Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport 04/2017 EN

Methodological aspects of labour market data

Skill-relatedness matrices for Germany

Data method and access

Frank Neffke,
Anne Otto,
Antje Weyh

Bundesagentur für Arbeit

Skill-relatedness matrices for Germany

Data method and access

Frank Neffke (Harvard University)
Anne Otto (IAB)
Antje Weyh (IAB)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Abstract

This document provides a brief overview on the construction of skill-relatedness matrices for Germany. The datasets for these matrices are available for free online download on the homepage of the Research Data Centre at the IAB. We explain the statistical procedure to derive the skill-relatedness measure and describe the resulting skill-relatedness matrices with network graphs. The downloadable datasets contain estimates of skill-relatedness between all pairs of industries using four different industrial classification systems and among all pairs of occupations (KldB88). More detailed information on these issues is to be found in the companion paper by Neffke et al. (2016).

Zusammenfassung

In diesem Dokument wird ein kurzer Überblick über die Erstellung von Skill-relatedness-Matrizen gegeben. Die Datensätze, welche diese Matrizen enthalten, können auf der Homepage des Forschungsdatenzentrums am IAB frei heruntergeladen werden. Wir erläutern die statistische Vorgehensweise zur Berechnung des Maßes der Skill-relatedness und beschreiben diese Matrizen mit Hilfe von Netzwerkgraphen. Die zum Download verfügbaren Dateien enthalten für vier verschiedene Wirtschaftszweigklassifikationen jeweils das berechnete Maß der Skill-relatedness zwischen allen Paarkombinationen unter den Wirtschaftszweigen (WZ73, WZ93, WZ03 und WZ08). Dieses Maß wurde ebenfalls zwischen den verschiedenen Paarungen aus allen einzelnen Berufen (KldB88) ermittelt und steht als Matrize zur Verfügung. Detailliertere Informationen zu diesen Aspekten finden sich im Aufsatz von Neffke et al. (2016).

Keywords: Labor flows, industry, occupations, skill-relatedness, matrices, network graphs.

1 Introduction

This document provides a brief overview on the construction of skill-relatedness matrices for Germany. At the Research Data Centre of the IAB, the datasets for these matrices are available for free online download under the following link: http://doku.iab.de/fdz/reporte/2017/MR_04-17_EN_data.zip. We explain the statistical procedure underlying these estimates and describe the resulting skill-relatedness matrices. The downloadable datasets contain estimates of skill- relatedness between all pairs of industries using four different industrial classification systems and among all pairs of occupations. More detailed information on the methodology used to measure skill-relatedness and a description of empirical patterns can be found in the companion paper by Neffke et al. (2016). When using these data, please cite the companion paper with the following reference:

Neffke, Frank; Otto, Anne; Weyh, Antje (2016): Inter-industry labor flows. (IAB-Discussion Paper, 21/2016), Nürnberg.

Skill-relatedness aims to provide a quantitative estimate of similarities between two industries or occupations in terms of their human capital requirements. Measuring the relatedness among industries in a direct way using skill-use information would require an exhaustive description of the human capital needs of each industry or occupation that would allow for a pairwise comparison of these human capital needs. Furthermore, a full set of weights for each item (e. g., skill) used to describe these human capital requirements would be needed. Due to these obvious difficulties to gather this information accurately and exhaustively for all industries in the economy, we approach the measurement of skill-relatedness in an indirect way. For this purpose, we rely on information on the ease of movement among industries or occupations revealed in job switches.

2 Data

2.1 Data source

To measure skill-relatedness we use the Employee History (German: Beschäftigten-Historik – BeH)[1], which is based on the social security records of Germany (see Bender/Möller 2010). This dataset contains anonymized sociodemographic and comprehensive employment information (i. e., workers's daily wage, occupation, employment status) for every individual covered by Germany's social security system. Furthermore, the industry and location of each individual's work establishment are known. The Employee History includes information about each worker on 30th of June in every year between 1975 and 2014.

2.2 Labor force definition

From the Employee History database, we select workers who meet a set of conditions that ensure focusing on workers who have a regular full-time job. Workers are included if they are:

- between 18 and 65 years of age on 30th of June;

[1] We denote the Employee History as Historic Employment and Establishment Statistics (HES) in our corresponding paper: Neffke et al. (2016).

