Bhattacharya, Kunal; Berg, Venla; Ghosh, Asim; Monsivais, Daniel; Kertész, János; Kaski, Kimmo; Rotkirch, Anna

Network of families in a contemporary population

Published in:
EPJ Data Science

DOI:
10.1140/epjds/s13688-018-0137-9

Published: 01/12/2018

Document Version
Publisher's PDF, also known as Version of record

Bhattacharya et al. *EPJ Data Science* (2018) 7:9
https://doi.org/10.1140/epjds/s13688-018-0137-9

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Network of families in a contemporary population: regional and cultural assortativity

Kunal Bhattacharya[1*], Venla Berg[2], Asim Ghosh[1], Daniel Monsivais[1], János Kertész[1,3], Kimmo Kaski[1] and Anna Rotkirch[2]

*Correspondence:
kunal.bhattacharya@aalto.fi
[1]Department of Computer Science, School of Science, Aalto University, Aalto, Finland
Full list of author information is available at the end of the article

Abstract

Using a large dataset with individual-level demographic information of almost 60,000 families in contemporary Finland, we analyse the regional variation and cultural assortativity by studying the network between families and the network between kins. For the network of families the largest connected component is found to consist of around 1000 families mostly originated from one single region in Western Finland. We characterize the networks in terms of the basic structural properties. In particular, we focus on the *k*-cores and the presence of transitive triangles. Clustering in the networks is found to result from homophily by language and religious affiliations. The large network fragments appear to be small-worlds. We also compare the fragments in the kin network with respect to the average coefficient of relationship. The measures of assortativity are able to distinguish the families in terms of their regions of origin. Overall, we distinguish between two patterns of regional effects, the 'metropolitan' and the 'cultural' pattern.

Keywords: Complex networks; *k*-core; Transitivity; Kinship; Homophily

1 Introduction

Human families include parents, children, grandchildren and lasting pair bonds between usually unrelated spouses, so that families typically encompass at least three family generations and two kin lineages [1]. This complexity of familial ties allows for various kinds of associations between different extended families, for example, through marriage and intermarriage within a kin group. Family members usually help each other by providing emotional, practical and financial support [2]. Even in the cases of contemporary wealthy and globalised societies it is known that parents, children, grandchildren and siblings stay close to each other [3], which can lead to genetic homogeneity in certain geographical regions. At the same time, the effective fertility differences and migration can affect the patterns of familial ties in the social structure [4].

Here we study a network of families, using a unique and nationally representative register dataset from contemporary Finland. We investigate the overall network characteristics of the connected components as well as the roles played by the following factors: (i) spatial proximity, as measured at a regional level, (ii) language preferences, as indicated by language (Finland has two official languages: Finnish and Swedish), (iii) genetic relatedness, as measured through assumed biological relatedness for kinships that result from the net-

work. Focussing on the connected components we identify which geographical regions most of the observed network stems from. We investigate "clustering" in the network by examining transitivity within the largest connected components by type of spoken language. We also investigate the influence of the structure of the network of families on the patterns of biological relatedness by studying a network of kins derived from the network of families.

Our approach relies on constructing a network by joining a large number of independent genealogical networks using the information on marriage ties. In studying the network of families we treat the genealogical networks for extended families as the units, whereas in the case of kinship we examine a network between individuals. In both these cases we study the structures in terms of quantities that are generally used in characterizing social and technological networks [5]. Several studies in the past with a wide range of perspectives have analyzed kinship data and networks obtained from historical records [6]. There has been a series of studies on statistical properties related to finding of common ancestors and distribution of family names [7–9]. Several studies rooted in ethnography have focused on the precise role of marriage ties and migration on the reshaping of network properties like distances, presence of cyclic paths and community structure [4, 10–12]. These studies have also discussed the role of network structure in the generation of conflict or cooperation among different communities. Genealogical networks have also been extensively used in genetic studies focusing on the issues of fertility and genetic diseases, among other topics [13–19].

2 Data

The FINNFAMILY dataset [20] is a nationally representative and anonymised dataset of multiple generations of individuals of the late 20th century population of Finland, a country with around 5.5 million inhabitants, derived from the National Population Register of Finland through Statistics Finland [21]. The data contains information on extended families of randomly selected Finns (index persons) from six birth cohorts (1955, 1960, 1965, 1970, 1975, and 1980). From each cohort, 10,000 index persons are selected so that altogether there is information on 60,000 extended families. As shown in Fig. 1, a single extended family in the dataset consists of the following four generations: the zeroth generation comprising of mothers and fathers of the index persons; the first generation comprising of index persons and their siblings and half-siblings; the second generation comprising of the children; and the third generation comprising of the grandchildren. In the case of half-siblings, the data includes the half-sibling's other parent, either mother or father (randomly selected, to avoid including two half-siblings that are not genetically related). After merging the families for which the index person is a member of another family, we were left with 48,750 different families. Aggregated over all the extended families the dataset consists altogether of around 700,000 individuals.

For each individual, the data has demographic information including the date of birth (actually, month and year), the place of birth (administrative regions called "Maakunta" in Finnish), the date of death, the date of marriage and divorce, and the yearly information of the region of residence. The currently demarcated administrative regions in Finland (Fig. 2) exhibit a substantial degree of cultural and economic similarity including recognizable regional dialects, symbols and local food traditions [22]. For our analysis we consider here 18 out of the 19 regions, excluding Ahvenanmaa (the Åland Islands) region owing to its small population and being separated from the mainland of Finland.

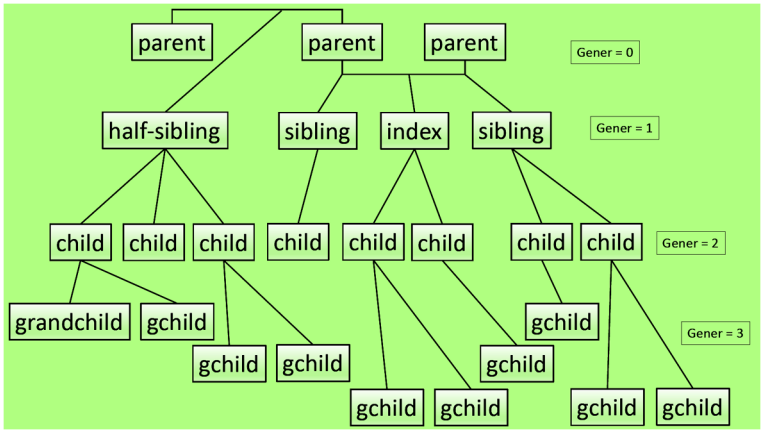
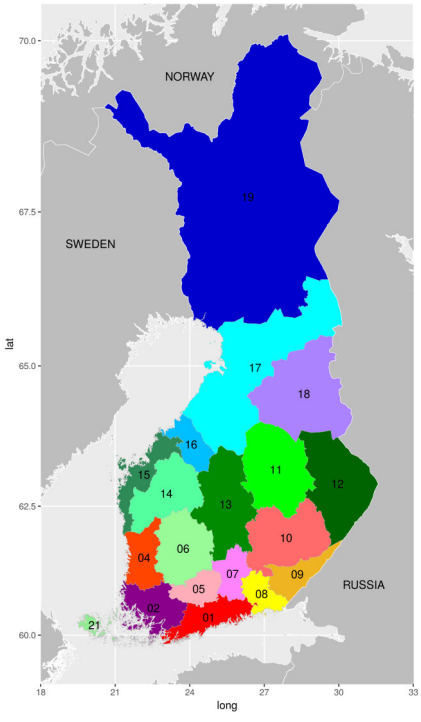


