Discussion Paper

A Person Network of the Netherlands

D.J. van der Laan

May 2022

As part of its research programme Statistics Netherlands has derived a network of the entire Dutch population. It contains relationships between the 17 million inhabitants of the Netherlands and consists of five so called layers (main types of relations): family, household, neighbours, work and school. Combined these layers contain over 1.4 billion relationships. This paper describes the methods used to derive the network and gives an overview of some of its properties.

1 Introduction

Many important social and economic processes take place along networks: ideas spread, money and beliefs are transferred. Therefore, network analysis has become a important tool to understand society (Borgatti et al. 2009; Leij and Buhai 2008; Verdery et al. 2012; Newman and Park 2003). Recognising this, Statistics Netherlands has, using data present at Statistics Netherlands (Bakker, Van Rooijen, and Van Toor 2014), built a network covering the entire population and covering multiple types of relations: family, household, neighbours, work and school. It should, therefore, give more inclusive overview of social structure than many of the networks currently used in social research.

Statistics Netherlands collects data, mostly from various government agencies, in order to support government and companies in making policies and assisting politics, media researchers, and civilians in evaluating the results of those policies. In accordance with the law which governs Statistics Netherlands conduct[1)] and the General Data Protection Regulation[2)], the data are protected by extensive rules and regulations with respect to privacy to ensure that no information about individual persons is disclosed in any output from these data sets, and individuals can never be recognized in the output. This applies to all analyses presented here[3)]. The data collected by Statistics Netherlands data was used to derive the network. The vertices in the network are all persons that were registered in the official population register on October 1st 2018. The edges are defined by family relations, household membership, neighbours, co-workers and class-mates. The resulting network contains 17.3 million vertices and 1.4 billion edges.

This paper describes the method with which the network has been derived and the exact definitions of the relations present in the network and gives an overview of some of the properties of the network.

1) https://wetten.overheid.nl/BWBR0015926/2018-07-01
2) Regulation (EU) 2016/679 (General Data Protection Regulation)
3) Full and detailed information about the security measures and data privacy protections in place for the analysis of microdata for research can be found here: https://www.cbs.nl/en-gb/our-services/customised-services-microdata/

2 Derivation of the network

2.1 Terminology

The network consists of persons with relations between those persons. We will use the term *vertex* (plural: *vertices*) to denote the persons and *edge* to denote the relation between two *vertices*. The edges in the network are directed and labelled. An edge labelled '<label>' from A to B should be read as 'A has a '<label>' B', for example, 'A has a 'parent' B. Directed means that the relation is one-way: an edge between A and B labelled 'parent' (B is the parent of A) does not mean that the same relation exists between B and A (A is not also the parent of B). However, many of the edges will have a corresponding edge in the other direction. For example, when there is an edge between A and B labelled 'parent', there will also be a corresponding edge labelled 'child' between B and A. Sometimes the corresponding edge will also have the same label as with 'sibling' and 'household member'.

The out-degree of a vertex is the number of out going edges of a vertex. The in-degree is the number of incoming edges. It is possible that both or one of them are zero: a person does not have any (known) relation.

The network is divided in a number of layers. Each layer contains all vertices but a disjunct subset of edge types. For example, the family layer contains all edges that are based on family relations.

2.2 Population

The vertices of the network are all 17.3 million persons registered on October 1st 2018 in the official Dutch Population register (Basisregistratie Personen, BRP). All persons are included in the network even when a person or vertex does not have any edges.

2.3 Family Layer

We use the parent-child-register of Statistics Netherlands as the basis of the family layer. This register contains all (legal) parent-child relations since January 1st 1995. A parent-child edge can only exist if both parent and child have been recorded in the official Dutch population registers (formerly GBA, or currently BRP). The relation between child A and parent B can be schematically visualised as:

$$A \xrightarrow{\text{parent}} B \implies B \xrightarrow{\text{child}} A. \tag{1}$$

