Using heterogenous information networks for integrative discourse mapping

The Covid19 example

Alexander Brand
Institute of Social Sciences
University of Hildesheim
Hildesheim, Germany
alexander.brand@uni-hildesheim.de

Tim König
Institute of Social Sciences
University of Hildesheim
Hildesheim, Germany
tim.koenig@uni-hildesheim.de

Wolf J. Schünemann
Institute of Social Sciences
University of Hildesheim
Hildesheim, Germany
wolf.schuenemann@uni-hildesheim.de

ABSTRACT

This short paper presents a novel way of mapping knowledge communities in discourse by utilizing heterogenous information networks (HINs) and a two-stage grouping procedure. After laying out the theoretical foundations of a discourse analytical framework grounded in the sociology of knowledge, it will demonstrate the applicability of the framework on the platform Twitter. In exploratively analysing a sample of 6.317.324 tweets on the Covid19 pandemic, we will show how clustered HINs can make visible the social embeddedness of knowledge production in digital environments.

CCS CONCEPTS

•Applied computing~Law, social and behavioral sciences~Sociology•Information systems~Information systems applications~Data mining~Clustering

KEYWORDS

Network Analysis, Covid19, Heterogenous Information Networks, Discourse Analysis, Sociology of Knowledge

1 Introduction: Scalable and adaptive mapping tools for discourse analysis

In this short paper, we present ongoing work that seeks to combine the theoretical foundations of knowledge-oriented discourse analysis with the application of heterogenous information network (HIN) analysis. Given its long tradition of investigating the politics of knowledge and meaning-making, social science discourse research can help to avoid the common pitfalls of studying insulated information elements without reflecting the relevant "social relations of knowledge and knowing" [1, p. 18]. With its toolboxes for the inquiry of the social construction of reality, the discourse analytical approach can help to go beyond facts, especially when studying online communication. These toolboxes, however, need renewal and extension in order to live up to the newly accessible data worlds of digital societies. One of the generic operations of discourse analysis, and quite often the best way to reduce the complexity of results, is discourse mapping. Discourse mapping is to be understood as an umbrella term for visualisation techniques that allow for synoptical integration of relational or knowledge structures dissected through discourse analysis. The short paper puts emphasis on this particular task.

Qualitative discourse research and other approaches within the interpretive paradigm of social sciences have developed a great multitude and rich variety of mapping strategies for illustrative and instructive syntheses of empirical findings [2, 3, 4, 5]. Such maps can be regarded as a type of small-sized, non-standardised, interpretation-loaded knowledge graphs, making visible the heterogenous knowledge communities that shape discourses. Established discourse mapping strategies, though highly flexible, cannot easily cope with the large-scale data of online communication and their need for standardisation and automation. Due to their requirements in interpretive work, they are not scalable, thus not extendable to wider contexts of meaning-making or transferable to other subjects. In this paper, we argue that the flaws in established mapping strategies can be - at least partly - overcome by using HINs for the integrative and adaptive mapping of discourses. HINs are defined as a directed graph consisting of multiple types of objects or multiple types of relations between objects [6]. This mirrors the assumptions of discourse maps, which relate different actors to different kinds of information in order to make visible their specific knowledge communities.

In the next section, we present the theoretical foundations of our work, informed by a sociology of knowledge perspective on discourse. This is followed, in section 3, by a description of our methodology. In section 4, we present an exemplary online discourse analysis of a publicly available, large-scale dataset of Covid19 Tweets. We chose the pandemic as context, as we expect the respective dataset to indeed represent a complex network of discursive formations and structures of knowledge production.

