Social Science Computer Review

http://ssc.sagepub.com

Analysis of Kinship Relations With Pajek

Vladimir Batagelj and Andrej Mrvar

Social Science Computer Review 2008; 26; 224 originally published online Dec 3, 2007;
DOI: 10.1177/0894439307299587

The online version of this article can be found at:
http://ssc.sagepub.com/cgi/content/abstract/26/2/224

Published by:
SAGE Publications
http://www.sagepublications.com

Social Science Computer Review
Volume 26 Number 2
May 2008 224-246

10.1177/0894439307299587
http://ssc.sagepub.com
hosted at
http://online.sagepub.com

Analysis of Kinship Relations With Pajek

Vladimir Batagelj
Andrej Mrvar
University of Ljubljana, Slovenia

In the article, two general approaches to analysis of large sparse networks are presented: fragment searching and matrix multiplication. These two approaches are applied to analysis of large genealogies. Genealogies can be represented as graphs in different ways: as Ore graphs, p-graphs, or bipartite p-graphs. We show that p-graphs are more suitable for searching for relinking patterns, whereas Ore graphs are better for computing kinship relations using network multiplication. Algorithms described in this article are implemented in the program Pajek.

Keywords: *genealogy; Ore graph; p-graph; bipartite p-graph; calculating kinship relations; relinking marriages; relinking index; large networks; social network analysis; Pajek*

Introduction

People collect genealogical data for several different reasons or purposes:

- Researchers in history, sociology, and anthropology (Hamberger, Houseman, Daillant, White, & Barry, 2005; White, Batagelj, & Mrvar, 1999) use genealogies to compare different cultures. In these researches, they consider kinship as a fundamental social relation.
- Individuals collect records about their families or about people living in a longer period on a selected territory, for example:
  - Mormon genealogy (MyFamily.com, 2004)
  - Genealogy of the Škofja Loka District (Hawlina, 2004)
- Genealogy of American presidents (Tompsett, 1993)
- There exist special genealogies where the relation is "nonbiological":
  - Students and their Ph.D. thesis advisors (Theoretical Computer Science Genealogy; Johnson & Parberry, 1993; Mathematics; Coonce, 1999)
  - Genealogies of gods of antiquity (Hawlina, 2004).

Many programs for data entry and maintenance of genealogical records can be found on the market (GIM, Brother's Keeper, Family Tree Maker, etc.), but only few analyses can be done using

Authors' Note: This work was partially supported by the Slovenian Research Agency, Project J1-6062-0101. It is a detailed version of a part of the talks presented at "Dagstuhl Seminar 05361: Algorithmic Aspects of Large and Complex Networks," September 4–9, 2005, Dagstuhl, Germany; and at the meeting "Algebraic Combinatories and Theoretical Computer Science," February 12–15, 2006, Bled, Slovenia.

these programs. This was the reason to expand the program Pajek (Batagelj & Mrvar, 2006; White et al., 1999) with some procedures for the analysis and visualization of large genealogies. Pajek is a general program for the analysis and visualization of large networks. It is free for noncommercial use.

GEDCOM Standard

GEDCOM (Family History Department, 1996) is a standard for storing and exchanging genealogical data which is used to interchange and combine data from different programs which were used for entering the data. The lines in Table 1 are extracted from the GEDCOM file of European royal families (Royal Genealogies, 1992).

From data represented in the described way, we can generate several networks, as explained in the following section.

Table 1
Part of 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 Hospital, 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 using networks in different ways: as an *Ore-graph*, a *p-graph*, and a *bipartite p-graph*.

Ore-graph

In an Ore graph of genealogy, every person (INDI tag in the GEDCOM file) is represented by a vertex, and they are linked with relations: Marriage relation *is a spouse of* (FAMS or FAM + HUSB + WIFE) is represented with edges; and relation *is a parent of* (FAMC or FAM + CHIL) is represented by arcs pointing from each of the parents to their children—partitioned into the relations *is a mother of* (red dotted) and *is a father of* (blue solid) (see Figure 1).

p-graph

In a p-graph, vertices represent individuals (INDI not in FAM + HUSB + WIFE) or couples (FAM + HUSB + WIFE). In case that person is not married, she or he is represented by a

Figure 1
Ore Graph

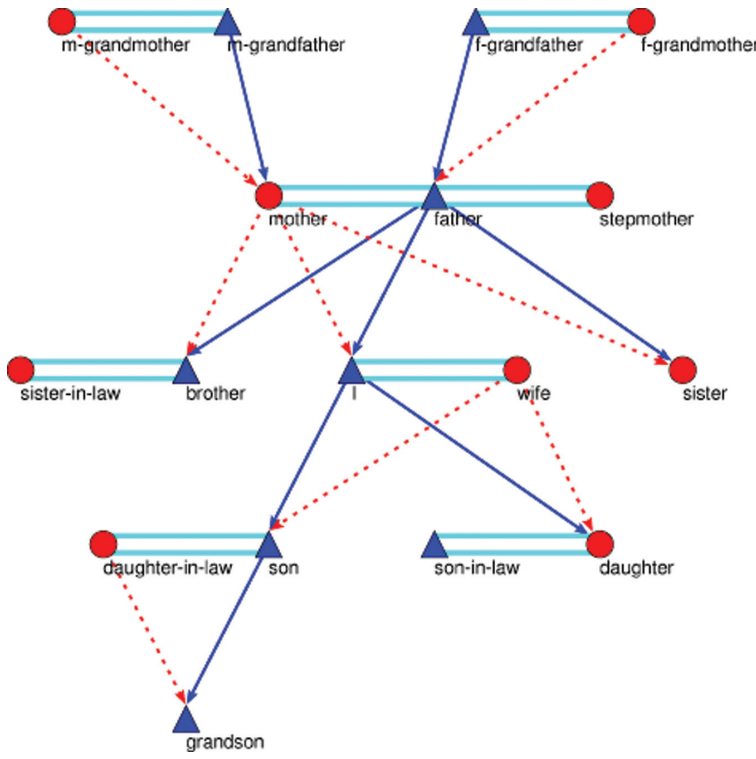
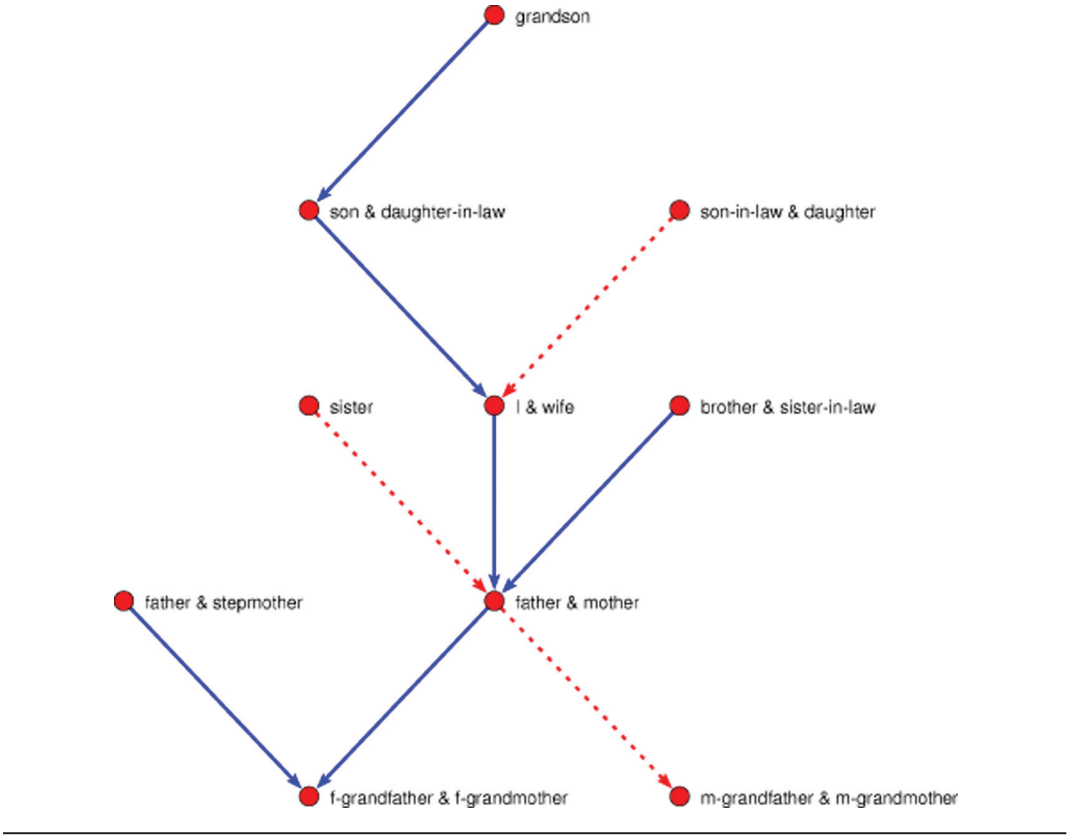