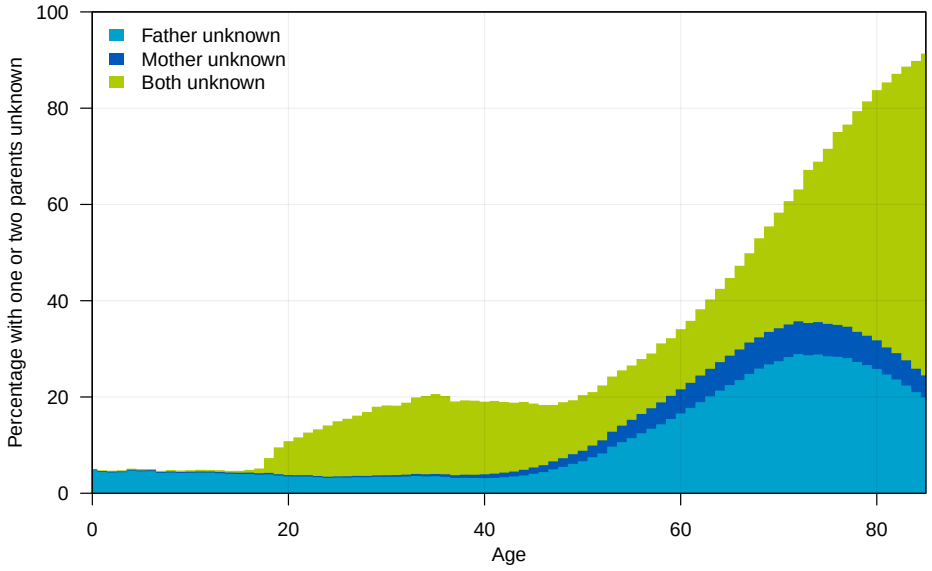
An arrow between two persons indicates an edge between the two persons. The label above the arrow indicates the type of edge. From these two edge types we can derive a large number of other edge types. These are summarised below.

In deriving the edge types below we use the complete child-parent-file, which includes persons which are no longer part of the population at the reference date of the social network. After having derived all family edges based on the whole parent-child register, we remove vertices with all in and outgoing edges that are not part of the population on the reference date of October 1st 2018. This way, even if the parents are deceased, we can, for example derive the sibling

Table 2.1 Information on parents in the population

Information on parents	Number of persons	Fraction
Both parents known	12 580 092	0.73
Mother unknown	404 989	0.02
Father unknown	1 712 359	0.10
Both parents unknown	2 559 767	0.15

Figure 2.1 Percentage of persons in the population with one or more unknown parents as a function of age.



relationships between their children, given that the parent-child relationships were present in the register. However, even then not for all persons are both parents known. Table 2.1 shows the information present in the child-parent-file for persons in the population. Figure 2.1 shows how the fraction of missing parents changes as a function of age: this fraction increases for older persons. The main reason for this is that both the child and parent have had to be registered in the population register after January 1st 1995. Therefore, parents that have deceased before that date are not registered. Another group with where information on the parents is often missing are immigrants as their parents often do not live in the Netherlands. Missing parents also affects other relations as these are based on the parent-child relations. When the parents are missing sibling relations cannot be derived, for example. Also when the parents of the parent (the grand-parent) are missing, uncles, aunts and cousins cannot be derived.

Grand parent The parents of the parents of a person are its grand parents:

$$A \xrightarrow{\text{parent}} B \xrightarrow{\text{parent}} C \implies A \xrightarrow{\text{grand parent}} C. \tag{2}$$

Grand child The children of the children of a person are its grand children.

$$A \xrightarrow{\text{child}} B \xrightarrow{\text{child}} C \implies A \xrightarrow{\text{grand child}} C. \tag{3}$$

Full sibling For siblings we distinguish three different types depending on the number of parents they share. In principle two persons are siblings when they share at least one parent.

When they share two parents they are full siblings:

$$A \xrightarrow{\text{parent}} B \wedge C \xrightarrow{\text{parent}} B \wedge A \xrightarrow{\text{parent}} D \wedge C \xrightarrow{\text{parent}} D \wedge A \neq C \implies A \xrightarrow{\text{full sibling}} C, \quad (4)$$

with ∧ the logical and. The $A \neq C$ (A is not the same person as C) is necessary as a person cannot be its own sibling.

