Appendix

A *regnet* python package

The *regnet* package is divided into two central data structures. The first is a simple single year, unweighted, multi-layer structure suitable for determining connections, neighborhoods, and degree. The second structure is multi-layer, can span multiple years, and allows edge-weights to be tailored to the research question at hand. A proposal for a weighting scheme that reasonably resembles interaction probabilities is presented in Subection **??**. This structure is designed to efficiently find shortest paths between nodes, and presents the layer types through which they are created. A detailed instruction guide on how to use the different structures and methods is provided in the *Readme* section of the *regnet* repository[1].

B Visualization of the local higher-order clustering coefficient

The intuition between the higher order clustering coefficient is illustrated in Figure 1. The blue dots denote nodes belonging to a clique of size two (1A) or three (1B). The higher-order local clustering coefficient of the large blue node gives the probability of an adjacent node – as the yellow one – to be connected to all the other nodes in the clique, i.e., the probability that the dotted edges exist[2].

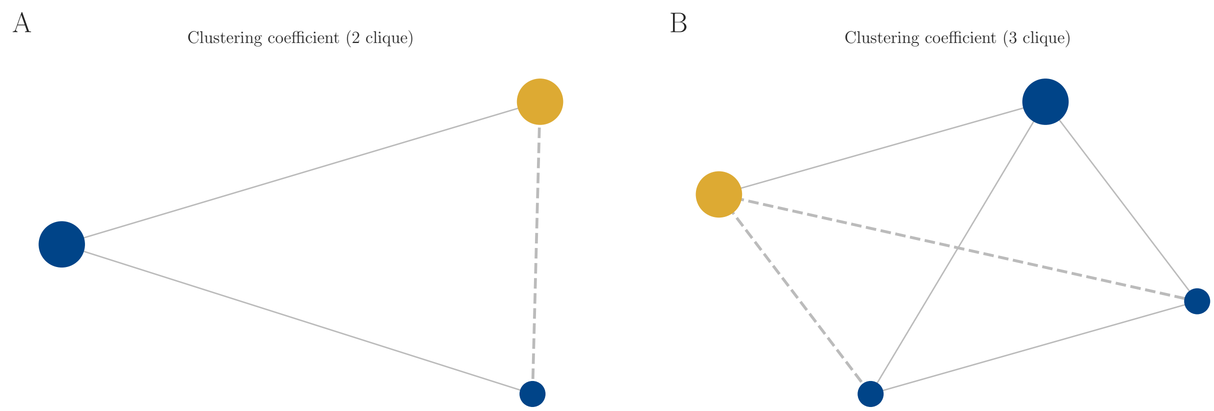

Figure 1. Visualization of the local higher-order clustering coefficient[2]. The blue nodes denote nodes belonging to a clique. The local clustering coefficient of the large blue node can be computed as the share of adjacent edges (as the one to the yellow node) that also have edges to all the other nodes in the clique (i.e., the dotted edges) relative to all adjacent edges.

C Computation of the local higher-order clustering coefficient

Higher-order clustering can be efficiently measured in the bipartite representation of a network with container nodes. The ℓ-th order local clustering coefficient for a node u can be computed via

$$C_\ell(u) = \frac{\ell|\mathscr{K}_{\ell+1}(u)|}{(d_u - \ell - 1)|\mathscr{K}_\ell(u)|} \tag{1}$$

where d_u denotes the degree of u and $\mathscr{K}_\ell$ the set of ℓ-cliques containing u[2]. In the bipartite network, we compute the person-degree of u by taking the length of the set of all person-neighbors in all layers. To compute the total number of the ℓ-sized cliques containing u, it is key to realize that all individuals connected to a container node are completely clustered, i.e., they are all connected to each other. The number of cliques created via a container node can, therefore, be computed using simple combinatorics.

