Data & Policy (2024), 6: e5
doi:10.1017/dap.2023.48

RESEARCH ARTICLE

CAMBRIDGE UNIVERSITY PRESS

Inferring social networks from unstructured text data: A proof of concept detection of hidden communities of interest

Christophe Malaterre[1,2] and Francis Lareau[3]

[1]Département de philosophie, Université du Québec à Montréal (UQAM), Montréal, Québec, Canada
[2]Centre interuniversitaire de recherche sur la science et la technologie, Université du Québec à Montréal (UQAM), Montréal, Québec, Canada
[3]Département d'informatique, Université du Québec à Montréal (UQAM), Montréal, Québec, Canada
Corresponding author: Christophe Malaterre; Email: malaterre.christophe@uqam.ca

Received: 24 May 2023; **Revised:** 30 October 2023; **Accepted:** 30 November 2023

Keywords: hidden colleges; hidden communities of interest; philosophy of science; social networks; text-mining; topic modeling

Abbreviations: HCoI, hidden communities of interest; LDA, latent Dirichlet analysis; SNA, social network analysis

Abstract

Social network analysis is known to provide a wealth of insights relevant to many aspects of policymaking. Yet, the social data needed to construct social networks are not always available. Furthermore, even when they are, interpreting such networks often relies on extraneous knowledge. Here, we propose an approach to infer social networks directly from the texts produced by actors and the terminological similarities that these texts exhibit. This approach relies on fitting a topic model to the texts produced by these actors and measuring topic profile correlations between actors. This reveals what can be called "hidden communities of interest," that is, groups of actors sharing similar semantic contents but whose social relationships with one another may be unknown or underlying. Network interpretation follows from the topic model. Diachronic perspectives can also be built by modeling the networks over different time periods and mapping genealogical relationships between communities. As a case study, the approach is deployed over a working corpus of academic articles (domain of philosophy of science; N=16,917).

Policy Significance Statement

Mapping groups of actors sharing similar interests is crucial for understanding and addressing their specific claims. This occurs, for instance, when targeting groups of voters during political campaigns or sorting out claims of sets of stakeholders during negotiations, or still when mapping the balance of pros and cons around large infrastructure public projects. Yet, social network data, such as who-follows-who, that would be necessary to delineate these groups of actors and identify their claims are not always available. In this study, we propose a computational approach that makes use of the textual contents produced by actors to infer underlying social networks. Indeed, since actors' claims manifest themselves in the topical content of the text data they produce, we show that these texts can be used to identify latent groups of individuals sharing similar interests (i.e., hidden communities of interests).

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

1. Introduction

Analysis of networks of relationships between different actors is well-known to provide a wealth of insights, be it about the identification of key members and their relationships, the interdependence and flows of influence among individuals, groups, and institutions, or still the very structure of the networks themselves and their evolution over time. This is what social network analysis (SNA) is all about: developing and applying network methods to investigate social structures wherever they may occur: friendship and acquaintance networks, business networks, knowledge networks, and so forth. In social networks, nodes typically refer to persons, organizations, or generally speaking actors, while edges or links represent some form of connection between the nodes. The underlying idea is that the network formed by these nodes and edges can be understood as a kind of structure that captures crucial aspects of a particular social or political phenomenon.

In political science, network theory and methods have been increasingly applied in the past decades to a broad range of questions (Lazer, 2011; Ward et al., 2011; Victor et al., 2017). For instance, questions about voting, political participation, interest groups, and legislative networks have been addressed with SNA (Fowler, 2006; Huckfeldt, 2009; Battaglini and Patacchini, 2019; Praet et al., 2021), as have numerous issues in public policy and public administration (e.g., health policy) (Luke and Harris, 2007; Shearer et al., 2014), or in comparative politics (Siegel, 2011). Maybe one of the best known areas where network analyses have been applied concern international relations (Hafner-Burton et al., 2009), notably on questions of terrorism, trade networks, global governance, and advocacy networks (Krebs, 2002; Ressler, 2006; Knoke, 2015; Varone et al., 2017).

Due to the richness that SNA affords, this type of approach is now ubiquitous across disciplines, from the natural sciences to the human and social sciences. Unsurprisingly, SNA also plays a significant role in the very study of science itself. In science studies indeed, the social networks that scientists form have been examined under numerous perspectives, from the identification of latent social structures and "hidden colleges" (Crane, 1969) to broader investigations about the general dynamics of science (Small, 1999; Boyack et al., 2005), including the role of networks of scientists and institutions over such issues as problem selection, discovery, collaboration, or even career dynamics (Tang et al., 2008; Fortunato et al., 2018; Kong et al., 2019).

In any case, before being submitted to analysis, social networks first need to be built, which presupposes access to data about actors and their relationships. Sources of information are very diverse and depend on the cases at hand. In political science, data can, for instance, be manually collated from press articles, extracted from pieces of legislature, reconstructed from email exchanges, or still, mined from social media such as X (previously Twitter), Facebook, and the like. Likewise, in science studies, major sources of network data, including academic social media and citation data have been mined to identify researchers' profiles, collaborations, and trajectories (Tang et al., 2008; Kong et al., 2019). Often, the interpretation of social networks requires supplementary information from which to infer the meaning of the relationships between actors or the specificity of groups of actors or communities. In other words, having data about nodes and edges is not enough when it comes to understanding some of the specific structural features of social networks. For instance, one will want to gain insights into the key properties of specific clusters of actors that become apparent through SNA: What is special about that particular advocacy group compared to the others? What do these political representatives have in common that may explain their higher success in passing pieces of the legislature? Such additional data may have been collected, typically as metadata, jointly with actors and interaction data. Yet, often, it is assembled in a second step. For instance, to make sense of scientific co-citation networks, data about author research specialties are often needed in addition to author names (nodes) and co-citation frequencies (edges), and are typically obtained by examining key publications or keywords through other means (e.g., Raimbault et al., 2016, Réale et al., 2020). With such approaches, networks are built from relational data (links between nodes), and the meaning of the resulting communities (set of nodes which may share similar content) is inferred with the help of supplementary data.

To address this network interpretability issue, some researchers have proposed to computationally mine the textual data of actors and develop specific topic models that could incorporate such actor-related

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

data (Steyvers et al., 2004), notably in the case of social media and directional networks (McCallum et al., 2007; Pathak et al., 2008), or in the case of co-authorship data (Zhou et al., 2006; Zhang et al., 2007). Others have proposed to further develop community detection algorithms so as to include not only topological information but also prior constraining data on nodes, as in semi-supervised community detection algorithms and graph neural network approaches (Yang et al., 2014; Ye et al., 2018). On the other hand, specific topic modeling techniques have been built upon to elaborate approaches aimed at more than simply extracting topics from textual data. This is, for instance, the case when topic models are used as a first step to identify product opportunities using social media mining (Ko et al., 2018). Here too, we propose to use topic models not only to extract topics from texts but also to infer actor networks.

The present work tackles the question of social networks in extreme contexts where semantic or textual information produced by actors is abundant but relational data are scarce or even inexistent. In other words, contexts where data about the nodes are present, as well as textual data attributable to each node, but where no data about the edges between nodes exist. Such situations could be seen as even precluding the very notion of social network. Yet, as we propose to show, underlying communities of actors can still be identified based on their shared semantic content. We call such communities "hidden communities of interest" (HCoI), that is, groups of actors sharing similar semantic contents but whose social relationships with one another may be unknown. HCoI's reflect the existence of underlying latent social networks whose study can nevertheless be pursued to gain insights, for instance, into their structure and evolution.

To a certain extent, semantic network approaches may be used in such contexts. Semantic networks typically map relationships between terms used in a given corpus by measuring their co-occurrences (Carley, 1993; Danowski, 1993; Doerfel and Barnett, 1999). These networks have been used in a wide variety of contexts to analyze word association patterns, for instance, in the context of information and communications technology policy (Danowski et al., 2023), in cognitive science for understanding semantic memory (Siew et al., 2019; Kumar et al., 2022; Christensen and Kenett, 2023), in business and management to analyze public–private partnerships (Castelblanco et al., 2021), and in many domains of the social sciences (Segev, 2021). When specifically targeting named entities, semantic networks can be used to infer relationships between these named entities by mapping their co-occurrences in texts. This has been done, for instance, to identify covert networks (Diesner and Carley, 2004) or reconstruct the social networks of cabinet members of past US presidents (Danowski and Cepela, 2010). Semantic network approaches can also be used jointly with social network approaches, for instance, to better understand the semantic content of specific clusters of authors within a social network, as has been done in health policy by examining the tweets of communities of vaccine-hesitant influencers inferred from the graph of their social relationships (Ruiz et al., 2021). Furthermore, when texts are attributed to actors, semantic networks can then be built for all actors and compared with one another using a matrix similarity measure, resulting in an overall actor network (Danowski, 2011).