Half-sibling A half-sibling are two persons that share only one parent.

$$A \xrightarrow{\text{parent}} B \wedge C \xrightarrow{\text{parent}} B \wedge A \xrightarrow{\text{parent}} D \wedge C \xrightarrow{\text{parent}} E \wedge A \neq C \wedge D \neq E \implies A \xrightarrow{\text{half-sibling}} C. \quad (5)$$

Sibling (unknown) In principle this should cover all types of siblings. However, as for some persons one or both of the parents are not known, there is also a group of persons for which we cannot determine if they are full or half-siblings. When two person share a parent, but when for at least one of them one of the parents is unknown, we cannot determine their exact edge type.

$$A \xrightarrow{\text{parent}} B \wedge C \xrightarrow{\text{parent}} B \wedge (A \xrightarrow{\text{parent}} ? \vee C \xrightarrow{\text{parent}} ?) \wedge A \neq C \implies A \xrightarrow{\text{sibling (unknown)}} C, \quad (6)$$

where ? indicates that this person is unknown and ∨ is the logical or (not exclusive).

When we refer to the type ‘sibling’ this refers to either of the relations ‘full sibling’, ‘half-sibling’ or ‘sibling (unknown)’.

Co-parent Persons that have a child together are co-parents:

$$A \xrightarrow{\text{child}} B \wedge C \xrightarrow{\text{child}} B \wedge A \neq C \implies A \xrightarrow{\text{co-parent}} C. \quad (7)$$

This does not mean that the two persons are currently living together. This is covered by the relation ‘partner’ defined below.

Aunt/uncle

$$A \xrightarrow{\text{parent}} B \xrightarrow{\text{sibling}} C \implies A \xrightarrow{\text{aunt/uncle}} C. \quad (8)$$

Niece/nephew

$$A \xrightarrow{\text{sibling}} B \xrightarrow{\text{child}} C \implies A \xrightarrow{\text{niece/nephew}} C. \quad (9)$$

Cousin

$$A \xrightarrow{\text{uncle/aunt}} B \xrightarrow{\text{child}} C \implies A \xrightarrow{\text{cousin}} C. \quad (10)$$

2.4 Household layer

Statistics Netherlands also determines to which household each person belongs. A household is a group of people having common arrangements for sharing expenses and daily needs, living in a shared residence. It is possible that on a given address multiple households are living. Households are derived mainly from information from the population register, namely the addresses of each person and information on marriages and children. For example, a married couple with children living on one address with no other persons present on that address are considered to form one household. Information from the tax office and information on institutional households (e.g. care facilities) are used to derive household for other addresses. For the vast majority of addresses the household composition can be determined in this way. For a small percentage of addresses (approx. 5%) the household composition is imputed. Based on the properties of the persons living on the address, probabilities for the different possible household compositions are estimated and one composition is randomly selected using these probabilities.

Not only is it know to which household each person belongs, also the position in the household is determined for each person. In this case the positions are recoded to the following three types: household member ('hh member in'), partner in household ('partner in'), institutional household member ('inst hh member in'). Therefore, a household consisting of a couple with child, will result in the following set of relations:

$$\begin{aligned} A &\xrightarrow{\text{partner in}} h \\ B &\xrightarrow{\text{partner in}} h \\ C &\xrightarrow{\text{hh member in}} h \end{aligned}$$

From the relations between persons and households, relations between persons are derived. Note that a given household can only have two persons defined as 'partner in' even when there other persons are living in the same household that are, for example, married. For example, in a household where a couple also houses the married parents of one them, the youngest couple are labelled as partners and the parents as other household members.

Partner

$$A \xrightarrow{\text{partner in}} h \wedge B \xrightarrow{\text{partner in}} h \wedge A \neq B \implies A \xrightarrow{\text{partner}} B. \tag{11}$$

Household member

$$A \xrightarrow{\text{hh member in/partner in}} h \wedge B \xrightarrow{\text{hh member in/partner in}} h\wedge$$

$$A \neq B \implies A \xrightarrow{\text{hh member}} B \tag{12}$$

Note that when A and B have the edge type 'partner' they also have and edge with type 'hh member'. When they also share a child, they also have an edge with type 'co-parent'. It was decided to not make a choice between these types as it depends on the type of analysis and the goal of the analysis if having multiple edges between two persons is a problem and, if so, which type is more relevant.

