The Socio-Semantic Space of John Mohr

Camille Roth, Nikita Basov

▶ **To cite this version:**

Camille Roth, Nikita Basov. The Socio-Semantic Space of John Mohr. Poetics, 2020, pp.101437. 10.1016/j.poetic.2020.101437 . halshs-02558518

HAL Id: halshs-02558518
https://shs.hal.science/halshs-02558518

Submitted on 29 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This paper is intended

for the **Special Issue** of Poetics on "Discourse, Meaning, and Networks: **Advances in Socio-Semantic Analysis**"

edited by **Jennifer Lena**

The Socio-Semantic Space of John Mohr

Camille Roth

CNRS (French National Centre for Scientific Research)

*Nikita Basov**

Centre for German and European Studies
St. Petersburg State University

*corresponding author:
Universitetskaya 7/9 St Petersburg
Russia
007 812 324 08 85
n.basov@spbu.ru

Acknowledgements

We would like to thank all those who helped in spreading the word about our – rather short – call for self-nominations for the survey about relationships with John and/or for their advice on how to disseminate the call broader and the overall support throughout the process: Jan Fuhse, Jennifer Lena, Allison Pugh, Geoffrey Raymond, Oleg Komlik, Craig Rawlings, Clayton Childress, Juan Lejarraga, and Elena Tsumarova. We also thank Paul DiMaggio, Ron Breiger, and Robin Wagner-Pacifici for their advice on conducting the survey and the feedback on different versions of the visualization.

Funding

The related work of Camille Roth is supported by the "Socsemics" Consolidator grant funded by the European Research Council (ERC) under the European Union Horizon 2020 research and innovation program (grant agreement No. 772743).

The related work of Nikita Basov is supported by Russian Foundation for Basic Research (project No. 18-011-00796).

The Socio-Semantic Space of John Mohr

Abstract

This small essay is a part of a back-to-back pair of tributes to John W. Mohr. It is meant as a complement to the essay by Robin Wagner-Pacifici, Ronald Breiger and Paul DiMaggio focused on their personal experiences of work and friendship with John. Our contribution, in contrast, presents a computationally produced map of the sociocultural space of John, molded by his numerous projects and initiatives in his crucial role as an editor, thinker, and boundary spanner. Our map utilizes a combination of online publication data and a survey of John's social alters across the world. This map, we hope, illuminates the broad engagement of John into diverse research areas and scholarly communities, which definitely stimulated both his bright mind and the dwellers of these communities. We also believe this visualization highlights the meaningful social and cultural connections, inspired by John and molding the socio-semantic network analysis of today.

John W. Mohr, Professor of Sociology at the University of California—Santa Barbara, a valuable collaborator, an inspired researcher, a guest co-editor of this Special Issue, and a friend, has passed away on August 24, 2019. In his role as a thinker and communicator, John both played a major intellectual role in a number of ideational spaces and spanned a number of academic communities. Hence, he not only played a crucial role in molding the present landscape of socio-semantic analysis, but also makes an exciting case for a socio-semantic analysis himself. To both give due to John's legacy and to have an outlook on the academic space for emergence of which he did so much, we collected data on his social and semantic surroundings and mapped the socio-semantic space he was engaged in.

In this, we based on a combination of online publication data and a small survey of John's collaborators and colleagues, co-authors and co-editors, students and teachers, mentors and mentees,—from all over the world, which was conducted in September, 2019. We aimed to show the broad engagement of John into diverse research areas, which definitely stimulated both his exciting ideas and us, the individuals who were lucky to know him in person. This small study, which is by no means exhaustive or fully robust, is also illustrative in pulling together many of the connections inspired by John and focal for the becoming of socio-semantic network analysis (see Basov, Breiger and Hellsten, 2020): the mixed method approach blending qualitative data and interpretation with computational analytical techniques and automatic data retrieval, interdisciplinarity combining of social sciences and humanities with natural and formal sciences, the cross-scale junction bringing together 'Big' and '**thick**' Data—to name a few.

Below, we outline how we collected the data and mapped the network, provide a brief qualitative comment on the resulting map, and reflect on the methodological implications for socio-semantic network analysis suggested by our small project.