Figure 2
p-Graph



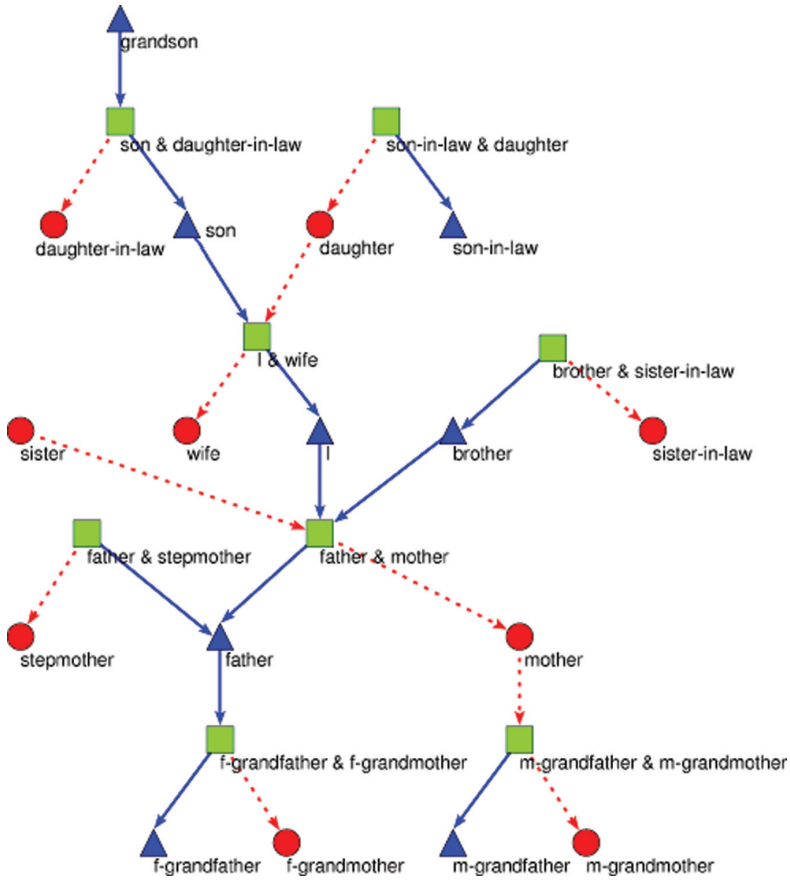
vertex; otherwise, the person is represented with the partner in a common vertex. There are only arcs in p-graphs—they point from children (CHIL) to their parents (FAM). (See Figure 2.) The solid arcs represent the relation *is a son of*, and the dotted arcs represent the relation *is a daughter of*.

p-graphs are usually also used for visual representation of genealogies. Because they are acyclic graphs, the vertices can be assigned to levels. Special algorithms for drawing genealogies are included in Pajek. As an example, part of the Bouchard genealogy (Beauregard, 1995) and its most relinked part are presented in Figures 4 and 5.

Bipartite p-graph

A bipartite p-graph has two kinds of vertices—vertices representing couples (rectangles) and vertices representing individuals (circles for women and triangles for men)—therefore, each married

Figure 3
Bipartite p-graph



person is involved in two kinds of vertices (or even more if he or she is involved in multiple marriages). Arcs, again, point from children to their parents (see Figure 3).

Genealogies Are Sparse Networks

We shall call a genealogy *regular* if every person in it has at most two parents. Genealogies are *sparse* networks—the number of lines is of the same order as the number of vertices. In this section, some bounds on the number of lines in different kinds of regular genealogies are given (Mrvar & Batagelj, 2004).

In a *regular Ore genealogy* (V, E, A), the set of vertices V is partitioned into two subsets: V_i – set of individuals – single persons, and V_m – set of married persons. Therefore,