KnOD'21 Workshop - April 14, 2021

2 Theoretical Foundations

Knowledge is an ambiguous concept. Even in empirical social sciences, it is frequently understood as a measurable resource of individual people. This conception is especially prevalent in the fields of political studies, including political psychology and political sociology, wherein - somewhat surprisingly - knowledge is mostly conceived as an individual instead of a collective resource [7, 8]. Theoretically and methodologically, this goes along with a widely shared pre-occupation with the micro-level foundations of social action in political studies at the expense of the relational dimension [9, 10] essential for knowledge production. In contrast to such prevalent conceptions, we root ourselves in a sociology of knowledge tradition [11, 12, 13, 14]. Moreover, the accompanying methodological re-orientation towards a social science research tradition of discourse analysis helps to avoid individualist misconceptions of knowledge and provides research methods that allow to go beyond facts in the empirical inquiry of the social construction of knowledge. This seems particularly helpful these days, as so-called disinformation is increasingly gathering academic interest and the scholars involved are running the risk of neglecting the social dimension in the production of knowledge [15].

Delving into processes of collective meaning-making by applying discourse analytical methods is essential, as "information by itself usually has no value: it is a raw material that gains value if further processed in specific ways and if meaning and a certain quality are attached to it"[16, p. 15]. Thus, knowledge cannot belong to the features of an individual (a user of digital media in our case) but is produced, processed in and obtained from discourses. Information or facts are 'consumed' by users only through these collectively built filters of perception. While emanating from a Foucauldian, post-structuralist tradition, discourse analysis does not necessarily mean to neglect the crucial relevance of agency. It is our social-constructivist conception that makes us attribute a hub-like role to actors (here: users of online media) in the basic design of our complex networks instead (see analysis section below). As discourses "are performed through social actors' (often competing or conflictual) discursive practices" [17, p. 3] it is actors that performatively produce the linkages that we can map in a network, be it links to other entities that are identifiable in online discourses such as URLs, hashtags, named entities, or other users. Taking such entities not only as linkages in communicative networks but as constitutive elements of issue publics or even communities of discourse and knowledge, we can rely on theoretical and analytical assumptions developed in the field of digital communication studies [18, 19, 20].

These relational patterns can be made visible by studying digital trace data at a large scale. The social media platform Twitter provides a particularly well-suited test case for our methodology. Twitter, with its characteristics of both a social network and an information network, makes visible the formation of knowledge communities through the curation of information flows by its users. While user influence on the platform follows a power-law distribution, all users are free to distribute, share and comment on information with their own followers, effectively providing the tools to collectively shape information environments in a network-based manner [21]. By looking at Twitter, we can make visible the processes which filter and curate the information environments of its users without neglecting the role of these very users and their networks.

3 Methodology

Methodologically, we use a two-stage grouping procedure. First, we obtain a mesoscopic representation of the network. Following Bar-Hen et al [22], such a representation of the network is obtained by grouping together nodes of the same entity and the same cluster and displaying them as blocks. This representation is very similar to a general block model with the notable difference of additional separation by node type. The choice to use a block model method in the first step was made with regard to the good performance of such models for large networks. Additionally, it allowed us to draw on applied research on combined clustering of multiple types of entities, such as documents and text in the case of Gerlach et al [23]. In the second step, a simpler clustering procedure can be carried out, which takes the edge weights into account. In the following chapters we refer to the clustered mesoscopic view of the network as the macroscopic view and to its clusters as macroclusters.

4 Exemplary Analysis

For our analysis we used the TweetsCOV19 dataset [24]. We chose the pandemic as context not just due to its current relevance, but because we expect the respective dataset to indeed represent a complex network of discursive formations and knowledge communities. This would include, among others, various special discourses of scientific experts, governmental communicative discourses, as well as general public discourse co-constituted by mainstream and social media. TweetsCOV19 is an annotated publicly available Twitter corpus of more than 8 million tweets on Covid19, including data from October 2019 - April 2020. For our analysis, we used a restricted version starting with the first public appearance of a Covid19 case in the general media on 12/31/2019, preventing false positive matches. After this we build our sample consisting of 6.317.324 tweets. A timeline of the number of tweets can be found in Appendix 1.