To gather the data on John's *social network*, we globally disseminated a request for self-nominations in order to collect information on John's collaborators, including (but not limited to) mail lists and webpages of American Sociological Association (including internal lists of the Sociology of Culture, History of Sociology, and Theory sections), European Sociological Association, International Sociological Association (including also the list of the Junior Sociologists Network of the ISA), Socnet listserv of the International network for social network analysis and corresponding groups there, University of California at Santa Barbara administration, and personal communications of John's recent students. We asked them, would they have any relationship with John, whether studentship, mentoring, research collaboration, editing, organizing, work on a board/committee, co-chairing, or any other type of professional academic relationship. And, irrespective of its perceived strength/frequency, we invited them to send us their name, affiliation, and a few words on the kind of relationship they had with John. The information on John's ties obtained from the responses was then combined with information about all types of John's collaborations retrieved from his most recent CV. Overall, this retrieved a list of 135 people.

Furthermore, we aimed to show which thematic areas associated with communities of researchers in his social network John engaged. We decided to do it by producing the *socio-semantic network* around John, connecting people to each other (in itself, a purely social network), key concepts[1] they used in titles of their works—to each other (a purely semantic network), and concepts—to people using them (a hybrid network). To this end, we began by attempting to extract exhaustive bibliographical data for the people in John's social network for their whole career until today—principally, publication titles and lists of co-authors from among these people—using their verified Google Scholar profiles[2], CVs, and other open-access online data. For six of the initial 135 people, no information about any publications could be found online, hence, only the remaining 129 people were included in the final socio-semantic network (together with John—130 social network nodes).

Defining connections between people is straightforward. We weighted links connecting people in John's social network based on numbers of their co-authored publications: *x* co-authored items means a weight of *x*. Plus a forced connection from John to all target people with a fixed weight of 0.1, to account for relations other than co-authorship. This resulted in 287 links with various weights.

[1] In semantic network analysis, the term 'concept' usually refers to word lemmas.

[2] As Google Scholar is largely automatically crawled, it often gets noisy to the end of the publications lists. To deal with this, we used essentially all entries associated to a profile up to the point where dates were missing, which we considered was obviously noisy. Hence, the limitations and the lists of publications may be incomplete.

To define concepts, we focused on publications with English titles (9,997 articles out of 11,216) and applied Spacy, a natural language processing library for Python, to extract n-grams called 'noun chunks'. These include up to 3 words considering sub-noun chunks included in larger chunks, focusing on right-hand decompositions (e.g., the 'social network model' noun chunk also yields 'network model' and 'model'; yet 'world food security' does not entail 'world food' or 'world', but rather 'food security' and 'security'). This provided an initial list of 20,518 distinct lemmas, far too many concepts to produce a tractable visualization. Hence, we applied several sequential filters to extract a much smaller list of concepts that are meaningful for John's epistemic network. We first filtered only lemmas used by at least one disconnected triad, i.e. 3 people who were not co-authors (two by two). The idea is that we focus on lemmas that were not just used in two different social contexts (bridging at least two people) but in three distinct contexts (thus, 'thridging' at least three people). This reduced the list to 1,548 lemmas. Further, we defined a notion of 'relevant lemmas for a person': we focused on the k most-used lemmas of each person (term frequency *tf*)[3], k being their social degree (i.e. number of co-authors) and $k \geq 3$ by force. This made it possible to associate more lemmas to people who were more present in John's landscape. If an actor has 15 co-authors, they will be connected to 15 lemmas. If it is 0, then 3 lemmas, still. This reduced the global list of lemmas to 303, from which we manually removed a remarkably low number of 28 stopwords[4], further justifying the fact that this selection is very relevant.

At this stage, we have a selection of 130 actor nodes and 265 lemma nodes, and a certain number of lemma neighbors for actors, among the most co-appearing, with a cut-off that aims at respecting each actor's or lemma's respective importance in this socio-semantic network. Links from people to lemmas were such as defined above, using the notion of 'relevant lemmas for a person', which are 559 links. Links from lemmas to people were defined similarly, but not identically. We introduced the notion of 'relevant people for a lemma'. We kept the k' most-using people for each lemma, where k' is a measure that aims at providing a sense of importance of a lemma similarly to the sense of importance of people conveyed by k above (it is the $\log(1+h(c))$ where $h(c)$ is the Hirschman-Herfindahl index of equivalent lemmas c is connected to in the original full n-gram co-occurrence network). This retrieved 693 links. Edges between lemmas were weighted based on 'relevant lemmas for a lemma', i.e. the k' most co-appearing lemmas where k' is $\log(1+h(c))$, which yielded 404 links. Hence, each step adds a roughly equivalent number of links, a total of 1943 links for 395 nodes. A post-hoc cut-off for socio-semantic links on one side, and semantic links on the other side, was applied, such that the number of times a given node is a target is similarly bounded. This principally improves readability and clarity but

