extract main paths from the network of traversal weights. The method which we follow here consists of choosing the source vertex (or vertices) incident with the arc(s) with the highest weight, selecting the arc(s) and the head(s) of the arc(s), repeating this step until a sink vertex is reached. In the example of Figure 10, the main paths start with vertex  $v_1$  and vertex  $v_5$  because

both source vertices are incident with an arc carrying a traversal weight of 0.25. Both arcs point toward vertex v6, which is the next vertex on the main paths. Then, the paths proceed either to vertex v2 and on to vertex v4 or directly from vertex v6 to vertex v4. We find several main paths, but they lead to the same sink, so we conclude that the network represents one research tradition.

A **main path component** is extracted in a way which is similar to the slicing procedure used for *m*-slices. Choose a cutoff value between zero and one, and remove all arcs from the network with traversal weights beneath this value. The components in the extracted networks are called main path components. Usually, we look for the lowest cutoff value which yields a component which connects at least one source vertex to one sink vertex. This value is equal to the lowest traversal weight on the main paths. In our example, this cutoff value is 0.25 and we obtain a main path component which includes all papers except v3, which is a marginal paper in the research tradition represented by this data set.

Of course, paper v3 may be very important in another research tradition. The choice of the papers to be included in the data set restricts the number and size of research traditions which can be found. Like a genealogy, a citation network is virtually endless so it cannot be captured entirely in a research project. The researcher has to set limits to the data collection, but this should be based on sound substantive arguments.

### Application

*Net>Partitions>Core>All*

In Chapters 3 and 7, we discussed the commands for detecting components, bi-components, and *k*-cores, which identify cohesive subgroups in a network. A citation network is directed and acyclic, so you should search weak components instead of strong components and find *k*-cores on input and output relations (command *All* in the *Net>Partitions>Core* submenu).

*Net>Citation Weights*

*Info>Network  
>Line Values*

Main path analysis is very easy in Pajek. The commands in the *Net>Citation Weights* submenu compute the traversal weights for lines and vertices in an acyclic network. There are two commands: *Source - Sink* and *Vertex - Sink*. The *Source - Sink* algorithm counts the paths between all source and sink vertices as explained above. The *Vertex - Sink* command traces paths from all vertices to the sink vertices. In the latter procedure, citations of early papers receive lower weights because they can not be part of paths emanating from later papers, so we advise to use the *Source - Sink* command. The traversal weights of the papers (vertices) are stored in a vector and the weights of the citations (lines) are saved as line values in a new network, which can be inspected with the *Info>Network>Line Values* command.

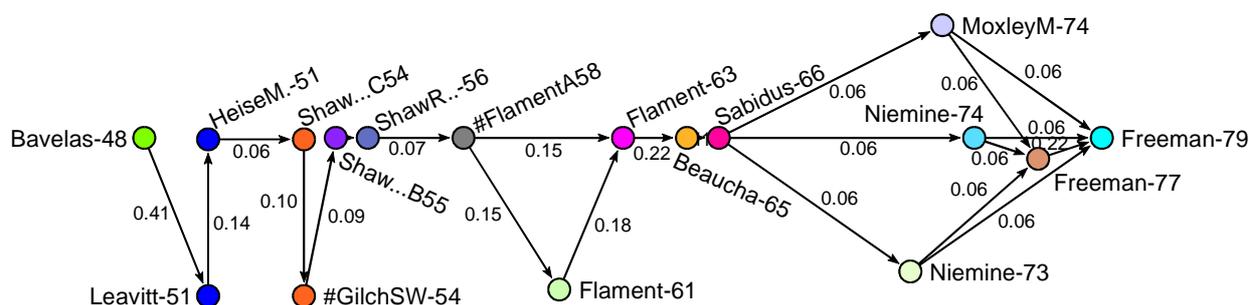
When we apply the *Source - Sink* command to the centrality literature network, about 90 percent of the lines have a traversal weight of 0.05 or less, and thirteen lines have a value which exceeds 0.103 (Table 1: be sure the network labeled 'Citation weights (Source-Sink)' is selected in the drop list). Clearly, one citation is very important to the development of the centrality literature: it has an extremely high traversal weight of 0.41. This is the citation of Bavelas' 1948

paper by Leavitt in 1951. Bavelas (1948) and Leavitt (1951), as well as Freeman (1979) and Flament (1963) are the vertices with the highest traversal weights. These are the crucial papers in the centrality literature.

