Changing Societies & Personalities, 2023
Vol. 7, No. 4, pp. 71–92
https://doi.org/10.15826/csp.2023.7.4.252

ARTICLE

Grouping of Occupations Based on Intragenerational Mobility

Sergey A. Korotaev, Elena N. Gasiukova
National Research University Higher School of Economics, Moscow, Russia

ABSTRACT
Occupation is a key factor in human thinking, feeling, and behavior. Theoretically derived occupational groupings or classes are typically used to transform occupations into a variable suitable for statistical manipulations. We argue that such groupings are unlikely to produce groups that are homogeneous across a broad set of attributes. Instead, we offer a data-driven approach to identify groups of occupations based on respondents' mobility data using network analysis. The vertices of the network are codes of occupations, and the edges reflect the number of transitions between them. Using modularity maximization, we identify four communities and evaluate the stability of the resulting partition. As an example demonstrating the efficiency of the resulting grouping, we present a comparison of the predictive power of this grouping and one of the generally accepted groupings of occupations, that is ESeG (European Socio-economic Grouping), in relation to the human attitudes and values found in previous publications. The results indicate the preference of our grouping.

KEYWORDS
occupations, network analysis, ISCO codes, intra-generational mobility, career trajectories, modularity networks, class

ACKNOWLEDGEMENT
The research was supported by the Russian Science Foundation (project No. 22-28-20426).

Received 19 November 2023
Accepted 21 December 2023
Published online 27 December 2023

skorotaev@hse.ru, egasyukova@hse.ru

Occupation is a key variable in sociology and other social sciences. As a class, it is often used as a predictor of a vast number of indicators ranging from political behavior to health, including values (Booth, 2020; Kohn, 1969), tolerance (Adriaenssens et al., 2022), moral position, and sentiments (De Keere, 2020; Sayer, 2010). Occupation influences (directly and indirectly) thinking, feeling, and behavior throughout the life course, which different authors (Stephens et al., 2014) extensively discuss. However, regardless of the chosen subject of analysis and the theoretical approach used, every research inevitably implies the homogeneity of the analyzed professional groups in terms of properties essential for the work. A necessary (but not necessarily sufficient) condition for this is the existence of boundaries for mobility between such groups. Indeed, it is hardly possible to talk about specific cultures or personality types inherent in occupations (groups of occupations) *a* and *b* if their representatives regularly move from *a* to *b* or from *b* to *a*. Identifying such relatively closed groups presents a certain methodological complexity and requires a large amount of data. Not long ago, it was suggested to use network analysis to solve this problem (Toubøl & Larsen, 2017; Toubøl et al., 2013). On the basis of occupation transition data, occupational groupings have been proposed for a number of countries via network analysis (Cheng & Park, 2020; Schmutte, 2014; Toubøl & Larsen, 2017; Toubøl et al., 2013). In Russia, for the implementation of this task data from the Russian Longitudinal Monitoring Survey of Higher School of Economics (RLMS HSE) suitable; however, to the best of our knowledge, there are no studies dedicated to creating groups of occupations based on mobility data. The aim of this research is to solve the above problem.

Occupation as a Variable

To become a variable in the model, an “occupation” must go a long way. First, the interviewer asks a respondent to describe in a free form what they do at work. Next, an encoder selects a code for this description according to a certain system of occupation codes (for example, ISCO08, see below). Then a researcher works with coded occupations and no longer refers to the respondents’ initial answers (as a rule, they do not have access to those answers). In this research, we do not intend to problematize the use of existing occupational classifiers; however, it is worth paying attention to one aspect important for the present study.

Allocating a code to an occupation can be represented as a translation from the language of the agents involved—people who describe their own and other people’s occupations, based on their own practical interests—into the language of universal classifiers. With such (as with any) translation, there is probably a discrepancy between differences that are significant in the local context (e.g., those allowing to find a proper vacancy or an employee for a position), with those differences that seemed important to the members of the commission that create international classifiers. A possible empirical manifestation of such a scenario is the frequent change of occupation codes when re-surveying the same respondent with the same job, because two codes, which are different from the point of view of classification, equally correspond to a “real” occupation. For authors who study the level of mobility, such changes in occupation

codes present an obstacle because they overestimate the level of mobility (see Cheng & Park, 2020). Our research has a primarily methodological focus; we are interested in the relationships between occupational codes (ISCO08), so such artificial transitions are a legitimate source of data for us.

In their original form, occupation codes can hardly be used in analysis because of their large number. Therefore, the next step is to convert the occupation codes into a variable suitable for inclusion in the model. One of the options for such a transformation is to create a numerical variable, for example, a scale of prestige or social status, which associates each occupation code with a certain value at this scale.[1] In this research, we are interested in another approach that is more common, namely, combining occupational codes into several groups, usually called classes.

Typically, researchers use a priori, that is theoretically driven, schemes, primarily the Goldthorpe class scheme and its derivative European socio-economic classification, or ESeC (Rose & Harrison, 2007). The shortcoming of using such a grouping is that the connection of the principles underlying the grouping with the phenomena of interest to the researcher remains indirect (at best). For example, the Goldthorpe scheme is based on an assumed difference in the types of contracts between employees and employers. Although there may indeed be significant differences in culture (practices, dispositions, personalities) between the resulting classes, it is easy to doubt whether such a grouping best explicates these differences.

An alternative is data-driven approaches that group occupations in an empirical manner. If the targeted groups/classes have some internal homogeneity (and therefore intergroup differences) in the traits of their members, we can expect the composition of these groups to be relatively stable over time. Indeed, intensive population exchange between groups would lead to the disappearance of intergroup differences. Composition stability means a relatively high level of mobility (change of occupations) within groups and a low level of mobility between them. Thus, mobility tables (inter- or intra-generational) are the starting point for many studies of this kind. A mobility table is a square table, the rows and columns of which are the names of occupations/groups of occupations (e.g., origin by row and destination by column), and the values in the cells are the number of respondents with the corresponding initial and final occupations. Log-linear analysis has been long used as the main method of such tables analyzing. It models the logarithm of frequency in a cell as a linear combination of the coefficients of the row, column, and their interaction. Typically, log-linear analysis is used to study the nature of transitions between given categories, and not to aggregate these categories.[2] However, criteria have been proposed that allow testing of the statistical admissibility of certain category combinations (Breiger, 1981; Goodman, 1981). The criteria allow testing of a small number of theory-driven hypotheses about possible combinations of categories. In fact, these criteria are applied to a very small initial list of occupational groups subject to consolidation, and often serve to update or adapt already existing groups. Thus, the log-linear analysis of mobility tables does not allow implementing a truly data-driven approach to the grouping of occupations.