Figure 1 A typical genealogical network in the dataset for a single extended family. The structure of an extended family is shown and it consists of the four generations: the zeroth generation comprising of the parents and a step-parent of the index person; the first generation comprising of the index person and his or her siblings and half-siblings; the second generation comprising of the children; and the third generation comprising of the grandchildren



Code	Region
01	Uusimaa
02	Varsinais-Suomi
04	Satakunta
05	Kanta-Häme
06	Pirkanmaa
07	Päijät-Häme
08	Kymenlaakso
09	Etelä-Karjala
10	Etelä-Savo
11	Pohjois-Savo
12	Pohjois-Karjala
13	Keski-Suomi
14	Etelä-Pohjanmaa
15	Pohjanmaa
16	Keski-Pohjanmaa
17	Pohjois-Pohjanmaa
18	Kainuu
19	Lappi
21	Ahvenanmaa

Figure 2 Map showing the 19 different administrative regions ("Maakunta") of Finland. The names are provided on the right with the corresponding codes [21]. The numbers missing (03 and 20) from the sequence for codes correspond to regions that were historically merged with other neighbouring regions. The colouring scheme used above to denote the regions is the same as the one used to denote the nodes in the network of families based on regions of origin

A few regions stand out historically and culturally: Uusimaa is the region of Finland's capital Helsinki, the largest urban settlement in the country: 30 per cent of the total population lives in Uusimaa. The former capital Turku is now the third largest urban settlement

and is located in the Varsinais-Suomi region, while the second largest urban settlement Tampere is situated in the region of Pirkanmaa. The geographical region in the central and Western Finland (comprising the various Pohjanmaa regions) is known for a history of agriculture and entrepreneurship, and also for its comparatively high fertility. A religious sect within the Protestant church, called the Laestadians lives in this area, especially in the Pohjois-Pohjanmaa region. Laestadians do not approve of modern contraception, which has contributed to a larger proportion of large families with four or more children both among the members of this sect and also among their non-Laestadian neighbours [23]. Also many regions and especially the Northern and Eastern regions of Finland have witnessed emigration to the Southern regions, especially to Uusimaa region [24]. Finally, Finland has a national minority of Swedish-speaking Finns, comprising around 6 per cent of the total population. Swedish-speaking Finns are typically living in the Western and coastal areas of the country.

3 Methods

From the set of individual genealogical networks for extended families we construct two types of networks, one between families and the other between kins. We construct the network of families using the data of the extended families of the index persons. A node in this network is a 'family' comprising of an index individual and his or her parents, full siblings and half-siblings, children and grandchildren, reflecting the four family generations in our data as described above. A link in the network between two families is a 'parental union' between a male and a female having at least one common child through marriage or out of wedlock. We identify the links by searching for individuals that belong to multiple families. Note, that the presence of such a person in a given family would also mean the presence of either the mother or the father, with whom the person is related genetically. Once such a link was found we attributed the year of birth of the first offspring to this link. Together, the set of parental unions and the set of families constitute the network of families (see Additional file 1).

To each family we assign a region of origin. Since parental unions involve reproducing individuals, to assign regions we focus on the birth regions of those individuals who have children. We exclude the zeroth generation mothers and fathers, as by definition they belong to the same family. However, not all individuals in a given family have the same birth region. In cases when the birth regions of the reproducing individuals in a single family are extremely diverse, the assumptions with regard to the regional influences become weak. In contrast, the number of families where all the reproducing individuals are born in the same region is expected to be smaller. Therefore we calculate the number of families in which at least a fraction θ of its reproducing individuals were born in the same region (see Additional file 2). We choose $\theta = 0.6$ which allows to include 81% of the original number of families, while also fulfilling the criterion of having a large majority of the reproducing individuals in a given family being born in the same region. We assume that the region assigned to a family has over the time influenced the different generations at multiple levels including social, cultural and genetic inheritance, so that transitivity and assortativity may further intensify the cultural and genetic density.

We construct the network of kins or the kin network in the following way. A node in the network of kins represents a firstborn offspring resulting from a parental union belonging to the network of families. Thus, a node in the network of families having links to two

different families results into a pair of nodes in the network of kins that are linked via kinship. For example, in a case when two sisters from a given family marrying into two different families, the two firstborns become linked as first cousins (see Additional file 1). Assuming that all families are distinct in terms of the genetic material that is inherited by its members, we only include those kinship relations where the sets of families of the two individuals do not completely overlap. Thus, kinships such as the parent-offspring relationship are excluded from this study. A more general description of a kin network would require the inclusion of all the types of kinship [25]. To characterize the kin network we use the information on the assumed genetic relatedness or coefficient of relationship (r) that is attributed to each type of kinship. The coefficient of relationship is defined as the probability that two individuals share an allele due to having descended from a common ancestor, such that $r = 100\%$ for identical twins, and $r = 0$ for unrelated individuals [26, 27]. With our approach and the limits on the number of available generations, the types of kinship entering the study are, half-siblings ($r = 25\%$), first-cousins ($r = 12.5\%$), half-aunt/uncle-nephew/niece ($r = 12.5\%$), first-cousins-once-removed ($r = 6.25\%$), half-first-cousins ($r = 6.25\%$), second-cousins ($r = 3.125\%$), and half-second-cousins ($r = 1.5625\%$).

The index persons were chosen randomly from the whole population of Finland. In this sense, the network that we construct is a sampling of the actual network in place. However, the features that emerge from the sampled network seem to have resulted from the influences of diverse regional factors. In this sense the sampled network provides us with the "lower limits" on the structural characteristics present in the actual network.

4 Results

4.1 Connected components and regions of origin

First, we discuss our findings related to the structure of the network of families. Of the 48,750 families in the data, for 14,406 we could detect linkages to other families in the network. The total number of links (parental unions) is 9766. In Fig. 3(a) corresponding to each region of origin we show a comparison between the number of families that could be linked to those for which any linkage was not detected. The families with linkages produce a set of connected components with the largest connected component (LCC) being made up of 957 families and 1211 links. Figure 4 illustrates the first seven components, in descending order of size. We can see that the LCC is much larger than the following smaller components, with 957 nodes in the LCC compared to 68 nodes in the two components next in size. The distribution of size of the connected components (clusters) and degrees (number of connections for a given node) are shown in Fig. 3(b). The degree distribution decays very fast terminating at a maximum degree of 17. The distribution of clusters in the network is similar in shape but indicates the presence of the LCC and other larger clusters. The composition of the clusters can be analyzed to show how the families from any given region are distributed among them. For a given cluster size we plot the number of families from a given region. The Fig. 3(c) shows the results for four different regions. Interestingly, this figure reveals that within the LCC (extreme right on the domain) region 17 (Pohjois-Pohjanmaa) contributes most, followed by an almost equal presence of regions 01 (Uusimaa) and 19 (Lapland).