**Table 1** - Traversal weights in the centrality literature network.

Line Values	Frequency	Freq%	CumFreq	CumFreq%
( ... 0.0000]	90	14.68	90	14.68
(0.0000 ... 0.0515]	465	75.86	555	90.54
(0.0515 ... 0.1030]	45	7.34	600	97.88
(0.1030 ... 0.1545]	8	1.31	608	99.18
(0.1545 ... 0.2059]	2	0.33	610	99.51
(0.2059 ... 0.2574]	2	0.33	612	99.84
(0.2574 ... 0.3089]	0	0.00	612	99.84
(0.3089 ... 0.3604]	0	0.00	612	99.84
(0.3604 ... 0.4118]	1	0.16	613	100.00
Total	613	100.00		

The *Citation Weights* commands automatically identify the main paths in the citation network. The commands create a partition identifying the vertices on the main paths (cluster one) in the original citation network and it produces a new network which contains the main paths (see Figure 11). In the centrality literature, the main paths start with Bavelas (1948), proceed to Leavitt (1951), and, finally, end with Freeman (1977 and 1979).

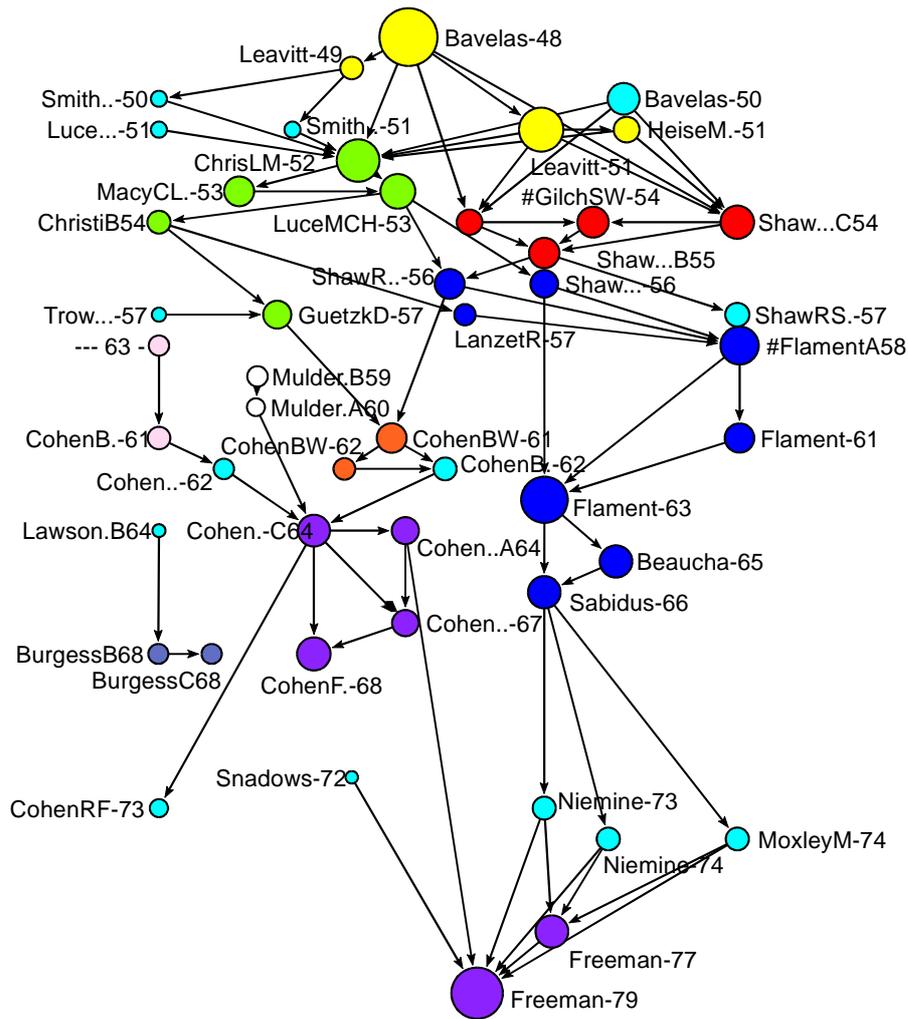


**Figure 11** - A main path in the centrality literature network.

*Net>Transform>Remove*  
*>lines with value*  
*>lower than*

The lowest traversal weight of the arcs in the main path is 0.05, but it is interesting to use a slightly lower cutoff value to obtain the main path component here. Let us delete all arcs with traversal weights lower than 0.03. This can be done with the *Remove>lines with value>lower than* command in the *Net>Transform* submenu. The 78 arcs which remain in the network (Figure 12) group the non-isolated vertices into two weak components: one large component with 46 papers and a small component with three papers by Lawson and Burgess.

In Figure 12, vertex size indicates the traversal weight of a paper and vertex colors indicate weak components if arcs with traversal weight up to 0.06 would have been removed. This figure reveals that the literature on network centrality was split into two lines between 1957 and 1979. One line was dominated by Cohen and the other by Flament and Nieminen. In 1979, Freeman integrated both lines in his classic paper.



**Figure 12** - Main path component of the centrality literature network.

### 11.7 Summary

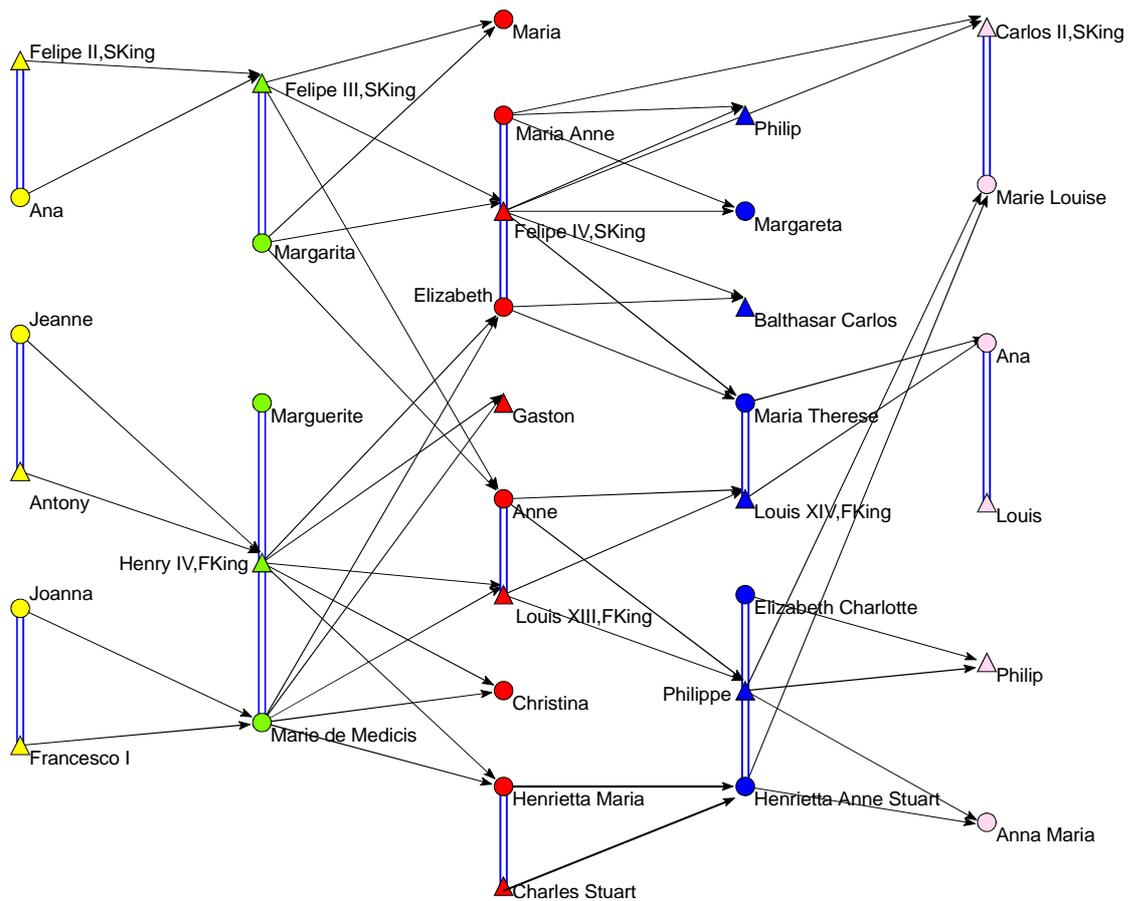
This is the last chapter presenting methods which cope with the dynamics of time in network analysis. Over time, social relations branch off into a gamut of independent strands. Kinship relations, for instance, create family trees which expand rapidly over generations. Sometimes, however, these strands merge after some time, e.g., people with common ancestors marry. This is called structural relinking, which is a measure of social cohesion over time. A social system with much relinking is relatively cohesive because relinking shows that people are oriented towards members of their own group or family.