[1] See Lambert & Griffiths (2018); also, see Bessudnov (2012) for an example of analysis using Russian data.

[2] As an example of log-linear analysis of mobility in Russia, see Yastrebov (2016).

Network analysis is an approach that has recently gained popularity among mobility researchers. It allows overcoming the above limitations of log-linear analysis. Cheng and Park (2020) present a more detailed comparison of various approaches. The source data is still the mobility table described above. The vertices of the network are occupations connected by an edge if the value in a table cell at the intersection of the corresponding row and column is greater than zero. There are different approaches to analyze such mobility networks: a clique-based algorithm (Toubøl & Larsen, 2017; see also Toubøl et al., 2013), which can cluster only those vertices that form a clique, that is a subset of vertices, each pair of which is connected by an edge; Infomap algorithm used by Cheng and Park (2020); community detection based on modularity maximization (e.g., Schmutte, 2014), which is used in this work and will be discussed below. Before moving on to the description of the research methodology, it is necessary to consider the structure of the occupation codes considered in this research, namely, ISCO08.

International Standard Classification of Occupations (ISCO08)

In 1952, the International Labour Organization (ILO) published the first International Classification of Occupations for Migration and Employment (ICOMEP), with descriptions of 1,727 occupations based on national reference materials (International Labour Organization, n.d.). In 1957, the first complete version of the International Standard Classification of Occupations was presented at the International Conference of Labor Statisticians (ICLS), and ISCO-58 was officially published in 1958. Later, ISCO was modified several times to comply with the changing realities of the labor market and its use in cross-country comparisons. The codes were last updated in 2008.

The ISCO structure is hierarchical, with occupational codes corresponding to the characteristics of a job, which is defined as a set of tasks and duties, and the required skill level for its fulfillment where a skill is defined as the ability to perform tasks and duties (International Labour Organization Department of Statistics, n.d.). Each occupation has a four-digit code: the first digit reflects the level of qualification and the scope of activity (major groups), each subsequent digit details the scope of activity and tasks performed by workers, the second digit determines sub-major groups, and the third one identifies minor groups. In total, the codebook contains 436 four-digit codes, 130 three-digit codes, 43 two-digit codes, and 10 single-digit codes. If an encoder cannot assign one of the four-digit codes to the respondent based on the available data, then they may limit themselves to assigning only the first three, two, or even one digit(s). Thus, the total number of unique codes found in the database is extremely large, and many of them are few in number.

Since the position of sparse codes in the network can be largely random, we needed to take a number of steps to exclude numerically insignificant categories from the analysis. First, ISCO08 codes were consolidated to the first three digits (minor groups). Furthermore, a numerically insignificant three-digit code could be attached to a code of a more numerous group within a common sub-major group (two-digit code), in which case the combined category received the name of the more numerous group

code. In some cases, when no logical grouping of three-digit codes would achieve the required group size, all codes within that sub-major group were combined into one group, which received the name of the sub-major group. However, three of the sub-major groups were still too small in number (95, 94, 73); therefore, they were attached to the more numerous sub-major groups within the corresponding major groups. Some codes could not be aggregated in this way, so they were assigned to a separate category "other" and excluded from the created network. Thus, agricultural workers were excluded (major group 6 and sub-major group 92). In addition, the "others" category included one- and two-digit codes assigned when the encoder could not reliably attribute the respondent's occupation to a certain minor group (three-digit code). The exception was group with code 83, which, for some reason, turned out to be quite numerous. The total share of codes classified as "others" turned out to be insignificant. The unemployed also formed a separate category, which was excluded from the network creation. The ISCO08 source codes, descriptions of the corresponding occupations, and the categories to which they were assigned are shown in Table 1. In total, 48 categories were obtained (excluding "other" and "unemployed"); hence, the created network would have 48 vertices.

Table 1

Correspondence of Aggregated Categories and ISCO08 Codes, Membership of Categories in Communities and ESeG-Groups

Community	ESeG	Categories (network vertices)	ISCO08 three-digit codes	Description
1 (blue)	1	11	11_	Chief executives, senior officials, and legislators
1 (blue)	1	12	12_	Administrative and commercial managers
1 (blue)	1	132	131	Manufacturing, mining, construction, and distribution managers
			132	Manufacturing, mining, construction, and distribution managers
			133	Information and communications technology service managers
1 (blue)	1	134	134	Production and specialized services managers
1 (blue)	1	14	14_	Hospitality, retail and other services managers
1 (blue)	2	21	21_	21 science and engineering professionals
3 (brick)	2	22	22_	22 health professionals
3 (brick)	2	231	231	University and higher education teachers
3 (brick)	2	233	232	Secondary education teachers
			233	Vocational education teachers

Table 1 Continued

Community	ESeG	Categories (network vertices)	ISCO08 three-digit codes	Description
3 (brick)	2	234	234	Primary school and early childhood teachers
3 (brick)	2	235	235	Other teaching professionals
1 (blue)	2	24	24_	Business and administration professionals
1 (blue)	2	25	25_	Information and communications technology professionals
1 (blue)	2	261	261	Legal, social, and cultural professionals
			262	Librarians, archivists, and curators
1 (blue)	2	263	263	Social and religious professionals
			264	Authors, journalists, and linguists
1 (blue)*	2	265	265	Creative and performing artists
2 (green)	3	311	311	Physical and engineering science technicians
			314	Life science technicians and related associate professionals
2 (green)	3	312	312	Science and engineering associate professionals
2 (green)	3	313	313	Process control technicians
3 (brick)	3	32	32_	Health associate professionals
1 (blue)	3	331	331	Financial and mathematical associate professionals
1 (blue)	3	332	332	Sales and purchasing agents and brokers
1 (blue)	3	333	333	Business services agents
1 (blue)	3	334	334	Administrative and specialized secretaries
1 (blue)	3	335	335	Government regulatory associate professionals
1 (blue)	3	34	34_	Legal, social, cultural, and related associate professionals
1 (blue)	5	41	41_	General and keyboard clerks
1 (blue)	5	42	42_	Customer services clerks
1 (blue)	5	43	43_	Numerical and material recording clerks
1 (blue)	5	44	44_	Other clerical support workers