The approach we propose here for identifying latent networks of actors is somewhat akin to this last case: the main idea is to infer actor networks from similarities in the content of their textual data. Yet, instead of building actor semantic networks and comparing these networks with one another, we compute author topic profiles from a topic model fitted to the complete textual data, something Danowski (2011) somehow tried but inconclusively with the tools available at the time (to the best of our knowledge, this is the only attempt we have been able to identify post hoc). Here, we show that applying a combination of topic modeling and community detection approaches can indeed help reconstruct underlying communities of authors—or HCoI's as we propose to name them. An advantage of using topic models in this context is the additional possibility of easily gaining insight into the semantic specificity of each community. Furthermore, by segmenting the corpus into time periods, diachronic analyses can be conducted to examine the temporal evolution of the communities and their genealogies.

Such an approach should be of interest to a diversity of political science contexts where textual data is accessible jointly with actors that produce them, but the relationships between these actors are unknown. This could be the case with textual data gathered from a variety of Internet sources (e.g., advocacy blogs, political websites, newsfeeds) as well as varied published materials (e.g., newspaper articles, books, pamphlets etc.) or even transcripts of recorded data (e.g., interviews). Such textual data, of course, may

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

concern any policy-related topics, be they in public policy and administration (e.g., textual data expressing opinions about public health policies), international relations (e.g., textual data produced by various extremist groups), or voting and political participation (e.g., textual data originating from political blogs). The identification of HCoI's in these varied contexts should in turn provide valuable insights at all stages of the policy lifecycle, from agenda setting and policy formulation, to decision-making, policy implementation, and evaluation (Howlett et al., 2020).

In the present contribution, we use as a test case a non-policy-oriented corpus of 16,917 full-text academic articles in the domain of the philosophy of science. Besides its simply being available—from previous studies that led to its topical analysis (Malaterre and Lareau, 2022)—, the corpus has the advantage of containing a large amount of textual data (in English language) produced by numerous authors over a period of nearly 90 years, thereby allowing diachronic analyses.

The proposed approach starts by fitting a latent Dirichlet allocation (LDA) topic model to this corpus, thereby resulting in topic probability distributions for all full-text articles. Having split the corpus into four broad time periods, these topic probability distributions were averaged out per author and per time period, depending on the author's contribution to each article. This resulted in author topic profiles for each time period. Correlation analyses between author topic profiles then led to the construction of author correlation networks for each time period, which were submitted to Louvain community detection. In turn, the topic profile of each community was quantified by averaging out author topic profiles per community. These community topic profiles provide immediate insights into the semantic specificities of each community. Furthermore, measuring pairwise distances between community topic profiles across time periods provides a means of understanding the diachronic evolution of communities and their genealogies.

In what follows, we first describe the data and methods in more detail (Section 2). We then present the results, notably the networks of communities that were detected and their temporal evolution (Section 3). These findings are then discussed (Section 4).

2. Data and Methods

Since the proposed identification of HCoI's relies on the textual content produced by actors, the approach comprises two main steps: first, the assembly of the working corpus and its preparation for computational analyses and second, the computational analyses per se (Figure 1).

2.1. Step 1—Corpus assembly and preprocessing

For this case study, we used a corpus of full-text academic articles that had been assembled in (Malaterre and Lareau, 2022). The corpus spans from 1930 (the first issue of the earliest published journal) to 2017 and includes 16,917 research articles from eight of the most significant philosophy of science journals in English language (Table 1). In the present case, there are good reasons to consider the corpus as a representative sample of all the texts produced by the authors assembled here. Of course, philosophy of science is published in numerous other venues, including general philosophy journals, disciplinary focused journals, science journals or books. It is also published in many non-English languages. Nevertheless, the corpus includes the most authoritative journals of the field. It also includes the journals that started the field some ninety years ago and are still flagship journals today. These are therefore good reasons for accepting the corpus as offering a representative perspective of the discipline.

The corpus was cleaned and preprocessed in a standard way. To reduce the size of the lexicon, and therefore computation time, only nouns, verbs, adverbs and adjectives were kept following part-of-speech (POS) tagging and lemmatization (TreeTagger package (Schmid, 1994) with Penn TreeBank tag sets (Marcus et al., 1993)) and words occurring in fewer than 50 sentences in the corpus were removed. In parallel, author names were checked and disambiguated, ensuring similar spellings were used throughout the corpus. All authors (N=8,009) were assigned publication weights based on their respective number of articles (coauthored articles were evenly split). Four main time periods of 21 years each were then defined (1930–1951, 1952–1973, 1974–1995, 1996–2017).

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

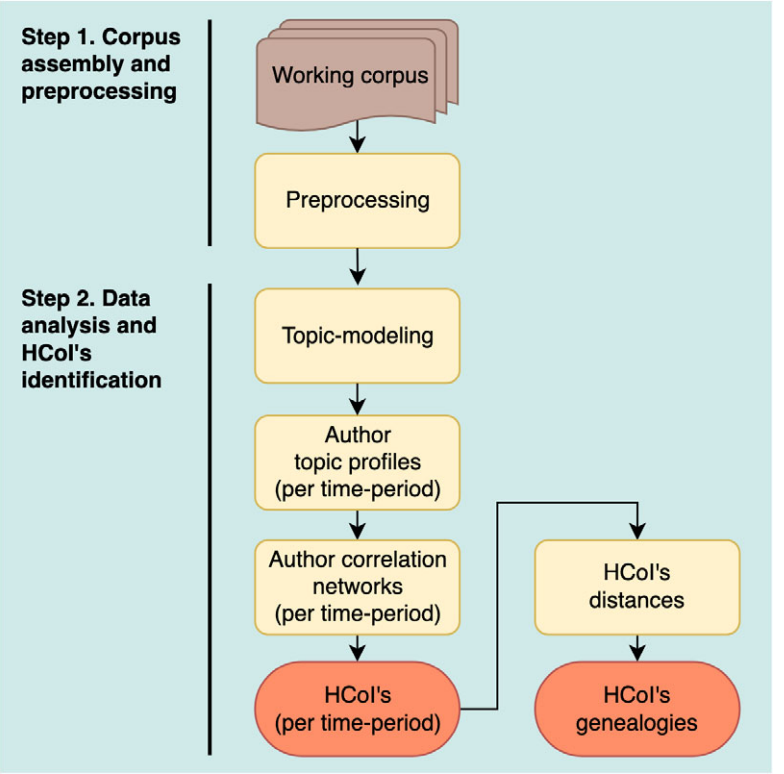

Figure 1. *Approach to identify HCoI's from a corpus of text data.*

Table 1. *Document distributions depending on sources (journals)*

Journals (alphabetic order)	Publication periods	Articles
British Journal for the Philosophy of Science	1950–present	1,862
Erkenntnis	1930-1940; 1975–present	2,127
European Journal of Philosophy of Science	2011–present	156
International Studies in the Philosophy of Science	1986–present	560
Journal for General Philosophy of Science	1970–present	929
Philosophy of Science	1934–present	4,605
Studies in History and Philosophy of Science A	1970–present	1,421
Synthese	1936–1939, 1946–1949, 1955–present	5,257
Total	1930–2017	16,917

As shown in Figure 2, the volume of articles has significantly increased over the past eight decades, from 1,575 for the 1930–1951 period to 8,300 for the 1996–2017 period, which is a 5.3-fold increase. Meanwhile, the number of authors has incurred an eightfold increase. Knowing that the number of articles per author has roughly remained constant throughout all four periods at about 2, the increase in authors denotes an increase in co-authorship. Indeed, the number of multiauthored articles has increased fourfold, from 4% in the first period to 16% in the last. Although this share is significantly lower than that in the sciences where single-authored articles are now virtually non-existent (e.g. in ecology, Barlow et al., 2018), or even in some areas of the humanities (e.g. in economics, Kuld and O'Hagan, 2018), multiple-authorship has been steadily rising in the philosophy of science. Note that the proportion of authors who