In a genealogy, the amount of structural relinking can be assessed provided that we use a special kind of network: the P-graph. In contrast to an Ore graph, which represents each person by a vertex, parenthood by arcs, and marriage by (double) lines, couples and bachelors are vertices and individuals are arcs in a P-graph. Because symmetric marriages are not represented by lines in the P-graph, each bi-component is an instance of structural relinking.

Methods for analyzing citation networks handle the time factor in a slightly different way. Here, we want to identify the publications which are the crucial links in the literature on a particular topic. Scientific papers contain knowledge, and citations indicate how knowledge flows through a scientific community. Each flow follows a path of citations and citations which occur in a lot of paths are important to the transmission of knowledge: they have high traversal weights. Citations with high traversal weights are linked into main paths, which represent the main lines of development in a research area. The papers and authors connected by citations of some minimum traversal weight constitute main path components, which are hypothesized to identify scientific specialties or subspecialties.

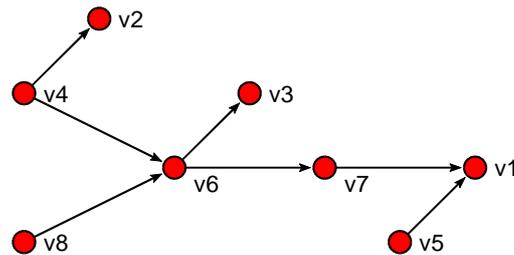
11.8 Exercises

- 1 The Ore graph depicted below shows a part of the family relations of Louis XIII, king of France (1601-1643). Calculate the remove of his relation with Henrietta Anne Stuart.



- 2 Which people constitute the family of orientation of Louis XIII and what is his family of procreation?
- 3 What is a generation jump? Indicate one in the Ore graph of Exercise 1.
- 4 Draw a P-graph which contains the same information as the Ore graph of Exercise 1.

- 5 How can we distinguish between a blood-marriage and a relinking non blood-marriage in a P-graph? Give an example of both types of relinking in the genealogy of Louis XIII.
- 6 Explain why the relinking index of a tree is zero.
- 7 List all paths from sources to sinks in Figure 10 and show that the citation weight of the arc from v2 to v4 is correct.
- 8 Identify the source and sink vertices, the paths between them, and the traversal weight of the arcs in the citation network depicted below. What is the main path?



### 11.9 Assignment 1

The GEDCOM file `Isle_of_Man.ged` contains the combined genealogies of approximately 20 families from the British Isle of Man. Describe the overall structure of this network and the sections with structural relinking. Which types of relinking do occur?

### 11.10 Assignment 2

Publications and citations pass on scientific knowledge and traditions, so do advisors to their students. The file `PhD.net` contains the relations between Ph.D. students and their advisors in theoretical computer science; each arc points from an advisor to a student. The partition `PhD_year.clu` contains the (estimated) year in which the Ph.D. was obtained. Search for separate research traditions in this network and describe how they evolve.