$$V = V_i \cup V_m \text{ and } V_i \cap V_m = \varnothing$$

Figure 4
The Bouchard Genealogy

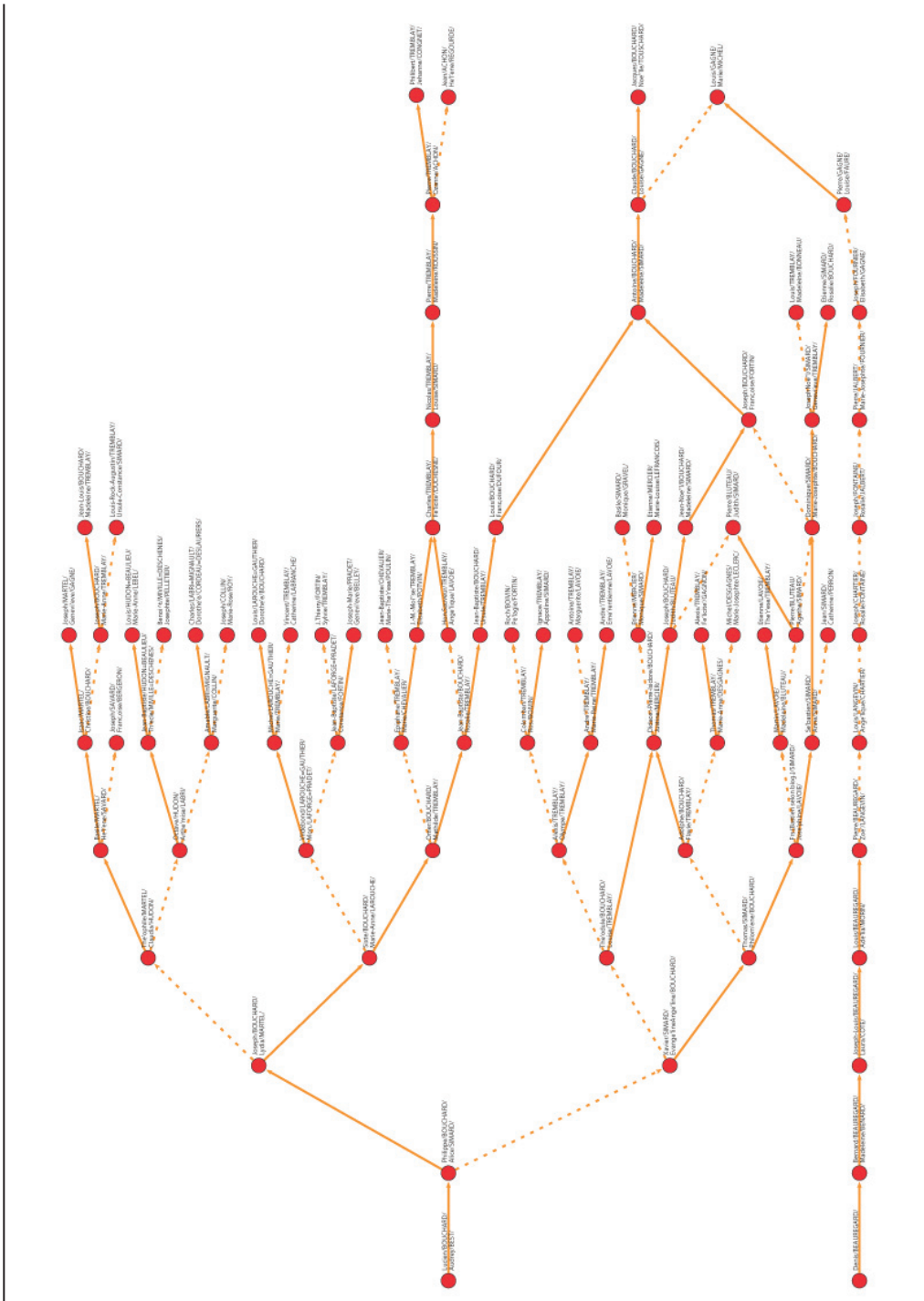
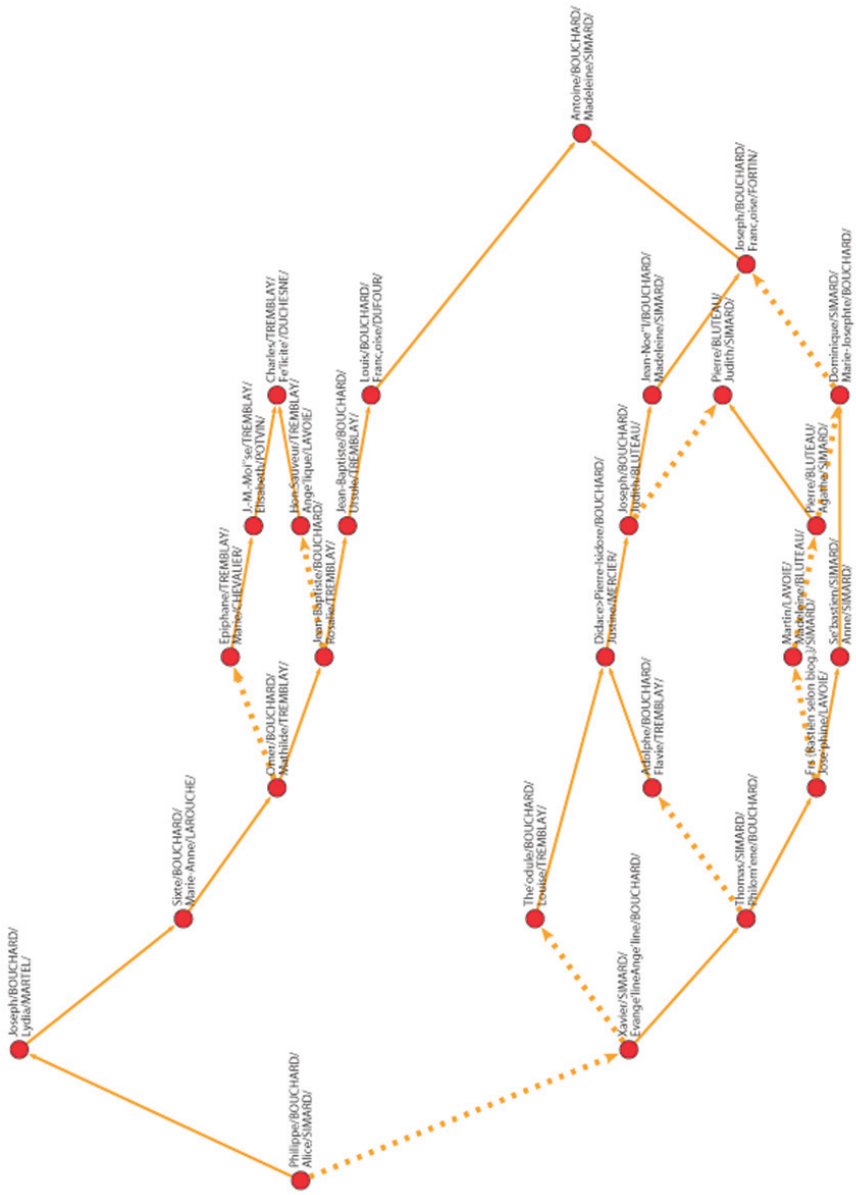


Figure 5
The Most Relinked Part of the Bouchard Geneology



For marriage links E, we have

$$2|E| = \sum_{v \in Vm} \deg_E(v) = \sum_{v \in Vm} 1 + \sum_{v \in Vm:\deg_E(v) > 1} (\deg_E(v) - 1)$$

Denoting the second term in the last expression by M (multiple marriages surplus), we get

$$2|E| = |V_m| + M$$

In most real-life genealogies, it holds that $|V_i| \geq M$ (most of the married people are married only once, and some are not married—the number of single persons outnumbers the multiple marriages surplus), and therefore $|V| = |V_i| + |V_m| \geq 2|E|$, and finally

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

We shall say that such genealogy is *usual*. All genealogies in Table 2 are usual.

Note that there exist genealogies in which $|V_i| \geq M$ doesn't hold. For example, for a complete bipartite graph $K_{3,3}$ (three men and three women married to each other, without children), we have $|E| = 9$, $|V| = 6$, $|V_i| = 0$, and $M = 12$.

Because in regular Ore genealogy, for every vertex $v \in V$ for its input degree on directed part holds $\text{indeg}_A(v) \leq 2$, the upper bound for the number of arcs A is as follows:

$$|A| = \sum_{v \in V} \text{indeg}_A(v) \leq 2|V|$$

Table 2
Number of Vertices and Number of Lines in Ore Graphs and p-Graphs for Some Large Genealogies

Data Set	$\|V\|$	$\|E\|$	$\|A\|$	$\|L\|/\|V\|$	$\|V_i\|$	M	$\|V_p\|$	$\|A_p\|$	$\|A_p\|/\|V_p\|$
Loka	47,956	14,154	68,052	1.71	21,074	1,426	35,228	6,192	1.03
Silba	6,427	2,217	9,627	1.84	2,263	270	4,480	5,281	1.18
Ragusa	5,999	2,002	9,315	1.89	2,374	379	4,376	5,336	1.22
Tur	1,269	407	1,987	1.89	549	94	956	1,114	1.17
Royal92	3,010	1,138	3,724	1.62	1,003	269	2,141	2,259	1.06
Franklin[a]	294,964	103,290	362,190	1.58	100,543	12,159	203,833	195,650	0.96
Drame	29,606	8,256	41,814	1.69	13,937	843	22,193	21,862	0.99
Hawlina	7,405	2,406	9,908	1.66	2,808	215	5,214	5,306	1.02
Marcus	702	215	919	1.62	292	20	5,070	496	0.98
Mazol	2,532	856	3,347	1.66	894	74	1,750	1,794	1.03
Presidents	2,145	978	2,223	1.49	282	93	1,260	1,222	0.97
Royale	17,774	7,382	25,822	1.87	4,441	1,431	11,823	15,063	1.27

a. From Franklin.ged (2000).