Table 1 Continued

Community	ESeG	Categories (network vertices)	ISCO08 three-digit codes	Description
4 (yellow)	7	51	51_	Personal service workers
1 (blue)	7	52	52_	Sales workers
3 (brick)	5	53	53_	Personal care workers
4 (yellow)	5	54	54_	Protective services workers
2 (green)	6	71	71_	Building and related trades workers
2 (green)	6	72	72_	Metal, machinery, and related trades workers
2 (green)	6	74	74_	Electrical and electronics trades workers
2 (green)	6	75	73_	Handicraft and printing workers
			75_	Food processing, woodworking, garment, and other craft and related trades workers
2 (green)	6	81	81_	Stationary plant and machine operators
2 (green)	6	82	82_	Assemblers
4 (yellow)	6	83	83	Drivers and mobile plant operators
2 (green)	6	831	831	Locomotive engine drivers and related workers
4 (yellow)	6	832	832	Car, van, and motorcycle drivers
4 (yellow)	6	833	833	Heavy truck and bus drivers
2 (green)*	6	834	834	Mobile plant operators
			835	Ships' deck crews and related workers
3 (brick)*	7	91	91_	Cleaners and helpers
			94_	Food preparation assistants
2 (green)	7	93	93_	Laborers in mining, construction, manufacturing, and transport
2 (green)	7	96	95_	Street and related sales and service workers
			96_	Refuse workers and other elementary workers
Others			0, 2, 3, 7, 8, 9, 13, 23, 33, 35, 61, 62, 92, 310, 315	

Note. * unstable membership, _ any number or lack thereof. Source: developed by authors on the basis of International Labour Organization Department of Statistics (n.d.).; Meron et al. (2014).

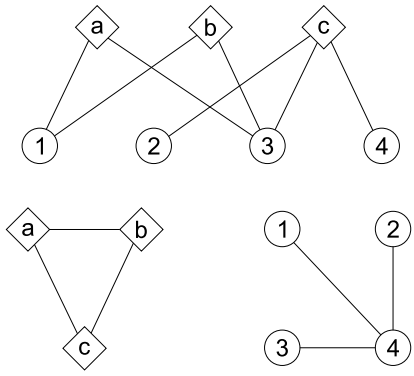
Methodology

In this article, networks are used to model the relationships between occupations and people. A network with two types of vertices is called a bipartite network. It connects only the vertices of different types, for instance, people and occupations in our study. However, neither occupations nor people may be connected with each other. We also analyze networks consisting of uniform vertices: only workers or only occupations. If a connection between vertices is seen as a binary indicator (either there is a connection, or there is not), the network is unweighted. If the edges can connect vertices with greater or lesser "strength," then they should have a corresponding "Weight" value; such a network is called weighted. Finally, connections between vertices may be directed, e.g., we can separately analyze the flow of people moving from occupation *a* to occupation *b,* and the reverse flows from *b* and *a*; such a network is called directed. In this study, we ignore the direction of movement and consider only its intensity; hence, we only deal with undirected networks.

Let us look at an example of the network type considered in this article. We have four respondents and know their occupations for four consecutive years: 1—*abba*, 2—*cccc*, 3—*abcc*, 4—*cccc*, where the numbers represent the respondents, and the letters represent their occupations in chronological order. Figure 1 shows a bipartite network built on these data. The edges connect occupations and people who worked during the period under review. The network is not weighted; therefore, repeated tenure of a person in the occupation does not affect the image in any way. Hence, if we discard the last year of observations for our four respondents, the figure will not change. The figure shows that respondents 1 and 3 have highly overlapping work experience (occupations *a* and *b*), whereas the same people match occupations *a* and *b*. We may assume that these occupations require similar skills or attract people with similar educational backgrounds.

Figure 1

An Example of a Bipartite Network



Note. Diamonds mean occupations and circles mean people (above), and its weighted projection (below). Bold edges have a weight of 2, thin edges have a weight of 1.

In a bipartite network, only vertices of different types can be connected, but we are primarily interested in the relationships between occupations. To explicate them, we can consider two projections: occupations and people (see Figure 1, lower part). The edges connect those occupations, in which the same people are found. Similarly, in another projection, the edges connect respondents with experience in the same occupation. People who have experience in only one occupation (e.g., 2 or 4) do not affect the occupation network in any way. The projections are weighted, and the weight of the edges is equal to the number of common occupations/people. We can see that occupations *a* and *b*, and people 1 and 2, are connected by the edge with a greater weight than others are.

Assume that we have data (see the next section) on the successively occupied occupations for k years of a large number of respondents N, the number of occupations O was determined above. Based on these data, an "occupation–person" network can be built. In this case, its projections are networks of occupations and networks of persons. The next stage is the identification of the groups in the network. Within these groups, occupations are more closely connected with each other than with representatives of other groups. Such groups are called communities; one of the popular strategies for finding them is modularity maximization. The concept of modularity was introduced by Newman and Girvan (2004) for unweighted networks and was further extended to weighted ones (Newman, 2004a). The value is the difference between the proportion of edges connecting vertices within the same community, and the same value if we preserve the degrees of vertices (the number of edges emanating from a vertex) in our network but otherwise connect vertices together at random.

Let us denote the adjacency matrix by **A**, each of its elements A_{ij} corresponds to the weight of the edge between vertices i and j (if any) or 0 (if none). Since the networks considered in this article are undirected, then $A_{ij} = A_{ji}$. The sum over a row or column is equal to the degrees of the corresponding vertex. For a partition into l groups, we denote the membership of the i vertex as $c_i \in \{1, 2, \ldots, l\}$. Then the number of vertices that do not cross the boundaries of communities will be:

$$\frac{1}{2}\sum_{i,j} A_{ij}\ \delta(c_i, c_j), \tag{1}$$

where $\delta(x, y)$ is the Kronecker symbol, equal to 1 if $x = y$, otherwise equals 0; the coefficient 1/2 is due to the fact that in the matrix **A** each edge occurs twice, like A_{ij} and A_{ji}. Next, for a random placement of edges, we calculate the mathematical expectation of the weight of the edge between vertices i and j that have degrees k_i and k_j. The sum of cells in **A** squares gives the total number of edges taking into account their weight m. An edge with weight f will be considered as f edges of unit weight, and then in total we have m edges with $2m$ endpoints. If one endpoint belongs to vertex i, then its second endpoint can be any of $2m - k_i$, we subtracted the number of ends lying on i, because in our network, edges leaving and entering the same vertex are impossible. The probability that the second endpoint hits vertex j is then $k_j/(2m - k_i)$. The mathematical expectation of the weight of the edge between vertices i and j is equal to $k_i k_j/(2m - k_i)$. Replacing A_{ij} in formula (1) with this value, we obtain

the number of edges that do not cross the boundaries of communities in the case of a random placement of edges (at fixed degrees of vertices):

$$\frac{1}{2}\sum_{i,j}\frac{k_j k_i}{2m-k_i}\,\delta(c_i,c_j). \tag{2}$$

Subtracting (2) from (1) and dividing by m to obtain the fraction, we arrive at the modularity formula:

$$\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_j k_i}{2m-k_i}\right)\delta(c_i,c_j).$$