- in a full-time employment;
- without missing information on industry, occupation or region of work;
- not coded in one of the following occupational categories (KldB88): family workers (agricultural or non-agricultural), disabled, in re-employment programs (German: Rehabilitanden), caregivers (German: Pflegepersonen), in small employment relations (German: mit Haushaltscheckverfahren gemeldete Arbeitnehmer), trainees with unspecified occupation, interns, job seekers, workers in early or partial retirement, and recipients of disability payments (German: Ausgleichsgeldbezieher) (also see Table 1).

  They are not included if the skill-content of workers in specific industries is unclear. Therefore we exclude certain industries. These industries are listed in Table 1 for each of the classification systems.

2.3 Inter-industry labor flows

Using the above labor-force definition, we focus on workers who change jobs from one establishment to another between two consecutive years. Hence, we only use direct job-to-job flows and omit job switches where a worker has periods of non-employment between two jobs. Such periods of non-employment can only be observed between the 30th June of a given year and the 30th June of subsequent year(s).

These job-to-job switches are tracked in the Employee History using an establishment identifier. Establishments in the Employee History may refer to physical addresses or to a number of different work places that belong to the same firm, but that are located in the same municipality. To prevent that spurious changes in establishment identifiers are mistaken for job switches, we drop all workers who move in larger blocks from one establishment to another. A detailed description of this procedure is described in Neffke et al. (2016, Appendix B) and also in Hethey-Maier/Schmieder (2013).

Inter-industry labor flows are the sum of workers who switch between jobs in two different industries. Inter-occupational flows are defined in an analogous fashion. The Employee History contains two occupational classification systems. The KldB88 covers the period from 1975 to 2010, whereas the new occupational classification (KldB2010) was introduced in 2012. In addition, the Employee History comprises four different industry classification systems. The Classification of Economic Activities 1973 (WZ73) existed from 1975 to 1998. This classification system has no clear correspondence to international classification systems. From 1998 to 2003, the German Classification of Economic Activities 1993 (WZ93) is used. At the 4-digit level, this classification is harmonized with the European Union's Nomenclature statistique des activités économiques dans la Communauté européenne, Revision 1.0 (NACE 1.0). The German Classification of Economic Activities 2003 (WZ2003) covers the years from 2003 to 2008 at the 5-digit level and is derived from the European NACE 1.1 classification. Finally, from 2008 to the present year, the German Classification of Economic Activities 2008 (WZ08) at the 5-digit level is available. This system is equivalent to the European NACE Revision 2.0 (NACE 2.0) classification.

3 Method

In order to infer how related two industries are in terms of their human capital requirements, we compare the observed labor flows between the industries to a null model. In this null model, workers switch industries at random. In particular, we take each industry's labor inflow and outflow as given and assume that a worker that leaves an industry chooses a new industry with a probability equal to the observed inflow ratio of the destination industry. That is, $\hat{F}_{ijt,t+1}$, the number of workers moving from industry i to industry j between the years t and $t+1$ expected under the null model, is given by:

$$\hat{F}_{ijt,t+1} = \sum_j F_{ijt,t+1} \frac{\sum_i F_{ijt,t+1}}{\sum_i \sum_j F_{ijt,t+1}}$$

where $F_{ijt,t+1}$ is the observed number of workers switching from i to j. The first term represents the total outflows from industry i, whereas the second is the total inflow of workers into industry j as a ratio of the sum total of all inter-industry labor flows in the economy.

We use these expected flows as a benchmark to which we compare observed flows. In particular, we calculate the following ratio of observed-to-expected flows:

$$SR_{ijt,t+1} = \frac{F_{ijt,t+1}}{\hat{F}_{ijt,t+1}}$$

Values of $SR_{ijt,t+1}$between 0 and 1 indicate that observed flows are below expected flows, whereas values from 1 to infinity indicate that observed flows are above expected flows. One disadvantage of the $SR_{ijt,t+1}$ quantity is that it has a highly skewed distribution. As a consequence, averages of this variable will be disproportionally affected by observations in the variable's right tail. Therefore, we transform the $SR_{ijt,t+1}$ variable as follows:

$$SR^*_{ijt,t+1} = \frac{SR_{ijt,t+1} - 1}{SR_{ijt,t+1} + 1}$$

This transformation maps the $SR_{ijt,t+1}$ symmetrically around 0. As a consequence, a given degree of overrepresentation of labor flows has the same, yet opposite value as the same degree of underrepresentation of such flows. Finally, we define that each industry (occupation) is perfectly related to itself:

$$\mathrm{SR}^*_{\mathrm{ijt,t+1}} = \frac{\mathrm{SR}_{\mathrm{ijt,t+1}}\text{-}1}{\mathrm{SR}_{\mathrm{ijt,t+1}}+1}$$

$$SR^*_{ii} = 1$$

We calculate $SR_{ijt,t+1}$ for every pair of two consecutive years in the period 1975 to 2014. To improve the precision of the estimates, we average $SR^*_{ijt,t+1}$ across all years in which a given classification system is in use. For confidentiality reasons, we only keep industries in which the inflow or outflow of workers summed across the entire period exceeds 10 industry switchers. Moreover, we round the average $SR^*_{ijt,t+1}$ to the nearest fourth decimal.