$$\left|\mathscr{K}_\ell^{\mathscr{C}_i}(u)\right| = \binom{|\mathscr{C}_i| - 1}{\ell - 1} \tag{2}$$

where $\mathscr{C}_i$ is the set of person-neighbors of a u-neighboring container node. There is a correction by (-1) because we want to obtain the number of cliques containing u, not all cliques in $\mathscr{C}_i$. As cliques can overlap between different layers and containers (e.g., a person can be both members of the same household and family or be a colleague at two different firms), we have to correct for the overlaps using the principle of inclusion-exclusion when computing the total number of ℓ-sized cliques.

$$|\mathscr{K}_\ell(u)| = \sum_{\mathscr{S}\subseteq\{1,\ldots,|\mathscr{C}|\}} (-1)^{|\mathscr{S}|+1} \binom{|\cap_{s\in\mathscr{S}}\mathscr{C}_s|-1}{\ell-1} \tag{3}$$

where $\mathscr{C}$ is the set containig all $\mathscr{C}_i$'s.

D Weighting scheme

D.1 Computation of weights for colleague and classmate layer

To compute the edge-weights for the colleague and classmate layer, we assume a k-regular-random network structure, i.e., every colleague/classmate is connected to exactly k other random colleagues/classmates. For this, we need to define how many individuals k an average individual is connected to through their class or workplace container. A suitable k in our case is $k = 20$, roughly corresponding to the long-term average classroom size in Denmark[3]. While assuming random, independent formation of edges is comparatively unrealistic, the average shortest path length of a k-regular random networks scales similarly to those of small-world networks of more complex, realistic structures, e.g., Watts-Strogatz or modular hierarchical networks[4,5]. Importantly, the k-regular-random network provides a useful equation for analysis, intuitively relating increased degree of connectivity k to shorter path lengths. Consider an employer/education program size n. For every employer of size n, the average k-random-regular distance is defined as:

$$d = \max\left(1, \frac{\log(n)-\gamma}{\log(k-1)} + \frac{1}{2} + \frac{\log(1-2/k)}{\log(k-1)}\right),$$

where γ is the Euler-Mascheroni constant. We set the weight as the maximum of the average distance in the k-regular-random network and 1 to assure that in small companies the distance between colleagues is of unit length. In a company of 500 employees the estimated distance between two individuals is then 2.13, for the largest workplace present in the dataset it is 3.2.

D.2 Robustness checks

Figure 2 illustrates the reasonability of our sampling approach for shortest paths. The plot presents the distribution as a kernel density function of a varying number of source nodes to 10,000 random target nodes. We observe quick convergence of the distribution with increasing size of sampled source (target) nodes (cf. Figure **??**). Figure 3 depicts changes in the distribution of shortest paths when the edge-weights of individual layers are either halved or doubled and all other weights are left constant compared to the scheme suggested in Subection **??**. The adjustments of the weights lead to minor shifts in both the overall distribution and increased dispersion, especially for the family layer. Overall the distribution remains in a reasonable range compared to other social networks[6–8]. The robustness check demonstrates the necessity to adjust the weights to the research question at hand. Figure 4 demonstrates the distribution of distances in the bipartite network when there are either no weights (4A) or distances between classmates and colleagues are not adjusted for their size (4B). With all edge-weights equal to one the distances become more concentrated. Note that here, almost none of the shortest paths lead through the neighborhood layer as not setting any weights implies a distance of 5 between neighbors. Adjusting the classmate and colleagues weights to uniform length leads to significantly smaller average shortest paths. Similarly, the distribution of distances when not adjusting any weights utilizing the individual-centered view becomes more concentrated around 4 and smaller on average (mean distance in the random sample is 4.3) (4C). Compared to other social networks these are implausibly small, underlining the need for a meaningful adjustment of weights.