Real modularity maximization algorithms ignore the subtraction of k_i in the denominator, setting $k_i << m$ (for criticism, see, e.g., Cafieri et al., 2010):

$$\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_j k_i}{2m}\right)\delta(c_i,c_j). \tag{3}$$

One way to maximize modularity is to use a fast greedy algorithm (Newman, 2004b). In the first step, each of the O occupations is assigned to a separate cluster. At each subsequent step, the algorithm chooses which two communities to combine into one to maximize the increase or minimize the decrease in modularity because of this action. The algorithm operation can be represented as a dendrogram, depicting the steps at which groups merged. This work uses the implementation of the algorithm in the igraph R package (Csárdi et al., 2023).

Given a community structure resulting from modularity maximization, it would be highly desirable to know the extent to which this structure reflects the properties of the population and is not an artifact of a particular sample. Typically, in statistics, numerical tests are used to identify the significance of a result; however, such tests can hardly be adapted for our purposes. The most obvious solution is to check the stability of the resulting partition by multiple repetitions of the analysis on a set of bootstrap samples. If it is true for each pair of vertices that they (almost) always end up together in the same community or (almost) always end up in different communities, then such a partition should be considered stable. Based on this logic, a numerical measure of a partition robustness was proposed (Shizuka & Farine, 2016). However, we are interested not so much in an aggregate indicator of a partition quality, but in the stability of each specific occupation occurrence in the cluster, to which it is assigned. Such information will allow us to identify the core and periphery of communities, which is important from the point of view of both substantive interpretation and methodology.

Data

We used data from 15 waves of the RLMS HSE from 2007 to 2021 for analysis. The age of the respondents was limited to 25–60. An earlier stage of a career may be associated with casual part-time jobs for students, and therefore is not indicative for our analysis.

When using several waves of panel data, the first task is to arrange them into a representative sample for the planned study. The simplest and most obvious

strategy is to use a balanced panel, which means selecting those respondents who, at t_1 time were included in a representative sample (designated as S_{cs1}, where cs1 is cross-sectional in t_1) and were interviewed at all subsequent moments t_2, t_3, … t_k (designated as S_{bp1}). If the probability of missing one or more waves from the 2nd to k is the same for all members of S_{cs1} (i.e., they have the same probability of getting into S_{bp1}), then the S_{bp1} sample may also be considered representative. This simplest case is usually referred to as missing completely at random, or MCAR (Rubin, 1976). A more plausible assumption is that the probability of attrition is not equal, but can be predicted based on the observed (at t_1) characteristics of the respondent (missing at random, MAR). In this case, the representativeness of S_{bp1} can be ensured by selecting the following weights:

$$w_{\text{a}i} = \frac{1}{1 - p(r_i = 1|x_{1i})},$$

where i is the respondent's identifier, r is a binary indicator of attrition from S_{bp1} (1 = attrition), x_1 is a vector of the respondent's characteristics at t_1 time, respectively $p(r = 1| x_1)$ is the probability that the respondent missed one or more waves, provided that in the first wave they had characteristics of x_1. This probability can be calculated using logistic regression, then

$$\log \frac{p(r = 1|x_1)}{1 - p(r = 1|x_1)} = \beta x_1,$$

where β is the set of estimated coefficients, and the vector x_1 starts at 1, the respondent's index is omitted. Therefore,

$$w_{\text{a}} = 1 + \exp(\beta x_1). \quad (4)$$

Considering the task, the expected assumption is that the labor market in the period under review is homogeneous in time. This allows us to take several *k*-year samples, the beginning of which fall on different waves (S_{bp1}, S_{bp2}, ...), and create a pooled sample, which will also be a representative sample of the *k*-year labor trajectories of Russians. Let us note the following points:

- S_{bp1} member must be included in the representative sample S_{cs1}, but is not required to be included in S_{cs2}, although it must be included in the full RLMS sample in year t_2;
- when pooling samples S_{bp}, the same record of the same person can be duplicated several times. For example, when a respondent's set of occupations for four years ($abcd$) is known and they were in a representative sample for at least the first two years (window size k = 3), S_{bp1} contains the sequence of occupations abc, S_{bp2} contains the sequence of occupations bcd. As a result, the combination bc will appear twice in the final sample;
- the larger k, the more unique combinations of occupations are included in the final sample, so for the mentioned respondent with k = 3, five combinations will be recorded such as ab, bc (twice), ac, cd, bd; with k = 4, the six combinations will be those listed earlier + ad.

Based on the above, we can describe the trade-off associated with the choice of k parameter. On the one hand, with large k, we are more efficient in using the data of each real respondent (who did not miss any of the k waves), since we do not lose the combination of their occupations, as well as data from respondents who left the representative sample, but were interviewed in subsequent years. On the other hand, with k increasing, the proportion of respondents to be excluded from the pooled sample also increases, along with the duplication of data.

In order to discuss further, we should clarify the meaning of original and duplicated data. Consider a respondent with an $abab$ career trajectory of four years who has been in a representative sample for all that time. If we assume k = 4, he will be in our pooled sample once, on the bipartite network he will appear as a vertex with two edges leading to a and b. If k = 3, it will fall into the pooled sample twice, in the first year with the aba trajectory, and in the second one with the bab trajectory. So, this implies two vertices with two edges each; if k = 2, there are three such vertices. In fact, this is equivalent to assigning a weight of 2 (k = 3) or 3 (k = 2) to the initial "original" information that the respondent's career combines occupations a and b. The weights themselves are not a big problem; however, when they reach large values, they can lead to instability in estimates.