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

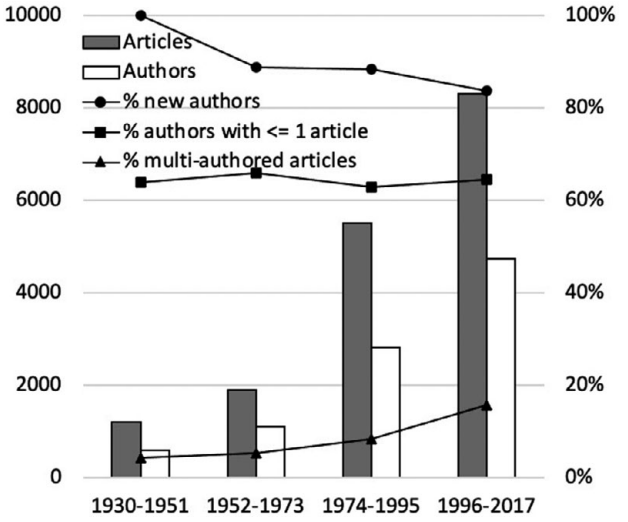

Figure 2. *Authors and articles.*

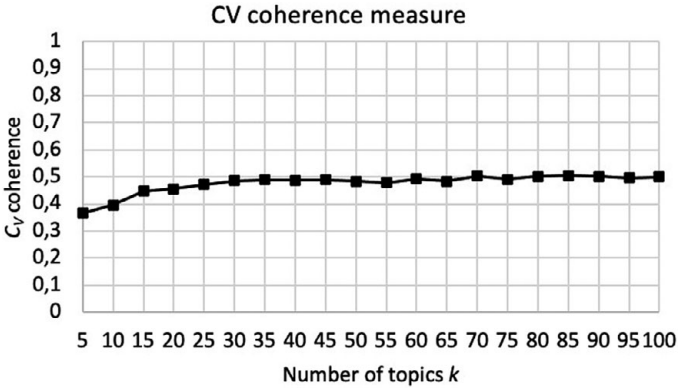

Figure 3. *Cv coherence topic models as a function of the number k of topics.*

only publish once (or "transients," see Crane, 1969) is relatively stable at about 65%. This not significantly different from what is observed elsewhere, a partial explanation being the share of doctoral students and postdoctoral researchers (e.g. in synthetic biology, Raimbault et al., 2016). Moreover, while the number of new authors (from one period to another) is above 80%, this proportion has been decreasing over time. This means that, although many authors in any given period were not present in the previous period, new authors now tend to represent a smaller share of authors compared to what it used to be.

2.2. Step 2—Data analysis and HCoI's identification

Following (Malaterre and Lareau, 2022), topic modeling was carried out with the well-known LDA algorithm, following (Blei et al., 2003) and (Griffiths and Steyvers, 2004). Units of analyses were complete articles. The topic modeling operation resulted in 25 probability distributions over the lexicon of the corpus terms (each probability distribution considered to represent a topic), and the probability distributions of these topics in each one of the 16,917 articles. The number of topics $k = 25$ was chosen as a compromise between an optimal coherence measure (Röder et al., 2015) for a variety of models from $k = 5$ to 100 (see Figure 3) and upon manual inspection of top-words (in particular for models below 35, since higher-k models led to no increase in coherence). In the end, the topics of the model with $k = 25$ were found to be more meaningfully

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

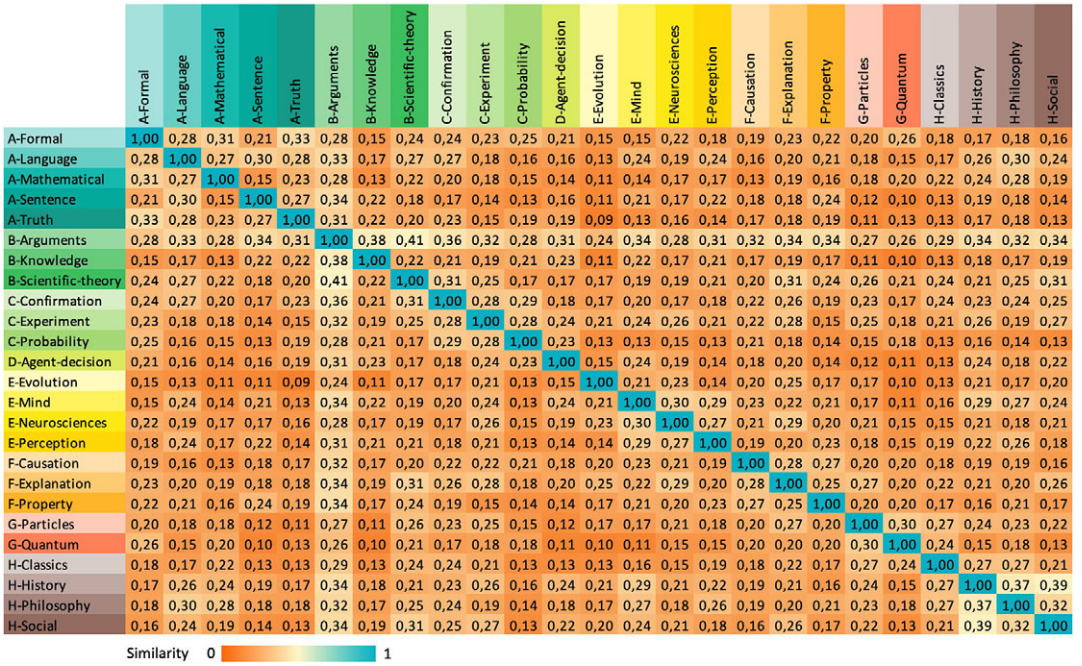

Figure 4. *Inter-topic similarity measures (1-Hellinger distance).*

interpretable compared to the others, while being characterized by a relatively high coherence (though not the highest).

Inspection of the most probable terms within each topic and of selected text excerpts made it possible to carefully interpret and label all topics. For ease of handling, topics were also grouped into categories based on their correlation within corpus documents, and Louvain community detection performed on the graph of topic correlations in Gephi (Bastian et al., 2009). These categories were interpreted based on expert knowledge of the field (the categories are denoted by a capital letter in front of the topic name). To give a sense of the 25 topics, their top 10 words are listed in Table 2, sorted by categories and alphabetic order.

In this particular case study, the topics correspond to well-known research themes in the philosophy of science (Malaterre and Lareau, 2022). Group A of topics denotes research questions that are characteristic of the philosophy of language and logic. Group B includes topics in epistemology and theory of knowledge (including questions about realism), while group C relates more specifically to induction, confirmation, and the use of probabilities. Group D is about rational decisions and game theory. Topics in the philosophy of biology and the neurosciences are found in group E. Group F includes a set of traditional topics that concern the process of scientific explanation, the nature of causation, and the status of natural kinds. Topics in the philosophy of physics are in group G, with thermodynamics, electromagnetism, chemistry in one topic, and relativity and quantum theory in the other. Finally, group H gathers topics that are characterized by a more historical or social discourse. These include research themes in the history of science and in the history of philosophy, but also investigations on the social dimensions of science.

Note that a measure of similarity between topics was also calculated as 1-d, with d = Hellinger distance —which is appropriate for probability distribution—between topic probability distributions over the corpus lexicon, using the Gensim implementation (Rehurek and Sojka, 2010). The corresponding heat map (Figure 4) confirms that the topics are overall very dissimilar, the only slight exceptions being the topic B-Arguments, whose similarity values are among the highest, reaching about 0.4 with the two other topics of cluster B (hence denoting quite a generic topic), and topic H-History, also with similarity values near 0.4 with two other topics of cluster H.

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

Table 2. *Topics and keywords*

Topic name	Top 10 words
A-Formal	set; function; relation; define; definition; structure; order; model; theory; class
A-Language	language; sentence; term; meaning; concept; use; statement; logical; mean; word
A-Mathematical	mathematical; mathematics; number; proof; axiom; geometry; theory; object; point; line
A-Sentence	sentence; context; use; say; reference; content; name; true; semantic; speaker
A-Truth	logic; truth; true; proposition; sentence; logical; formula; follow; rule; world
B-Arguments	argument; claim; say; question; make; view; reason; fact; case; point
B-Knowledge	belief; knowledge; epistemic; believe; know; case; evidence; reason; justification; true
B-Scientific-theory	theory; scientific; theoretical; empirical; realism; realist; truth; science; true; claim
C-Confirmation	law; hypothesis; statement; evidence; theory; condition; inductive; problem; confirmation; fact
C-Experiment	datum; experiment; value; use; test; result; experimental; model; hypothesis; method
C-Probability	probability; measure; value; give; chance; case; function; distribution; degree; frequency
D-Agent-decision	agent; action; decision; game; choice; act; utility; strategy; moral; preference
E-Evolution	selection; population; organism; evolutionary; gene; biological; individual; group; evolution; specie
E-Mind	behavior; state; mental; action; psychological; human; function; psychology; person; child
E-Neurosciences	system; information; process; cognitive; level; mechanism; state; representation; structure; function
E-Perception	object; experience; perception; see; color; perceptual; visual; content; red; image
F-Causation	causal; cause; event; effect; causation; condition; case; variable; time; occur
F-Explanation	model; explanation; explain; account; explanatory; phenomenon; use; case; system; provide
F-Property	property; world; object; physical; relation; kind; entity; part; identity; exist
G-Particles	theory; energy; law; particle; electron; atom; physical; physic; chemical; system
G-Quantum	time; state; space; quantum; system; theory; particle; physical; field; point
H-Classics	motion; body; force; newton; law; Galileo; earth; move; light; time
H-History	work; time; man; history; new; year; make; life; century; write
H-Philosophy	world; nature; knowledge; concept; experience; kant; sense; thing; idea; reality
H-Social	science; scientific; social; research; scientist; philosophy; knowledge; problem; history; practice

Obviously, the topics identified through such topic models depend on the textual data that are being modeled. A policy-related corpus will exhibit topics about policy matters as expressed by the texts. Note that the size and structure of the corpus also matter, notably with respect to the number k of topics that will ultimately be chosen: this is the case for the units of analysis (e.g., entire documents vs. sections of paragraph), and the overall number of these units of analysis. Metrics can help in the choice of k yet, though we have found that the interpretability of the topics by human judgment and their relevance for the research questions at stake also matter much (Grimmer and Stewart, 2013).