Institutional household member As institutional households (e.g. care homes) can be large and because the relation between institutional household members will generally differ from that between regular household members, it was decided to separately label institutional household members.

$$A \xrightarrow{\text{inst hh member in}} h \wedge B \xrightarrow{\text{inst hh member in}} h \wedge A \neq B \implies A \xrightarrow{\text{inst hh member}} B \quad (13)$$

2.5 Neighbour layer

Each person in the population register is assigned to a household. The address of each household is known. This is used to calculate for each household h , the closest ten households with a distance less than 50 metres, $\mathcal{N}_h = \{g_1, \ldots, g_k\}$ $(k \leq 10)$. When there are multiple households with the same distance (at the tenth position), households are randomly selected. Household members of $\mathcal{N}_h$ are neighbours of household members of h.

As institutional households can be very large, it was decided to treat these slightly different. Institutional households with four persons or less are treated as above. They are households just like any other household and can therefore have neighbours and be neighbours of other households. In institutional households with more than four persons, each person is considered a separate household when determining neighbours. Therefore, each person in these institutional households can have ten neighbouring households (which can be other persons from the institutional household) and non-institutional households can have up to ten persons from neighbouring institutional households as neighbours.

Neighbour

$$A \xrightarrow{\text{any hh member in}} h \wedge B \xrightarrow{\text{any hh member in}} g \text{ (with } g \in \mathcal{N}_h) \implies A \xrightarrow{\text{neighbour}} B. \quad (14)$$

Note that it not necessary that if household g is one of the ten closest households to h $(g \in \mathcal{N}_h)$, household h is also one of the ten households closest to g. Therefore, is B is a neighbour of A, that does not imply that A is also a neighbour of B.

2.6 Work layer

Using income tax data (*Polisadministratie*), it is possible to derive for each working person the company that person is working at on October 1st for the year for which the network is derived.

$$A \xrightarrow{\text{works at}} q \quad (15)$$

From this we can derive all co-workers of a given person. For a company q with n_q employees, this would result in $(n_q - 1)^2$ directed edges in the network. There are large differences in the sizes of companies (which also include institutions such as government agencies and universities). It was decided to put a limit of one hundred on the number of co-workers of a person. First, it is unrealistic that a person actually knows more than a certain number of co-workers. Second, without a limit on the number of co-workers persons know, the work layers dominates the social network (over 80% of the relations in the network would be of the type co-worker). Third, many of the large companies have different locations (for example retail

Table 2.2 Variables used to define classes.

Type of school	Variables defining a class
Primary education (*Basisschool*)	School id, location id, year
Secondary education (*Voortgezet onderwijs*)	School id, location id, type of education, year
Secondary Special education (*Speciaal voortgezet onderwijs*)	School id, location id, number of years followed
Vocational (*MBO*)	School id, type of education, number of years followed
Higher education (*HBO*, *University*)	School id, location id, type of education, number of years followed

chains, supermarket chains, government agencies). It is likely that employees mainly have contact with other employees from the some location. By limiting the number of co-workers, we can select co-workers which are geographically close thereby increasing the likelihood that we link co-workers that actually know each other.

Co-worker Let $\mathcal{C}_A$ be the 100 geographically closest persons of A working at the same company. When there are multiple persons at the same distance at the 100$^{\text{th}}$ position persons are randomly chosen. When a company has 101 employees or less $\mathcal{C}_A$ are all employees of the company excluding A. Then:

$$A \xrightarrow{\text{works at}} q \wedge B \xrightarrow{\text{works at}} q \wedge B \in \mathcal{C}_A \implies A \xrightarrow{\text{co-worker}} B \tag{16}$$

For companies with 101 employees or less, this results in a network that is symmetric — a edge from 'A' to 'B' implies a edge from 'B' to 'A' — and each person is connected to each other person in the company: they form a clique. For larger companies the network is no longer symmetric. The out-degree will always be a 100 while the in-degree can vary, but will on average be a 100.