We proceeded as follows: In the first step, we constructed a poly-partite network from the tweets with the username, mentions, hashtags, URLs and named entities as node types and edges of one type which symbolize references (e.g., User X uses Hashtag Y in a tweet). Named entities were extracted using scores from the Fast Entity Linker Core library and URLs were expanded when necessary [24]. Furthermore, we removed stop words from user mentions, hashtags and named entities. This led to a poly-partite network with the following properties:

Table 1: Basic metrics of the constructed poly-partite network

Metric	Value
Total sum of nodes	708.352
Sum of unique usernames	130.997
Sum of unique user mentions	176.991
Sum of unique hashtags	158.438
Sum of unique URLs	145.041
Sum of unique named entities	96.885
Total sum of edges	111.399.912

In the next step, an agglomerative collapsing algorithm [25, 26] was used to block the nodes in the network. Following our agent-centric theoretical assumptions that knowledge is produced by a community of users (see above), interblock connections consist of user-user, user-hashtag, user-URL, and user-named entity relations. Due to the large amount of edges an agglomerative heuristic was applied, which iteratively tries to find a better configuration of blocks by progressively merging blocks together [25]. The final model selected via the lowest entropy criteria consists of 16 named entity blocks, 28 hashtag blocks, 19 user mention blocks, 23 URL blocks and 14 username blocks. An overview of the number of nodes in each block can be found in Appendix 2. In the next step, the macroclusters were computed via a simple greedy clustering algorithm, clustering the blocks in the mesoscopic network. In the last step a qualitative coding of the blocks and clusters was performed. To ensure the interpretability of the results, we calculated the PageRank of each node in the original poly-partite network and considered the top 10 nodes per block for the coding, similar to the evaluation of structural topic models and in line with Twitter's power-law distribution. For the coding of the usernames, we further included the account description into the coding step. URLs were coded via clues in the URL title. Generally, we used simple heuristics for content coding. For example, clusters containing actors from the fields of music, art and film were coded as "Cultural", while URLs coded with "Protection from Covid19" contain reports on different levels of protection in relation to aspects like ethnicity.

5 Results

Our results indicate a heterogenous discursive space. For the evaluation of the results, we present two novel visualizations: A full macroscopic view of the poly-partite network and a qualitatively annotated visualization of each macrocluster.

The macroscopic representation allows to visualize the general structure of the network in a representation similar to discourse maps commonly used in social science discourse research. As can be seen in the Figure 1, four of the five groups are rather strongly separated from each other, while a fifth (grey) is more torn apart. However, we also observe some outliers. For example, three blocks of the third macrocluster are located relatively apart from the rest of their cluster.

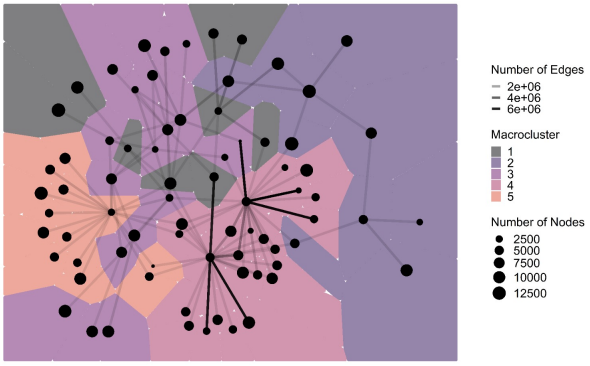


Figure 1: Full macroscopic view of the network with a stress-based layout. The background colour symbolizes the assignment of the nodes to the respective macrocluster (Dirichlet tessellation). The node size represents the number of nodes in the blocks of the poly-partite network. The visibility of the edges symbolizes the number of connections between blocks. Blocks with less than 0.1% of all out-going edges were cut from the representation to support the visualization.