The probability distributions of topics inside documents were then used to identify the semantic signature of authors in terms of their contributions to particular topics. First, articles were sorted

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

depending on their publication year into the four time periods defined above, and equally split between coauthors (i.e., that an article with two authors counted for ½ for each author in terms of weight). Then, article topic distributions were averaged out per author for each one of these periods (taking into account the co-authorship weights). This step resulted in topic profiles for each author based on their publications during any given time period. In other words, for each time period, probability distributions over the 25 topics were computed for all authors having published during that time period. These distributions are what we call "author topic profiles".

For each time period, Pearson correlations among these author topic profiles were calculated. Correlation networks were built in Gephi (Bastian et al., 2009), using Louvain community detection (with default parameters). To reduce noise, only authors with weighted publication above 2 were retained (thereby filtering out "transient authors"), and a correlation threshold was set to 0.6 (this resulted in keeping all significant author communities connected to the network main component across all four time periods while removing clutter). To facilitate the interpretation of each author community, topic profiles (i.e., topic probability distributions) were calculated at the community level by averaging out their author topic profiles.

To get further insights into the genealogy of communities over time, we calculated the Hellinger distances between community topic profiles across time periods and focused on closest pairings. These sets of distances made it possible to identify which communities persisted over time, which other ones appeared or disappeared, bifurcated or merged, thereby generating a diachronic picture of the evolution of author communities and their main topics.

3. Results

3.1. HCoI's and their evolution through time

The approach we described makes it possible to identify groups of actors sharing similar semantic contents as revealed by the texts they produce, but whose social relationships with one another may be initially unknown or underlying (HCoI's). Technically speaking, HCoI's are groups of actors whose topic profiles (obtained through a weighted average of their document topic profiles) are highly correlated with one another. Adding a temporal dimension makes it possible to map the evolution of the different communities through time. In the present case study, we chose to investigate the different HCoI's though four successive time windows so as to shed light not only on the structure of the resulting networks of actors, but also on its evolution and the relative importance of the different communities through time.

As with any network, key structural features of HCoI networks can be analyzed with the help of descriptive measures such as density (ratio of actual edges to the total possible number of edges), betweenness (the extent to which nodes lie between other nodes), modularity (probability that two associates of a node are themselves connected), cohesion or degree (number of edges per node) etc. (e.g., Wasserman and Faust, 1994; Borgatti et al., 2013; Yang et al., 2016). Table 3 summarizes some of these network statistics in the present case study (calculated with Gephi on each graph). The density of the networks tends to decrease over time, while measures of betweenness, cohesion, and modularity increase: this indicates that the HCoI's gradually differentiated from one another over time, resulting in more distinct and compact communities in the 2000s compared to the 1930s. This trend toward more numerous and highly distinctive communities also visually stands out when looking at the sequence of network graphs over time (Figures 5–8). These figures include a network representation of the communities present during the corresponding time period (Figures "a," where nodes represent actors; actor name size and node size proportional to actor weighted number of publications; node color corresponding to community dominant topic), and the topic profiles of the communities (Figures "a").

As is apparent, the overall field of interest—the discipline of philosophy of science—has significantly grown in terms of both domains of interest and actors. In particular, the number of specialized communities has incurred a threefold increase over the past eight decades. In the 1930s–1940s (Figure 5a), the field comprised just a handful of communities. A clearly identifiable cluster (1a) consists in the community of

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

Table 3. *Network statistics per time period*

Network statistics	1930–1951	1952–1973	1974–1995	1996–2017
Density	0.167	0.057	0.027	0.030
Betweenness centrality (average path length)	2,551	3,463	4,226	4,161
Cohesion (average degree)	31,558	18,876	25,465	41,115
Modularity	0.427	0.674	0.779	0.834
Number of communities	4	7	13	13
Node coverage (giant component)	96.45%	93.50%	98.83%	98.77%

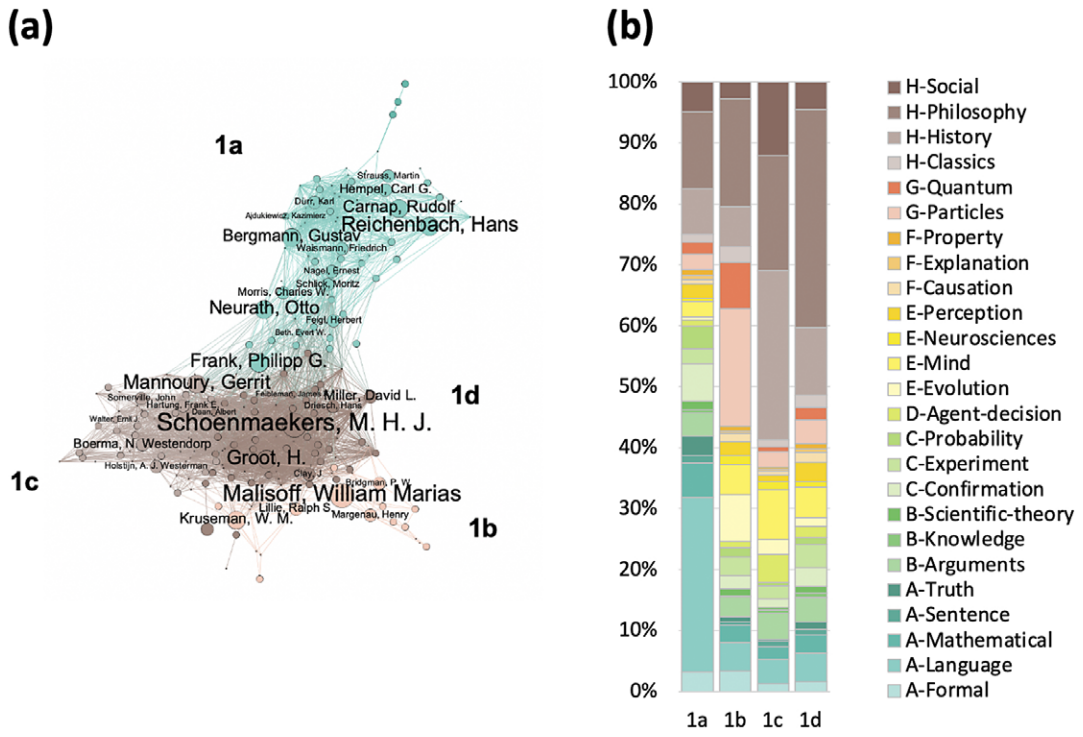

Figure 5. *(a) Actor communities and (b) their topic profiles 1930–1951.*

the logical positivists and members of the Vienna circle (e.g., Neurath, Reichenbach, Carnap, Hempel), distinctively focused on the philosophy of language and logic (as can be seen on the topic profile of that community on Figure 5b). As is well-known to experts in the field, the subsequent development of the philosophy of science owes much to these authors. The other half of the network consists of two closely interconnected communities, with, on the one hand (1c), a group of actors somehow at the border between philosophy and other humanities (e.g., history, anthropology, economics, psychology), and on the other (1d), authors engaging in more traditional metaphysics or ontology (e.g., realism, subjectivity etc.). Although still engaging with science, these two groups remained much anchored to a classical philosophical discourse. A distinct and much smaller community (1b), somehow at the fringe of the network, consists of philosophers focusing more on physics, and discussing issues related to matter, energy, or physical theories (e.g., electromagnetism or quantum mechanics; note the presence of Malisoff, founder of *Philosophy of Science*).