Thus the LCC is dominated by families from a particular region. Overall, for 13% of the families that have links, the region of origin is Pohjois-Pohjanmaa. This is a relatively large contribution since the largest percentage (14%) of the families belong to the Uusimaa region around the capital Helsinki. On the extreme left in Fig. 3(c), dominance by

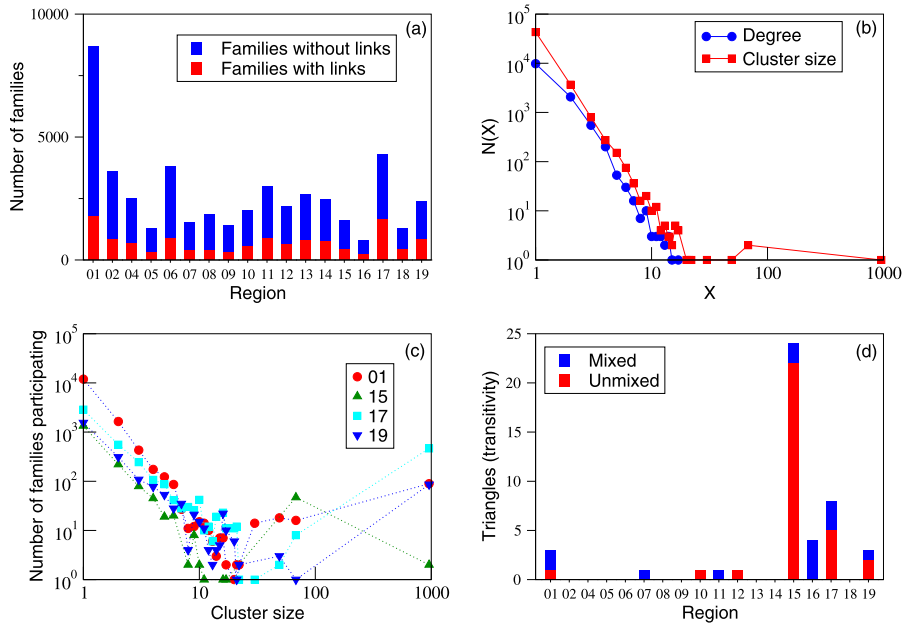


Figure 3 Characteristics of the network. (**a**) Number of families distributed over the 18 regions of mainland Finland. The plot shows the number of families that could be linked to other families in comparison with the total number of families. (**b**) The degree distribution of the network (circles) and the distribution of sizes of connected components or clusters (squares). (**c**) Participation in the different connected components by families from different regions (four cases illustrated). (**d**) Number of transitive triangles that each region participates in. In an 'Unmixed' case, the triangle is solely composed of families belonging to the particular region of origin; and in a 'Mixed' case the triangle comprises of families from different regions with at least one belonging to the particular region

the Uusimaa region can be observed in the case of smaller clusters including families that could not be linked. As observed from Fig. 3(a) the total number of families (with or without links) that originate from Uusimaa region constitute a much larger fraction of the total set of families when compared to any other region. This not surprising, since the Uusimaa region includes the capital area and it has the highest population size and density of all the regions of Finland.

4.2 Transitivity

Next, we investigate the presence of triangles that may reflect the transitivity with regard to family relations (see Additional file 3) [28]. In Fig. 3(d) we plot the number of different transitive triangles that families from each region participates in. Here, in addition to counting triangles where all the participating families belong to a single region, we also count triangles in which the participating families belong to different regions. The largest number of transitive triangles is found to occur in region 15 (Pohjanmaa). Expecting that the presence of transitive triangles would lead to an increase in the number of linkages in the neighbourhood of the corresponding nodes, we analyse the strongly connected parts of the network.

We perform a k-core decomposition [29] which extracts the tightly connected parts of the network (see the Appendix for the definition). For the LCC, we find that $k_{\max}$, the maximum value of the degree (k) for which a core exists is 3 (i.e. a family belonging to the core is connected to three or more families). Therefore, the full LCC with 957 nodes could be

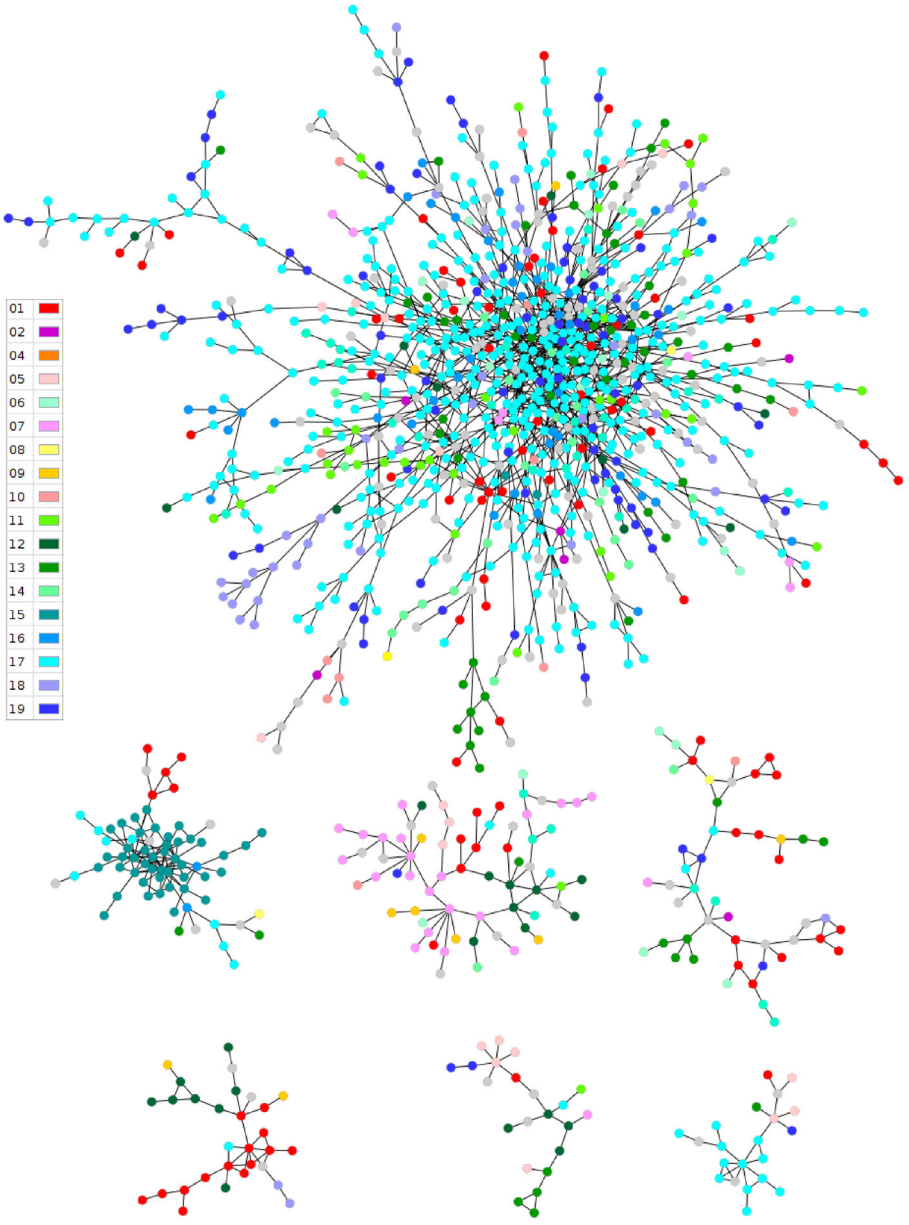


Figure 4 The network of Finnish families. Nodes represent families, links between two nodes are associations with offspring common to both families. The first seven connected components of the network are shown in descending order of sizes. The largest connected component (LCC) has 957 nodes and 1211 links. The next in the series have 68, 68, 49, 30, 22 and 21 nodes, respectively. The different colours represent different regions of origin for the families and are indicated in the legend. The families that could not be associated with a particular region are denoted in grey

partitioned into 3 shells. The outermost shell (each family in it has at least one connection) has 600 nodes, the second shell (each family in it has at least two connections) has 310 nodes, and the inner core has 47 nodes. While as a whole the LCC has an average degree of 2.5, the value at the core is 3.7. In addition, the concentration of families belonging to region 17 (Pohjois-Pohjanmaa) increases from the outermost shell to the inner core (outermost shell: 44%, second shell: 49%, core: 58%).