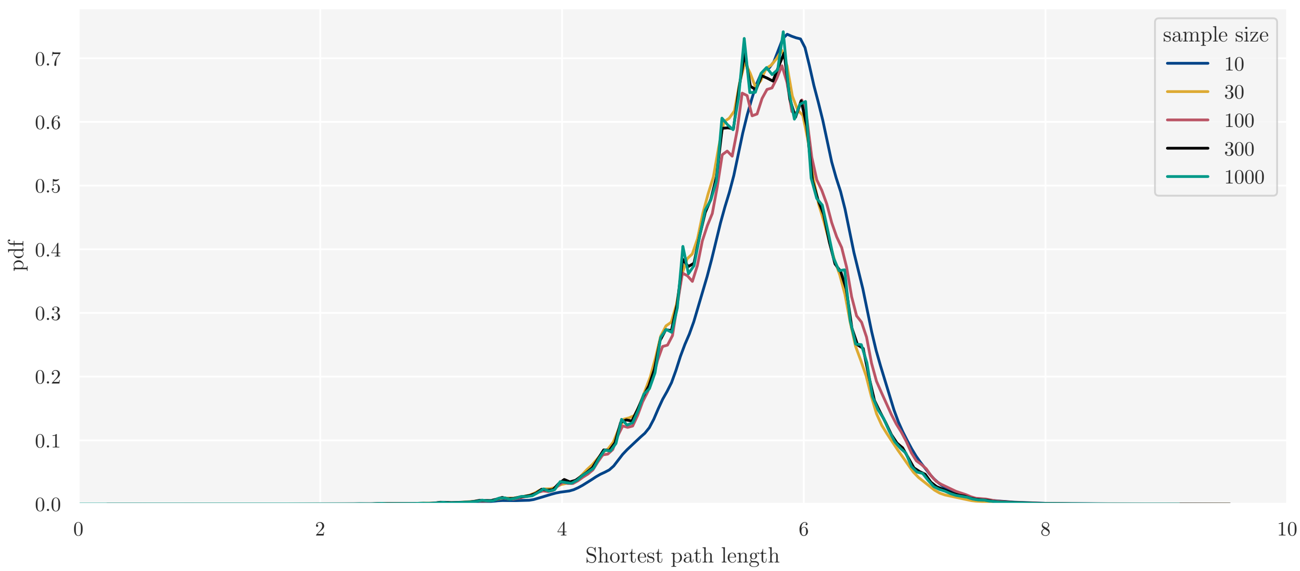

Figure 2. **Shortest Path Distribution with Varying Sample Sizes.** Density of the distribution of shortest paths of a varying number of random nodes to 10000 other random nodes in a single year network.