### 11.11 Further Reading

- The genealogical data of the Ragusan nobility example were coded from the Ph.D. thesis of Irmgard Mahnken (1960): *Das Ragusanische Patriziat des XIV. Jahrhunderts*. For an analysis of a part of the genealogy, see V. Batagelj, 'Ragusan families marriage networks' in A. Ferligoj & A. Kramberger (Eds.), *Developments in Data Analysis* (Ljubljana: FDV, 1969, 217-228) and P. Doreian, V. Batagelj & A. Ferligoj, 'Symmetric-acyclic decompositions of networks' in *Journal of Classification*, 17 (2000), 3-28.
- For the collection and storage of genealogical data, we advise to use the GEDCOM 5.5 standard (<http://www.gendex.com/gedcom55/55gcint.htm>). Good free software is the Genealogical Information Manager, available at

[http://www.mind.spring.com/~dblaine/gim\\_home.html](http://www.mind.spring.com/~dblaine/gim_home.html), and Personal Ancestral File, which is produced and distributed by the Church of Jesus Christ of Latter-day Saints ([www.familysearch.org](http://www.familysearch.org)). This organization compiles a large database of genealogical information from which downloads can be made. The genealogies from the Isle of Man (Assignment 1) were downloaded from

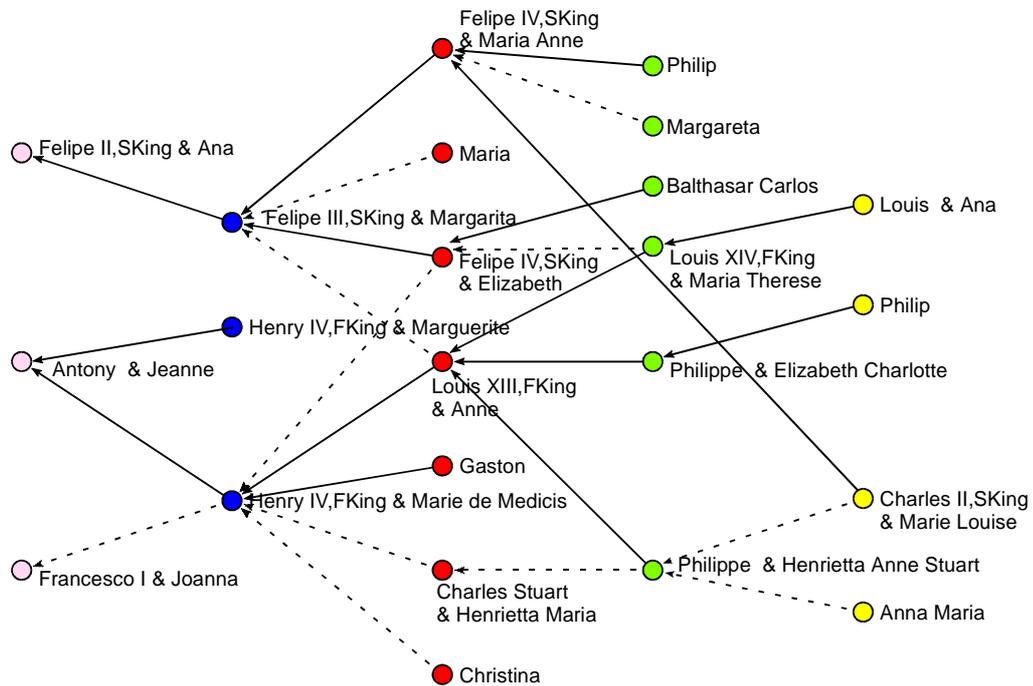
<http://www.isle-of-man.com/interests/genealogy/gedcom/index.htm>.

- For additional reading on the analysis of kinship relations in the social sciences, we refer to T. Schweizer & D.R. White, *Kinship, networks, and exchange* (Cambridge: Cambridge University Press, 1998).
- The centrality literature example was taken from N.P. Hummon, P. Doreian, & L.C. Freeman, 'Analyzing the structure of the centrality-productivity literature created between 1948 and 1979' (in: *Knowledge-Creation Diffusion Utilization*, 11 (1990), 459-480), which also introduces main path analysis. E. Garfield, *Citation Indexing. Its Theory and Application in Science, Technology, and Humanities* (New York: John Wiley & Sons, 1979) is a classic text on citation analysis.