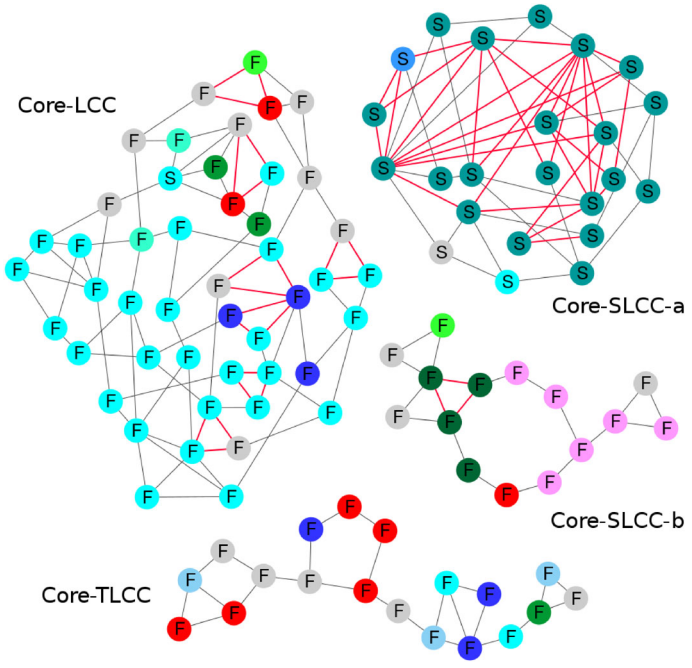


Figure 5 Presence of transitivity. Cores derived from the first four connected components in the network of families shown in Fig. 4. Links that are part of transitive triangles are coloured in red. The label on a node indicates the language that is spoken by the majority of the individuals in the family—Finnish ('F') or Swedish ('S'). Core-LCC: The $k_{\max} = 3$ core of the LCC ($N = 47$, $L = 86$). This subgraph is dominated by nodes from region 17 (Pohjois-Pohjanmaa). Core-SLCC-a: The $k_{\max} = 3$ core extracted from the first of the two second largest connected components ($N = 23$, $L = 55$). Most of the nodes in this subgraph belong to region 15 (Pohjanmaa). Core-SLCC-b: The $k_{\max} = 2$ core extracted from the second of the two second largest clusters ($N = 15$, $L = 19$). Core-TLCC: The $k_{\max} = 2$ core extracted from the third largest cluster ($N = 19$, $L = 22$)

We obtain the cores of the four largest components of the network, as depicted in Fig. 5. While the core of the LCC (Core-LCC) is dominated by families from region 17 (Pohjois-Pohjanmaa), the core of one of the two second largest clusters (Core-SLCC-a, $k_{\max} = 3$) is found to be composed of families mainly from region 15 (Pohjanmaa). The presence of a large number of transitive triads is observed in the latter core with links depicted in colour red. Such triangles are also present in the Core-LCC, but in a small number. The other two cores (Core-SLCC-b and Core-TLCC, having $k_{\max} = 2$) are relatively smaller in size, and transitive triangles are mostly absent.

To understand the possible reasons behind the simultaneous presence of such large number of triads in a specific subgraph (Core-SLCC-a) we probe the cultural similarities between the families. We labelled the nodes based on the language spoken by the majority of the individuals in such families. Results show (Fig. 5) that the Core-LCC is dominated by Finnish speakers, featuring only one family with a majority of Swedish speakers whereas, the Core-SLCC-b and the Core-TLCC have none at all. By contrast, the Core-SLCC-a has exclusively families with a majority of Swedish speakers. This indicates a very high degree of linguistic attraction within a population among both Finnish and Swedish speakers, and a higher degree of intermarriage and genetic relatedness among the Swedish-speaking cluster (Core-SLCC-a), as reflected in the frequency of triangles.

4.3 The network of horizontal kin ties

We study the pattern of genetic relatednesses within the available generations of the extended families by constructing the network of kins as described in Methods (Section 3). The fact that birth cohorts of the individuals in the network of kins are restricted so that around 80% of the links appear in the span of last 20 years and 50% within the last 10 years of the period of investigation, allows us to call this a network of horizontal kin ties (see Additional file 4).

First we extract the kin networks shown in Fig. 6 from the four cores shown in Fig. 5. The kin networks corresponding to the cores of the network of families reveal compositions very similar to the network of families themselves in terms of the birth regions of the individuals. The dense linking found between the families in the Core-SLCC-a is converted into a dense linking between kins in the Core-SLCC-a-kin. We characterize the kin networks by measuring the global clustering coefficient (CC) which provides a normalized measure for the frequency of transitive triangles in the kin network (this measure could not be used for the network of families due to the presence of triangles of two different types) [28]. We also calculate the average shortest path length (d) that characterizes the distance between two individuals measured on the kin network [30]. These are summarized in Table 1. Additionally, we provide the values corresponding to a Erdős–Rényi