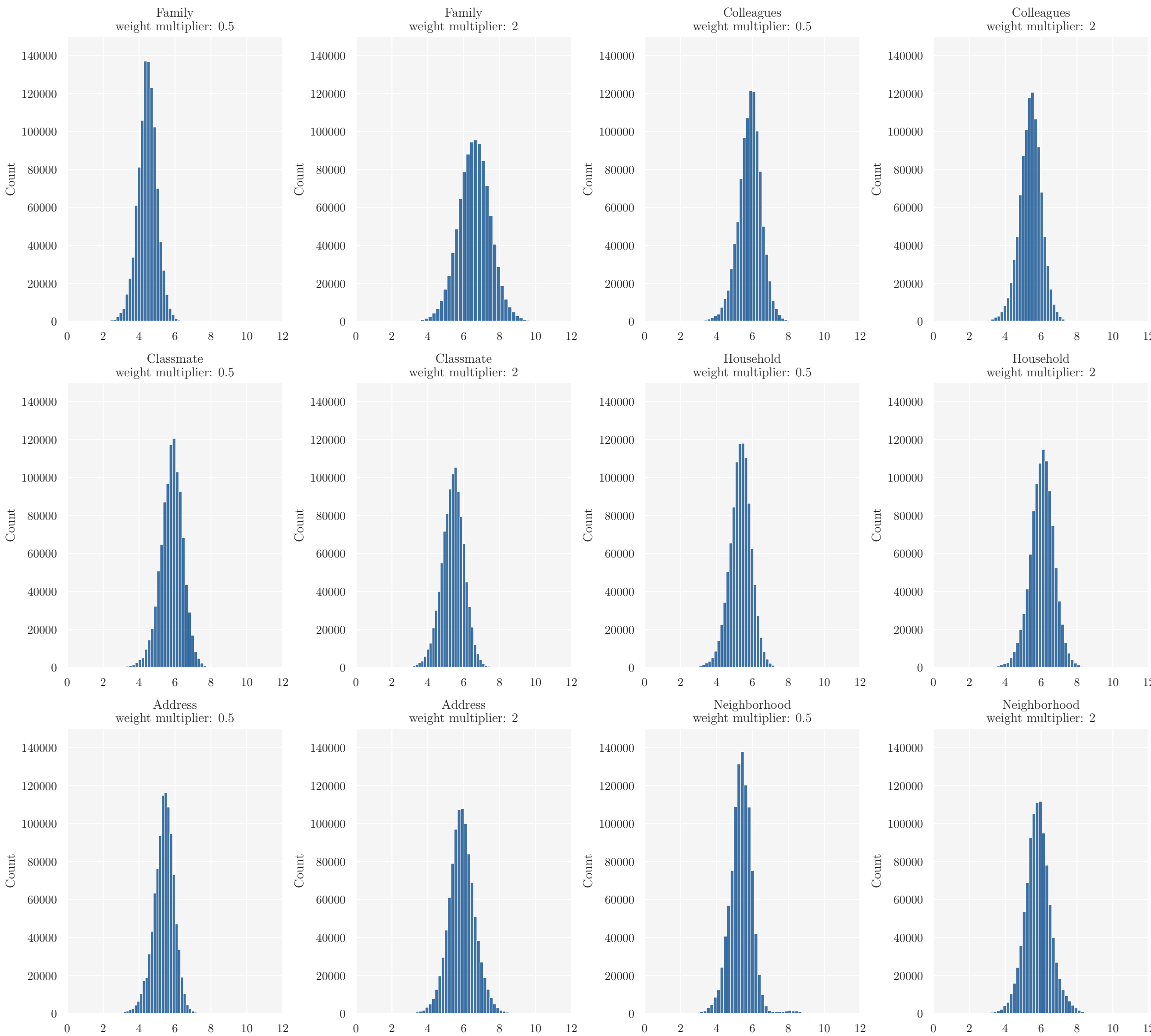

Figure 3. Shortest Path Distribution with Varying Edge Weights. Distribution of shortest paths from 100 random nodes to $1,000$ other random nodes in a single year network, varying the weights on a single layer and leaving the other edge-weights as described in the Subsection **??** and Subection D. The title of each plot indicates the layer with changed edge-weights and the factor with which the weights described in the above mentioned sections are multiplied (left plot: half the weight, right plot double the weight).