Therefore, we get in usual genealogies for the upper bound for the set of lines L (arcs and edges):

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

Connected components of *p-graphs* are almost trees—deviations from trees are caused by relinking marriages. For a p-graph (V_p, A_p), we have

$$|V_p| = |V_i| + |E|$$

Using also the equality $2|E| = |V_m| + M$, we get

$$|V| = |V_i| + |V_m| = |V_i| + 2|E| - M = |V_p| + |E| - M$$

and finally:

$$|V_p| = |V| - |E| + M$$

Because in most real-life genealogies, it holds also that $|E| \geq M$, we get for V_p for usual genealogies the following bounds:

$$|V| \geq |V_p| \geq |V| - |E| \geq 1/2|V|$$

In p-graphs for every vertex $v \in V_p$ for its output degree, it holds that $\text{outdeg}(v) \leq 2$. Therefore, the number of arcs in a p-graph has the following upper bound:

$$|A_p| = \sum_{v \in V_p} \text{outdeg}(v) \leq 2|V_p|$$

For the number of vertices V_b in a bipartite p-graph (V_b, A_b), we have

$$|V_b| = |V| + |E|$$

from where we get for usual genealogies the following bounds:

$$|V| \leq |V_b| \leq \frac{3}{2}|V|$$

For the number of arcs A_b, we have

$$|A_b| = |A_p| + 2|V_m|$$

and using $|A_p| \leq 2\,|V_b|$, $|V_m| = 2|E| - M$, and $|V| = |V_p| + |E| - M$, we get the bound

$$|A_b| \leq 2(|V_p| + |V_m|) = 2(|V_p| + 2|E| - M) = 2(|V| + |E|) = 2|V_b|$$

which for usual genealogies simplifies to

$$|A_b| \leq 3|V|$$

Some Datasets

To check the results, let us take several large genealogies and look at the corresponding Ore and p-graphs. A comparison is given in Table 2. In the table, the following notation is used:

Ore Graph

$|V|$: number of vertices
$|E|$: number of edges
$|A|$: number of arcs
$|L| = |E| + |A|$: total number of lines

p-Graph

$|V_i|$: number of individuals
M: multiple marriages surplus
$|V_p| = |V_i| + |E|$: total number of vertices
$|A_p|$: number of arcs

We can see that all genealogies are really very sparse.

Because the first five genealogies from Table 2 are used in the following examples, let us introduce them in more detail first.

- **Loka.ged** is a genealogy of people who were living (or are still living) in Škofja Loka District in the western part of Slovenia. The number of records in this data set is still growing. The genealogy was collected by P. Hawlina (2004).
- **Silba.ged** stores the genealogy of Silba Island. Silba is one of the middle-sized islands in Croatia, close to Zadar. These records were also collected by P. Hawlina (2004). Here, we expect high relinking because of special geographical position (isolation).
- **Ragusa.ged** is a genealogy of Ragusan noble families living between the 12th and 16th centuries (Dremelj, Mrvar, & Batagelj, 2002; Mahnken, 1960). Ragusa is an old name for Dubrovnik, Croatia. High relinking is expected because of Ragusa's geopolitical position in the past, and very restrictive marriage rules that were taken into account (e.g., a member of a noble family is supposed to marry another member of a noble family).
- **Tur.ged** is a genealogy of Turkish nomads (White et al., 1999). Among nomads, a relinking marriage is a signal of commitment to stay within the nomad group; therefore, again, high relinking is expected.
- **Royal.ged** is a public domain GEDCOM file containing information on 3,010 individuals of European royalty and their marriages (Royal Genealogies, 1992).

Comparison of Different Presentations

p-graphs and bipartite p-graphs have many advantages (see White et al., 1999):

- There are fewer vertices and lines in p-graphs than in corresponding Ore graphs.
- p-graphs are directed, acyclic networks (what enables us to draw p-graphs in layers).
- Every semicycle of the p-graph corresponds to a relinking marriage. There exist two types of relinking marriages:
  - Blood marriage: in which the man and woman from the couple have a common ancestor (e.g., marriage between brother and sister).
  - Nonblood marriage: For example, two brothers marry two sisters from another family.
- p-graphs are more suitable for most analyses.

Bipartite p-graphs have an additional advantage: We can distinguish between a married uncle and a remarriage of a father (see Figures 2 and 3). They enable us, for example, to find marriages between half-brothers and half-sisters. Some examples are given in the following sections.

Relinking Index

The relinking index is a measure of relinking by marriages among persons belonging to the same families.

Let n denotes number of vertices in a p-graph, m the number of arcs, k the number of weakly connected components, and M the number of maximal (or last) vertices (vertices having output degree 0, $M \geq 1$).

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

In a *regular* genealogy, $m \leq 2(n - M) = 2n - 2M$. Thus, $0 \leq k + m - n \leq k + n - 2M$, or

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

This quotient is called *the relinking index* (*RI*):

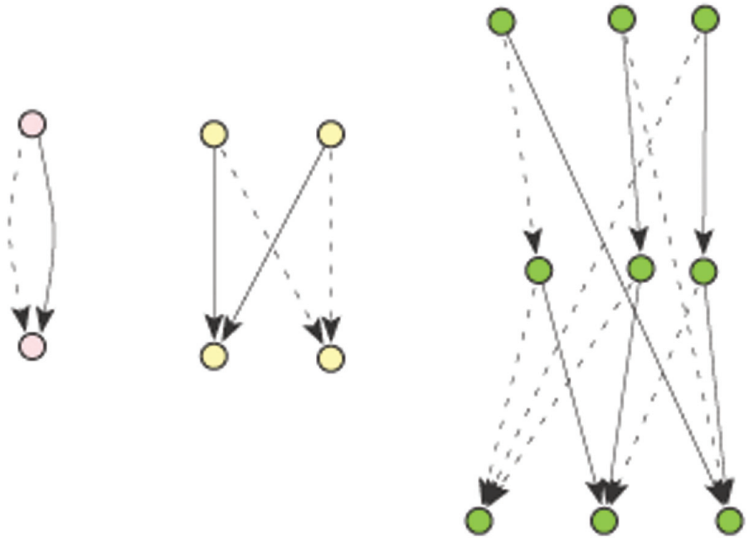
$$RI = \frac{k+m-n}{k+n-2M}$$

If we take a connected genealogy (selected weakly connected component) $k = 1$, we get

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

For a trivial graph (having only one vertex), we define $RI = 0$. (See also White et al., 1999.) *RI* has some interesting properties:

Figure 6
Patterns With Relinking Index 1 (p-Graph)



- $0 \leq RI \leq 1$
- $RI = 0$ (no relinking) if and only if the network is a forest/tree ($m = n - k$).
- For a cycle of depth, $h = \lfloor m/2 \rfloor = \lfloor n/2 \rfloor$, $RI = 1/(2h - 1)$ (the higher the depth, the weaker the relinking). For a cycle of depth 3 (6 vertices), $RI = 1/5$.
- There exist genealogies having $RI = 1$ (the highest relinking). Figure 6 presents such situations.
  - Marriage between brother and sister ($n = 2, m = 2, k = 1, M = 1$)
  - Two brothers married to two sisters from another family ($n = 4, m = 4, k = 1, M = 2$)
  - A more complicated situation ($n = 9, m = 12, k = 1, M = 3$)
- Arbitrary large genealogies with $RI = 1$ exist.

Often, we determine the relinking index for the largest biconnected component in a given genealogy (see the last rows in Table 3).

Relinking Patterns in p-Graphs

In Figure 7, all possible relinking marriages in p-graphs containing from 2 to 6 vertices are presented (subtypes and variants as to sex are not included). Patterns are labeled in the following way:

- First character: A is a pattern with a single first vertex (vertex without incoming arcs), B is a pattern with two first vertices, and C is a pattern with three.
- Second character: number of vertices in pattern (2, 3, 4, 5, or 6).
- Last character: identifier (if the first two characters are identical).