4 Skill-relatedness matrices

The downloadable files containing skill-relatedness matrices have the following naming convention:

<name_classification_system>.dta

Classification system names correspond to the headers (WZ73, WZ93, WZ03, WZ08, KldB88) of Table 1. All files have the following structure:

< name_classification_system>_1 – < name_classification_system>_2 – SRt:

- <name_classification_system>_1: industry or occupation of origin
- <name_classification_system>_2: industry or occupation of destination
- SRt: transformed skill-relatedness estimate (SR^*)

The skill-relatedness matrices can be downloaded under the following link: http://doku.iab.de/fdz/reporte/2017/MR_04-17_EN_data.zip. Note that, because labor flows have a direction, typically, SR_ij^*≠SR_ji^*, that is, SR^* is not symmetric.

Moreover, we provide additional information for the labor-flows and the skill-relatedness matrices:

- Table 2 provides an overview of the total inter-industry and inter-occupational flows, as well as the time period covered for each classification system.
- Figures 1 to 5 depict the SR^* matrices graphically for each classification system. To construct these figures, we first symmetrize the SR^* as follows:

$$SR_{ij}^{sym} = \frac{SR_{ij} + SR_{ji}}{2}$$

Next, we calculate the maximum spanning tree of the resulting graph and add the strongest 3N links, where N represents the number of classes in a classification system. The resulting network connects industries (or occupations), the nodes in the network, whenever these industries (occupations) are connected over one of the strongest 3N links among industries (occupations) in the economy. Color codes represent higher-level sectoral aggregates in the classification systems. The size of a node is proportional to the industry's average employment over the entire period.

References

Bender, Stefan; Möller, Joachim (2010): Data from the Federal Employment Agency. In: German Data Forum & Rat für Sozial- und Wirtschaftsdaten (eds.), Building on progress. Expanding the research infrastructure for the social, economic, and behavioral sciences. Vol. 2, Opladen: Budrich UniPress, pp. 943-958.

Hethey-Maier, Tanja; Schmieder, Johannes F. (2013): Does the use of worker flows improve the analysis of establishment turnover? Evidence from German administrative data. In: Schmollers Jahrbuch, Vol. 133, No. 4, pp. 477-510.

Neffke, Frank; Otto, Anne; Weyh, Antje (2016): Inter-industry labor flows. (IAB-Discussion Paper, 21/2016), Nürnberg.