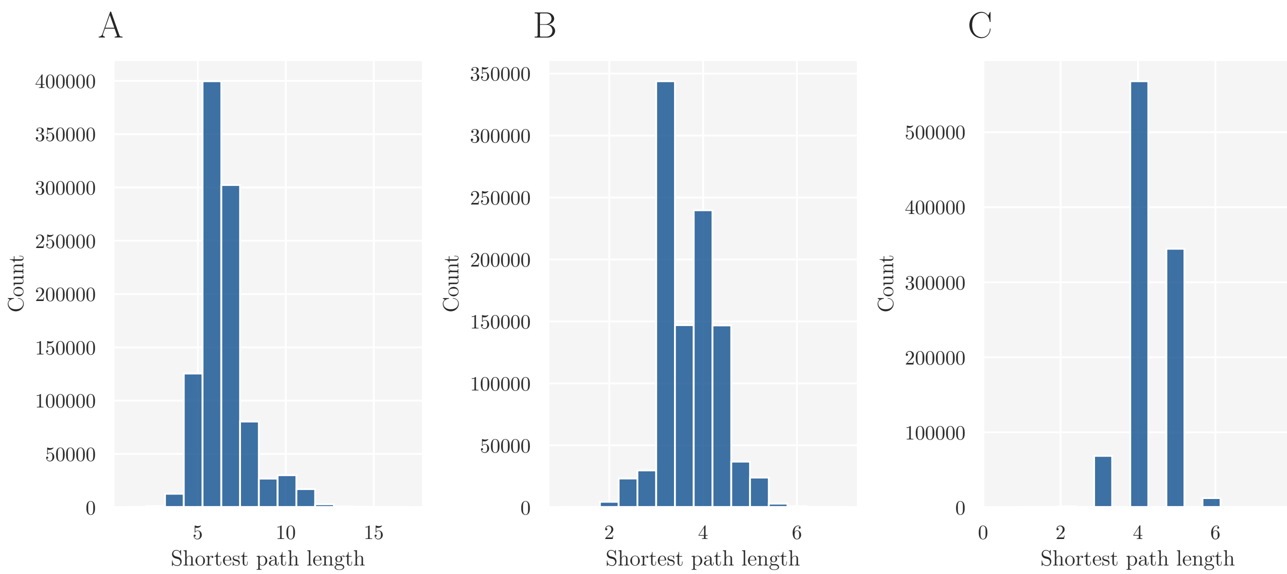

Figure 4. Shortest Path Distribution With Naive Weights. Distribution of shortest paths from 100 random nodes to $1,000$ other random nodes, when not assigning any weights in the bipartite view(Panel A) or assigning uniform weights to the colleagues and classmate layers of $1/2$ in the bipartite view (Panel B). Not assigning any weights in the bipartite view implies that neighbors have distance 5 (see Figure **??**, while setting uniform distance means that the distances between colleagues and classmates are 1 independent of the size of the workplace or class. Panel C illustrates not assigning any weights in the individual-centered view (with sampling for larger workplaces and classes).

E Number of unique edges table

View	Bipartite		Individual	
Time-span	0	10	0	10
Layer				
Classmate	2.64	13.26	105.11	404.02
Colleagues	3.09	11.34	125.07	800.55
Family	-	-	33.65	37.48
Household	11.68	24.75	4.58	7.58
Neighborhood	12.49	19.57	49.420	180.76
Total	29.90	68.91	311.22	1411.54

Table 1. Total Number of Unique Edges by Time-span, Layer and View (in Million). Note that the number of family edges are identical between the views.

F Detailed layer definitions

F.1 Family layer

The family relations from which this layer is created are recovered from the Danish population registry[9] which includes all individuals alive and living in Denmark from 1968 onward. This registry is augmented with archival data on marriages and births from the Danish national archives to reconstruct as many relations between parents and children alive in 1968 for individuals born in or after 1953.

All family relations are extrapolated from the parent-child relations in the registries. If a new family relation is created, e.g. via the birth of a child, this relation is included in the edge list for a specific year independently of the day of the year the relation is created. Note that to recover as many relations as possible deceased individuals were included in the data, as well.To include additional family relations, for example family by marriage, the family layer can be linked to the population register which includes information on marital status. Note however that this is not a simple process of generating an edge-list, but will also involve changing the source code for generating the family layer as well as requires access to raw-data from the registries from which the family layer is created.

Concerning coverage, all individuals in the data born in 2019 and 2020 have known parents and the percentage of individuals with known parents is over 73% for all individuals born in 1952 or later. For individuals born before 1952 the pro-portion of individuals with known parents varies from 10% for those born in 1940 to 27% for those born in 1951.

F.2 Household layer

Statistics Denmark computes a household identification number using information on address, age, and marital status from the population registry. The household information included in the household layer concerns households as registered on the first of January in a specific year. We do not allow for individuals to belong to more than one household in the same calendar year. The household identification number used in the current version of the network is available from 1986 and is updated every three months. Individuals in a household live at the same address, and a household consists of an individual or a couple with or without children.

Children have the same household identification number as their parents if

- they live at the same address as at least one of the parents,
- are under 25 years of age,
- have never been married or in a registered partnership,
- don't have children of their own that are registered in the central person registry (CPR),
- are not part of a cohabiting couple.

A couple is defined as two individuals that live together and are one of the following types:

- married couple
- registered partnership (from October 1st, 1989)
- living together and have at least one child that is registered in the CPR
- cohabiting defined as two individuals of opposite sex with an age-difference of under 15 years that do not have any children registered in the CPR and are not close family (brother-sister / parent-child) according to the CPR. In addition, the household has to contain only adults.

