exeley
Article | DOI: 10.21307/connections-2019.010

CONNECTIONS
Issue 1 | Vol. 40

Academic Collaboration via Resource Contributions: An Egocentric Dataset

Michał Bojanowski,[1,*]
Dominika Czerniawska[2]
and Wojciech Fenrich[3]

[1]Kozminski University, Warsaw, Poland.

[2]University of Manchester, Manchester, UK and University of Warsaw, Warsaw, Poland.

[3]University of Warsaw, Warsaw, Poland.

*E-mail: michal2992@gmail.com

Abstract

In order to understand scientists' incentives to form collaborative relations, we have conducted a study looking into academically relevant resources, which scientists contribute into collaborations with others. The data we describe in this paper are an egocentric dataset assembled by coding originally qualitative material. It is 40 multiplex ego networks containing data on individual attributes (such as gender, scientific degree), collaboration ties (including alter–alter ties), and resource flows. Resources are coded using a developed inventory of 25 types of academically relevant resources egos and alters contribute into their collaborations. We share the data with the research community with the hopes of enriching knowledge and tools for studying sociological and behavioral aspects of science as a social process.

Keywords
Collaboration networks, Resources, Sociology of science, Ego networks.

Scientometric studies report steadily increasing trend in multi-authored scientific publications. It is clearly an evidence that contemporary science requires cooperation and is not anymore a traditionally individualistic activity (Moody, 2004). The presented data set comes from a study in which our overarching research goal was to understand why some scientists collaborate, but some others do not. In particular, our approach was to think about incentives that might lead them to do so. Inspired by Coleman (1994) and, among others, Laudel (2001), Lewis et al. (2012) as well as our earlier results (Czerniawska et al., 2018), we assume that the incentives to collaborate come from academically relevant resources the scientists possess or control and the interests they might have in resources possessed or controlled by others. For example, a theorist and an experimentalist might be interested in each other's resources – ability to develop theoretical model of the studied problem and skills in conducting experiments, respectively. Unequal distribution of these resources across academic community and the necessity of pooling them to get ahead in contemporary science results in incentives to collaborate.

Current state of knowledge still lacks a universally accepted behavioral understanding of the scientific process, let alone standardized tools for measuring academically relevant resources. Hence, we conducted a qualitative study among Polish scientists with the goal to:

1. collect egocentric data on collaborative relations;
2. develop an inventory of academically relevant resources from respondents' reports; and
3. measure what resources (Item 2) collaborating parties (ego and alters) engage in their collaboration ties (Item 1).

The data we hereby share are based on transcriptions and coding of the originally qualitative material. The second section provides some brief background information on science in Poland and details our contribution. The presented study involved 40 interviews conducted on a sample of Polish scientists,

which we describe further in the third section. In the fourth section, we describe the way in which the inventory of resources was constructed. A complete list with example quotes is provided on the associated website.[1] The fifth section describes the structure of the data set. The sixth section provides illustrative examples. The seventh section provides the details where the data can be found and how it can be accessed. Finally, in the eight section, we discuss limitations and potential uses of the data.

Background and contribution

The presented data come from a study, which was conducted in Poland among Polish researchers. Polish scientific community is among the largest in Europe: according to OECD (2019) statistics, there were 132,000 researchers in Poland in 2016. At the same time, the funding and material resources are only average (cf. Czerniawska, 2018; Kwiek and Szadkowski, 2018). These conditions, next to some others, keep Polish science largely outside of the strict core of international scientific collaboration (Leydesdorff et al., 2013).

The organization and institutions of Polish scientific system resemble "Continental" systems (e.g. German scientific system). A typical scientific career requires a four year PhD program, a habilitation which is expected within eight years after a PhD. Obtaining a habilitation is perceived as the final step to becoming an independent scholar. Polish scientific community, similarly to many other scientific communities in Europe, is rather diverse. It is a mix of modern, competitive, internationalized disciplines and groups, and more conservative locally oriented areas (Kwiek, 2018).