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

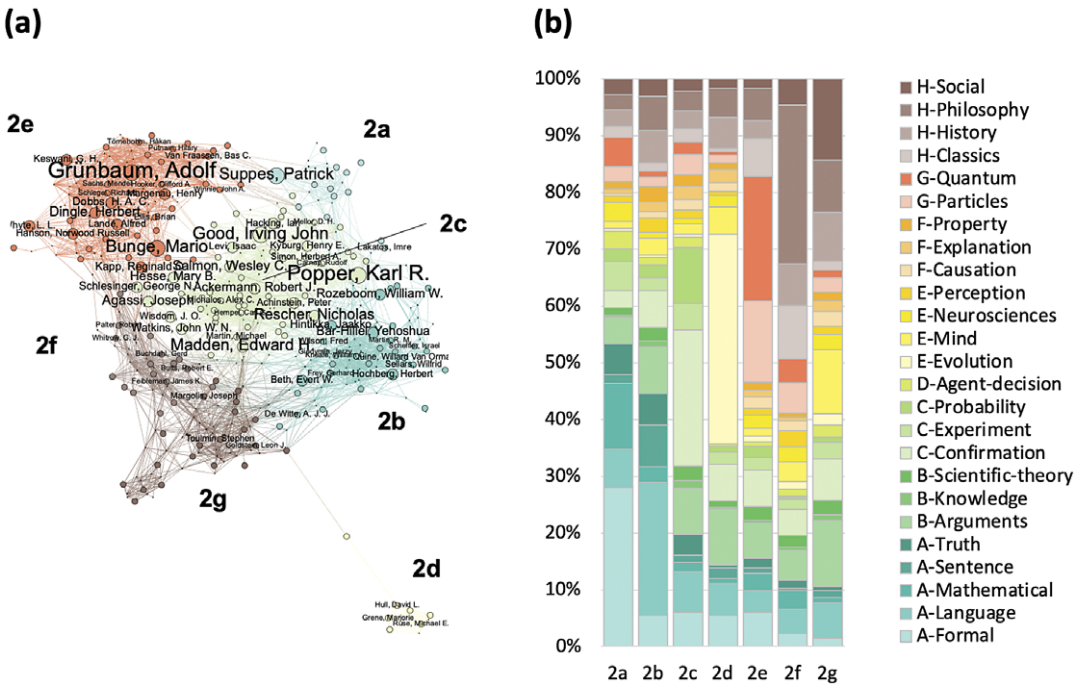

Figure 6. *(a) Actor communities and (b) their topic profiles 1952-1973.*

The actor network developed substantially in the 1950s throughout the early 1970s, with an increasing number of interconnected communities (Figure 6). Community (2a) includes logicians and philosophers of mathematics with a distinctively formal vocabulary. Philosophers of language constitute a separate community (2b). Occupying a central position in the network, (2c) is a community targeting specific issues related to confirmation and the status of scientific theories (e.g., induction, verifiability, corroboration, or refutation; note the presence of Popper). A small and peripheric group of actors (2d) consists of the nascent community of philosophers of biology, with a notable focus on evolutionary theory. On the contrary, philosophers of physics constitute a larger community (2e), addressing a diversity of epistemic issues related, for instance, to relativity theory or quantum mechanics. In continuity with the previous period, a distinctive community is constituted by actors at the border with traditional philosophy (2f), while a nearby community appears to address more sociological aspects of science (2g).

In the 1970s throughout the 1990s, the philosophy of science continued to grow in terms of actors but also in terms of topic communities (Figure 7). Community (3a) consists of logicians and philosophers of mathematics, somehow in continuity with a second community (3b) more centered on semantics and the philosophy of language (note the presence of Hintikka known for his work on formal epistemic logic and game semantics for logic). A specific community consists of actors addressing epistemology and theory of knowledge questions (3c). Questions about confirmation and the status of scientific theories characterize a community somehow at the center of the network (3d). Note the appearance of a specific community focused on probability theory and its relevance for science and knowledge (3e). Another new community consists of actors interested in decision and game theories, and their applications in science (3f). The community of philosophers of biology remains at the margin of the rest of philosophy of science but has significantly grown in size (3g; note the presence of Sober). A novel community has appeared around researchers more specifically targeting the philosophy of mind and the neurosciences (3h). Yet, another community of actors distinctively focuses on causation (3i). Two communities are characterized by topics related to the philosophy of physics: the first one with a clear focus on quantum mechanics and relativity (3k), and the second smaller one more oriented toward thermodynamics, chemistry, and electromagnetism

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

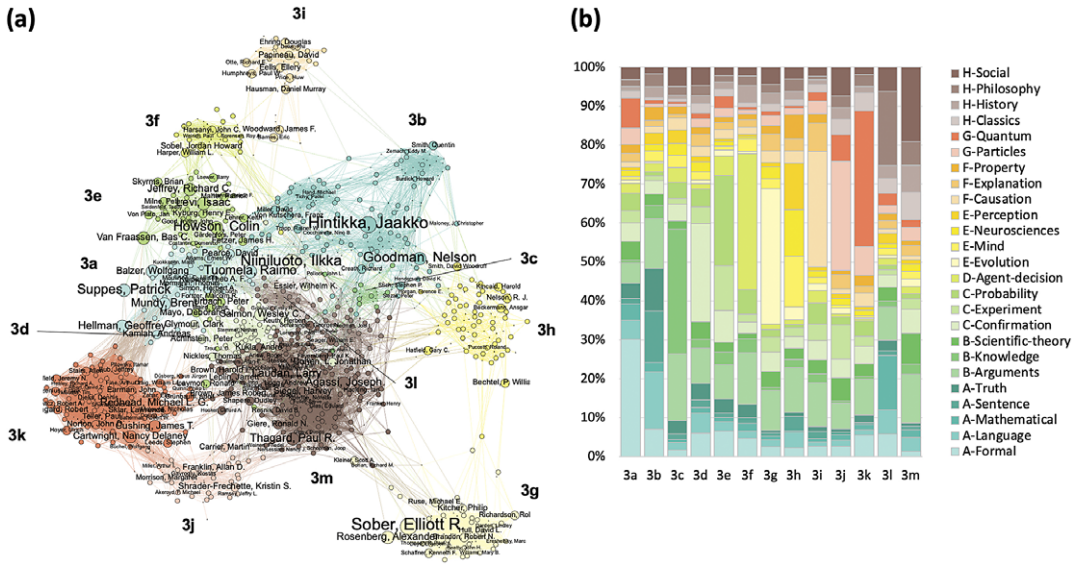

Figure 7. *(a) Actor communities and (b) their topic profiles 1974-1995.*

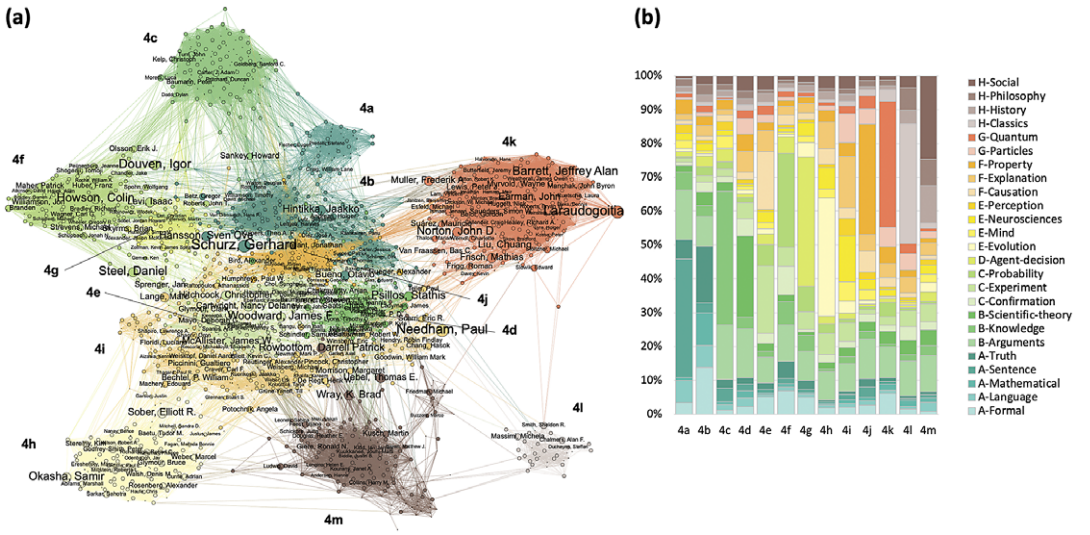

Figure 8. *(a) Actor communities and (b) their topic profiles 1996–2017.*

(3j). A relatively diffuse community gathers actors who tend to have a more traditional philosophical standpoint (3l). Finally, a large community consists of a diverse set of actors who tend to target some social dimensions in science (3m).