Table 3
Comparison of Genealogies According to Distribution of Patterns

Pattern	Loka	Silba	Ragusa	Tur	Royal	Total
A2	1	0	0	0	0	1
A3	1	0	0	0	3	4
A4.1	12	5	3	65	21	106
B4	54	25	21	40	7	147
A4.2	0	0	0	0	0	0
A5.1	9	7	4	15	13	48
A5.2	0	0	0	0	0	0
B5	19	11	47	19	8	104
A6.1	28	28	2	65	13	140
A6.2	0	2	0	0	1	3
A6.3	0	0	0	0	0	0
C6	10	12	19	15	5	61
B6.1	0	1	2	0	0	3
B6.2	27	39	63	54	12	194
B6.3	47	30	82	46	13	218
B6.4	0	0	53	0	8	
Blood marriages (total A)	51	42	9	149	51	302
Nonblood marriages (total B and C)	157	118	239	176	45	735
Number of individuals (Ore graph)	47,956	6,427	5,999	1,269	3,010	
Number of vertices (p-graph)	35,228	4,480	4,376	956	2,141	
Number of couples (p-graph)	14,154	2,217	2,002	407	1,138	
Number of biconnected components (p-graph)	29	4	2	3	5	
Size of largest biconnected component	4,095	1,340	1,446	250	435	
RI (largest biconnected component)	0.55	0.78	0.74	0.75	0.37	

It is easy to see that patterns denoted by *A* are exactly the blood marriages. All others are nonblood marriages. Also, in every pattern, the number of first vertices (vertices with property $\text{indeg}(v) = 0$) equals the number of last vertices (vertices with property $\text{outdeg}(v) = 0$).

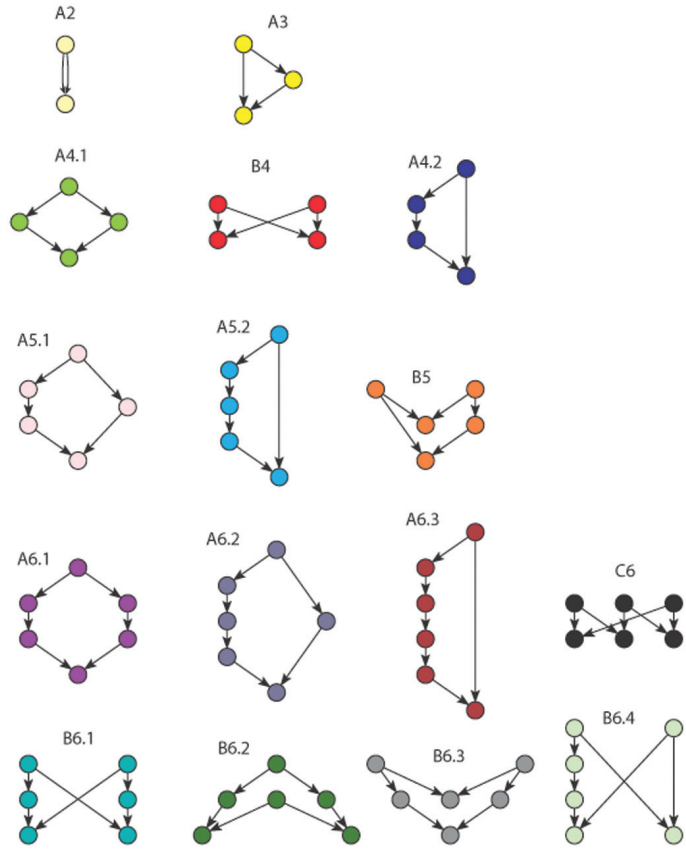
In Pajek, searching for relinking marriages can be performed using general fragment searching, which was included in Pajek in June 1997. For this purpose, we define a fragment (e.g., one of the graphs in Figure 7) and execute the command for searching for all occasions of the fragment in selected genealogy. We can use Macro Language or Repeat Last Command to search for all fragments in Figure 7.

Comparing Genealogies

Using frequency distributions for different patterns, we can compare different genealogies. As examples, let us take five genealogies mentioned in Table 2. Frequency distributions are given in Table 3.

The number of individuals in the Tur genealogy is much lower than in others, the Silba and Ragusa genealogies are approximately the same size, and Loka is a much larger genealogy, all of which we must also take into account. We take this into account in Table 4 with normalized

Figure 7
Relinking Marriages (p-Graphs With 2 to 6 Vertices)



frequencies for number of couples in the p-graph × 1,000. It can be easily noticed that most of the relinking marriages happened in the genealogy of Turkish nomads, the second highest incidence is in the Ragusa genealogy, and relinking marriages in other genealogies are much less frequent.

Several other characteristics can be found by looking at Tables 3 and 4:

- Probability of generation jump for more than one generation is very low (patterns A4.2, A5.2, and A6.3 do not appear in any genealogy; pattern A6.2 appears twice in the Silba genealogy and once in the royal genealogy; and pattern B6.4 appears five times in Ragusa and three times in Tur).
- In Tur, there are many marriages of types A4.1 and A6.1 (marriages among grandchildren and great-grandchildren). Such marriages are allowed among nomads but not in the other four genealogies.

Table 4
Frequencies Normalized With Number of Couples in p-Graph X 1,000

Pattern	Loka	Silba	Ragusa	Tur	Royal
A2	0.07	0.00	0.00	0.00	0.00
A3	0.07	0.00	0.00	0.00	2.64
A4.1	0.85	2.26	1.50	159.71	18.45
B4	3.82	11.28	10.49	98.28	6.15
A4.2	0.00	0.00	0.00	0.00	0.00
A5.1	0.64	3.16	2.00	36.86	11.42
A5.2	0.00	0.00	0.00	0.00	0.00
B5	1.34	4.96	23.48	46.68	7.03
A6.1	1.98	12.63	1.00	169.53	11.42
A6.2	0.00	0.90	0.00	0.00	0.88
A6.3	0.00	0.00	0.00	0.00	0.00
C6	0.71	5.41	9.49	36.86	4.39
B6.1	0.00	0.45	1.00	0.00	0.00
B6.2	1.91	17.59	31.47	130.22	10.54
B6.3	3.32	13.53	40.96	113.02	11.42
B6.4	0.00	0.00	2.50	7.37	0.00
Sum	14.70	72.17	123.88	798.53	84.36

- For all genealogies, the number of relinking "nonblood" marriages (e.g., patterns B4, B5, C6, B6.1, B6.2, B6.3, and B6.4) is much higher than the number of blood marriages (see the middle part of the table). That is especially true for Ragusa, where for "critical" marriages, special permission from the pope was needed. There were also economic reasons for nonblood relinking marriages: to keep the wealth and power within selected families.

Overall patterns of kinship relations reflect cultural norms for marriage: Who are allowed to marry? Property is handed down from one generation to the next along family ties, so marriages may serve to protect or enlarge the wealth of a family; family ties parallel economic exchange (de Nooy, Mrvar, & Batagelj, 2005).

In Figure 8, an example of nonblood relinking marriage in Ragusan nobility genealogy is shown. In this case, one couple (Junius Zrieva and Margarita Bona) belongs to three relinking marriages of type B4 (brothers and sisters exchanging partners from the same families).

In Figure 9, an example of two connected blood relinking marriages is shown. In this case also, generation jumps are present.

Using p-graphs, we cannot distinguish persons married several times from those married once. In this case, we must use *bipartite p-graphs*.

Using bipartite p-graphs, we can find marriages between half-brothers and half-sisters (as seen in the pattern on the left side of Figure 10). In the five genealogies, we found only one such example in Royal.ged (see the right side of Figure 10).

There exist marriages between half-cousins (Figure 11, left). We found one such marriage in the Loka genealogy (right side of Figure 11) and four in the Turkish genealogy.