[3] We ruled out using *tf.idf*, a classical weighting approach which consists in multiplying *tf* (term frequency for an actor) by *idf* (the inverse term frequency in the whole corpus) and typically aims at reducing the importance of frequent terms in the corpus. By contrast, it primarily emphasizes terms that are particular to specific actors. In our case, however, we rather wanted to focus on terms important to an actor while being not too particular to them, in order to emphasize commonality across the socio-semantic network. In this respect, using *tf* rather than *tf.idf* led us to keep fewer peculiar concepts.

[4] Including terms such as 'analysis', 'book review', 'introduction', 'issue', 'oxford handbook', 'understanding', 'university', 'whom', or 'workshop'.

does not change network topology much. This further removed 581 links, yielding the final network of 1362 links for 395 nodes, which was visualized below.

The visualization of the network was achieved by applying a force-directed layout to the graph, i.e., automatically positioned nodes in a way that links tend to attract nodes while nodes repel one another, in such a way that cohesive groups of nodes are going to be located in spatially cohesive clusters of nodes. The specific procedure was 'Force Atlas 2', with parameters resulting from trials-and-errors based on aesthetics and clarity (we ended up with scaling 4, gravity 0.1, link weight influence 0.1, in order to expand distances; lin-log mode to make clusters tighter; prevent node and label overlap, to improve readability). Link weights were not displayed as link thickness to assist readability but were accounted for in the computation of the layout. Reckon that we took the whole of the publications associated with each of the people through time, not only works they co-authored. So, not only co-authorships with John were taken into account, but all co-authorships of his social network alters.

The outcome of the described procedures is presented in Figure 1.

What do all these elaborate procedures show us? On the social side, the visualization highlights the diversity of communities comprising John's space, both within and across the socio-semantic clusters. In the socio-semantic network we find sociologists, philosophers, historians, political scientists, management and organization scholars, computer and communication scientists, health scientists and ecologists, and many others. We also find both prominent academics and junior researchers, both American and Eurasian scholars. Crucially, these are not scattered into separate social subgroups or thematic semantic clusters along disciplinary, geographical, political or status boundaries, but mix together along the major themes of the semantic side. These themes include networks / structures / data, power / politics / capital, sociology / theory / culture and science / evolution / health, management / organization / institution, communication / system / technology—the thematic areas John has been contributing to throughout his career and the researchers of which he was pulling together to collaborate (see also Wagner-Pacifici, Breiger and DiMaggio, 2020). These are—hardly a coincidence—also the main thematic areas where socio-semantic network analysis prospers—as the papers of this Special Issue show (see Basov et al., 2020).