The trend toward an increase in terms of number of actors and a specialization of discursive topics continued in the 1990s throughout the 2010s (Figure 8). A community of philosophers of language (4a) can be seen quite tightly connected to a second community of philosophers of logic (including modal and intuitionistic logic) notably interested in notions of truth (4b). A nearby community consists of epistemologists, philosophers specializing in theory of knowledge (4c). Toward the center of the network, a community focuses on the status of scientific theories notably with respect to realist and anti-realist

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

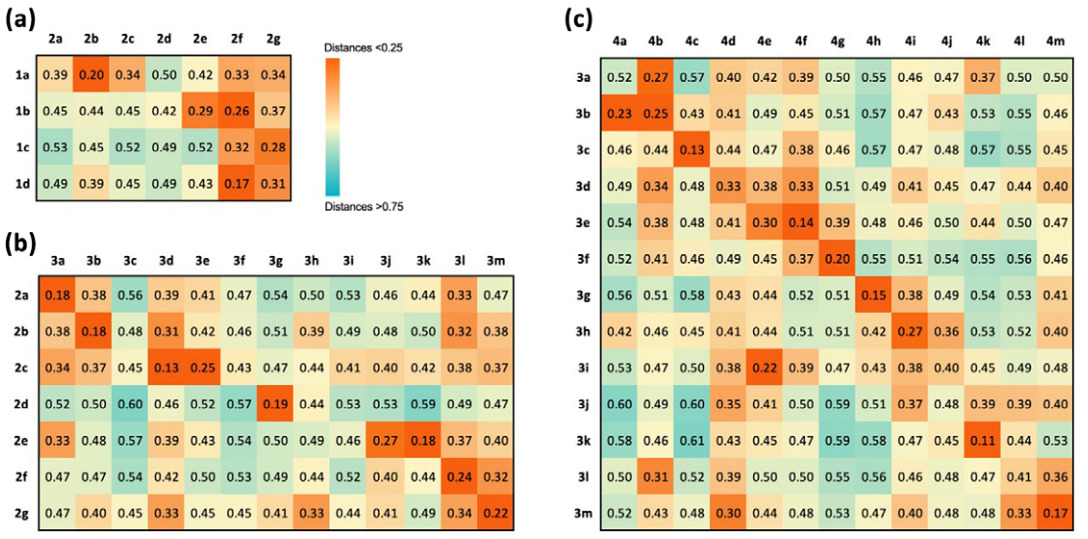

Figure 9. *Community distances (Hellinger distances between community topic probability distributions): (a) between the communities of the first period (1930-1951) and those of the second (1952-1973); (b) between the communities of the second period and those of the third (1974-1995); (c) between the communities of the third period and those of the fourth (1996-2017).*

stances (4d). A nearby and more diffuse community gathers actors interested in topics that relate to data, experiments, and modeling, but also somehow to causation (4e). The community of philosophers of probability, which had appeared in the previous period has grown in size and individuated itself (4f). A community of researchers somehow bridging philosophers of probability and of logic consists of actors focusing on game theory and various aspects of rational choice theory (4g). The community of philosophers of biology (4h) has significantly grown and is somehow more integrated with the rest of the network, notably with the community of philosophers of the neurosciences and others interested in scientific explanation (4i). At the center of the network lies a community generally interested in ontology (4j), addressing issues about properties or kinds among others. A large group of philosophers of physics constitutes a relatively well distinct community that tends to focus on relativity and quantum mechanics, with related issues such as the structure of space-time (4k). A noticeably distinct category of actors appears to mobilize classical philosophical works in their discussion of science (4l). Finally, a community of authors focuses on the social dimensions of science and various aspects of the practice of science (4m).

What was done here with this case-study corpus of academic articles could be done in any other policy-related context where textual data are abundantly produced by a set of actors whose relational data may be unknown. As soon as texts can be associated with actors, the topics extracted from these texts can be used to construct actor-specific topical profiles that can in turn be used to assess the relative proximity of these actors from one another. The resulting HCoI's then correspond to groups of actors who produce texts with similar semantic content. So to speak, HCoI networks capture actor relationships in terms of who-talks-about-the-same-things-as-who.

3.2. Retracing community genealogies

Measuring the pairwise distances between the topic profiles of any two communities from two different periods provides insights on the transformation of HCoI's into one another through time: the shorter the distances, the closer the communities in terms of their thematic interests (Figure 9). This makes it possible to map the evolution of HCoI's though time, and in particular their genealogies as communities from one time period split into two or more communities in the subsequent period, or the other way around. When

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

considered in conjunction with community topic profiles, the genealogical relationships between communities also make it possible to understand the relative shifts in significance of the different thematic interests through time among related HCoI's.

In the present case study, transitioning from the first period (1930–1951) to the second (1952–1973) (Figure 9a), one sees a reasonably good filiation between communities focused on philosophy of language and logic (1a to 2b). However, the other three communities tend to consolidate into one, and possibly into two (1b–d to 2f–g): the early philosophy of physics (1b) tends to bifurcate into a community still centered on similar physics-related topics (2e) and another community closer to traditional philosophy (2f), the latter being in the continuity of actors engaging in more metaphysics or ontology (1d). Note how the communities 2f and 2g appear to be relatively close to all the communities of the previous period, indicating a reconfiguration of actors and their topics of interest. Given the increase in the number of communities, this also denotes a form of marginalization of what once constituted the core of the philosophy of science. Note how the philosophy of biology community of the second period (2d) shows little continuity with previous communities, indicating the emergence of a novel HCoI.

The transition from the second period (1952–1973) to the third (1974–1995) also shows an increase in the number of communities, yet filiations tend to be stronger, indicating a form of stabilization of research communities with novel themes still emerging (Figure 9b). Philosophy of language and logic communities map well onto one another (2a-b to 3a-b). The community about confirmation and scientific theories (2c) persisted into 3d, while giving rise to a distinct community focused on probability theory and its relevance for knowledge (3e). The philosophy of biology community also persisted as a well identified set of actors and topics (2d to 3g). The community of philosophers of physics (2e) appears to have grown and split into one community more focused on relativity and quantum theory (3k) and another on the rest of physics (3j). The two socio-historico-philosophical communities (2f–g) somehow persisted (3l-m), though one notes a relative proximity of the latter communities to many of the communities of the previous period, indicating multiple reconfigurations. Four novel communities appeared in the 1970s–1980s without any clear filiation from communities of the previous period: a community focusing on knowledge theory (3c), another exploring game theory and rational choice (3f), yet another on philosophy of mind and the neurosciences (3h), and finally a community distinctively focusing on causation (3i).

The number of communities stabilized during the last decades of the 20th century. The transition from the third (1974–1995) to the fourth period (1996–2017) shows a relatively good continuity (Figure 9c). Communities of philosophers of language and logic slightly reorganized themselves depending on topic alignments but remained stable as a group (3a-b to 4a-b). Epistemologists persisted as a specific community, while gaining in momentum and autonomy (3c to 4c). The community focusing on probability theory and knowledge also persisted (3e to 4f), as well the communities on game theory (3f to 4g), on the philosophy of biology (3g to 4h), on the philosophy of mind and the neurosciences (3h to 4i), on the philosophy of relativity and quantum theory (3k to 4k), and on the social dimensions of knowledge (3m to 4m). On the other hand, some communities tend to have somehow dissolved into several. This is notably the case for the community on confirmation and scientific theories (3d) denoting a detachment from these topics in the 1990s–2000s. Similarly, the community focusing on chemistry, electromagnetism, or thermodynamics (3j) has somehow dissolved in subsequent decades, as well as the one which was centered on more traditional philosophical issues (3l). Finally, philosophers focusing on causation (3i) appear to have joined a broader community also interested in data, experiments, and modeling (4e).

Similar genealogical analyses can be carried out in any other context where actors and their HCoI's are similarly characterized by topic profiles. While an actor topic profile can be interpreted as a distinctive marker for each actor and used to cluster actors into specific interest-driven communities, the overall topic profile of each community can in turn be interpreted as a distinctive trait of that community and used to assess its relative proximity to other communities across time. This then makes it possible to trace back community genealogies and shed light on the origination of specific HCoI's.

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

4. Discussion

As we have seen, the findings result from a combination of topic modeling and community detection approaches. The main objective of these approaches is to identify HCoI, that is, groups of actors sharing similar semantic contents but whose social relationships with one another may be unknown. The methods make it possible to identify such communities from their semantic content and in the absence of known social connections. They also make it possible to assess the relative topic proximity of these communities at a given time and diachronically.

In the present case study, the identification of HCoI's in the academic domain of the philosophy of science highlight semantic reconfigurations in the field, with some communities dissolving into others (e.g., the community of confirmation and scientific theories of the 1970s–1990s), and others subsequently emerging (e.g., the communities of philosophers of biology in the early 1970s, or of epistemologists in the 1980s). Overall, the evolution of actor communities shows a phase of growth and diversification as the number of actors (and interests) increased, followed by a later phase of stabilization characterized by a form of intellectual entrenchment of larger and usually well delineated communities. These results concur with known episodes of the field, for instance the role of logical positivism in the constitution of the philosophy of science in the early 20th century (Giere and Richardson, 1996) or the emergence of a philosophy of biology in the 1970s as can be reconstituted by examining dedicated anthologies (Sober, 2006; Rosenberg and Arp, 2009), which confers confidence to the approach.

The methods can be relevant to a broad range of policy-related contexts where textual data can be gathered alongside with corresponding author-actors, even in the total absence of relationship data between actors. The construction of topic profiles for all actors is sufficient for inferring the underlying content-oriented networks they form, thereby making it possible to identify HCoIs that these actors form (alongside with a topic-based interpretation of the distinctive semantic profile of each community). When a temporal dimension is added, HCoI genealogies can also be generated, providing additional insights into the appearance and evolution of communities.