Explaining the presence or absence of collaboration relations among scientists by referring to complementarities between them is not a new idea. For example, Qin et al. (1997) in their bibliometric analysis use institutional affiliation to capture different specialization of scientists. Moody (2004) approximates different types of contributions by analyzing subject codes put on articles indexed by Sociological Abstracts. Our goal was to collect a list of resources they believe are relevant when working as a scientist. We believe a genuine contribution of the presented data set lies in that detailed information on the flow of resources in scientific collaborations. The catalogue, which is a unique contribution in scientific collaboration studies, was constructed based on the extensive literature review and themes mentioned by our interviewees. The data have been used to study whether structurally non-redundant ties are more likely to be characterized with resource contributions not found in other ties (Bojanowski and Czerniawska, 2020).

Sample

Data come from 40 individual in-depth interviews (IDI) conducted between April and August 2016 by two interviewers. The quota sample consists of 20 female and 20 male scientists from six Polish cities. Respondents represented a broad range of disciplines: natural sciences, social sciences, life sciences, the humanities, engineering, and technology on different levels of career from PhD candidates to professors. The interviewees mentioned 334 collaborators in total. Interviews lasting between 24 and 90 min were recorded and later transcribed.

Measurement

Each interview consisted of several parts, three of which are of relevance here:

1. Respondents were asked to name up to 10 important researchers they have collaborated with during last five years. Each collaborator was discussed separately giving information about gender, scientific degree, nationality, and university department (if possible). See Section 5.1 below.
2. During the interview a network of collaboration among collaborators mentioned in item (1) was reconstructed using cork board, pins, and rubber bands. See the example in Figure 1. Cork boards were photographed and later digitized into edgelist data. See Section 5.2 below.
3. For each collaborator, the respondents were asked about academically relevant resources he/she contributed to the collaboration and what resources were contributed by the collaborator. Interviewees were provided with a broad framework, which would help them identify resources such as financial resources (e.g. funding), human resources (e.g. knowledge, skills), and social resources (e.g. collaborators).
4. Interviews were audio-recorded and later transcribed. The text of the transcripts was analyzed using QDA Miner Lite[2] in order to code

[1] At https://recon-icm.github.io/reconqdata/articles/resource_inventory.html.

[2] A product of Provalis Research, see https://provalisresearch.com/products/qualitative-data-analysis-software/.

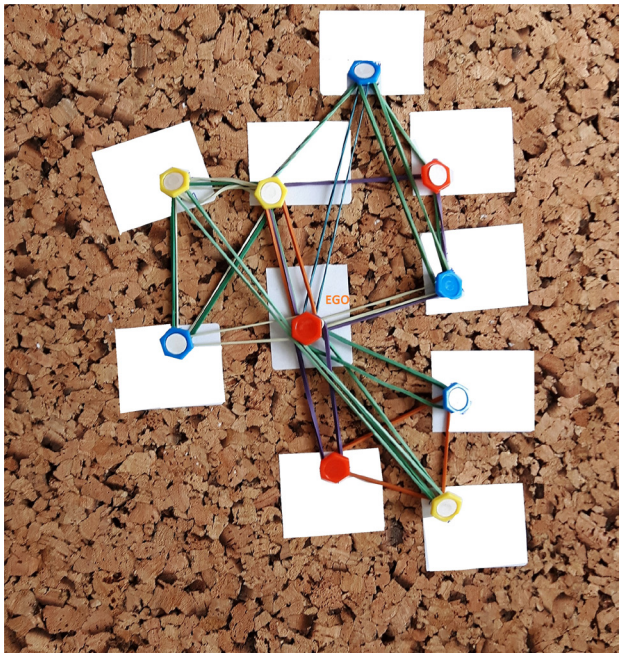

Figure 1: Using cork board, pins, and rubber bands to collect data on collaborations. Small cards contained names or nicknames which have been masked.