### 11.12 Answers

- 1 Louis XIII is the uncle (mother's brother) of Henrietta Anne Stuart, so she is a relative in the third degree if we restrict ourselves to blood relations. Louis XIII is also her step-father, so the degree is one if we include marital relations.
- 2 The family of orientation of Louis XIII include his parents Henry IV and Marie de Medicis, his brother Gaston, and his sisters Elizabeth, Christina, and Henrietta Maria. Marguerite, the other wife of Henry IV, may or may not belong to the family of orientation. His family of procreation contains his wife Anne and their children Louis XIV and Philippe.
- 3 A generation jump in a genealogy refers to a relinking marriage which connects people of different genealogical generations, which are calculated from the point of view of their common ancestor. The marriage between Carlos II and Marie Louise creates a generation jump, because Carlos is a grandson of Felipe III and Margarita (second remove) and Marie Louise is the granddaughter of the daughter (Anne) of Felipe III and Margarita (third remove).

- 4 The P-graph should look like the figure below. Do not forget to draw different arcs for men and women and to reverse the direction of arcs.



- 5 In a P-graph, the husband and wife involved in a blood-marriage share at least one ancestor: there are two paths from the blood-marriage to an ancestor, for instance, from Philippe and Henrietta Anne Stuart to Henry IV, king of France, and his spouse Marie de Medicis. Both Philippe and Henriette Anne Stuart are their grandchildren. A relinking non-blood marriage is a marriage between descendents of families which are already linked by intermarriage, for example, the Spanish king Felipe III and the French king Henry IV are linked by two marriages among their children: Felipe IV and Elizabeth, Louis XIII and Anne. In a P-graph, this type of relinking is characterized by two semipaths (or one path and one semipath) between couples.
- 6 Structural relinking involves semicycles: vertices are connected by two paths or semipaths. Since trees do not contain semicycles by definition, there is no relinking and the relinking index is zero.
- 7 The eight paths are: (1)  $v_1 \rightarrow v_3$ , (2)  $v_1 \rightarrow v_4$ , (3)  $v_1 \rightarrow v_2 \rightarrow v_4$ , (4)  $v_1 \rightarrow v_6 \rightarrow v_4$ , (5)  $v_1 \rightarrow v_6 \rightarrow v_2 \rightarrow v_4$ , (6)  $v_5 \rightarrow v_6 \rightarrow v_4$ , (7)  $v_5 \rightarrow v_6 \rightarrow v_2 \rightarrow v_4$ , and (8)  $v_5 \rightarrow v_2 \rightarrow v_4$ . Four paths include the arc  $v_2 \rightarrow v_4$ , viz., paths 3, 5, 7, and 8, which is half of all paths, so the traversal weight of this arc is 0.5.
- 8 The source vertices are  $v_4$ ,  $v_8$ , and  $v_5$ ;  $v_2$ ,  $v_3$ , and  $v_1$  are sink vertices. There are 6 paths from sources to sinks: (1)  $v_4 \rightarrow v_2$ , (2)  $v_4 \rightarrow v_6 \rightarrow v_3$ , (3)  $v_4 \rightarrow v_6 \rightarrow v_7 \rightarrow v_1$ , (4)  $v_8 \rightarrow v_6 \rightarrow v_3$ , (5)  $v_8 \rightarrow v_6 \rightarrow v_7 \rightarrow v_1$ , and (6)  $v_5 \rightarrow v_1$ . The arcs  $v_4 \rightarrow v_2$  and  $v_5 \rightarrow v_1$  are included in one of these paths, so their traversal weight is one divided by six: 0.167. The other arcs are included in two paths, so their traversal weights are 0.333. There are four main paths: (1) from  $v_4$  to  $v_3$ , (2) from  $v_4$  to  $v_1$ , (3) from  $v_8$  to  $v_3$ , and (4) from  $v_8$  to  $v_1$ .