2.7 School layer

DUO (*Dienst Uitvoering Onderwijs* [Implementation Service Education]) maintains registers on students for the various education types (see table 2.2). Depending on the type of education, it is possible to approximately determine which persons go to the same class. It is unfortunately not possible to determine exactly to which classes persons go. For example, in primary education, where classes are based on a combination of school id, location id and school year, there can be multiple classes with children in the same school year in the same location resulting in classes that are too large. At the same time, for small schools, multiple school years can be in the same physical class. Also in secondary and higher education different types of education can share classes. Table 2.2 shows for each type of education the variables used to define classes.

Class mate Persons going to the same class are class mates:

$$A \xrightarrow{\text{in class}} u \wedge B \xrightarrow{\text{in class}} u \wedge A \neq B \implies A \xrightarrow{\text{class mate}} B. \tag{17}$$

Table 3.1 Number of edges and summary statistics of the degree distribution of each of the edge types. The information on the degree distribution is shown for those persons that have the corresponding edge type with other persons.

Edge type	Number of persons (10^3)	Number of edges (10^3)	Mean degree	Quantiles of degree distribution							
				5%	10%	25%	50%	75%	90%	95%	99%
Child	8,835	19,161	2.17	1	1	2	2	3	3	4	6
Parent	10,965	19,161	1.75	1	1	1	2	2	2	2	2
Half-sibling	940	1,743	1.86	1	1	1	2	2	3	4	6
Sibling (unknown)	2,268	5,743	2.53	1	1	1	2	3	5	7	9
Full sibling	11,008	21,897	1.99	1	1	1	2	2	4	5	8
Co-parent	7,776	8,035	1.03	1	1	1	1	1	1	1	2
Grand parent	5,432	13,214	2.43	1	1	1	2	3	4	4	4
Grand child	3,524	13,214	3.75	1	1	2	3	5	7	9	15
Niece/nephew	7,761	40,273	5.19	1	1	2	4	7	11	15	24
Aunt/uncle	9,219	40,273	4.37	1	1	2	4	6	8	10	14
Cousin	9,106	87,522	9.61	2	2	4	7	13	20	26	42
Partner	8,533	8,533	1.00	1	1	1	1	1	1	1	1
Household member	13,982	31,931	2.28	1	1	1	2	3	4	4	6
Inst. household member	72	735	10.16	1	1	1	2	6	14	41	168
Neighbour	16,810	352,645	20.98	10	11	16	21	26	31	33	38
Co-worker	2,588	76,199	29.45	1	3	7	20	46	74	87	97
Co-worker (sampled)	4,898	489,823	100.00	100	100	100	100	100	100	100	100
Class mate (primary)	1,433	55,540	38.76	10	14	22	34	50	68	82	118
Class mate (special)	68	1,773	26.17	4	6	11	19	33	56	69	119
Class mate (secondary)	959	62,450	65.10	7	11	24	51	90	138	172	265
Class mate (vocational)	513	28,360	55.32	2	5	12	31	70	137	198	343
Class mate (higher)	701	85,727	122.35	6	12	34	85	168	286	377	591

Table 3.2 The sizes of the main layers in the network.

Layer	Number of edges (10^3)	Average degree	
		Excl. zeros	Incl. zeros
Family	270,235	15.66	16.43
Household	41,199	2.39	2.93
Neighbour	352,645	20.43	20.98
Work	566,022	32.80	75.61
School	233,849	13.55	63.66

3 Properties of the network

Table 3.1 summarises the main properties of the network. For each edge type, it shows the number of persons in the population that have the given edge type with another person, the total number of edges in the network and properties of the out-degree distribution. For example, looking at the first line, we see that 8.8 million persons have one or more children (that are alive) and persons that have children have on average 2.2 children (that are alive). Table 3.2 summarises the table even further. The family, school and neighbour layers are approximately of equal size with 270, 233 and 353 million relations respectively. The work layer is larger with a size of 566 million relations. Relations from businesses with a hundred employees or more dominate the work layer. The household layer is the smallest with 41 million relations.

Figure 3.1 shows the degree distribution for each of the four main layers in the network. The large degrees in the household layer are caused by institutional households where person can have a large number of household members.

Figure 3.1 Out-degree distribution of the five main layers. Persons with a out-degree of zero in the network are omitted from the distribution.