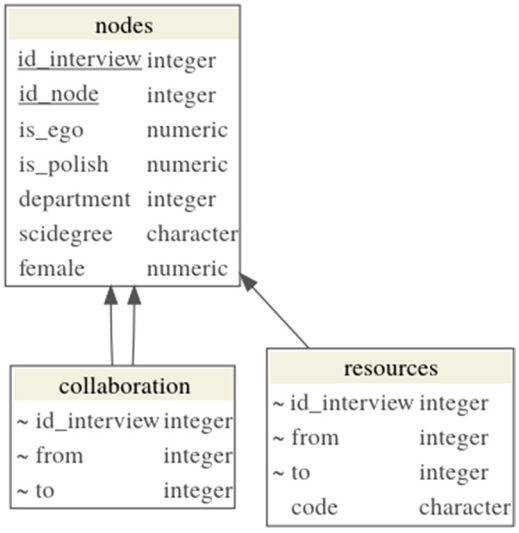

Figure 2: The data consist of three interrelated tables. Table 'nodes' contains information about all persons. Table 'collaboration' is an edgelist of collaboration ties. Table 'resources' is a multiplex edgelist of resource flows.

resources engaged by respondents (the egos) and their collaborators (the alters) to every collaboration. The coding was performed by two persons. Random sample of the interviews was double-checked by different researchers to ensure reliability. The data are available in table resources and described in detail below.

While collaboration networks assembled from part (2) include alter–alter ties, the data on resources from part (3) were acquired for ego-alter dyads only.

Structure of the data

The data are contained in three inter-related tables diagrammatically presented in Figure 2. Below we describe each table in detail.

Node attributes

The table nodes contain information about every person in the study – all egos and all alters. It has 374 rows and the following seven variables:

- `id_interview` – Interview identification number.
- `id_node` – Person identification number, unique within each interview.
- `is_ego` – Binary variable equal to 1 if person is the ego (respondent), 0 otherwise.
- `is_polish` – Binary variable equal to 1 if person is affiliated with a Polish academic institution, 0 otherwise.
- `department` – Marking scientists if they work at the same department. If `department` is not missing then all scientist within the same interview sharing the same value of department work at the same department at the same university.
- `scidegree` – Scientific degree of the scientist. One of "`mgr`"=MA, "`dr`"=PhD, "`drhab`"=habilitated doctor, or "`prof`"=full professor.
- `female` – Binary variable equal to 1 if person is female, 0 if male.

Pair of variables `id_interview` and `id_node` together constitutes a key uniquely identifying each row in the nodes table.

Collaboration networks

The table collaboration is essentially an edge list of collaboration ties. It has 1,732 rows and the following three variables:

- `id_interview` – Interview identification number.

- `from` and `two` – Person identification numbers referencing the `id _ node` variable from the nodes table.

In other words, a row consisting of values, say, `id _ interview = 1`, `from = 2`, `to = 3` indicates that researchers 2 and 3 where reported as collaborating in the interview 1.

Resource contributions

Data about resources engaged by respondents (egos) and their collaborators (alters) to every collaboration were coded based on transcripts. The data are provided in table `resources` having 1,761 rows and the following four columns:

- `id _ interview` – interview identification number.
- `from` and `two` – person identification numbers (within each interview) referencing the `id _ node` variable from the `nodes` table.
- `code` – a textual code identifying type of resource contributed by researcher `from` into the collaboration with researcher `to`.

Resources engaged in collaborations (variable `code`) were coded with a coding scheme covering different elements of a research process in different disciplines. The scheme consists of 25 codes such as:

- 'Conceptualisation' – coming up with an idea for a study, providing general theoretical framework; designing a general framework for a study.
- 'Methodology' – designing methodology for a study.
- 'Investigation' – conducting research, gathering data.
- 'Data analysis' – data analysis, quantitative as well as qualitative.
- 'Data curation' – managing and archiving data.
- 'Software creation' – writing software for research process.
- 'Prototype construction' – building a prototype that is used in research process.

Complete list of codes together with examples of coded interview fragments is available at the website.[3]

[3]https://recon-icm.github.io/reconqdata/articles/resource_inventory.html.

Table 1. Frequencies of gender and scientific degree for egos and alters. Symbol '–'(dash) corresponds to missing data.