We can select the hyperparameter k in a more formal way by taking two parameters as criteria, namely the total number of original transitions in the sample and the proportion of the number of original transitions from the total number of transitions between occupations in the reconstructed network. These values were obtained numerically for k from 2 to 7; the observed trends indicate that further search is impractical. The results are presented in Table 2. In general, with k = 4 we obtain the least loss of information relative to the full base. The ratio of original transitions and those used in the model is also acceptable. Large k's, as can be seen, do not provide any advantage. The presented calculation has been carried out for the aggregated categories presented above. However, for original ISCO codes k = 4 is also optimal (not presented).

Table 2

Sample Parameters for Different Lengths (k) of the Considered Career Trajectory

k	Number of original transitions	Total number of transitions	Share of original transitions
2	9,510	12,848	0.74
3	11,511	21,184	0.50
4	12,158	26,315	0.46
5	11,676	28,588	0.41
6	11,458	29,656	0.39
7	11,013	29,245	0.38

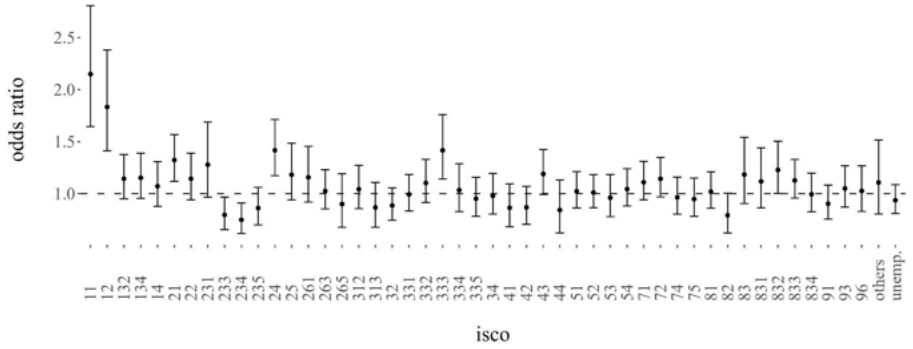
Note. Source: authors.

By choosing the value of k, we can calculate the attrition-related weights. To do this, we will use logistic regression, the predictors of which are age (linear and quadratic), as well as the aggregated categories. The dependent variable is a binary indicator equal to zero if the respondent's occupational status (ISCO code or information that

the respondent is unemployed) is known for all four years, otherwise, it is equal to one. The probability of dropping out of the sample decreases monotonically with age at the age interval in question (not presented), the odds ratio of ISCO categories is presented in Figure 2. We can see that, in general, representatives of groups with higher status have a greater chance of missing the survey, probably because of the higher spatial mobility of such people. The age effect can be explained in the same way. The final weight of an observation in the created sample is determined by two weights: the w_{cs} (cs is cross-sectional) weight associated with the sample design, which is contained in the RLMS database, and the w_a attrition weight calculated by formula (4): $w_t = w_{cs} \times w_a$.

Figure 2

Estimated Odds Ratios by ISCO Group for Attrition (95% CI)



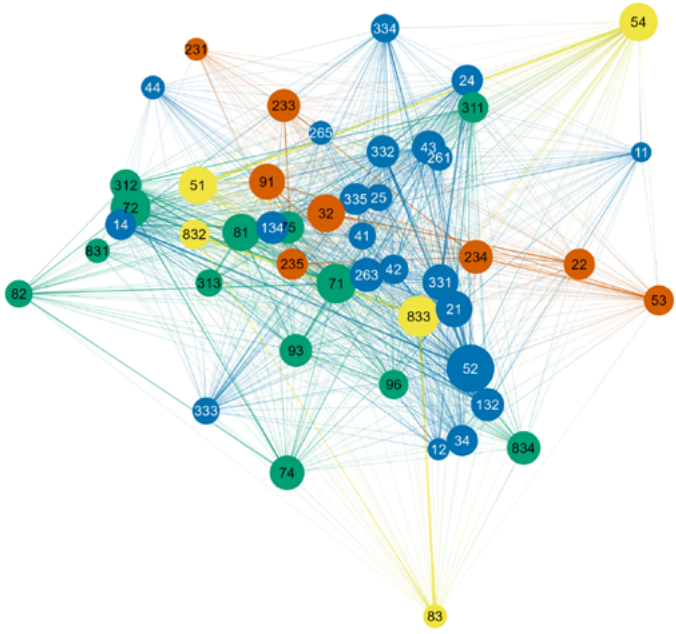
Note. The reference group is 311 in ISCO08. CI = confident interval.

If the MAR assumption is accepted, then the database consisting of 12 balanced samples (2007–2018), each of which is composed of four-year career trajectories of respondents from a representative part of the RLMS, is a representative sample of four-year career trajectories of Russians (taking into account the w_t weights). However, a worse scenario is also possible, when the probability of non-response is determined by the value of unobservable indicators (missing not at random, MNAR). For example, the transition to occupation g is often accompanied by a relocation, then in our sample the connections of this occupation with others will be underrepresented, despite the use of w_t weights. Working with MNARs (and MNAR testing itself) is a complex task. We can offer the following simple check. If we take $k = 2$ and compare the profile of respondents' occupations in the representative sample of RLMS vs. the occupations of the second year of respondents in our sample, then they should not differ significantly under MAR. If such differences are found, MNAR occurs. This check was carried out using the Chi-Square test of independence with Rao-Scott correction; the Rao-Scott correction is necessary because the samples in question have weights. The differences turned out to be insignificant even at the 10% level, which allows us to reject the MNAR hypothesis.

Results and Discussion

The network with communities obtained from modularity maximization is shown in Figure 3. The achieved modularity value is 0.32. The dendrogram of the vertex mergers resulting from the fast greedy algorithm application is shown in panel A of Figure 4. In panel B of Figure 4, for each pair of occupations, the share of bootstrap resamplings where they ended up in the same community is shown in heatmap format. Only a few occupations have unstable membership. For example, code 265 gravitates to both the first (bottom to top in the left panel, blue) and the third (brick) community. Code 834 is toward the second (green) or fourth (yellow) community. The yellow community tends to split into two, i.e. 51 and 54 versus 83, 833, 832, in cases where the algorithm finds five communities rather than four (in 24% of resamplings). If this scenario were to occur, 91 people could also be categorized in the first of these clusters (51 and 54). Note that all examples of unstable membership refer to occupations that were included in the corresponding communities during one of the last mergers, while most of them (except perhaps 265 and 83) cannot be regarded as small in number. Hence, we believe that the unstable membership of occupations is not a consequence of sample limitations but reflects the actual relationships between the ISCO codes.