The annotated version of the network (Figure 2) enables a more qualitative look at the different areas of discourse. Consistent with Figure 1, it is observable that macrocluster one is a mixed compilation with no clear identity. The second cluster, "Organizational Aspect and Early Response", contains blocks associated with aspects like the role of organizations in the pandemic and early response actions like the proposed usage of Hydroxychloroquine for the treatment of Covid19 patients. The third cluster, "Technology and Daily Life", takes a deeper dive into media, culture, and technological aspects. There is a certain proximity to macrocluster two, which is also notable in Figure 1. The aforementioned three outliers are more related to topics like weapons and US politics. The fourth cluster, "Culture and Safety", deals more with aspects like mental health and media, while the fifth cluster, "Uncertainty", focusses on the uncertainties of living through the pandemic. Following this differentiation, we can see that the chosen representation suggests a description of the Covid19 corpus along the lines of organizational aspects, reaction compulsion, cultural and technical adaptations, media use and general uncertainty. These aspects do not appear in isolation, but within the framework of a complex web of different emphases and affiliations.

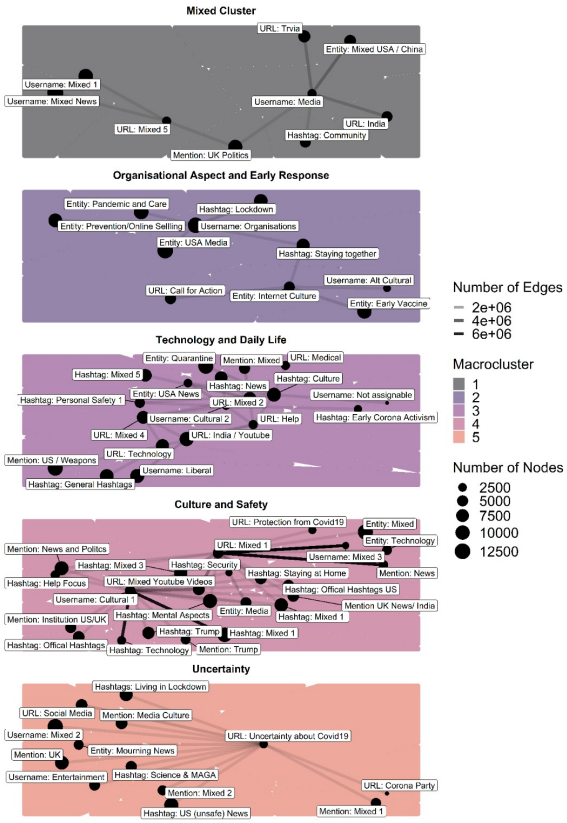


Figure 2: Subdivision of the macroscopic perspective of the network by macrocluster. Colours are consistent with the background colours in Figure 1. The node size represents the number of nodes in the blocks of the poly-partite network. The visibility of the edges symbolizes the number of connections between the different blocks. Blocks with less than 0.1% of all out-going edges were cut from the representation to support the visualization. Node types are username, mentions, hashtags, URLs and named entities (here shortened to "entities").

6 Conclusion

This paper aimed to showcase a methodology building on established discourse analytical assumptions about the social embeddedness of knowledge production with a scalable framework for heterogenous information network analysis. We showed that on Twitter, a platform affording user-based knowledge production and sharing, HINs can make these processes visible. Therefore, we demonstrated that HINs provide a powerful tool for social science discourse research to map large-scale online discourses. As such, they can help unearth the complex discourse formations in which knowledge is produced, especially in digital contexts where the amount of data often makes a qualitative approach unfeasible. Possible applications range from the mapping of issue-centred discourses to the identification of (mis-)information hubs on social media and the large-scale analysis of policy networks. In the explorative analysis of the Covid19 pandemic on Twitter, five macroclusters with differing users, URLs, hashtags, named entities and mentions became visible. As such, we can identify these clusters as knowledge communities, collectively shaping heterogenous information environments through their intra- and intercluster relations. In order to exhaust the possibilities of this approach, future analyses should consider utilizing even more diverse types of data to compute as clusters. The framework is highly flexible and able to incorporate multiple data sources and types of nodes. This flexibility can stretch to different types of data, such as textual or visual analyses, and even heterogenous data sources, such as different platforms. Furthermore, HINs allow for the specification of different edge types for an even more sophisticated model. This allows researchers to tailor their analysis around specific subjects without compromising neither theoretical foundations nor scalability. However, the selection of nodes should be theory-driven in order to avoid arbitrariness and remain economical with regard to computational resources. As such, our next steps to improve the information richness of the macroclusters would be the implementation of quantitative text analysis into the model, giving a more in-depth look into the knowledge communities surrounding the Covid19 pandemic on Twitter beyond facts.