Appendix

Table 1: Dropped industry and occupation categories

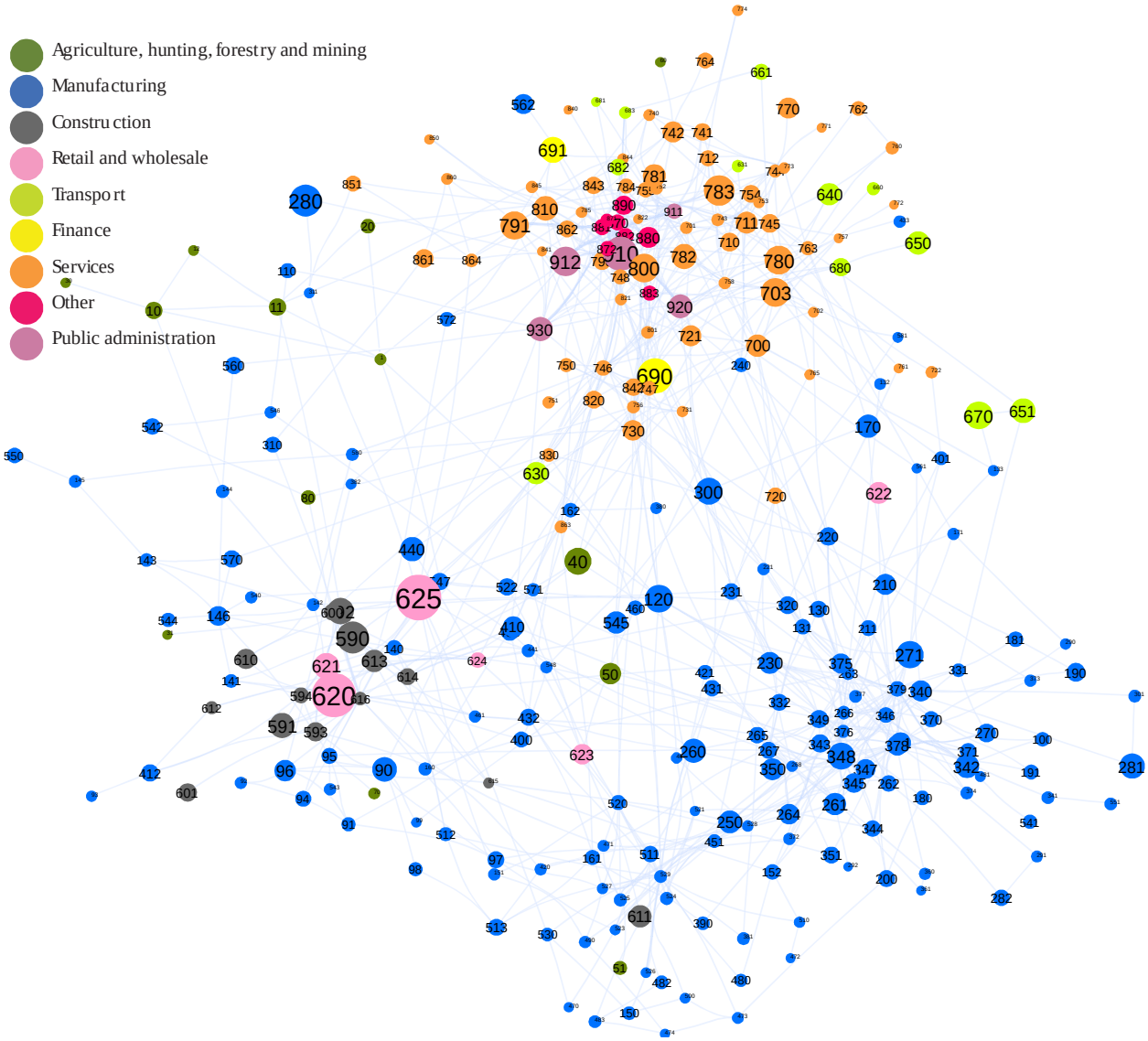
Classifications of Economic Activities				Occupations
WZ73	**WZ93 (NACE 1.0)**	**WZ03 (NACE 1.1)**	**WZ08 (NACE 2.0)**	**KldB88**
865: Labor recruitment and provision of personnel	7450: Labor recruitment and provision of personnel	7450: Labor recruitment and provision of personnel	7810: Employment placement agencies	43: Family workers (agricultural)
900: Private household production	9500: Households as employers	9500: Households as employers	7820: Temporary employment agencies and job pools	555: Disability
921: Foreign armed forces stationed in the FRG	9530: Re-employment agencies	9530: Re-employment agencies	7830: Payrolling	666: Re-employment trajectory
940: Representation of foreign countries	9540: Sheltered workshops	9540: Sheltered workshops	9530: Re-employment agencies	888: Caregivers
950: Employment abroad	9900: Activities of extraterritorial organisations	9900: Activities of extraterritorial organisations	9540: Sheltered workshops	924: Simplified registration In
951: Administrative backlog 1			9700: Households as employers	971: Family workers (non-agricultural)
952: Administrative backlog 2			9810: Goods production for private use	981: Trainees
953: Re-employment agencies			9820: Services production for private use	982: Interns
954: Sheltered workshops			9900: Activities of extraterritorial organisations	983: Job seekers
995: In early retirement				991: Unspecified occupation
996: Vocational training in schools				995: In early retirement
997: Other				996: In partial retirement
998: Disability payment recipients				997: Disability payment recipients
999: Unknown				999: Unknown

Industries and occupations that have been dropped from the skill relatedness matrices. The header row provides the German name of each classification system, with its European Union equivalent in parentheses where applicable.