Figure 8
Three Connected Nonblood Relinking Marriages B4 in Ragusa.ged

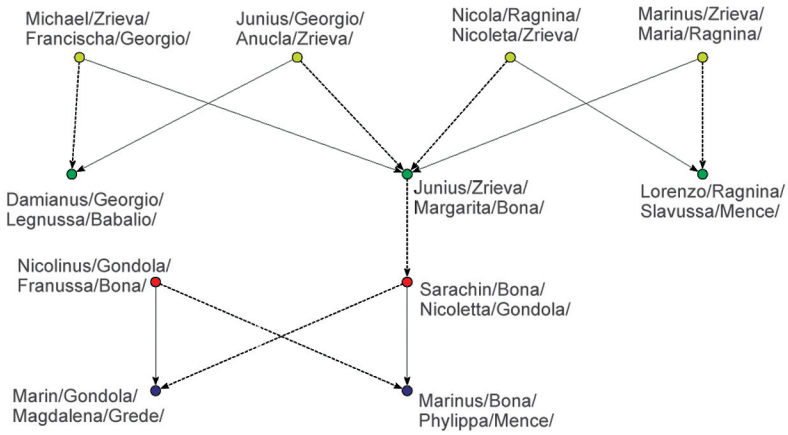
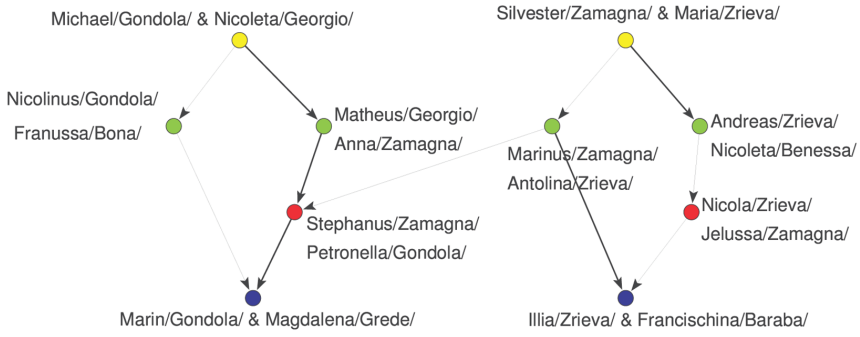


Figure 9
Two connected Blood Relinking Marriages A5.1 in Ragusa.ged



Network Multiplication

To a simple two-mode network $N = (I, J, E, w)$, where I and J are sets of *vertices*, E is a set of *edges* linking I and J, and w: $E \to \mathbb{R}$ is a *weight*, we can assign a *network matrix* $\mathbf{W} = [w_{ij}]_{I \times J}$ with the following elements: $w_{ij} = w(i, j)$ for $(i, j) \in E$, and $w_{ij} = 0$ otherwise.

Given a pair of compatible networks $N_A = (I, K, E_A, w_A)$ and $N_B = (K, J, E_B, w_B)$ with corresponding matrices $\mathbf{A}_{I \times K}$ and $\mathbf{B}_{K \times J}$, we call a *product* of networks N_A and N_B the network $N_C = (I, J, E_C, w_C)$, where $E_C = \{(i, j): i \in \boldsymbol{I}, j \in J, c_{ij} \neq 0\}$ and $w_C(i, j) = c_{ij}$ for $(i, j) \in E_C$. The product matrix $\mathbf{C} = [c_{ij}]_{I \times J} = \mathbf{A}^*\mathbf{B}$ is defined in the standard way:

$$c_{ij} = \sum_{k \in K} a_{ik} \cdot b_{kj}$$

Figure 10
Bipartite p-Graphs: Marriage Between Half-Brother and Half-Sister (Left), and Example of Such Marriage (Right)

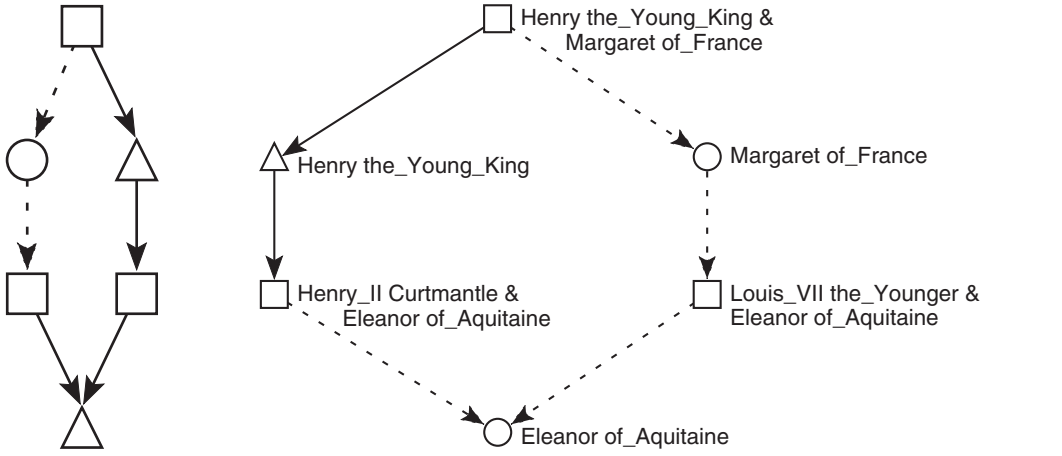
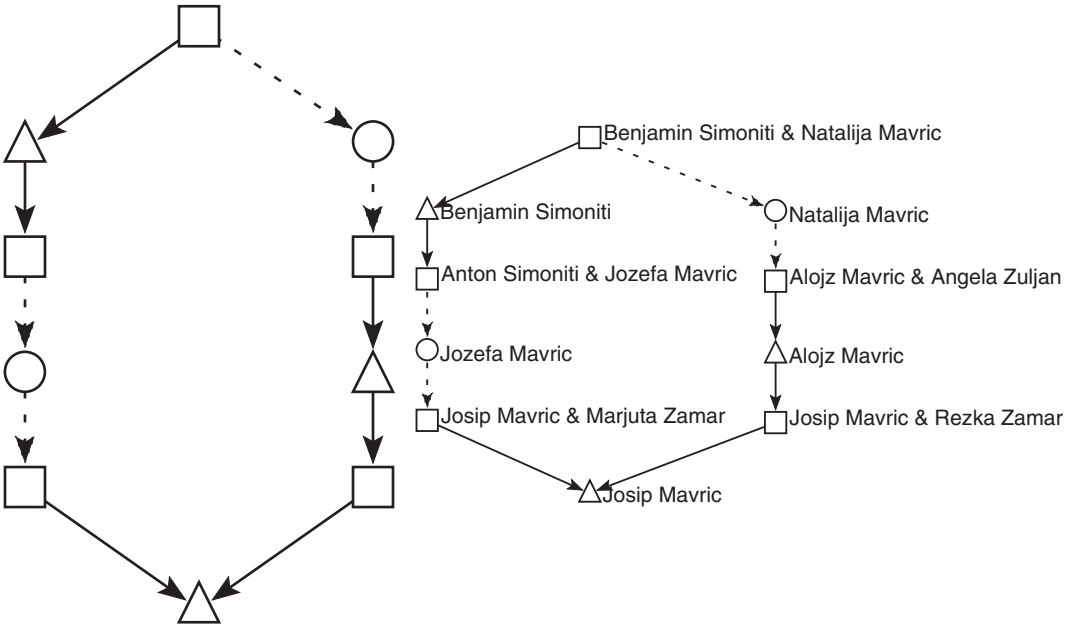


Figure 11
Bipartite p-Graphs: Marriage Among Half-Cousins (Left), and Example of Such Marriage (Right)



In the case when $I = K = J$, we are dealing with ordinary one-mode networks (with square matrices). In the case of large sparse networks, the main problem with the product is that it needs not to be sparse itself. It is easy to prove that if at least one of the sparse networks N_A and N_B has a small maximum degree on K, then also the resulting product network N_C is sparse and can be efficiently computed. For details about fast sparse network multiplication, see Batagelj and Mrvar (in press). The fast sparse network multiplication was included in Pajek in April 2005.