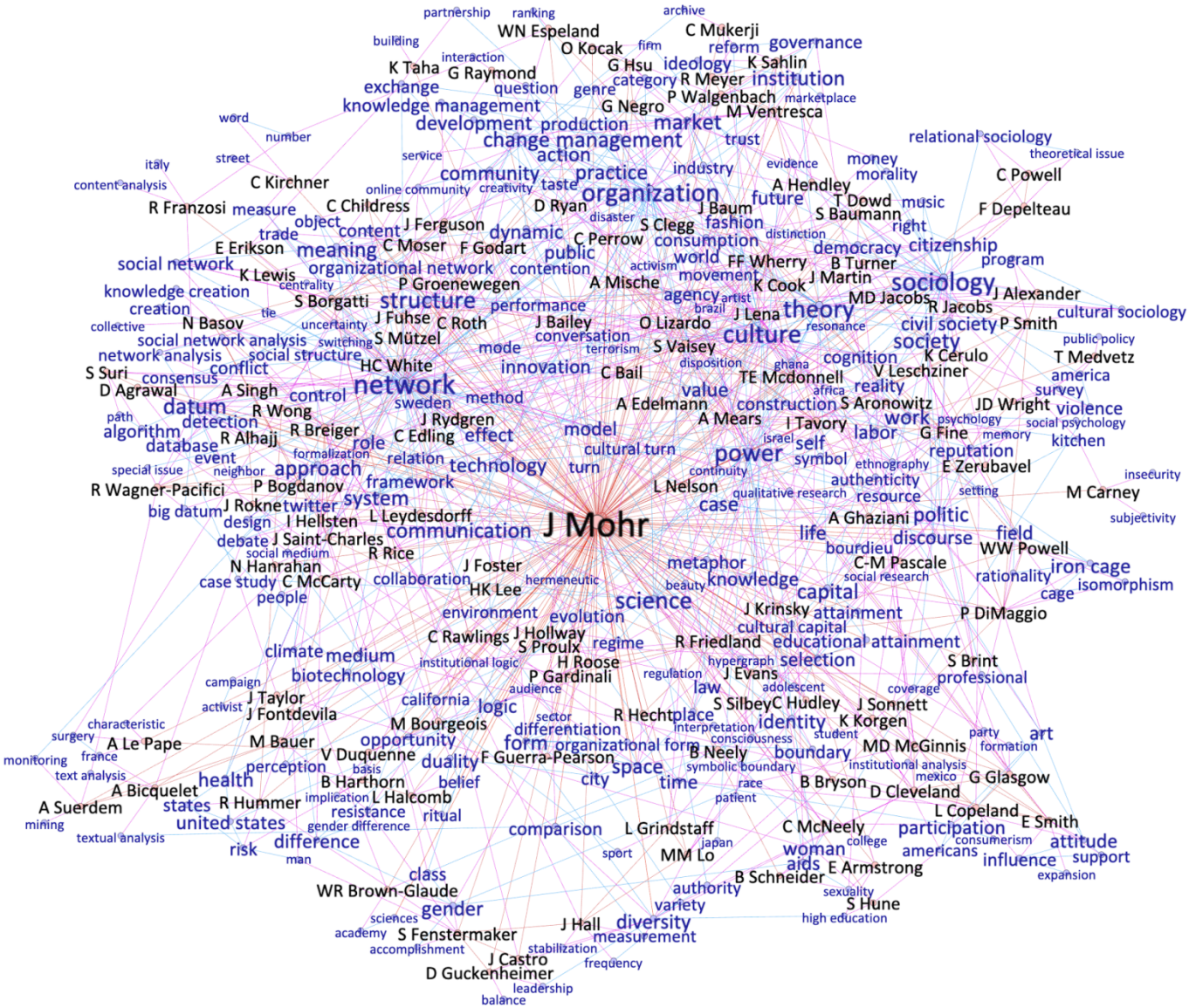

Fig. 1. Socio-semantic space of John Mohr

People are represented in black and concepts—in blue. Label sizes are proportional to total node degree (taking into account all types of links)

Visualizing John's socio-semantic space has not been merely about delivering an illustration. Rather, it also has been a learning process. In particular, this rather quantitative endeavor forefronted two key issues that arise when dealing with large socio-semantic networks, such as John's, which feature more than a hundred academics and their work.

The first issue is related to the set of concepts. As opposed to social actors whose boundaries are pretty well defined (they are, indeed, individuals), defining semantic nodes is a priori a much more open question. Deciding what counts as a concept and what describes it appropriately offers many alternatives: Should it be a word, a term, an n-gram, a category, a portion of utterance, a proposition, and how to delineate and select them? In this sense, the two fundamental ontologies of socio-semantic networks, actors and concepts, are not symmetric. This question assuredly arises when we deal with bibliographic data sets which typically consist of article titles and, sometimes, abstracts and keywords, as is also commonplace in scientometric endeavors. Traditionally, natural language processing tools are applied on this type of textual information to extract small linguistic units: Generally, words or noun chunks of a certain length. It however usually yields a potentially large set of concepts, of which many may be irrelevant—the so-called 'stop words', whose selection is partly automatic (based on general or contextual stop word lists) and partly manual (according to a more or less subjective appreciation of relevance and meaningfulness). Focusing on the most frequent terms generally does not reduce the proportion of stop words—noise happens to be everywhere. In other words, filtering focused on the most frequent words, even in a weighted fashion, often proves to be weakly efficient and to preserve a significant portion of noise which, in turn, has to be dealt with by hand. By contrast, we could observe that a filtering based on systemic criteria was much more successful. In particular, what we called ‘thridging concepts’ essentially aimed at conserving terms that were most useful to cover authors who potentially belonged to distinct social clusters i.e., focusing on terms that were likely to be common without necessarily being the most frequent. By doing so, we applied a filtering that was much more distributed across the term frequency spectrum—in effect, trying to remove noise a little bit everywhere, while keeping a remarkably small number of terms that we eventually deemed to be stop words (28 out of 303, less than 10%). This approach is not dissimilar to the notion of formal concept 'stability' (Roth, Obiedkov, and Kourie, 2008) developed in the context of Galois lattices, where 'unstable' formal concepts are filtered everywhere in the lattice, rather than in the bottom part, as used to be the case in the field.