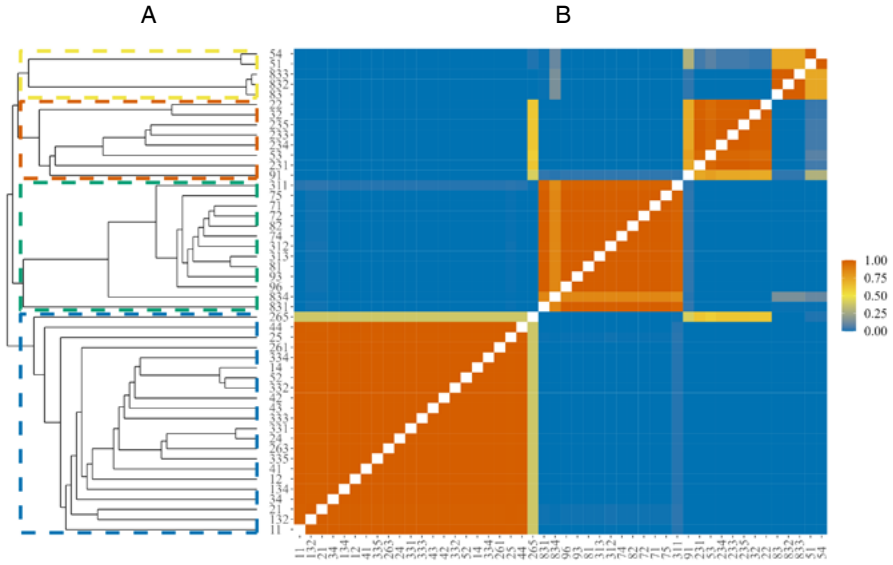
Figure 3
Network of Occupations



Note. The numbers correspond to occupational groups, the thickness of the lines reflects the weight of the connection, and the colors indicate community membership: blue—1, green—2, brick—3, yellow—4.

Figure 4

The Resulting Partition: Dendrogram of Community Merging and Stability Check



Note. Panel A: Dendrogram of community merging when running the fast greedy algorithm. The colors indicate community membership: blue—1, green—2, brick—3, yellow—4. Panel B: Heatmap of the probability of occupations belonging to the same community by row and column (among bootstrap reassemblies).

The most obvious question that arises when creating a new grouping is how different it is from those previously proposed. In the present paper, we will limit ourselves to comparing the resulting scheme with ESeG (European Socio-Economic Groups, see Meron et al., 2014), elaborated at the request of Eurostat to create a pan-European grouping of occupations based on ISCO08 (the previous ESeC grouping was originally created for ISCO88). The declared purpose of this scheme is to ensure "grouping of individuals with similar economic, social and cultural characteristics." To encode ESeG, the first two digits of the ISCO code and employment data (self-employed or employee) are used. Grouping is possible without considering employment data, in which case the fourth group, "small entrepreneurs," is not allocated. We assume that two groupings are consistent if the boundaries of one do not intersect the boundaries of the other, i.e., each group of one grouping is either completely included in a group of the other, or completely includes several groups of the other. The correspondence of the ESeG grouping to the resulting communities is shown in Table 1. If the first ESeG grouping is completely included in the first community, then the second group is divided between the 1st and the 3rd communities, each of the third, fifth, and sixth groups is included in three communities at the same time (1, 2, 3; 1, 3, 4; 2, 3, 4), and the sixth group is included in four communities at once. In general, we can conclude that ESeG does not reproduce the boundaries of occupational mobility that we identified. Therefore, if members of the ESeG group are similar in one cultural dimension, we cannot be sure that they will automatically be similar in another.

Of course, one study is not enough to doubt the applicability of ESeG to the analysis of Russian data; however, the obtained result indicates the need for further research and a critical attitude towards the application of existing schemes for ISCO code grouping.

Next, we move on to a meaningful interpretation of the resulting communities. Socio-demographic indicators of communities in 2007 and 2021 are shown in Table 3 and Figure 5. Because we excluded some occupations related to agriculture from the analysis, we provide tables that represent only the urban population. Let us start with the smallest community—the fourth one (yellow). It includes occupations such as service sector workers (51: cooks, waiters, hairdressers, etc.), security guards (54), and drivers (832, 833, drivers also statistically prevail in category 83). Among representatives of occupations in this group, men predominate (Table 3), as well as older respondents (Figure 5).

Table 3

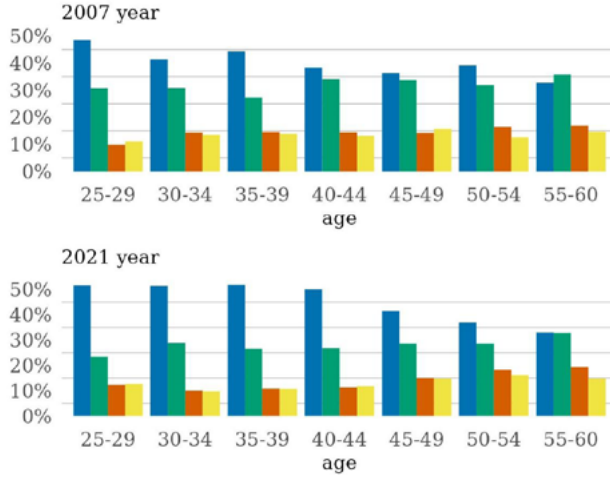
Socio-Demographic Characteristics of Urban Respondents from Founded Occupational Communities

Community	Women's share, %		Median personal income, rub.		Mean personal income, rub.	
	Year		Year		Year	
	2007	2021	2007	2021	2007	2021
1 blue	66.5	65.0	11,953	34,965	14,360	43,389
2 green	27.0	22.6	9,926	33,939	11,715	38,129
3 brick	91.7	89.3	6,948	29,887	8,543	33,266
4 yellow	21.0	22.5	8,911	34,919	10,771	37,959

Note. Source: authors.

Figure 5

Occupational Structure of Age Cohorts



Note. Sum of columns in each cohort is 100%. The colors indicate community membership: blue—1, green—2, brick—3, yellow—4.

The third community (brick) includes health (associate) professionals (22, 32), teachers (23), and personal care workers (53, e.g., birth assistants and nannies). The above occupations belong to *biudzhetniki* [public servants], the most feminized group, as approximately 90% of their representatives are women; this is the oldest and lowest-income group (Table 3, Figure 5).

The second community (green) includes skilled workers (7), workers and operators from ISCO08 major groups 8 and 9, as well as science and engineering associate professionals (31). Predictably, this category includes the largest proportion of men, with their predominance only increasing from 2007 to 2021 (Table 3).