Note that the resulting networks differ from social networks as usually construed. Indeed, social networks are typically understood as depicting actual relationships between actors, such as who-follows-who, who-is-friend-with-who, who-sends-a-message-to-who, and so forth. By contrast, in the HCoI approach that is proposed here, relationships between actors are inferred from similarities in the thematic content of their texts: the actor networks are based on the similarity of their topic profiles (averaged from their respective texts). In short, HCoI's are about who-talks-about-the-same-thing-as-who. Whereas typical social networks can be said to encapsulate specific social relationships between actors (e.g., that of sending messages from one another), HCoI's capture relationships between actors that are mediated by the texts produced these actors. Obviously, in some contexts, both social relational data and textual data may be available, offering the special opportunity to build multilayer or multiplex networks offering complementary perspectives on the same phenomenon. In any case, mapping out HCoI's can be used as a heuristic to identify actual social connections between actors.

As mentioned earlier (Section 1), semantic networks may also be used to identify HCoI's (e.g., Danowski, 2011). However, this requires first building semantic networks for all actors based on their texts (which means calculating square matrices of the dimensionality of the size of the lexicon of the corpus), then measuring the pairwise distances between all networks (i.e., between all matrices). In the case of corpora with numerous authors, this approach may prove computationally demanding. In the case of the present corpus, this would mean building 8,009 semantic matrices of dimensionality 23,672 × 23,672 (size of the lexicon after lemmatization and POS tag filtering), and then calculating 8,009 × 8,009 /2 matric distances. By comparison, the approach proposed here relies on only three matrices of smaller dimensionality: an author × topic matric (8,009 × 25), a document × topic matrix (16,917 × 25), and a topic × term matrix (25 × 23,672). Although semantic networks can provide more details in terms of pairwise relationships between terms as opposed to the bag-of-words approach of topic models, the latter should prove less computationally demanding and more feasible. Furthermore, the topic model approach to HCoI's also naturally leads to community topic profiles, facilitating their interpretation.

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

Of course, the quality of the HCoI networks obtained through textual analyses depends on the representativeness of the working corpus. This is crucial for the conclusions that will be drawn from the analyses. Note that the methods are agnostic as to the type of texts contained in the corpus. In the present case study, full-text academic articles were used (and we showed how multiple authorship could be handled). Yet, the methods can target any types of texts, be they posts on social media, blogs, letters, reports, (including textual transcriptions of audio content), but also surveys, voting intentions (Pekar et al., 2022), or e-petitioning (Harrison et al., 2022). A possible limitation that may be raised is in terms of languages: the methods we have described work best with a mono-language corpus. For a multilingual corpus, a simple approach is to machine translate the non-English texts into English so as to produce a monolingual working corpus (English is taken here as an example and could be replaced by any other language accepted by the POS tagging package). This is a strategy that has been successfully tested elsewhere (Vries et al., 2018; Malaterre and Lareau, 2022).

Note also that the overall approach for identifying HCoI's is agnostic as to which topic model is chosen. Here, LDA was used but other topic modeling algorithms are possible, provided the resulting topics are not crisp-associated with documents but are assigned via continuous measures amenable to renormalization as probability distributions. Indeed, a crisp clustering of documents into topics would also result in a crisp clustering of actors into topics, preventing the use of distance measures to assess the relative proximity of actors with one another, and the ultimate representation of actors within a network. As we explained, one of the main stages of the methods is to calculate author topic profiles which then serve as a basis for identifying HCoI's. In turn, topic profiles can be calculated at the community level, proving a topic chart for each community.

While other social network approaches (e.g., bibliometric approaches in the case of scientific networks) typically rely on supplementary investigations about actor profiles to interpret and make sense of the observed networks, the methods we propose here make it possible to automatically pin down the specific identity of the identified communities in terms of discursive topics. Overall, this is crucial to understand what these communities talk about and how they are related to one another: their concerns, their claims, or broadly speaking their interests. Furthermore, the relatedness of communities with one another in terms of their interests can also be assessed by examining the topology of the HCoI's networks and by identifying genealogical relationships between communities over time. Such an approach is relevant to many policy-related questions at all stages of the policy lifecycle, wherever textual data are available and attributable to specific actors. HCoI mapping can notably lead to a better understanding of the social forces in presence and their claims on specific issues, thereby contributing to agenda setting, for instance by collecting indirect inputs from citizens through social media analysis (Belkahla Driss et al., 2019; Ronzhyn and Wimmer, 2021). This can be done to understand a broad range of interests, from privacy issues in government AI deployment (Saura et al., 2022) to the extraction of social network information in the case of anti-corruption policy (Diviák and Lord, 2023). The approach can in turn increase political legitimacy during policy implementation, as generally the case with data-driven, evidence-based decision-making (Starke and Lünich, 2020). HCoI mapping can also facilitate policy a posteriori evaluation, for instance, by aggregating perceived quality of citizen-centric public service in complement to other social media text analytics (Reddick et al., 2017).

5. Conclusion

Combining topic modeling and community detection methods makes it possible to uncover HCoI and map their proximity in terms of semantic content both synchronically (through correlation networks) and diachronically (over time periods). Using, as case study, a working corpus of 16,917 full-text academic articles written by 8,009 philosophers of science from the 1930s up to the 2010s, this approach revealed how these actors constituted well delineated HCoI's characterized by specific topic profiles, and how these HCoI's evolved over time. Being consistent with what is otherwise known by experts in the field, these results lend credibility to the approach. The results notably show how it is possible to gain insights into the social structures underlying sets of texts through the characterization of their topic profiles. The

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

approach provides insights into author-based communities, notably their semantic content in the form of directly interpretable topic profiles, but also their relative proximity and temporal evolution. When data about actual social interactions are not available but textual data are, mapping such HCoI networks can provide very relevant insights about groups of social actors sharing similar interests. In cases where both textual and social data are available, HCoIs analyses could also lead to a complementary content-based perspective compared to usual SNAs.

Data availability statement. Code and datasets available on Zenodo.org: https://doi.org/10.5281/zenodo.7967417.

Acknowledgments. The authors thank the audience of the 56th Hawaii International Conference on System Sciences for comments on an earlier version of this work (see Malaterre and Lareau, 2023).

Author contribution. Conceptualization: C.M., F.L.; Data curation: F.L.; Formal analysis and investigation: C.M., F.L.; Funding acquisition: C.M.; Investigation: C.M., F.L.; Methodology: C.M., F.L.; Project administration: C.M.; Resources: C.M.; Software: F.L.; Supervision: C.M.; Validation: C.M., F.L.; Visualization: C.M.; Writing—original draft preparation: C.M.; Writing—review and editing: C.M., F.L. Both authors approved the final submitted manuscript.

Funding statement. C.M. acknowledges funding from Canada Social Sciences and Humanities Research Council (Grant 430-2018-00899) and Canada Research Chairs (CRC-950-230795). F.L. acknowledges funding from the Fonds de recherche du Québec Société et culture (FRQSC-276470) and the Canada Research Chair in Philosophy of the Life Sciences at UQAM.
Competing interest. The authors declare none.

References

Barlow J, **Stephens PA**, **Bode M**, **Cadotte MW**, **Lucas K**, **Newton E**, **Nuñez MA and Pettorelli N** (2018) On the extinction of the single-authored paper: The causes and consequences of increasingly collaborative applied ecological research. *Journal of Applied Ecology 55*(1), 1–4. https://doi.org/10.1111/1365-2664.13040

Bastian M, **Heymann S and Jacomy M** (2009) Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*. San Jose, CA: AAAI.

Battaglini M and Patacchini E (2019) Social networks in policy making. *Annual Review of Economics 11*(1), 473–494. https://doi.org/10.1146/annurev-economics-080218-030419

Belkahla Driss O, **Mellouli S and Trabelsi Z** (2019) From citizens to government policy-makers: Social media data analysis. *Government Information Quarterly 36*(3), 560–570. https://doi.org/10.1016/j.giq.2019.05.002

Blei DM, **Ng AY and Jordan MI** (2003) Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Borgatti SP, **Everett MG and Johnson JC** (2013) *Analyzing Social Networks*. Los Angeles, CA: SAGE.

Boyack KW, **Klavans R and Börner K** (2005) Mapping the backbone of science. *Scientometrics 64*(3), 351–374. https://doi.org/10.1007/s11192-005-0255-6

Carley K (1993) Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology 23*: 75–126.

Castelblanco G, **Guevara J**, **Mesa H and Sanchez A** (2021) Semantic network analysis of literature on public-private partnerships. *Journal of Construction Engineering and Management 147*(5), 04021033. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002041

Christensen AP and Kenett YN (2023) Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological Methods 28*(4), 860–879. https://doi.org/10.1037/met0000463

Crane D (1969) Social structure in a Group of Scientists: A test of the "invisible college" hypothesis. *American Sociological Review 34*(3), 335. https://doi.org/10.2307/2092499

Danowski JA (1993) Network analysis of message content. In Richards WD and Barnett GA (eds), *Progress in Communication Sciences*, Vol. *12*. Norwood, NJ: Ablex Publishing Corporation, pp. 197–221.