Table 2: Descriptive statistics of labor flows in Germany

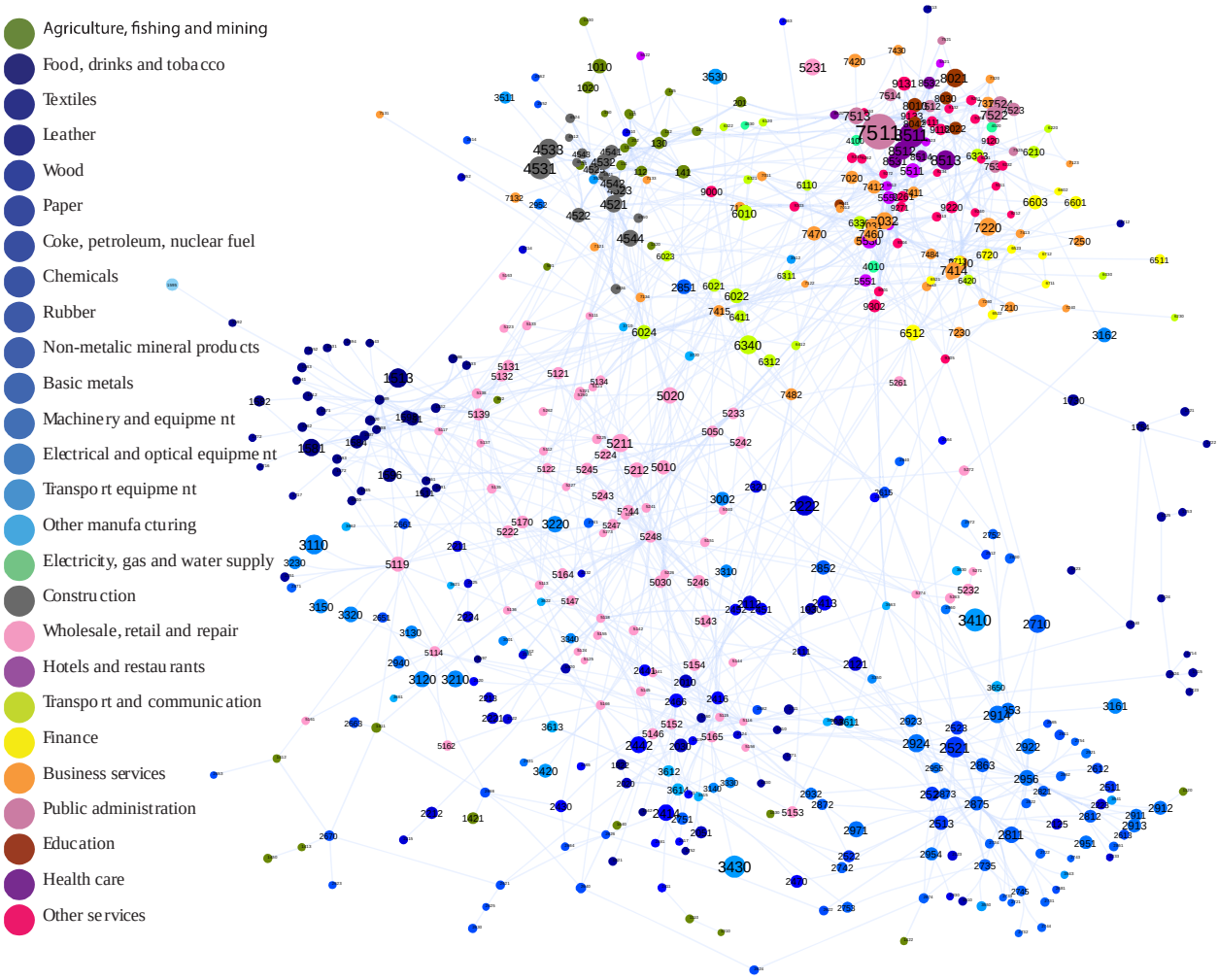
	Classifications of Economic Activities				Occupations	
	WZ73	**WZ93 (NACE 1.0)**	**WZ03 (NACE 1.1)**	**WZ08 (NACE 2.0)**	**KLDB88**	**KLDB88-SUF**
	3-digit-level	**4-digit-level**	**4-digit-level**	**3-digit-level**	**3-digit-level**	**3-digit-level**
Period	1975-1998	1998-2003	2003-2008	2007-2014	24,937,060	24,223,479
Total number of flows per period	20,397,609	4,982,365	3,418,760	5,529,890	712,487	692,099
Average number of flows per year	886,853	996,473	683,752	789,984	1975-2010	1975-2010
Total number of years	23	5	5	7	35	35
Total number of classes	288	499	506	597	329	120

Figure 1: Skill relatedness network – German Classification of Economic Activities WZ73, 1975-1998