Basic Kinship Types

Anthropologists typically use a basic vocabulary of kin types to represent genealogical relationships. One common version of the vocabulary for basic relationships (Fischer, 2005) is given in Table 5. At the bottom of the table, some derived relations are added (uncle, aunt, semisibling, grandparent, grandfather, and niece). The last three columns show additional properties of some relations (symmetric, transitive, and acyclic relation). In the table, a different character (○) is used for an "almost transitive relation," a relation which is transitive if the unit relation is added to it.

See also a song, "My Own Granpa" (Richards, 2001).

Calculating Kinship Relations

Pajek generates three relations when reading genealogy as an Ore graph:

- **M**: *is a mother of*
- **F**: *is a father of*
- **E**: *is a spouse of*

Table 5
Basic and derived kinship types and their properties

KinType	Symmetric	Transitive	Acyclic
P (parent)			•
F (father)			•
M (mother)			•
C (child)			•
D (daughter)			•
S (son)			•
G (sibling)	•	○	
Z (sister)		○	
B (brother)		○	
E (spouse)	•		
H (husband)			•
W (wife)			•
U (uncle)			•
A (aunt)			•
Ge (semi-sibling)	•	○	
gP (grandparent)			•
gF (grandfather)			•
Ni (niece)			•

To compute all other kinship relations, we additionally need two binary diagonal relations to distinguish between male and female:

- **J**: *female* (1: female; 0: male)
- **L**: *male* (1: male; 0: female)

Other *basic* relations can be obtained from relations **M**, **F**, **E**, **J**, and **L** by running given macros which perform the following network operations (most of them include network multiplication):

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

Several *derived* relations can be computed, for example:

- *is an aunt of* ($\mathbf{A} = \mathbf{Z} * \mathbf{P}$)
- *is an uncle of* ($\mathbf{U} = \mathbf{B} * \mathbf{P}$)
- *is a semisibling of* ($\mathbf{Ge} = (\mathbf{P}^T * \mathbf{P}) \setminus \mathbf{I}$)
- *is a grandparent of* ($\mathbf{gP} = \mathbf{P}^2$)
- *is a grandfather of* ($\mathbf{gF} = \mathbf{F} * \mathbf{P} = \mathbf{L} * \mathbf{gP}$)
- *is a niece of* ($\mathbf{Ni} = \mathbf{D} * \mathbf{G}$)

The macros mentioned are available in Pajek distribution. After loading a genealogy as an Ore graph, we run a selected macro (e.g., *is an uncle of*) and obtain as a result a network with the new relation (uncle) added to the list of already existing relations (by reading, only the relations *spouse*, *father*, and *mother* are generated).

Sizes of Kinship Relations in Genealogies

As an example, we took the five genealogies mentioned in previous sections and computed the sizes of their kinship relations (Table 6). We added the number of individuals in the bottom row. To make comparison easier, we normalized the numbers by the cardinality of the parent (or child) relation. The result is shown in Table 7. We can see that all obtained relations are sparse. The densest relation is uncle, but still its density is less than two times that of the parent relation.

Other Analyses

People collecting data about their genealogies are interested in several other "standard" analyses. Let us look at some other analyses that can be performed in Pajek and give us some interesting results. Bear in mind, however, it is true that some of these analyses are interesting only from the perspective of individuals collecting the data.

Table 6
Sizes of Kinship Relations in Genealogies

Relation	Loka	Silba	Ragusa	Tur	Royal
P (parent)	68,052	9,627	9,315	1,987	3,724
M (mother)	33,722	4,629	4,359	965	1,714
F (father)	34,330	4,998	4,956	1,022	2,010
C (child)	68,052	9,627	9,315	1,987	3,724
D (daughter)	32,647	4,518	3,577	857	1,589
S (son)	35,405	5,109	5,738	1,130	2,135
G (sibling)	69,347	7,803	8,782	2,485	2,858
Z (sister)	66,874	7,314	6,949	2,256	2,634
B (brother)	71,820	8,292	10,615	2,714	3,082
E (spouse)	14,154	2,217	2,002	407	1,138
W (wife)	14,154	2,217	2,002	407	1,138
H (husband)	14,154	2,217	2,002	407	1,138
A (aunt)	80,995	10,564	10,644	3,477	2,973
U (uncle)	81,695	11,372	16,665	3,816	3,453
Ge (semisibling)	76,746	8,972	10,763	2,926	3,372
Number of individuals	47,956	6,427	5,999	1,269	3,010

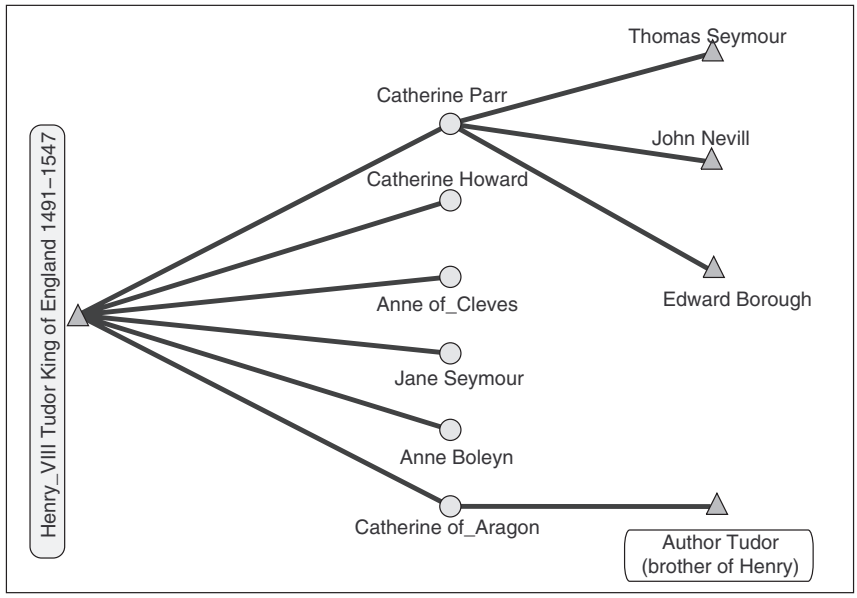
Table 7
Normalized Sizes of Kinship Relations in Genealogies

Relation	Loka	Silba	Ragusa	Tur	Royal
P (parent)	1.000	1.000	1.000	1.000	1.000
M (mother)	0.496	0.481	0.468	0.486	0.460
F (father)	0.504	0.519	0.532	0.514	0.540
C (child)	1.000	1.000	1.000	1.000	1.000
D (daughter)	0.480	0.469	0.384	0.431	0.427
S (son)	0.520	0.531	0.616	0.569	0.573
G (sibling)	1.019	0.811	0.943	1.250	0.767
Z (sister)	0.983	0.760	0.746	1.135	0.707
B (brother)	1.055	0.861	1.140	1.366	0.828
E (spouse)	0.208	0.230	0.215	0.205	0.306
W (wife)	0.208	0.230	0.215	0.205	0.306
H (husband)	0.208	0.230	0.215	0.205	0.306
A (aunt)	1.190	1.097	1.143	1.750	0.798
U (uncle)	1.200	1.181	1.789	1.920	0.927
Ge (semisibling)	1.128	0.932	1.155	1.473	0.905

Tracking changes in relinking patterns over time would give us insight whether the rules regarding "what is allowed and what is not" in different cultures are changing over time.

Special situations (outliers) can be found very easily, for example individuals married several times and individuals having the highest number of children. In some genealogies, we can find a lot of interesting multiple marriages. Figure 12 shows several multiple marriages in Royal.ged. We can

Figure 12
Several Multiple Marriages in Royal.ged



see that Henry VIII was married six times, and one of his wives (Catherine Parr) was married four times. Henry VIII had been betrothed to his brother's (Arthur Tudor) widow, Catherine of Aragon.