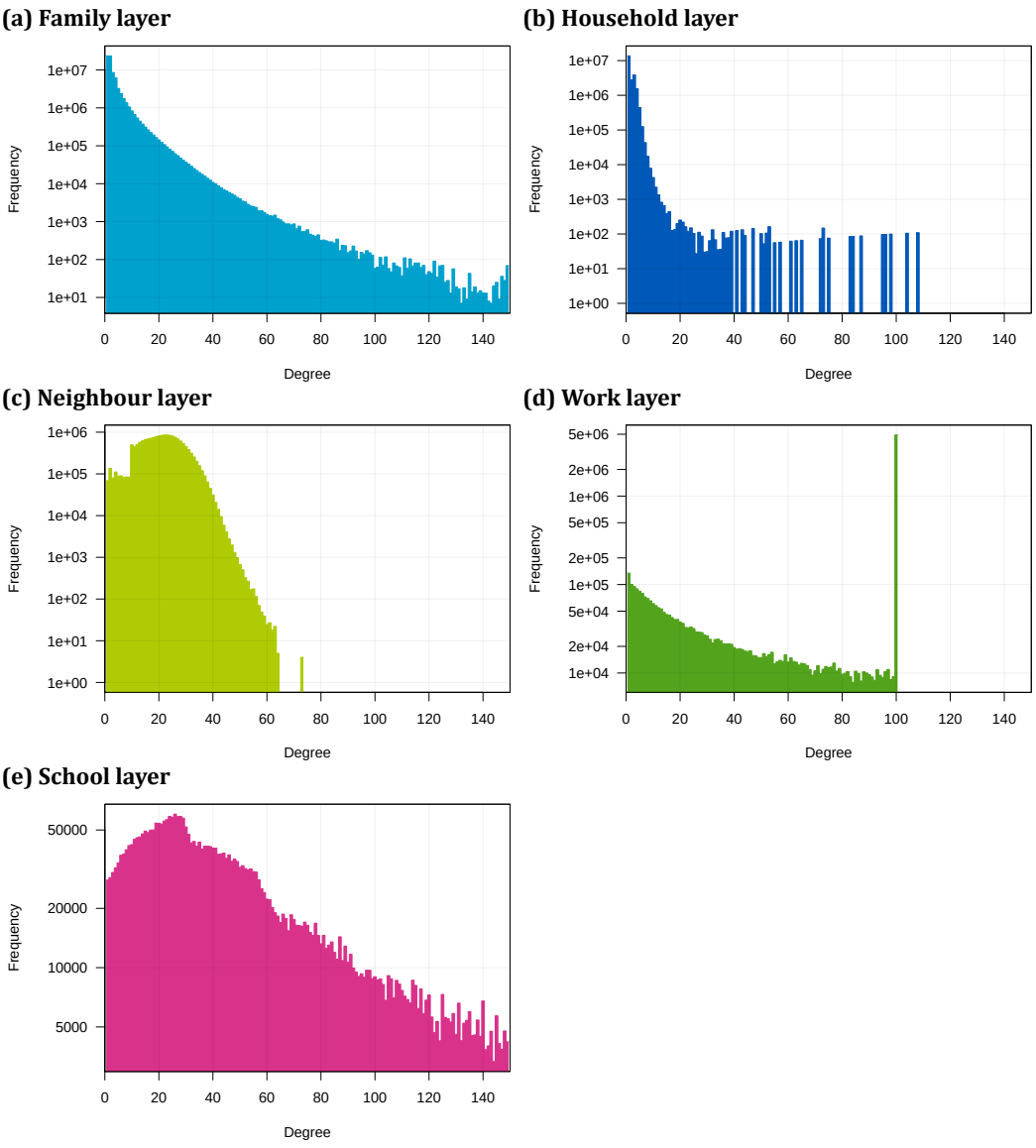


Table 3.3 Inverse relation types for each relation type. A hyphen indicates that this relation type is not symmetric.