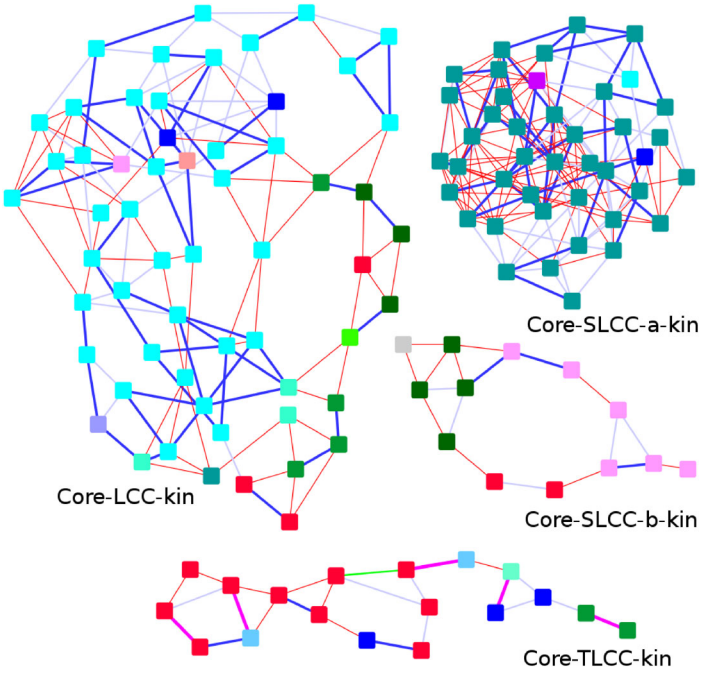


Figure 6 Network of kins. The subnetworks represent patterns of connection between individuals who are horizontal kin and are extracted from the network of families illustrated in Fig. 5. Each node in the above subgraphs is an individual (offspring) resulting from a link (parental union) in the subnetworks in Fig. 5. The colour of a node represents the birth region of the individual. The colour of a link indicates the nature of kinship—magenta: half siblings, blue: first-cousins, violet: first-cousins-once-removed, green: half-first-cousins, red: second-cousins. We denote the kin network corresponding to the core of the LCC as Core-LCC-kin, that corresponding to the core of first of the second largest connected components as Core-SLCC-a-kin, the kin network corresponding to the core of the other second largest connected component as Core-SLCC-b-kin, and the core for the third largest connected component as Core-TLCC-kin. The characterization of these subgraphs in terms of the usual topological parameters is provided in Table 1

Table 1 Parameters describing the structure of the subgraphs in the network of kins

Subgraph	N	$\langle k \rangle$	d	d/d_{random}	CC	CC/CC_{random}	$\langle r \rangle$	$\langle r_{\text{sum}} \rangle$
Core-LCC-kin	58	4.5	3.6	1.3	0.43	5.4	7.2	32.4
Core-SLCC-a-kin	42	8.6	2.0	1.2	0.53	2.5	5.7	49.0
Core-SLCC-b-kin	13	2.8	2.7	1.1	0.32	1.4	5.4	15.1
Core-TLCC-kin	18	2.6	3.8	1.2	0.21	1.4	10.0	26.0
LCC-kin	1052	5.5	6.9	1.7	0.54	108.0	7.2	39.6
SLCC-a-kin	93	8.1	3.0	1.4	0.57	6.5	6.3	51.0
SLCC-b-kin	67	4.1	4.9	1.7	0.46	7.7	6.8	27.9
TLCC-kin	44	3.4	5.0	1.6	0.53	6.6	8.1	27.5

Number of nodes (N), average degree ($\langle k \rangle$), average shortest path length (d), clustering coefficient (CC), the average relatedness on links ($\langle r \rangle$, in %), and the average aggregated relatedness for nodes ($\langle r_{\text{sum}} \rangle$). The average relatedness is calculated from the fact that each link corresponds to one of the following values (%) of assumed genetic relatedness: 25.0 for half siblings, 12.5 for first-cousins, 12.5 for half-aunt/uncle-niece/nephew, 6.25 for first-cousins-once-removed, 6.25 for half-first-cousins, 3.125 for second-cousins, and 1.5625 for half-second-cousins. We also provide ratios d/d_{random} and CC/CC_{random} where, d_{random} and CC_{random} are the average shortest path length and clustering coefficient for the Erdős–Rényi model having the same number of nodes and links.

model for random linkages with similar values for number of nodes and links [31]. We also use the information of the types of kinship to measure the average coefficient of relationship ($\langle r \rangle$) by summing over all the assumed genetic relatednesses (r) [26, 27] for all the links in a given subnetwork and then dividing by the total number of links in the subnetwork. We also provide $\langle r_{\text{sum}} \rangle = \langle k \rangle \langle r \rangle$, which is the average of aggregated genetic relatedness at nodes.

Among the four kin networks corresponding to the cores in the network of families, the CC appears to be the highest in the Core-SLCC-a-kin, and as such it results from the excess of transitive triangles observed in the Core-SLCC-a, composed of Swedish-speaking families. In the Core-SLCC-a-kin, in contrast to the other three, the fact that a random individual could be found linked to the highest number of close kins is evidenced from the high values of the average degree $\langle k \rangle$ and the average aggregated genetic relatedness $\langle r_{sum} \rangle$. Interestingly, the average relatedness in the network appears to be high in Core-TLCC-kin, due to the presence of half-sibling relationships. Under the criterion, $d/d_{\text{random}} \gtrsim 1$ and $CC/CC_{\text{random}} \gg 1$, all the four networks shown in Fig. 6 appear to be small-worlds in terms of structure [30].

Additionally, we include in the analysis the kin networks directly derived from the four largest clusters in the network of families without being restricted to their cores (see Additional file 5). Remarkably, for the LCC-kin, the kin network derived from the LCC in the network of families, which is far more larger in size compared to the LCC-core-kin, the small-world feature appears to be preserved if not enhanced as observed from the amplification of the ratio CC/CC_{random} with only marginal increase in the value of d/d_{random}.

In Fig. 7 we show the frequencies of different types of kinships that are found in the entire kin network. The relationships that are most frequent (around 30% in each case) are the first-cousin, first-cousin-once-removed and second-cousin. Relationships that mainly originate from family ties formed due to multiple marriages of individuals are present in smaller number. For each kind of relationship we also provide the fraction of cases where the kins are born in different regions. This fraction is found to increase as the tie strength (characterized by the genetic relatedness) decreases. A possible cause behind this is the migration of individuals from extended families into the different regions during several generations.

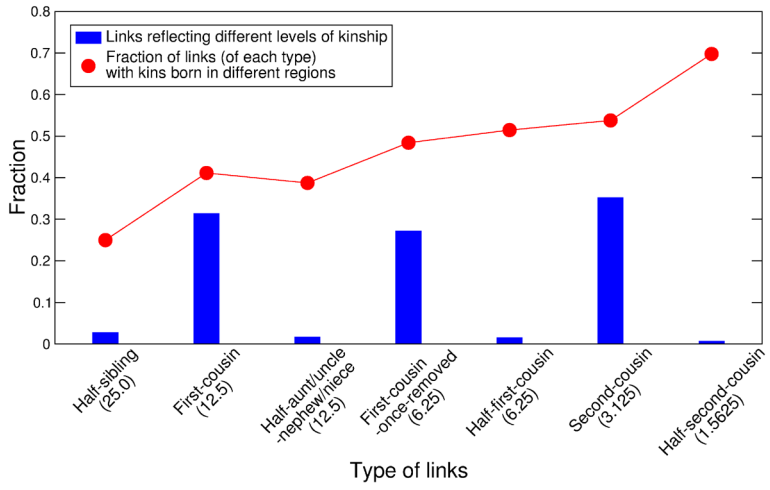


Figure 7 Nature of kin ties. The fraction of different types of links present in the network of kins is indicated by the bars. For a given type of kinship, the circles indicate the fraction of cases (out of the total in each case) when the individuals in the pair were born in different regions. The coefficients of relationship (%) are indicated inside the brackets