REFERENCES

[1] R. Keller, 2018. The sociology of knowledge approach to discourse. An introduction. In *The sociology of knowledge approach to discourse*, R. Keller, A.-K. Hornidge, and W. Schünemann, Eds. Abingdon, Oxon and New York, NY: Routledge, 16–47.

[2] A.E. Clarke, C. Friese, and R. Washburn. 2018. Situational analysis: Grounded theory after the interpretive turn (Second edition). Los Angeles, London, New Delhi, Singapore: Sage.

[3] A. Luther. 2017. The Entity Mapper: A Data Visualization Tool for Qualitative Research Methods. *Leonardo*, vol. 50, no. 3, 268–271. DOI: 10.1162/LEON_ a_ 01148.

[4] R. Keller, 2013. *Doing discourse research: An introduction for social scientists*. London: SAGE Publications.

[5] A. Luther and W. J. Schünemann, 2018. From analysis to visualisation: Synoptical tools from SKAD studies and the entity mapper. In *The sociology of knowledge approach to discourse*, R. Keller, A.-K. Hornidge, and W. J. Schünemann, Eds. Abingdon, Oxon and New York, NY: Routledge, 274–299.

[6] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu. 2015. A Survey of Heterogeneous Information Network Analysis. *arXiv:1511.04854 [physics*]. Available: http://arxiv.org/abs/1511.04854.

[7] A. Downs. 1968. *Ökonomische Theorie der Demokratie*, vol. 8. Tübingen: Mohr.

[8] A. Lupia and M. D. McCubbins, 1998. *The democratic dilemma. Can citizens learn what they need to know?* Cambridge: Cambridge Univ. Press.

[9] D. Lazer and S. Wojcik, 2018. Political Networks and Computational Social Science. In *The Oxford handbook of political networks*, J. N. Victor, A. H. Montgomery, and M. Lubell, Eds. New York, NY: Oxford University Press, 115–130.

[10] J. N. Victor, A. H. Montgomery, and M. Lubell, 2018. Introduction: The Emergence of the Study of Networks in Politics. In *The Oxford handbook of political networks*, J. N. Victor, A. H. Montgomery, and M. Lubell, Eds. New York, NY: Oxford University Press, 3–57.

[11] P. L. Berger and T. Luckmann, 1969. *Die gesellschaftliche Konstruktion der Wirklichkeit. Eine Theorie der Wissenssoziologie*. Frankfurt and Main: Fischer.

[12] R. Keller, 2005. Analysing Discourse. An Approach From the Sociology of Knowledge. *Forum: Qualitative Social Research (FQS)*, vol. 6, no. 3, Art. 32.

[13] K. Mannheim, 1964. *Wissenssoziologie Auswahl aus dem Werk*, vol. 28. Berlin Neuwied: Luchterhand.

[14] S. Maasen, 2009. *Wissenssoziologie* (2., komplett überarb. Aufl.). Bielefeld: Transcript-Verlag.

[15] W. L. Bennett and S. Livingston, 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, vol. 33, no. 2, pp. 122–139, 2018, DOI: 10.1177/0267323118760317.

[16] M. Dunn Cavelty, 2008. *Cyber-security and threat politics: US efforts to secure the information age*. London: Routledge.

[17] A.-K. Hornidge, R. Keller, and W. Schünemann, 2018. Introduction. The sociology of knowledge approach to discourse in an interdependent world. In *The sociology of knowledge approach to discourse*, R. Keller, A.-K. Hornidge, and W. Schünemann, Eds. Abingdon, Oxon and New York, NY: Routledge, 1–15.