Relation	Corresponding inverse relation
Child	Parent
Parent	Child
Half-sibling	Half-sibling
Sibling (unknown)	Sibling (unknown)
Full sibling	Full sibling
Co-parent	Co-parent
Grand parent	Grand child
Grand child	Grand parent
Niece/nephew	Aunt/uncle
Aunt/uncle	Niece/nephew
Cousin	Cousin
Partner	Partner
Household member	Household member
Inst. household member	Int. household member
Neighbour	-
Co-worker	Co-worker
Co-worker (sampled)	-
Class mate (primary)	Class mate (primary)
Class mate (special)	Class mate (special)
Class mate (secondary)	Class mate (secondary)
Class mate (vocational)	Class mate (vocational)
Class mate (higher)	Class mate (higher)

3.1 Symmetry

As was mentioned in section 2.1 the network is in principle directed. However, some of the edge types in the network have a corresponding edge in the other direction. Table 3.3 shows for each edge type the corresponding inverse edge type if there is one. As can be seen in the table, the network is largely symmetric with the exception of the neighbour layer and the work layer for large companies (more than 101 employees).

4 Conclusion

Using data present as Statistics Netherlands, a network covering the entire Dutch population has been built. The vertices of the network are all persons in the official Dutch population register on October 1st 2018. The edges are family relations, household membership, neighbours, co-workers and class-mates, also on October 1st 2018. Since it cannot be inferred whether people actually interact with others in their network, the network is a reservoir of potential contacts and exposure rather than a realized network. Although the network covers a wide range of types on contact, a lot of important types of contact are not included. Friends and other contacts via church, sports or hobby clubs are not included, and neither is subjective information such as motivations, attitudes and quality of relationships. Nevertheless, it should give a more complete overview of the social environments of persons than many of the sources currently used. The possible applications of the network are numerous. At the moment we are using the network so study segregation (Laan and Jonge 2019), and the network has already been used to

study the exposure to crime of young persons living in the Hague (Posthumus et al. 2020) and to study the self-reliance of elderly (Das and Jonge 2020).

As part of the research programme and in collaboration with the POPNET project (POPNET 2022) we are looking to improve on the current version of the network. For example, at the moment the network is observed only for one moment in time, a longitudinal network might give more information on how society is changing. An other example is the work layer, where there is uncertainty in which co-workers actually know each other. Adding information on company locations might improve on this. Although the current version of the network is already a big step forward for investigations of the Dutch society, improvements are still possible.

Acknowledgements

We thank Eszter Bokányi, Edwin de Jonge, Yuliia Kazmina and Frank Takes for their useful feedback and discussions on the network.

References

[1] Bart F.M. Bakker, Johan Van Rooijen, and Leo Van Toor. "The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics". In: *Statistical Journal of the IAOS* 30.4 (2014), pp. 411–424.

[2] Stephen P Borgatti et al. "Network analysis in the social sciences". In: *science* 323.5916 (2009), pp. 892–895.

[3] Marjolijn Das and Edwin de Jonge. *Zelfredzaamheid van ouderen en gebruik van Wmo*. CBS, 2020. URL: https://www.cbs.nl/nl-nl/longread/statistische-trends/2020/zelfredzaamheid-van-ouderen-en-gebruik-van-wmo?onepage=true.

[4] Dingeman Jan van der Laan and Edwin de Jonge. *Measuring segregation using a network of the Dutch population*. Paper presented at The 5th International Conference on Computational Social Science. 2019.

[5] Marco van der Leij and Ioan-Sebastian Buhai. "A social network analysis of occupational segregation". In: *FEEM Working Paper* 31 (2008). DOI: 10.2139/ssrn.1117949.

[6] Mark EJ Newman and Juyong Park. "Why social networks are different from other types of networks". In: *Physical review E* 68.3 (2003), p. 036122.

[7] POPNET. *Population-Scale Network Analysis - A new research data infrastructure in computational social science*. 2022. URL: https://popnet.io (visited on 04/20/2022).

[8] Hanneke Posthumus et al. *Criminaliteit in netwerken van jongeren uit Den Haag Zuidwest*. CBS, 2020. URL: https://www.cbs.nl/nl-nl/longread/aanvullende-statistische-diensten/2020/criminaliteit-in-netwerken-van-jongeren-uit-den-haag-zuidwest?onepage=true.

[9] Ashton M Verdery et al. "Social and spatial networks: Kinship distance and dwelling unit proximity in rural Thailand". In: *Social networks* 34.1 (2012), pp. 112–127.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information