Finally, the first (blue) group is the largest, and it includes the highest status occupations: managers (1) and most professionals (except those who belong to the second community); this group also includes almost all white-collar semi-professionals and clerks. The group has the most balanced gender composition among all, although women are still the majority (Table 3). Unlike other groups, the proportion of representatives of this group is growing among younger generations, i.e., under 40 years old (Figure 5), and they gravitate most towards the largest cities (not shown). Collectively, the group brings together a diverse group of office staff living primarily in large cities. Our analysis did not reveal the existence of boundaries between career trajectories within this community.

In the introduction, we pointed to the use of occupational groupings as predictors of respondent personality attributes. The simplest demonstration of the effectiveness of the resulting communities in comparison with other groupings (ESeG) is their ability to better predict the characteristics of respondents. A few indicators found in the papers were considered: tolerance towards gays and lesbians[3] (Adriaenssens et al., 2022), attitudes towards redistribution,[4] which is one of the indicators of moral worldviews (De Keere, 2020; see also Kulin & Svallfors, 2013), attitudes towards migration[5] (Davidov et al., 2018); gender equality attitudes[6] (for example, Soboleva, 2019); wealth values[7] as the ESS human values indicator (Davidov et al., 2008).

The simplest models were used, in which the predictors were age, gender, and membership in an occupational group (see Table 4). Results show that older generation is less tolerant (to gays, lesbians, and migrants), do not strongly strive to make expensive purchases and more likely advocate for redistribution and gender equality. The communities of office workers, teachers, and doctors are more tolerant towards sexual minorities and greet gender rights of being employed for both sexes. Also, office workers have the least pronounced attitude towards redistribution.

[3] Question B34. *Gay men and lesbians should be free to live their own lives as they wish* (European Social Survey, 2016).

[4] Question B33. *The government should take measures to reduce differences in income levels*. The value corresponds to the degree of agreement (European Social Survey, 2016).

[5] Question B41. *Would you say it is generally bad or good for country's economy that people come to live here from other countries?* Values rise with approval (European Social Survey, 2016).

[6] Question B33a. *Men should have more right to job than women when jobs are scarce*. The value corresponds to the degree of agreement (European Social Survey, 2016).

[7] Question H1 CARD 76_B. *Important to be rich, have money and expensive things* (not like me at all/ not like me/ a little like me/ somewhat like me/ like me/ very much likely) (European Social Survey, 2016).

Biudzhetniki differ significantly from other communities in that they do not value wealth. Of greatest interest is the comparison of the predictive power of community-based and ESeG-based models. The R^2 for regressions with communities is higher in all models. Since our grouping has fewer categories than ESeG does, the gap in adjusted R^2 is even more significant between the models. In addition, *F*-statistics was used to compare the fits of different models. The check showed that the extended models (included both variables of ESeG and occupational communities) provided a better fit than a model with only ESeG occupational indicator.

Table 4

Prediction of Tolerance and Attitudes Towards Redistribution With Help of Founded Grouping and ESeG

	Tolerance	Attitudes towards redistribution	Attitudes towards migration	Gender equality attitudes	Wealth values
Age	−0.008*	0.007*	−0.013*	−0.008**	−0.024***
Gender (ref. = male)	0.063	0.128*	−0.144	−0.529***	0.075
Community (ref. = 1 blue):					
2 green	−0.25**	0.266***	−0.501**	0.203*	−0.17+
3 brick	0.069	0.226**	−0.212	0.057	−0.306**
4 yellow	−0.263*	0.214*	−0.355	0.354**	−0.166
R^2	0.019	0.022	0.011	0.088	0.052
Adj. R^2	0.016	0.018	0.008	0.085	0.048
ESeG R^2	0.018	0.014	0.007	0.085	0.049
ESeG adj. R^2	0.012	0.008	0.001	0.08	0.043
F-statistics (level of sig.)	5.0%	0.0%	5.0%	5.0%	10.0%
N	1,286	1,375	1,297	1,370	1,382

Note. Source: authors. $^+p < .1$, $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Conclusion

In the presented paper, we attempt to create a grouping of ISCO08 occupational codes based on data on respondents' mobility. Presumably, the relatively impenetrable boundaries between groups constitute a prerequisite for obtaining groups that are relatively homogeneous on attributes unknown in advance. Hence, the criterion for identifying groups is a high level of transitions between occupational codes within the group and a low level across group boundaries.

The RLMS data for 2007–2021 were used to realize this grouping, resulting in a network whose vertices are occupation codes while the edges are the number of transitions between these codes. Through modularity maximization, four communities

were identified. Their stability was assessed using a variety of bootstrap resamplings and proved to be quite high. The first community, being the largest, unites mainly office staff, the second one comprises workers, operators, and non-office associate professionals, the third one includes occupations in the fields of education and medicine, and the fourth community consists of occupations in the service sector, security guards, and drivers.

As an example demonstrating the efficiency of the resulting grouping, a comparison was made of the predictive power of this grouping and the European Socio-economic Grouping (ESeG) in relation to the indicators known in the literature. The results indicate that our grouping is preferable.

The proposed grouping may already be used as an alternative or an addition to the known occupational (class) schemes. However, the performed research has some limitations, overcoming which will allow us to obtain a more reliable grouping. First, only one of several possible ways to construct a network and identify communities was used (for other approaches, see Cheng & Park, 2020; Toubøl & Larsen, 2017). It is desirable to have a grouping that is consistently reproduced regardless of the method used. Different approaches are based on different initial assumptions; a comparison of the results obtained can serve as a diagnostic tool for the ISCO classification. Second, the resulting groups have an unbalanced gender composition, where some communities are predominantly female and others are male. In their research, Toubøl & Larsen (2017) obtained similar results. This may indicate the existence of gender-specific patterns of mobility, which requires further testing. Third, we examined four-year career spans ($k = 4$), but we can imagine a situation where the transition from profession *a* to *b* is quite common, but involves an intermediate stage of being in profession *c.* In this case, four years may not be sufficient to capture many transitions from *a* to *b* in the data; therefore, checking the stability of the graph partition with respect to different values of k is desirable. Finally, the question of the stability or change of these groups over time is important.