Children not living at home are under 18 years of age and each make up their own family. Technically they are counted as single. To be considered a child not living at home the following conditions need to be fulfilled:

- don't live at the same address as either of their parents
- be under 18 years of age
- never have been married or in a registered partnership
- do not have children registered in the CPR
- are not part of a cohabiting couple

Rules for changes in family identification number over time:

- When individuals in a couple move apart they each get a new id.
- When two individuals move together and create a couple they both get the same new id.
- When children move out of the home they get their own new id but the id for the rest of the family does not change.
- When children that previously moved out move back with their parents again before they turn 25 they get the same identification number they previously had.
- When one individual in a couple dies, the surviving part and their children living at home still get a new id. The deceased individuals children that are not also children of the surviving individual each get their own new id.
- When the couple type (see above) changes or the couple separates but still live together, the family id does not change.

Identification numbers that have been decommissioned are not re-used. Note that in some cases multiple identification numbers are registered for the same address. This is for example in case of three adults living at the same address where none are married or cohabiting. But also institutional households such as college dorms or care homes are examples of this.

Disambiguation: Note that the household definition used in the current version of the network corresponds to what in layman's terms may be regarded a family. The definition of a household used at Statistics Denmark may thus lead to cases in which one would expect individuals to be members of the same household but are not labelled as such according to our definition: e.g. children over 25 living at home do not form a household together with their parents and roommates do not form a household either. In the first case an edge between parent and child will however be present in the family layer, and in both cases an edge is likely in the neighborhood layer.

F.3 Neighborhood layer

In the neighborhood layer we include edges between an individual and the members of the geographically closest 10 households on the first of January of a specific year (as defined in the household layer) within a distance of 50 meters. This is the same approach as[10] take. Households for which the exact location is not known are disregarded at present. When multiple households exist with the same distance, households are selected uniformly at random. Note that when households are sampled this does not automatically lead to symmetry in the sense that if individual i is sampled as neighbor to individual j, individual j is not necessarily sampled as neighbor to individual i.

Note that the above definition implies that in sparsely populated areas, there may be under 10 households within a 50 meter radius. It may be worth considering and it is possible using the bipartite data to define other distances in the network to reflect other definitions of neighborhood. The address and exact geographical location of almost every household is known and included in the bipartite table for the neighborhood layer.

Disambiguation: Note that since a household in the network is what in layman's terms may be seen as a family, the neighborhood may for some individuals easily capture mainly other families in the same household, since their distance is 0 meters. E.g. for a person living in a collective, with 11 individuals, with no children and no married couples will not have any neighbors that are not part of the same address, as the individual's 10 closest households will all have distance zero. So in the colloquial meaning of the word "neighbor", such a person would have zero neighbors. Note also, that apart from such extreme edge cases, even 4-person collectives will then have at most 7 neighbors in the colloquial sense, even though the person will have 10 neighbors in the neighborhood-layer.

F.4 Colleague layer

In this layer, we use the information on work location from the labor market registry[9]. This registry contains a yearly status indicating the place(s) of work on the 30th of November of all individuals present in the Danish population on the 1st of January of the same year. For small workplaces (≤ 100 employees) we include edges between all employees in the yearly edge-list for the individual-centered view. For large workplaces (> 100 employees) we take a random sample of size 100 to create edges between colleagues. We choose such a limit to make sure that the edge lists do not become too large (the amount of edges grows at a rate of $\propto n^2$, and because it is unrealistic that an individual actually meets all colleagues at a large workplace. The network file can however be adapted to adjust the workplace size boundary that determines whether we sample colleagues or not. Note that when individual i is sampled as a colleague for individual j, we make sure to include an edge in the other direction as well, such that individual j is included as a colleague for individual i. This condition leads to some individuals from sampled workplaces having less than 100 colleagues (especially for those with < 200 employees). A final thing to consider regarding the sampling is the fact that if one is sampled as a colleague in one year this does not automatically mean one is also sampled for other years (even if both individuals still work at the same work-place). For the bipartite-view we do not sample as here the number of edges only grows linearly with the number of colleagues compared to the quadratic growth in the individual-centered view. Finally note that individuals that have a so-called 'fictional workplace', meaning individuals that work remotely or at multiple locations in the same municipality, are count-ed as having no colleagues. Independent business owners and spouses involved in their businesses are not included either.