4.4 Assortativity

Finally, we characterize the network of families as well as the network of kins in terms of the assortativity coefficient. In general, this coefficient is employed to statistically characterize the nature of ties in networks [32]. For example, in a large social network where individuals are characterized by their age, a positive assortativity would indicate that people of comparable age associate with each other, while a negative assortativity would indicate the opposite. The assortativity coefficient (a) is defined such that it lies between -1 and 1. When $a = 1$, the network is perfectly assortative, and when $a = -1$, the network is called completely disassortative (see in the Appendix for details). In our case, we use the region of origin for the families to calculate a for the network of families and the birth regions of the individuals for the case of the kin network. We obtain, $a = 0.535 \pm 0.023$ for the network of families and $a = 0.277 \pm 0.008$ for the kin network. These values indicate the role played by spatial or geographical factors in structuring the network and overall it reflects the fact that individuals are more prone to marry within the same region. However, the different regions have their own characteristics, which we investigate in the following fashion. We estimate the influence of each region by using jackknife resampling [33]. For a given network and a given region, all the nodes belonging to the region are removed, and the assortativity coefficient is recalculated (a^*). This calculation is repeated for all the regions, and the mean $\langle a^* \rangle$ is found. Then the difference $a^* - \langle a^* \rangle$ for a particular region indicates the effect of removal of the nodes belonging to the region from the network (shown in Fig. 8). For example, when the value of $a^* - \langle a^* \rangle$ is significantly greater than zero, it indicates that the overall assortativity increases when the nodes are removed from the network. Also implying that the nodes themselves participate in disassortative mixing. This phenomena appears to be stronger in cases of region 01 (Uusimaa) and 17 (Pohjois-Pohjanmaa). A contrasting observation is found for the case region 15 (Pohjanmaa) where we detected the presence Swedish-speaking families.

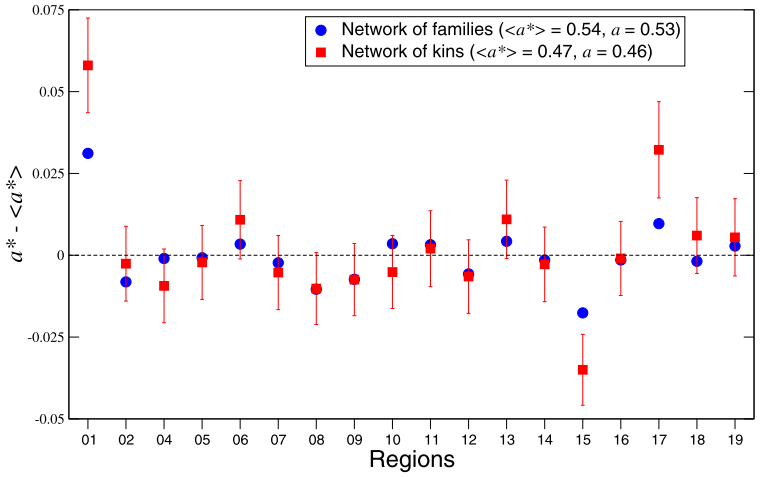


Figure 8 Assortativity coefficient. The plot shows the influence of the individual regions on the overall assortativity. For the network of families (circles) calculation is done using the region of origin of the families, and for the network of kins (squares), using the birth region of individuals. Here we estimate the influence of each region by using jackknife resampling [33]. For a given network, first the overall assortativity coefficient (a) is calculated using information on all the nodes (see Appendices). Then for a given region, all the nodes belonging to the region are removed, and the assortativity coefficient is recalculated (a^*). This calculation is repeated for all the regions, and the mean $\langle a^* \rangle$ is found. The difference $a^* - \langle a^* \rangle$ for a particular region (denoted on the horizontal axis) indicates the effect of removal of the nodes from the network. For example, a value of $a^* - \langle a^* \rangle$ being significantly greater than zero, indicates that the assortativity increases when the nodes are removed implying that the nodes themselves participate in disassortative mixing. For the kin network, the error bars in the quantity a^* are shown. For the network of families, we observe a clear correspondence of the values of $a^* - \langle a^* \rangle$ with the values in case of the kin network. However, the errors are much larger in magnitude (due to a lower link density) and therefore, are not shown in the figure

5 Summary and discussion

In this study we have investigated the patterns of families in a contemporary European population through a network constructed from data on extended families in Finland. We consider the families as nodes and the links result from joint parenting of children by individuals from the families. We have characterized the structural properties of the network of families and explored the transitivity and assortativity of the network with regards to the region of origin and linguistic identity. Using a large Finnish register data, we could link index persons with their parents, siblings, children, and grandchildren and further identified the links joining these extended families through marriage and reproduction. The results show that the sampled network is a collection of many disjoint components with the largest connected component including 8% of all linked families and 14% of all the detected parental unions.

As could be expected, the capital Helsinki and its surroundings in Uusimaa region (01), which has the highest population size and density of all the regions in Finland, dominates with regards to the frequency of families. The total number of families (with or without links) that originate from this region clearly constitute the largest fraction of the total set of families. However, in the case of the network composed of families linked with other families in the dataset, the pattern is different. Of families that have links to other families, the largest proportion (13%) originates from the region of Pohjois-Pohjanmaa (17). This region and its neighbouring regions are also known for their comparatively high fertility. The families with linkages produce a set of connected components with the largest

connected component being made up of 957 families and 1211 links. The number of individuals in the dataset which is almost 0.7 million is around 10% of the Finnish population that is 5.5 million. To our knowledge this is the first time the presence of one single connected network between families in such a representative population sample has been documented.

We also found that patterns of connectivity in the network are influenced by the regions in which the families are rooted. This finding is in line with a number of studies showing that the region of origin remains important for sociality of Europeans today [3]. Rates of internal migration are known to be higher in Finland and Scandinavia compared to Southern European countries [34], yet although a large proportion of Finns migrate to another region during the time of young adulthood, many eventually move back or closer to their region of origin once they have children themselves or after retirement [35]. Interestingly, the first and second largest connected components were predominantly populated by families rooted in a few specific regions. Furthermore, the kin networks corresponding to the cores of the four largest connected components were all dominated by one of the two national languages, Finnish or Swedish, the latter spoken by 6% of the total population but represented in much higher proportions. The fact that cultural homophily, in terms of religion and language, plays a major role becomes evident in our investigation of the presence of transitive relations between families. We found that the concentrated presence of a minority group of people with Swedish being their mother tongue, is reflected in the proliferation of triangles. Thus the majority of members in the families in the transitive core of the (one of the two) second largest connected component came from the region 15 (Pohjanmaa) and were Swedish-speaking. It is known that 40% of the Swedish-speaking population of Finland resides in this particular region. Furthermore, the kin cores of this particular connected component has the largest share of high degree nodes and clustering, as well as a higher estimate of the assumed genetic relatedness than the largest connected component has. The patterns revealed through the structure of the network is consistent with the genetic clustering found in the Swedish-speaking population of Pohjanmaa [36].