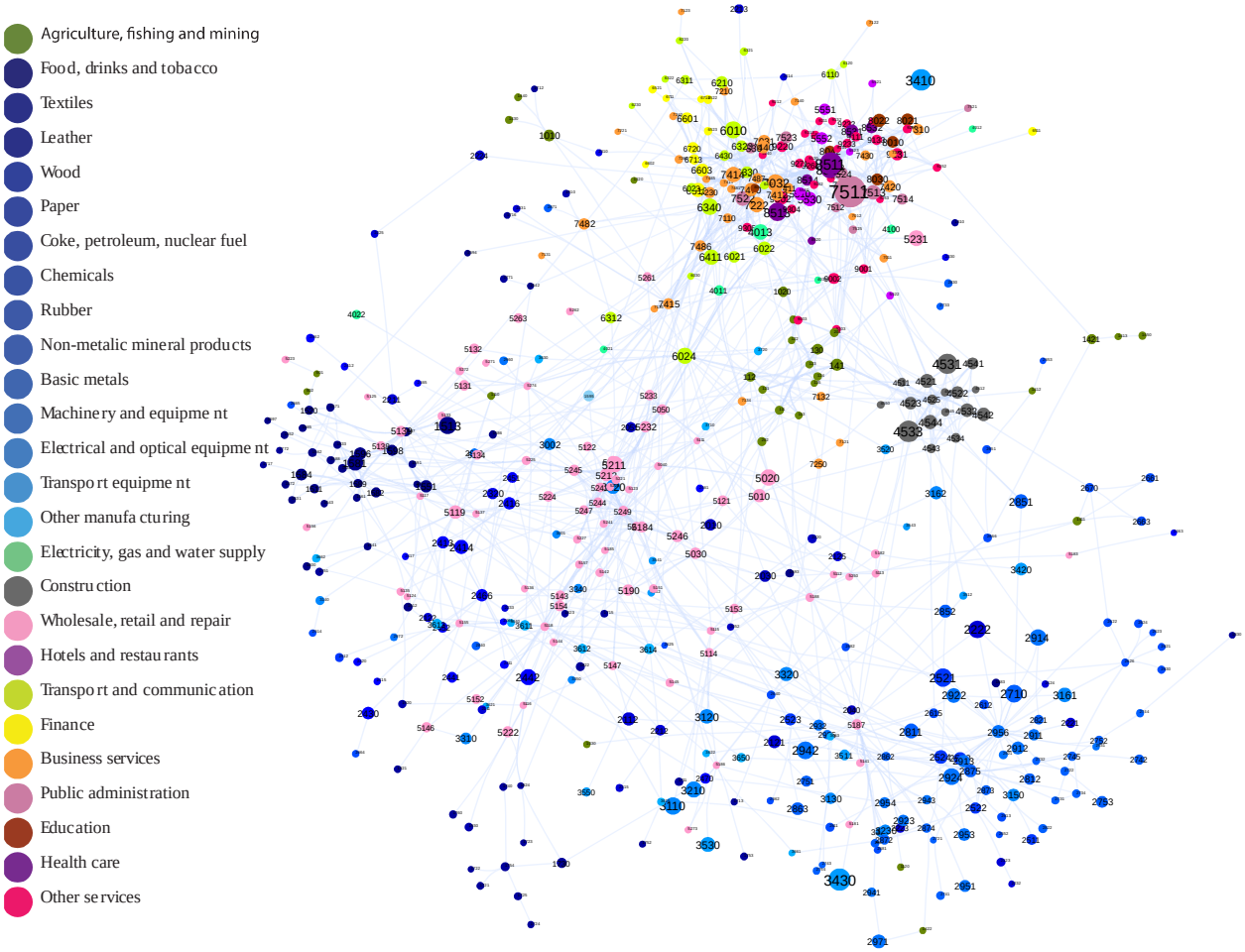
Skill-relatedness network using inter-industry flows. The network depicts the strongest $3N_i$ links, where N_i represents the total number of industries. Colors represent the aggregate sectors to which an industry belongs. Node sizes reflect the average yearly employment in the industry.