For large genealogies, some paths to famous people are "a wish" for several individuals. Checking whether two selected individuals are relatives and searching for the shortest genealogical connection between them can be performed using a simple shortest path search.

Searching for all ancestors or descendants of a selected person and searching for a person with the highest number of known ancestors or descendants are other easy tasks for network analysts.

Simple statistics, like the highest difference in age between husband and wife, the oldest or youngest person at the time of marriage, and the oldest or youngest person at the time of a child's birth, can also be found easily.

Searching for the longest matrilineage and especially patrilineage is important to find families with a long tradition, because family names are the father's surname in most Western societies.

Finally, we must say that often the special situations which we find in genealogies are just the result of errors made in data entry. In this case, we can still consider the results of analysis useful, namely, as a data consistency check.

Conclusion

Social network analysis turns out to be very useful in the research of genealogies. In the article, three different representations of kinship data were discussed: the Ore graph, p-graph, and bipartite p-graph. Several interesting results in large genealogies can be found by just using standard network analysis approaches, for example shortest paths, network multiplications, and fragment searching. For each application, suitable representation should be selected. In the article, we demonstrated that

p-graphs are more suitable for searching for relinking patterns, whereas Ore graphs are more suitable for computing additional kinship relations using network multiplication. Because some genealogies can be very large networks, only fast (i.e., subquadratic) algorithms can be used. Such algorithms have been developed and included in the program Pajek. Pajek was used to perform all calculations done in this article. It runs on Windows and is free for noncommercial use. Program and data can be obtained from its webpage (Batagelj & Mrvar, 2007).

References

Batagelj, V. (1996). Ragusan families marriage networks. In *Developments in Data Analysis, Metodološki zvezki* (Vol. 12, pp. 217–228). Ljubljana, Slovenia: FDV.

Batagelj, V., & Mrvar, A. (2007). Pajek. Retrieved February 26, 2007, from http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Batagelj, V., & Mrvar, A. (In press). Fast sparse matrix multiplication with Pajek. Submitted.

Beauregard, D. (1995). Bouchard genealogy. Retrieved February 26, 2007, from http://archiver.rootsweb.com/th/read/GEN-FF/1995-10/0813810937; http://ourworld.compuserve.com/homepages/lwjones/bouchard.htm; and http://www.genealogiequebec.info/testphp/info.php?no=25342.

Coonce, H.B. (1999) Mathematics Genealogy Project. http://genealogy.math.ndsu.nodak.edu/

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. New York: Cambridge University Press.

Dremelj, P., Mrvar, A., & Batagelj, V. (1999). Rodovnik dubrovniških plemiških družin med 12. in 16. stoletjem. *Drevesa, Bilten slovenskega rodoslovnega društva*, *1*, 4–11.

Dremelj, P., Mrvar, A., & Batagelj, V. (2002). Analiza rodoslova dubrovačkog vlasteoskog kruga pomoću programa Pajek. *Anali Zavoda povij. znan. Hrvat. akad. znan. Umjet, Dubr. 40*, 105–126.

Family History Department. (1996). Standard GEDCOM. Retrieved February 26, 2007, from http://homepages.rootsweb.com/ ~ pmcbride/gedcom/55gctoc.htm.

Fischer, M. D. (2005). Representing anthropological knowledge: Calculating kinship. Retrieved February 26, 2007, from http://www.era.anthropology.ac.uk/Era_Resources/Era/Kinship/prologTerm2.html.

Franklin.ged. (2000, November). Retrieved February 26, 2007, from http://www.roperld.com/gedcom/.

Hamberger, K., Houseman, M., Daillant, I., White, D. R., & Barry, L. (2005). Matrimonial ring structures. *Mathematiques et sciences humaines*, *168*, 83–121. Retrieved February 26, 2007, from http://www.ehess.fr/revue-msh/recherche_gb.php?numero = 168.

Hawlina, P. (2004): Slovenian genealogical society. Retrieved February 26, 2007, from http://genealogy.ijp.si/slovrd/rd.htm.

Johnson, D. S., & Parberry, I. (1993). Theoretical computer science genealogy. Retrieved February 26, 2007, from http://sigact.acm.org/genealogy/.

Krivošić, S. (1990). *Stanovništvo Dubrovnika i demografske promjene u prošlosti*. Dubrovnik, Croatia: Zavod za povijesne znanosti JAZU u Dubrovniku.

Mahnken, I. (1960). *Dubrovački Patricijat u XIV Veku*. Beograd, Yugoslavia: Naučno delo.

Mrvar, A., & Batagelj, V. (1997). Pajek: program za analizo obsežnih omrežij. Uporaba v rodoslovju. *Drevesa, Bilten slovenskega rodoslovnega društva*, 4–6.

Mrvar, A., & Batagelj, V. (2004). Relinking marriages in genealogies. *Metodološki zvezki: Advances in Methodology and Statistics*, *1*, 407–418.

MyFamily.com. (2004). Mormon genealogy. Retrieved February 26, 2007, from http://www.familytreemaker.com/00000116.html.

Ore, O. (1963). *Graphs and their uses*. New York: Random House.

Richards, B. (2001). Nutworks: My Own Granpa. *Connections*, 24(2), terminal notes.

Royal Genealogies. (1992). Royal genealogies. Retrieved February 26, 2007, from http://ftp.cac.psu.edu/ ~ saw/royal/royalgen.html.

Tompsett, B. (2004): American presidents GEDCOM file. Retrieved February 26, 2007, from http://www3.dcs.hull.ac.uk/public/genealogy/presidents/gedx.html.

White, D. R., Batagelj, V., & Mrvar, A. (1999). Analyzing large kinship and marriage networks with p-graph and Pajek. *Social Science Computer Review: SSCORE*, *17*, 245–274.

White, D. R., & Johansen, U. C. (2004). *Network analysis and ethnographic problems: Process models of a Turkish nomad clan*. Lanham, MD: Lexington.

White, D. R., & Jorion, P. (1992). Representing and computing kinship: A new approach. *Current Anthropology*, *33*, 454–462.

White, D. R., & Jorion, P. (1996). Kinship networks and discrete structure theory: Applications and implications. *Social Networks*, *18*, 267–314.

Vladimir Batagelj is professor of discrete and computational mathematics at the University of Ljubljana. He is a chair of the Department of Theoretical Computer Science, IMPM, Ljubljana. His main research interests are in mathematics and computer science: combinatorics with emphasis on graph theory, algorithms on graphs and networks, combinatorial optimization, algorithms and data structures, cluster analysis, visualization, and applications of information technology in education. With Andrej Mrvar, he has developed since 1996 a software program, Pajek, for the analysis and visualization of large networks. With coauthors, he recently published two books: *Generalized Blockmodeling* and *Exploratory Social Network Analysis With Pajek* (both Cambridge University Press, 2005). Address: Vladimir Batagelj, Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia. E-mail: vladimir.batagelj@fmf.uni-lj.si; http://vlado.fmf.uni-lj.si.

Andrej Mrvar finished his Ph.D. in computer science at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. He is associate professor of social science informatics at Faculty of Social Sciences. He has won several awards for graph drawings at competitions between 1995 and 2005. Since 2000, he has edited the statistical journal *Metodološki zvezki: Advances in Methodology and Statistics*. He is one of the coauthors of the software program Pajek (with Vladimir Batagelj) and one of the coauthors of the book *Exploratory Social Network Analysis With Pajek* (Cambridge University Press, 2005). Address: Andrej Mrvar, Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia. E-mail: andrej.mrvar@fdv.uni-lj.si; http://mrvar.fdv.uni-lj.si.