In general, each family has a number of reproducing individuals and some become part of the linkages observed in the sampled network when the family of the opposite sex partner is also present in the data. Therefore, under a simplistic description, the larger the family (and hence, the larger the number of reproducing individuals), the bigger the chance of this family to have a link. The plot of the number of links per family against the number of reproducing individuals for the different regions is shown in Fig. 9(a). Here we have taken into account all families, even those that could not be linked. This approach is expected to reduce the sampling bias. The linear correlation (ρ) is 0.85 and a fit suggests a linear relationship. As discussed, region 17 (Pohjois-Pohjanmaa) and its neighbours (15 (Pohjanmaa), 16 (Keski-Pohjanmaa), 18 (Kainuu) and 19 (Lappi)) on account of having a higher fertility are positioned towards the right on the horizontal axis. A critical view on contraception is likely to have contributed to this effect, which is a consequence of religious identities of families like being members of Lutheran sects such as the Laestadians in this region [23]. However, data about religious affiliation with this sect is not available from population registers. In contrast to this, region 01 (Uusimaa) and the other regions in the south are found to have lower fertility. Although, region 17 (Pohjois-Pohjanmaa) appears to have the largest of the families, its deviation from the linear relationship is negligible,

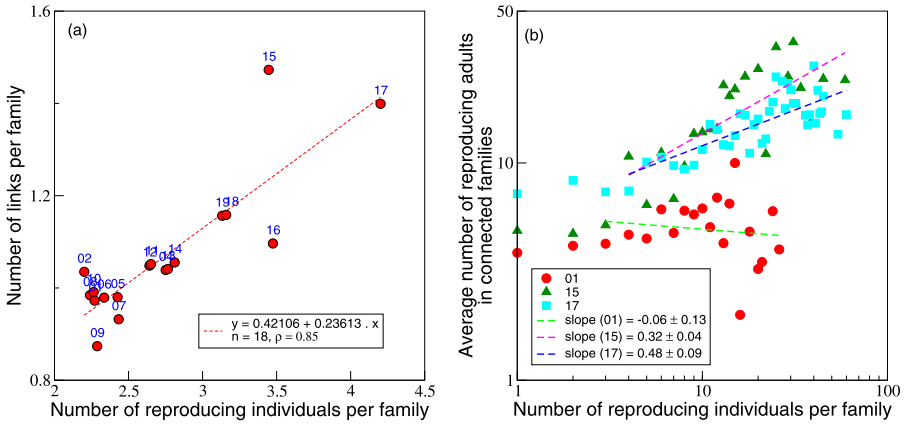


Figure 9 Aspects of the network as functions of the number of reproducing individuals in the family. (**a**) The variation of number of links per family in the network of families with number of reproducing individuals per family. The regions to which each of the points correspond are denoted by the labels. The dashed line is a linear regression fit. (**b**) For a family with a given number (x) of reproducing individuals in a given region, we count the total number of reproducing individuals in all the families (belonging to the same region) connected to it, and then divide by the number of connections (gives y). For each x we calculate the mean y. These values are shown as scatter plots for three different regions, namely, Uusimaa (circles), Pohjanmaa (triangles) and Pohjois-Pohjanmaa (squares). The dashed lines are linear regression fits for $\log(y)$ versus $\log(x)$, with slopes provided in the legend. Unlike in the case of Uusimaa, for Pohjanmaa and Pohjois-Pohjanmaa the slopes are significantly greater than zero

whereas region 15 (Pohjanmaa) and 16 (Keski-Pohjanmaa) show larger deviations. In the case of region 16 (Keski-Pohjanmaa) we observed families to be mostly linked to families from neighbouring regions.

Correlations in the connectivity pattern for families in region 17 (Pohjois-Pohjanmaa) can not be solely judged by the aspect of regional assortativity and large sized families resulting from higher fertility rate. There appears to be a tendency for the large families to get connect to each other. This is shown in Fig. 9(b). For a family of a given size measured in terms of the number of reproducing individuals, we calculate the average size of the connected families. This is similar to the nearest-neighbour average connectivity, which is used to quantify the degree correlations in networks [37]. A positive slope corresponding to region 17 (Pohjois-Pohjanmaa) indicates the presence of such correlations. Similar correlation is also observed in region 15 (Pohjanmaa). For the rest of the regions we did not find any significant correlation. The case of region 01 (Uusimaa) is illustrated where the slope is not different from zero. This kind of "degree assortativity" originating likely from religious reasons in addition to the regional assortativity could be the reason for them dominating in the largest connected component [38]. In fact it was demonstrated in [38] that when such assortativity is high a 'core group' is formed by high degree nodes on which a largest connected component grows but contrary to expectations does not grow steadily and does not extend into the rest of the network. The scenario is very similar to our case, and such high assortativity and resulting impedance in the growth of the largest component could additionally imply that the actual underlying network (from which the data is sampled) as a whole is not a small-world [39, 40]. This may be surprising to a certain extent as the fragments listed in Table 1 are small-worlds. Note, that we consider the small-world property in the context of networks between individuals with certain types of kinship (Fig. 7). As already discussed we base these findings on marriages between in-

dividuals belonging to families having similar linguistic or religious affiliations. A more direct imposition of the small-world character may occur through marriages between immediate kins, for instance in nomadic clans [4, 6]. In our data we found only six instances of marriages between cousins.

In sum, the general patterns of linkages found within this representative sample of a national population are indicative of a high assortativity in the network of families. Both the region of birth and the language appear to function as cultural attractors in the network and increase its clustering and transitivity. We can distinguish between two patterns of regional effects in this network, either showing 'metropolitan' family linkages or the 'cultural' family linkages. The metropolitan families are to be found in region around the capital and they are mostly part of smaller clusters and many of these families could not be linked to other families in this population sample. These families are present in large numbers and appear to be overwhelmingly linked to families originating from the other regions. Migration of population to the more industrialized southern regions of Finland results into a decrease of assortativity and network transitivity. The cultural linkages are found among families from the Pohjois-Pohjanmaa region as well as other western and northern regions of Finland. Here, the regional and linguistic identity seems to result in a strong regional connectivity in terms of family ties.

Appendix

A.1 *k*-core decomposition

The k-core of a given network is its largest subgraph (that is a network constituted by the nodes and links from the original network) which contains only those nodes that have degree larger than or equal to k. The core can be obtained from the original network by removing all the nodes having degree less than k, as well as the links associated with these nodes. This is followed by recalculating the degrees. This process is repeated until there are no nodes with degree less than k.

A.2 Assortativity coefficient

Following [32], we construct a symmetric matrix $\{e_{ij}\}$, where e_{ij} is the fraction of edges connecting nodes belonging to region i and region j. The edges connecting nodes of the same type are counted twice. The matrix satisfies, $\sum_{ij} e_{ij} = 1$, and $\sum_j e_{ij} = c_i$, where c_i is the fraction of edges with at least one end attached to nodes belonging to region i. The assortativity coefficient is given by:

$$a = \frac{\sum_i e_{ii} - \sum_i c_i^2}{1 - \sum_i c_i^2},$$

with the error estimate (σ_a) being,

$$\sigma_a^2 = \frac{1}{M} \frac{\sum_i c_i^2 + [\sum_i c_i^2]^2 - 2\sum_i c_i^3}{1 - \sum_i c_i^2},$$

where, M is total number of edges in the network.

Additional material

Additional file 1: Network construction. The figure illustrates the method of constructing the network of families and the network of kins from the set of genealogical networks for extended families. (PDF 83 kB)
Additional file 2: Assigning regions of origin to families. The figure shows number of families that could be assigned regions of origin as a function of the parameter θ, described in the Methods (Section 3). (PDF 39 kB)
Additional file 3: Types of triangles. The two different types of triangles that could be found in the network are shown in the figure (details in the caption). (PDF 167 kB)
Additional file 4: Network growth. Number of links that appear in a given year within the range 1965–2012 is shown in the figure, where the year corresponds to birth of the firstborn. (PDF 49 kB)
Additional file 5: Network of kins. The kin network derived from the four largest components of the network of families is shown in the figure. (PDF 981 kB)

Funding
This work was supported by the EU HORIZON 2020 FET Open RIA project (IBSEN) No. 662725 (KB, DM, KK), the Academy of Finland Research project (COSDYN) No. 276439 (AG and KK), the Academy of Finland research project No. 266898 (AR, VB), CONACYT (Mexico) grant No. 383907 (DM), and EU FP7 (MULTIPLEX) project No. 317532 (JK).