References

Adriaenssens, S., Hendrickx, J., & Holm, J. (2022). Class foundations of sexual prejudice toward gay and lesbian people. *Sexuality Research and Social Policy*, *19*(1), 63–84. https://doi.org/10.1007/s13178-020-00525-y

Bessudnov, A. (2012). A relational occupational scale for Russia. In P. Lambert, R. Connelly, B. Blackburn, & V. Gayle (Eds.), *Social stratification: Trends and processes* (pp. 53–68). Ashgate.

Booth, D. E. (2021). Post-materialism's social class divide: Experiences and life satisfaction. *Journal of Human Values*, *27*(2), 141–160. https://doi.org/10.1177/0971685820946180

Breiger, R. L. (1981). The social class structure of occupational mobility. *American Journal of Sociology*, *87*(3), 578–611. https://doi.org/10.1086/227497

Cafieri, S., Hansen, P., & Liberti, L. (2010). Loops and multiple edges in modularity maximization of networks. *Physical Review E*, *81*(4), Article 046102. https://doi.org/10.1103/PhysRevE.81.046102

Cheng, S., & Park, B. (2020). Flows and boundaries: A network approach to studying occupational mobility in the labor market. *American Journal of Sociology*, *126*(3), 577–631. https://doi.org/10.1086/712406

Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F., & Noom, D. (2023, August 12). *igraph for R: R interface of the igraph library for graph theory and network analysis* (v1.5.1). Zenodo. https://doi.org/10.5281/zenodo.8240644

Davidov, E., Cieciuch, J., & Schmidt, P. (2018). The cross-country measurement comparability in the immigration module of the European Social Survey 2014–15. *Survey Research Methods*, *12*(1), 15–27. https://doi.org/10.18148/srm/2018.v12i1.7212

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly*, *72*(3), 420–445. https://doi.org/10.1093/poq/nfn035

De Keere, K. (2020). Finding the moral space: Rethinking morality, social class and worldviews. *Poetics*, *79*, Article 101415. https://doi.org/10.1016/j.poetic.2019.101415

European Social Survey (ESS). (2016). *ESS Round 8 source questionnaire*. ESS ERIC Headquarters. https://stessrelpubprodwe.blob.core.windows.net/data/round8/fieldwork/source/ESS8_source_questionnaires.pdf

Goodman, L. A. (1981). Criteria for determining whether certain categories in a cross-classification table should be combined, with special reference to occupational categories in an occupational mobility table. *American Journal of Sociology*, *87*(3), 612–650. https://doi.org/10.1086/227498

International Labour Organization (ILO). (n.d.). *International standard classification of occupations: Brief history.* https://www.ilo.org/public/english/bureau/stat/isco/intro2.htm

International Labour Organization Department of Statistics (ILOSTAT). (n.d.). *International standard classification of occupations (ISCO): Classification.* https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/

Kohn, M. (1989). *Class and conformity: A study in values*. University of Chicago Press.

Kulin, J., & Svallfors, S. (2013). Class, values, and attitudes towards redistribution: A European comparison. *European Sociological Review*, *29*(2), 155–167. https://doi.org/10.1093/esr/jcr046

Lambert, P. S., & Griffiths, D. (2018). *Social inequalities and occupational stratification: Methods and concepts in the analysis of social distance.* Palgrave Macmillan. https://doi.org/10.1057/978-1-137-02253-0

Meron, M., Amar, M., Laurent-Zuani, A. C., Holý, D., Erhartova, J., Gallo, F., Lindner, E., Záhonyi, M., Váradi, R., Huszár, A., & Franco, A. (2014). *Final Report of the ESSnet on the harmonisation and implementation of a European Socio-economic classification: European Socio-economic Groups (EseG)*. National Institute of Statistics and Economic Studies. https://circabc.europa.eu/sd/a/519eafb9-186c-4e2c-a178-902f28501ba4/DSS-2014-Sep-08c%20ESSnet-ESeG%20-%20Final%20Report.pdf

Newman, M. E. (2004a). Analysis of weighted networks. *Physical Review E*, *70*(5), Article 056131. https://doi.org/10.1103/PhysRevE.70.056131

Newman, M. E. (2004b). Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*(6), Article 066133. https://doi.org/10.1103/PhysRevE.69.066133

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), Article 026113. https://doi.org/10.1103/PhysRevE.69.026113

Rose, D., & Harrison, E. (2007). The European socio-economic classification: A new social class schema for comparative European research. *European Societies*, *9*(3), 459–490. https://doi.org/10.1080/14616690701336518

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sayer, A. (2010). Class and morality. In S. Hitlin & S. Vaisey (Eds.), *Handbook of the sociology of morality* (pp. 163–178). Springer. https://doi.org/10.1007/978-1-4419-6896-8_9

Schmutte, I. M. (2014). Free to move? A network analytic approach for learning the limits to job mobility. *Labour Economics*, *29*, 49–61. https://doi.org/10.1016/j.labeco.2014.05.003

Shizuka, D., & Farine, D. R. (2016). Measuring the robustness of network community structure using assortativity. *Animal Behaviour*, *112*, 237–246. https://doi.org/10.1016/j.anbehav.2015.12.007

Soboleva, N. (2019). *Gender attitudes and achievement motivation across Europe (The evidence of ESS data)* [Research paper No. WP BRP 88/SOC/2019]. National Research University Higher School of Economics. https://wp.hse.ru/data/2019/10/18/1530677358/88SOC2019.pdf

Stephens, N. M., Markus, H. R., & Phillips, L. T. (2014). Social class culture cycles: How three gateway contexts shape selves and fuel inequality. *Annual Review of Psychology*, *65*, 611–634. https://doi.org/10.1146/annurev-psych-010213-115143

Toubøl, J., & Larsen, A. G. (2017). Mapping the social class structure: From occupational mobility to social class categories using network analysis. *Sociology*, *51*(6), 1257–1276. https://doi.org/10.1177/0038038517704819

Toubøl, J., Larsen, A. G., & Jensen, C. S. (2013, May 21–26). *A network analytical approach to the study of labour market mobility* [Paper presentation]. 33rd Sunbelt Social Networks Conference of the International Network for Social Network Analysis (INSNA), University of Hamburg, Hamburg, Germany.

Yastrebov, G. A. (2016). Sotsial'naia mobil'nost' v sovetskoi i postsovetskoi Rossii: Novye kolichestvennye otsenki po materialam predstavitel'nykh voprosov 1994, 2002, 2006 i 2013 gg. Chast' II [Social mobility in Soviet and post-Soviet Russia: A revision of existing estimates using representative surveys of 1994, 2002, 2006 and 2013. Part 2]. *Universe of Russia*, *25*(2), 6–36.