Figure 2: Skill relatedness network – German Classification of Economic Activities WZ93, 1998-2003



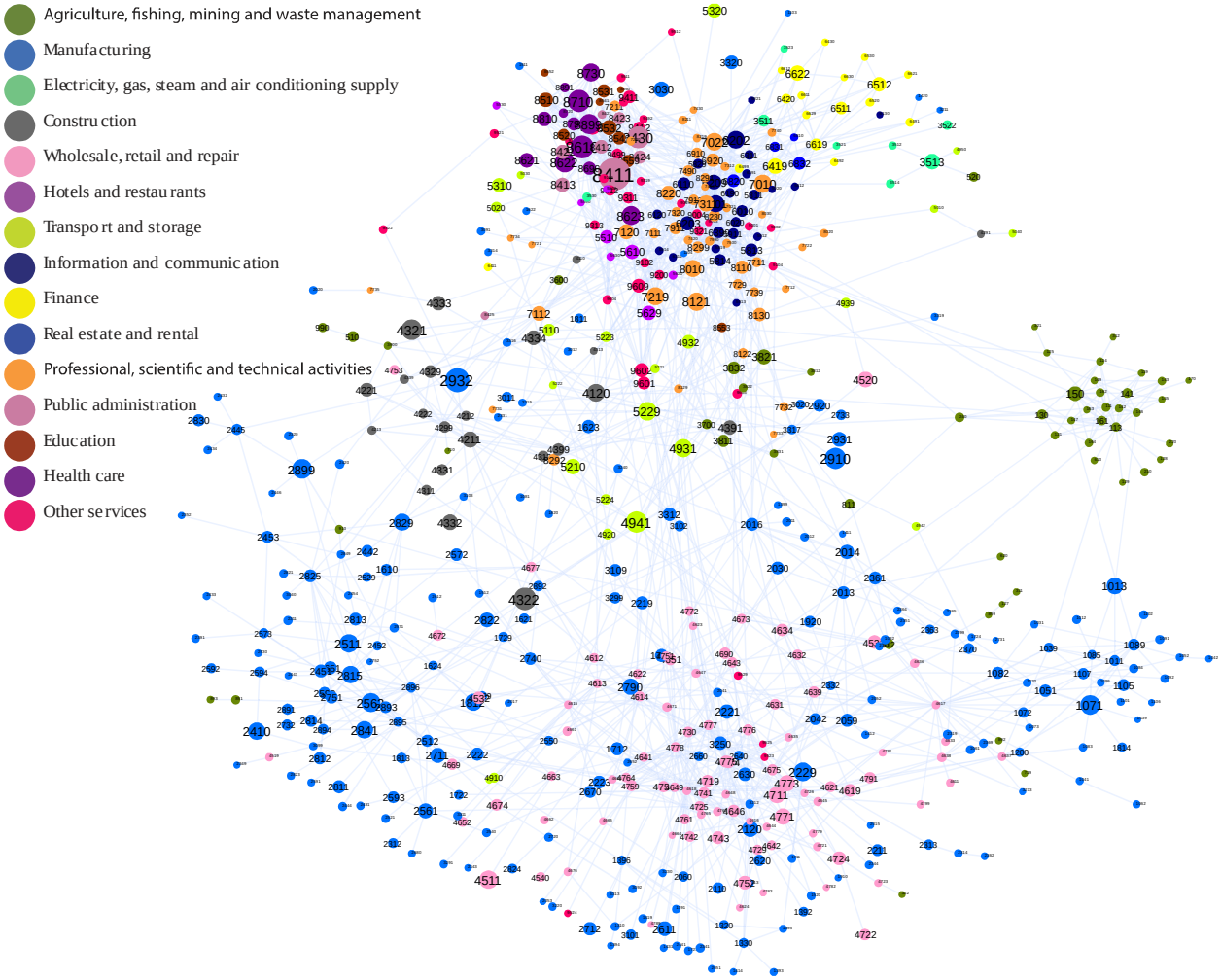
Skill-relatedness network using inter-industry flows. The network depicts the strongest $3N_i$ links, where N_i represents the total number of industries. Colors represent the aggregate sectors to which an industry belongs. Node sizes reflect the average yearly employment in the industry.

Figure 3: Skill relatedness network – German Classification of Economic Activities WZ03, 2003-2008