[18] L. A. Adamic and N. Glance, 2005. The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, 26-43. DOI: https://doi.org/10.1145/1134271.1134277

[19] A. Bruns and J. Burgess, 2011. The use of Twitter hashtags in the formation of ad hoc publics. In *6th European Consortium for Political Research General Conference*. University of Iceland, Reykjavik.

[20] M. Eriksson Krutrök and S. Lindgren, 2018. Continued Contexts of Terror: Analyzing Temporal Patterns of Hashtag Co-Occurrence as Discursive Articulations. *Social Media + Society*, vol. 4, no. 4. DOI: 10.1177/2056305118813649.

[21] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, 2014. Information network or social network? the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, Seoul, Korea, 493–498, DOI: 10.1145/2567948.2576939.

[22] A. Bar-Hen, P. Barbillon, and S. Donnet, 2020 .Block models for multipartite networks.Applications in ecology and ethnobiology. *arXiv:1807.10138 [stat]*. Available: http://arxiv.org/abs/1807.10138.

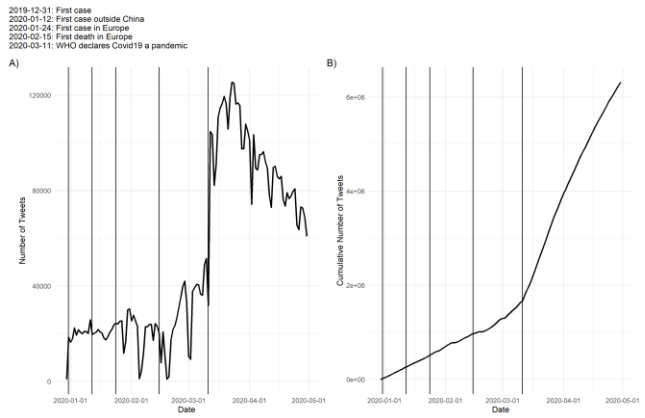
[23] M. Gerlach, T. P. Peixoto, and E. G. Altmann, 2018. A network approach to topic models. *Science advances*, vol. 4, no. 7, eaaq1360.

[24] D. Dimitrov *et al.*, 2020. TweetsCOV19 -- A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2991–2998. DOI: 10.1145/3340531.3412765.

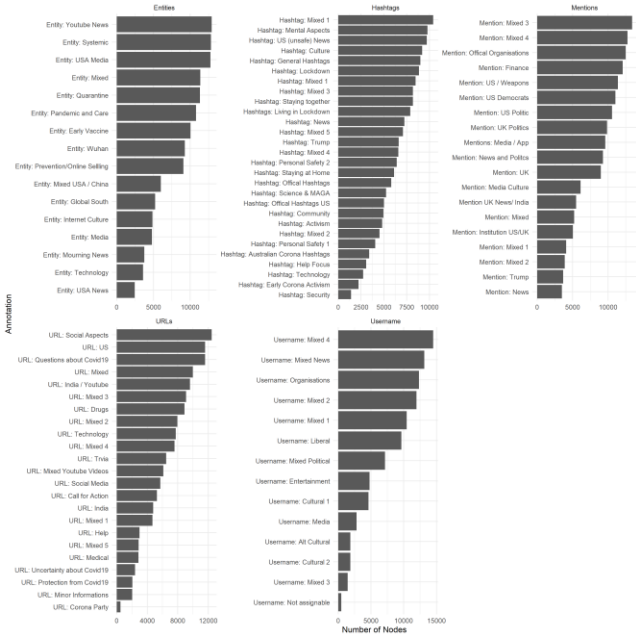
[25] T. P. Peixoto, 2014. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Phys. Rev. E*, vol. 89, no. 1, 012804. DOI: 10.1103/PhysRevE.89.012804.

[26] T. P. Peixoto, 2019. Bayesian stochastic blockmodeling. *arXiv:1705.10225 [cond-mat, physics:physics, stat]*, 289–332. DOI: 10.1002/9781119483298.ch11.

APPENDIX



Appendix 1: Number of Tweets over Time



Appendix 2: Number of Nodes in each Block