Availability of data and materials
The FINNFAMILY dataset is not publicly available. The dataset was acquired by Vaestoliitto, Helsinki from and with ethical permission from Statistics Finland, Helsinki. Professor Anna Rotkirch, Vaestoliitto may be contacted for details of the FINNFAMILY dataset; and Statistics Finland may be contacted to enquire about accessibility of data of similar or related nature.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceptualization: KB, AG, DM; methodology: All authors; formal analysis: KB; data curation: VB, AR; writing (original draft preparation): KB, VB, AR; writing (review and editing): All authors; visualization: KB; supervision: KK, AR; funding acquisition: AR, JK, KK. All authors read and approved the final manuscript.

Author details
[1]Department of Computer Science, School of Science, Aalto University, Aalto, Finland. [2]Population Research Institute, Väestöliitto, Finnish Family Federation, Helsinki, Finland. [3]Center for Network Science, Central European University, Budapest, Hungary.

Publisher's Note

Received: 11 September 2017 Accepted: 3 April 2018 Published online: 18 April 2018

References

1. Hughes AL (1988) Evolution and human kinship. Oxford University Press, New York, pp 42–47
2. Szydlik M (2016) Sharing lives: adult children and parents. Routledge, New York
3. Kolk M (2014) Multigenerational transmission of family size in contemporary Sweden. Popul Stud 68(1):111–129
4. White DR, Johansen U (2005) Network analysis and ethnographic problems: process models of a Turkish nomad clan. Lexington Books, Lanham
5. Newman M, Barabasi A-L, Watts DJ (2011) The structure and dynamics of networks. Princeton University Press, Princeton
6. White DR, Houseman M (2002) The navigability of strong ties: small worlds, tie strength, and network topology. Complexity 8(1):72–81
7. Derrida B, Manrubia SC, Zanette DH (1999) Statistical properties of genealogical trees. Phys Rev Lett 82(9):1987–1990
8. Derrida B, Manrubia SC, Zanette DH (2000) Distribution of repetitions of ancestors in genealogical trees. Phys A, Stat Mech Appl 281(1–4):1–16
9. Zanette DH, Manrubia SC (2001) Vertical transmission of culture and the distribution of family names. Phys A, Stat Mech Appl 295(1–2):1–8
10. White DR (2004) Ring cohesion theory in marriage and social networks. In: Mathématiques et sciences humaines (Mathematics and social sciences), vol 168
11. Hamberger K, Houseman M, Daillant I, White DR, Barry L (2004) Matrimonial ring structures. In: Mathématiques et sciences humaines (Mathematics and social sciences), vol 168
12. Johansen U, White DR (2006) Collaborative long-term ethnography and longitudinal social analysis of a nomadic clan in southeastern Turkey
13. Hodgson U, Laitinen T, Tukiainen P (2002) Nationwide prevalence of sporadic and familial idiopathic pulmonary fibrosis: evidence of founder effect among multiplex families in Finland. Thorax 57(4):338–342
14. Helgason A, Pálsson S, Guðbjartsson DF, Stefánsson K et al (2008) An association between the kinship and fertility of human couples. Science 319(5864):813–816
15. Alvarez G, Ceballos FC, Quinteiro C (2009) The role of inbreeding in the extinction of a European royal dynasty. PLoS ONE 4(4):e5174

16. Moreau C, Bhérer C, Vézina H, Jomphe M, Labuda D, Excoffier L (2011) Deep human genealogies reveal a selective advantage to be on an expanding wave front. Science 334(6059):1148–1150
17. Ceballos F, Álvarez G (2013) Royal dynasties as human inbreeding laboratories: the Habsburgs. Heredity 111(2):114–121
18. Wakeley J, King L, Wilton PR (2016) Effects of the population pedigree on genetic signatures of historical demographic events. Proc Natl Acad Sci USA 113(29):7994–8001
19. Kaplanis J, Gordon A, Wahl M, Gershovits M, Markus B, Sheikh M, Gymrek M, Bhatia G, MacArthur DG, Price A, Erlich Y (2017) Quantitative analysis of population-scale family trees using millions of relatives. bioRxiv. https://doi.org/10.1101/106427. https://www.biorxiv.org/content/early/2017/02/07/106427.1.full.pdf
20. Vaestoliitto. http://www.vaestoliitto.fi/in_english/population_research_institute/family_research/linked-lives/
21. Statistics-Finland. http://tilastokeskus.fi/meta/luokitukset/maakunta/001-2017/index.html
22. Virrankoski P (2001) Suomen Historia, vol 846. Suomalaisen Kirjallisuuden Seura, Helsinki
23. Terämä E (2010) Regional demographic differences: the effect of Laestadians. Finn Yearb Popul Res 45:123–141
24. Ghosh A, Berg V, Bhattacharya K, Monsivais D, Kertesz J, Kaski K, Rotkirch A (2017) Migration patterns across the life course of families: gender differences and proximity with parents and siblings in Finland. arXiv:1708.02432
25. Hamberger K, Houseman M, White DR (2011) Kinship network analysis. In: The SAGE handbook of social network analysis. Sage, Thousand Oaks, pp 533–549
26. Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56(645):330–338
27. Oliehoek PA, Windig JJ, Van Arendonk JA, Bijma P (2006) Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics 173(1):483–496
28. Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
29. Dorogovtsev SN, Goltsev AV, Mendes JFF (2006) *k*-core organization of complex networks. Phys Rev Lett 96(4):040601
30. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
31. Erdős P, Rényi A (1959) On random graphs. Publ Math (Debr) 6:290–297
32. Newman ME (2003) Mixing patterns in networks. Phys Rev E 67(2):026126
33. Efron B (1992) Jackknife-after-bootstrap standard errors and influence functions. J R Stat Soc, Ser B, Methodol 54:83–127
34. Bell M, Charles-Edwards E, Ueffing P, Stillwell J, Kupiszewski M, Kupiszewska D (2015) Internal migration and development: comparing migration intensities around the world. Popul Dev Rev 41(1):33–58
35. Tervo H (2000) Suomen aluerakenne ja siihen vaikuttavat tekijät. Kansantal Aikak 96(3):398–415
36. Hannelius U, Salmela E, Lappalainen T, Guillot G, Lindgren CM, von Döbeln U, Lahermo P, Kere J (2008) Population substructure in Finland and Sweden revealed by the use of spatial coordinates and a small number of unlinked autosomal SNPs. BMC Genet 9(1):54
37. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. Phys Rev Lett 87(25):258701
38. Newman ME (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701
39. Gomez-Gardenes J, Moreno Y (2004) Local versus global knowledge in the Barabási–Albert scale-free network model. Phys Rev E 69(3):037103
40. Small M, Xu X, Zhou J, Zhang J, Sun J, Lu J-a (2008) Scale-free networks which are highly assortative but not small world. Phys Rev E 77(6):066112