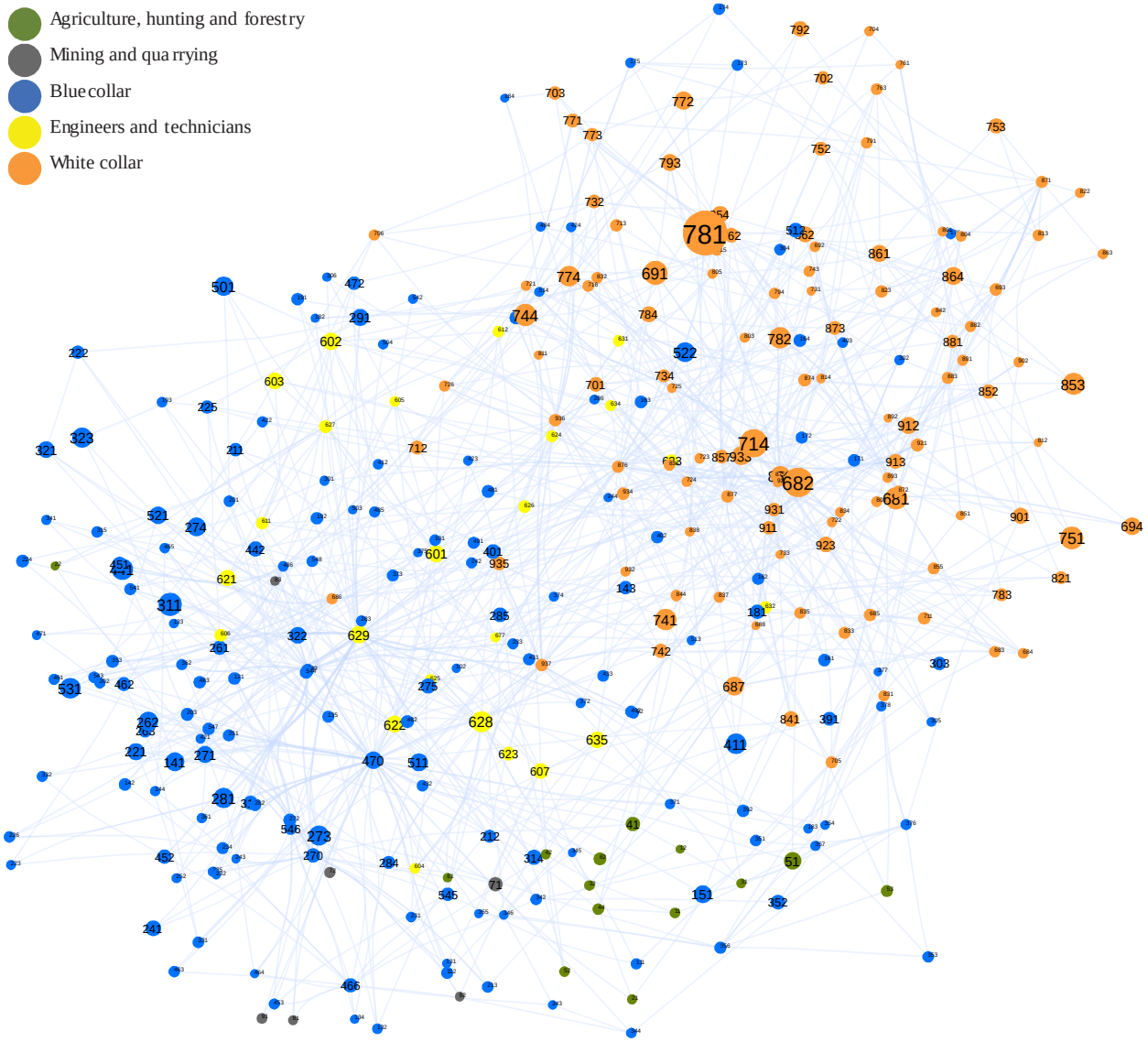
Skill-relatedness network using inter-industry flows. The network depicts the strongest $3N_i$ links, where N_i represents the total number of industries. Colors represent the aggregate sectors to which an industry belongs. Node sizes reflect the average yearly employment in the industry.

Figure 4: Skill relatedness network – German Classification of Economic Activities WZ73, 2007-2014



Skill-relatedness network using inter-industry flows. The network depicts the strongest $3N_i$ links, where N_i represents the total number of industries. Colors represent the aggregate sectors to which an industry belongs. Node sizes reflect the average yearly employment in the industry.

Figure 5: Skill relatedness network of the KldB88 classification, 3-digit-level (1975-2010)



Skill-relatedness network using inter-occupational flows. The top skill-relatedness values in the occupational flows are dominated by somewhat generic occupation, such as foremen, or general construction workers (“Bauhilfsarbeiter”). To prevent these nodes from dominating the network structure, the network depicts, unlike the inter-industry skill-relatedness networks, not the strongest links overall, but the strongest three links by node. Colors represent aggregate occupational segments. Node sizes reflect the average yearly employment in an occupation.

Imprint

FDZ-Methodenreport 4/2017

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Dana Müller, Dagmar Theune

Technical production
Dagmar Theune

All rights reserved
Reproduction and distribution in any form, also in parts,
requires the permission of FDZ

Download
http://doku.iab.de/fdz/reporte/2017/MR_04-17_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Anne Otto
Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nürnberg
Email: Anne.Otto@iab.de