Gender	Degree	Alter	Ego
Female	Full professor	20	3
Female	habilitated PhD	11	5
Female	MA	19	3
Female	PhD	57	7
Male	Full professor	87	8
Male	habilitated PhD	23	4
Male	MA	28	1
Male	PhD	63	9
–	MA	3	–
–	PhD	1	–
–	–	22	–

Selected descriptives

As a glimpse into the data, Table 1 shows frequency distribution of gender and scientific degree for egos and alters separately.

Figure 3 shows resource flow networks from one of the interviews:

Accessing the data

The data are available in a GitHub repository at https://recon-icm.github.io/reconqdata as an R package with accessible files in a CSV format. Users can use the data with R by installing the package or download the data files in CSV format using URLs provided in the README file.

Discussion

We close by discussing potential uses and limitations of the documented data set. We think that the data we share can be used in several contexts with substantive and methodological goals in mind. On the substantive side, the data can be used to address several research questions. For example to analyze co-appearance of different types of resources in collaboration ties – certain

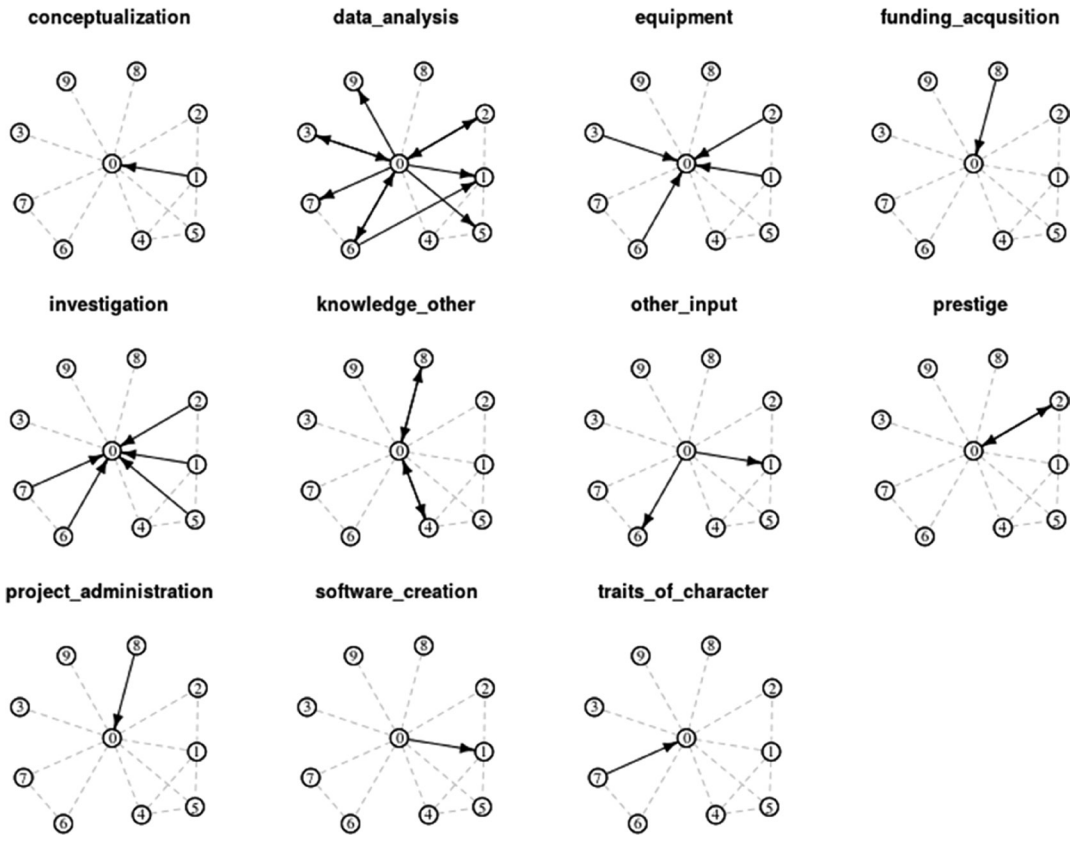

Figure 3: Collaboration (dashed, undirected) and resource flow (solid, directed) ties from one of the interviews.

types of resources tend to be contributed together. Further, the resource catalog could be improved and perhaps serve as a starting point for constructing a more standardized survey instrument.