The second issue relates to the definition of relationships. Again, co-authorship is rather straightforward and of relatively limited size—in John’s network, no one (except John) has more than a dozen of coauthors and fewer than 20 people have more than five. The situation is totally different in terms of actor-concept links (i.e., term usage) or concept-concept links (i.e., cooccurrence) where many authors published more than a hundred of items and where concepts have a lot of incoming links. A high network density could lead to a visualization where most of the nodes and links in the central area of the map are heavily overlapping and indiscernible. By construction, the most frequent concepts induce more links. The typical homogeneity in such networks means a few concepts are connected to many other concepts, contributing to the formation of a giant wool ball. Again, thresholding may be applied, but it only postpones the problem: It removes most of the less significant links on the periphery and still preserves the hairball in the center of the network, generally yielding a quite obfuscated bag of links and nodes in the middle. By applying a systematic rather than threshold-based data reduction approach, we could largely attenuate this effect; link clusters are distinguishable and spread out over the whole map. More broadly, this essay reinforced our belief that network visualizations must rely

on upstream data reduction approaches that deal with noise and redundancy in a systematic manner, rather than hope that downstream representation techniques, such as force-directed layouts, would palliate these issues if only they could be appropriately parameterized. This is especially true when the duality and, thus, asymmetry of social structure is taken into account as is the case with socio-semantic networks.

The experience of visualization John Mohr's socio-semantic space, thus, not only highlights the social and cultural connections, inspired by John and molding the landscape of socio-semantic network analysis today, but also allows for illuminating the challenges and even proposing solutions for visualizing socio-semantic networks. Apparently, John keeps contributing to the field despite the physical absence. Still here, giving us insights to discuss. Spanning the socio-semantic space.

References

Basov, N., Breiger, R., and Hellsten, I. 2020. 'Socio-semantic and other dualities'. Poetics. Special Issue on' Discourse, Meaning, and Networks: Advances in Socio-Semantic Analysis'.

Roth, C., Obiedkov, S., and Kourie, D. 2008. 'Towards concise representation for taxonomies of epistemic communities', Proc. of the 4th Intl. Conf. on Concept Lattices and their Applications (CLA 2006), Springer LNCS book series, LNAI 4923, 240–255.

Wagner-Pacifici, R., Breiger, R., and DiMaggio, P. 2020. 'John Mohr Appreciation'. Poetics. Special Issue on' Discourse, Meaning, and Networks: Advances in Socio-Semantic Analysis'.

Camille Roth has been holding a research professorship at CNRS since 2008 in computer science ("chercheur CNRS"). He also had a couple of tenured university positions in sociology, at Sciences Po as Associate Professor (“professeur”, 2016-18) and in Toulouse as Assistant Professor (“maître de conférences”, 2007-08). His research thus lies at the interface between social and computational sciences, featuring keywords such as socio-semantic systems, social cognition, algorithms and mathematical sociology. He founded in 2012 and currently leads the computational social science team at Centre Marc Bloch in Berlin (CNRS/Humboldt). He is currently the recipient of an ERC Consolidator grant on socio-semantic networks and over the past decade has been global or local PI for several multi-institution research projects, both at the French and European level, on blog networks, scientific communities, and peer-to-peer platforms, including Webfluence, Algopol and Algodiv (on informational dynamics of the digital public space) and QLectives (EU IP on quality collectives in socio-technical communities).

Nikita Basov is Senior researcher at the Faculty of Sociology, St. Petersburg State University and Scientific Manager of the Centre for German and European Studies (St. Petersburg State University—Bielefeld University). He investigates the fundamental principles of socio-cultural microdynamics. The main method is multivariate (socio-semantic and socio-material) network analysis, with a particular focus on mixing ethnographic methods with statistical modelling. His papers appeared in Social Networks, Poetics, and American Journal of Cultural Sociology. He is also the organizer of the conference series ‘Networks in the Global World’ in St. Petersburg and of ‘St Petersburg Summer School on Network Analysis’.