F.5 Classmate layer

In this layer we use information on enrolled students in primary, secondary, and tertiary education from the Danish education registers[11] to create edges between students enrolled at the same school/program in the same grade (primary and secondary education) or cohort (tertiary eduction). The data cannot be used to distinguish between multiple classes within the same grade/cohort. The data contain information from the 1st grade of primary school (from the year an individual turns six) onward. Note that an edge is created independent of how much overlap (in days) there was during the enrolled period. Also individuals that were enrolled in the same grade/cohort for just one day are considered classmates. For the bipartite-view, programs are included as container nodes to which individuals are connected. The different grades in primary and secondary education programmes are coded as their own study programmes, i.e. 1st grade and 2nd grade are coded as separate study programmes that have a 1-year length. For example, three students all starting in first grade in 2010 in the same school will be connected to one another in 2010. However when in 2011 one of the students skips a grade and therefore will start in 3rd grade in 2011, only two of the students (those that start 2nd grade in 2011) are connected in the network in 2011. Tertiary education study programmes can encompass multiple years, and we select only students from the same cohort (=start-year) to be connected.

Author contributions statement

J.C.: Designed the network and supervised its creation, B.K.: Responsible for the description and analysis of the network, B.F.M.: Developed the bipartite view of the networks including the edge-weighting scheme and wrote the *regnet* Python package, S.N.E., F.C.: Assisted in the creation of the network, S.L.J., A.B.N., L.H.M., D.D.L.: Conceived the original idea, B.K., J.C., J.E., L.H.M., A.B.N.: Wrote the paper with input from all authors A.B.N.: Supervised the description and analysis

Additional information

The authors declare no competing interests. DDL is deputy chairman of the Board at Statistics Denmark. This paper replaces a previous version presenting the dataset[12].

References

1. Maier, B. F. Regnet - A package for the creation and efficient shortest-paths computation in registry networks (2024).
2. Yin, H., Benson, A. R. & Leskovec, J. Higher-order clustering in networks. *Phys. Rev. E* **97**, 052306, DOI: 10.1103/PhysRevE.97.052306 (2018). ArXiv:1704.03913 [cond-mat, physics:physics, stat].
3. DST. Primary and lower secondary education (2023).
4. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442, DOI: 10.1038/30918 (1998).
5. Watts, D. J., Dodds, P. S. & Newman, M. E. J. Identity and Search in Social Networks. *Science* **296**, 1302–1305, DOI: 10.1126/science.1070120 (2002).
6. Stopczynski, A. *et al.* Measuring Large-Scale Social Networks with High Resolution. *PLoS ONE* **9**, e95978, DOI: 10.1371/journal.pone.0095978 (2014).
7. Leskovec, J. & Horvitz, E. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, 915–924, DOI: 10.1145/1367497.1367620 (Association for Computing Machinery, New York, NY, USA, 2008).
8. Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. The Anatomy of the Facebook Social Graph. *ArXiV 1111.4503* DOI: 10.48550/ARXIV.1111.4503 (2011). Publisher: [object Object] Version Number: 1.
9. Petersson, F., Baadsgaard, M. & Thygesen, L. C. Danish registers on personal labour market affiliation. *Scand. J. Public Heal.* **39**, 95–98, DOI: 10.1177/1403494811408483 (2011).
10. Van Der Laan, J., De Jonge, E., Das, M., Te Riele, S. & Emery, T. A Whole Population Network and Its Application for the Social Sciences. *Eur. Sociol. Rev.* **39**, 145–160, DOI: 10.1093/esr/jcac026 (2023).
11. Jensen, V. M. & Rasmussen, A. W. Danish education registers. *Scand. J. Public Heal.* **39**, 91–94, DOI: 10.1177/1403494810394715 (2011).
12. Bjerre-Nielsen, A. *et al.* Dataset Profile: A Danish Nation-Scale Network, DOI: 10.31235/osf.io/9wsx7 (2024).