On the methodological side, the value of the data set is that it is egocentric and multiplex at the same time. We see active development in statistical models for data collected through egocentric design (Krivitsky and Morris, 2017) as well as in modeling multilayer networks (Krivitsky et al., 2019). The data we share can be a useful test bed for such models.

The data have certain limitations. First, it comes from a qualitative study conducted on a quota sample. The obvious limitation is the lack of representativeness in the strict statistical sense. Nevertheless, it is representative in the loose sense – the respondents come from universities from different regions and of different size, from different disciplines and at different stages of scientific career. We believe it does cover the diversity of scientific positions pretty well.

Second, the data contain several instances of resource flows between the alters. However, the reliability of this data is rather low. Majority of respondents did not have enough information or were otherwise not confident enough in reporting the resource contributions. Consequently, these data were not collected systematically.

Acknowledgments

The authors thank (Polish) National Science Centre for support through SONATA grant 2012/07/D/HS6/01971 for the project Dynamics of Competition and Collaboration in Science: Individual Strategies, Collaboration Networks, and Organizational Hierarchies (http://recon.icm.edu.pl).

References

Bojanowski, M. and Czerniawska, D. 2020. Reaching for unique resources: Structural holes and specialization in scientific collaboration networks. *Journal of Social Structure*. Forthcoming. Preprint available on-line, available at: http://recon.icm.edu.pl/wp-content/uploads/2019/05/exchange_networks.pdf.

Coleman, J. S. 1994. *Foundations of Social Theory,* Harvard University Press, Cambridge, MA.

Czerniawska, D. 2018. Sieci współpracy i wymiany w centrach i na peryferiach. Przypadek polskiej nauki (PhD thesis). University of Warsaw, Warsaw, Poland.

Czerniawska, D., Fenrich, W. and Bojanowski, M. 2018. Actors, relations, and networks: Scholarly collaboration beyond bibliometric measures. *Polish Sociological Review*, 202: 167–185.

Krivitsky, P. N. and Morris, M. 2017. Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. *The Annals of Applied Statistics*, 11(1): 427–455.

Krivitsky, P. N., Koehly, L. M. and Marcum, C. S. 2019. Exponential-family random graph models for multi-layer networks. SocArXiv, available at: https://doi.org/10.31235/osf.io/dqe9b (accessed August 14, 2019).

Kwiek, M. 2018. *Changing European Academics: A Comparative Study of Social Stratification, Work Patterns and Research Productivity*. Routledge, London.

Kwiek, M. and Szadkowski, K. 2018. Higher education systems and institutions, Poland. In Teixeira, P., Shin, J. C., Amaral, A., Bernasconi, A., Magalhaes, A., Kehm, B. M. and Nokkala, T. (Eds), *Encyclopedia of International Higher Education Systems and Institutions,* Springer, pp. 1–10, available at: https://doi.org/10.1007/978-94-017-9553-1_375-1.

Laudel, G. 2001. Collaboration, creativity and rewards: why and how scientists collaborate. *International Journal of Technology Management*, 22(7–8): 762–781.

Lewis, J. M., Ross, S. and Holden, T. 2012. The how and why of academic collaboration: disciplinary differences and policy implications. *Higher Education*, 64(5): 693–708.

Leydesdorff, L., Wagner, C., Park, H. W. and Adams, J. 2013. International collaboration in science: the global map and the network, available at: http://arxiv.org/abs/1301.0801 (accessed August 10, 2019).

Moody, J. 2004. The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2): 213–238.

OECD. 2019. OECD science, technology and R&D statistics: main science and technology indicators, available at: https://data.oecd.org (accessed August 10, 2019).

Qin, J., Lancaster, F. W. and Allen, B. 1997. Types and levels of collaboration in interdisciplinary research in the sciences. *Journal of the American Society for Information Science*, 48(10): 893–916.