Danowski JA (2011) Counterterrorism Mining for Individuals Semantically-Similar to Watchlist members. In Wiil UK (ed.), *Counterterrorism and Open Source Intelligence*. Vienna: Springer Vienna, pp. 223–247. https://doi.org/10.1007/978-3-7091-0388-3

Danowski JA and Cepela N (2010) Automatic mapping of social networks of actors from text corpora: Time series analysis. In Memon N, Xu JJ, Hicks DL and Chen H (eds), *Data Mining for Social Network Data*, Vol. *12*. Annals of Information Systems. Boston, MA: Springer US, pp. 31–46. https://doi.org/10.1007/978-1-4419-6287-4_3

Danowski JA, **Van Klyton A**, **Tavera-Mesías JF**, **Duque K**, **Radwan A and Rutabayiro-Ngoga S** (2023) Policy semantic networks associated with ICT utilization in Africa. *Social Network Analysis and Mining 13*(1), 73. https://doi.org/10.1007/s13278-023-01068-x

Diesner J and Carley KM (2004) *Using Network Text Analysis to Detect the Organizational Structure of Covert Networks.* Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference. Carnegie Mellon: NACCSOS.

Diviák T and Lord N (2023) From text to ties: Extraction of corruption network data from deferred prosecution agreements. *Data & Policy 5*, e4. https://doi.org/10.1017/dap.2022.41

Doerfel ML and Barnett GA (1999) A semantic network analysis of the international communication association. *Human Communication Research 25*(4), 589–603. https://doi.org/10.1111/j.1468-2958.1999.tb00463.x

Fortunato S, **Bergstrom CT**, **Börner K**, **Evans JA**, **Helbing D**, **Milojević S**, **Petersen AM**, **Radicchi F**, **Sinatra R**, **Uzzi B**, **Vespignani A**, **Waltman L**, **Wang D and Barabási A-L** (2018) Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Fowler JH (2006) Connecting the congress: A study of cosponsorship networks. *Political Analysis 14*(4), 456–487.

Giere RN and Richardson AW (1996) *Origins of Logical Empiricism*. v. *16*. Minneapolis: University of Minnesota Press.

Griffiths TL and Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences 101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*(3), 267–297.

Hafner-Burton EM, **Kahler M and Montgomery AH** (2009) Network analysis for international relations. *International Organization 63*(3), 559–592.

Harrison TM, **Dumas C**, **DePaula N**, **Fake T**, **May W**, **Atrey A**, **Lee J**, **Rishi L and Ravi SS** (2022) Exploring E-petitioning and media: The case of #BringBackOurGirls. *Government Information Quarterly 39*(1), 101569. https://doi.org/10.1016/j.giq.2021.101569

Howlett M, **Ramesh M and Perl A** (2020) *Studying Public Policy: Principles and Processes*, 4th Edn. Oxford, New York: Oxford University Press.

Huckfeldt R (2009) Interdependence, density dependence, and networks in politics. *American Politics Research 37*(5), 921–950. https://doi.org/10.1177/1532673X09337462

Knoke D (2015) Emerging trends in social network analysis of terrorism and counterterrorism. In Scott RA and Kosslyn SM (eds), *Emerging Trends in the Social and Behavioral Sciences*, 1st Edn. Hoboken, N.J.: Wiley, pp. 1–15. https://doi.org/10.1002/9781118900772.etrds0106

Ko N, **Jeong B**, **Choi S and Yoon J** (2018) Identifying product opportunities using social media mining: Application of topic modeling and chance discovery theory. *IEEE Access 6*, 1680–1693. https://doi.org/10.1109/ACCESS.2017.2780046

Kong X, **Shi Y**, **Yu S**, **Liu J and Xia F** (2019) Academic social networks: Modeling, analysis, mining and applications. *Journal of Network and Computer Applications 132*, 86–103. https://doi.org/10.1016/j.jnca.2019.01.029

Krebs VE (2002) Mapping networks of terrorist cells. *Connect 24*(3), 43–52.

Kuld L and O'Hagan J (2018) Rise of multi-authored papers in economics: Demise of the 'lone star' and why? *Scientometrics 114*(3), 1207–1225. https://doi.org/10.1007/s11192-017-2588-3

Kumar AA, **Steyvers M and Balota DA** (2022) A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science 14*(1), 54–77. https://doi.org/10.1111/tops.12548

Lazer D (2011) Networks in political science: Back to the future. *PS: Political Science & Politics 44*(1), 61–68. https://doi.org/10.1017/S1049096510001873

Luke DA and Harris JK (2007) Network analysis in public health: History, methods, and applications. *Annual Review of Public Health 28*(1), 69–93. https://doi.org/10.1146/annurev.publhealth.28.021406.144132

Malaterre C and Lareau F (2022) The early days of contemporary philosophy of science: Novel insights from machine translation and topic-modeling of non-parallel multilingual corpora. *Synthese 200*(3), 242. https://doi.org/10.1007/s11229-022-03722-x

Malaterre C and Lareau F (2023) Identifying hidden communities of interest with topic-based networks: A case study of the community of philosophers of science (1930-2017). In *Proceedings of the 56th Hawaii International Conference on System Sciences*. Honolulu, HI: HICSS pp. 2473–2482. Available at https://hdl.handle.net/10125/102936. (accessed 12 January 2023)

Marcus MP, **Marcinkiewicz MA and Santorini B** (1993) Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19*(2), 313–330. https://doi.org/10.21236/ADA273556

McCallum A, **Wang X and Corrada-Emmanuel A** (2007) Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research 30*, 249–272. https://doi.org/10.1613/jair.2229

Pathak N, **DeLong C and Banerjee A** (2008) Social Topic Models for Community Extraction. In *The 2nd SNA-KDD Workshop'08 (SNA-KDD'08), August 24, 2008, Las Vegas, Nevada, USA.*, *10*. New York, NY: ACM.

Pekar V, **Najafi H**, **Binner JM**, **Swanson R**, **Rickard C and Fry J** (2022) Voting intentions on social media and political opinion polls. *Government Information Quarterly 39*(4), 101658. https://doi.org/10.1016/j.giq.2021.101658

Praet S, **Martens D and Van Aelst P** (2021) Patterns of democracy? Social network analysis of parliamentary twitter networks in 12 countries. *Online Social Networks and Media 24*, 100154.

Raimbault B, **Cointet J-P and Joly P-B** (2016) Mapping the emergence of synthetic biology. *PLoS One 11*(9), e0161522. https://doi.org/10.1371/journal.pone.0161522

Reddick CG, **Chatfield AT and Ojo A** (2017) A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use. *Government Information Quarterly 34*(1), 110–125. https://doi.org/10.1016/j.giq.2016.11.001

Réale, D., **Khelfaoui, M.**, **Montiglio, P.-O.**, & **Gingras, Y.** (2020). Mapping the dynamics of research networks in ecology and evolution using co-citation analysis (1975–2014). *Scientometrics*, *122*(3), 1361–1385. https://doi.org/10.1007/s11192-019-03340-4

https://doi.org/10.1017/dap.2023.48 Published online by Cambridge University Press

Rehurek R and Sojka P (2010) Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50.

Ressler S (2006) Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs*, *2*(2). Article 8 (July 2006). https://www.hsaj.org/articles/171. (accessed 23 October 2023)

Röder M, **Both A and Hinneburg A** (2015) Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. New York, NY: ACM Press. pp. 399–408. https://doi.org/10.1145/2684822.2685324

Ronzhyn A and Wimmer MA (2021) Research directions in policy modeling: Insights from comparative analysis of recent projects. *Data & Policy 3*, e13. https://doi.org/10.1017/dap.2021.8

Rosenberg A and Arp R (eds) (2009) *Philosophy of Biology: An Anthology.* Malden, MA: Wiley-Blackwell.

Ruiz J, **Featherstone JD and Barnett GA** (2021) *Identifying Vaccine Hesitant Communities on Twitter and their Geolocations: A Network Approach.* Available at http://hdl.handle.net/10125/71096. (accessed 10 October 2023)

Saura JR, **Ribeiro-Soriano D and Palacios-Marqués D** (2022) Assessing behavioral data science privacy issues in government artificial intelligence deployment. *Government Information Quarterly 39*(4), 101679. https://doi.org/10.1016/j.giq.2022.101679

Schmid H (1994) Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester: ACL. pp. 44–49.

Segev E (2021) *Semantic Network Analysis in Social Sciences*, 1st Edn. London: Routledge. https://doi.org/10.4324/9781003120100

Shearer JC, **Dion M and Lavis JN** (2014) Exchanging and using research evidence in health policy networks: A statistical network analysis. *Implementation Science 9*(1), 126. https://doi.org/10.1186/s13012-014-0126-8

Siegel DA (2011) Social networks in comparative perspective. *PS: Political Science & Politics 44*(1), 51–54.

Siew CSQ, **Wulff DU**, **Beckage NM and Kenett YN** (2019) Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity 2019*, e2108423. https://doi.org/10.1155/2019/2108423

Small H (1999) Visualizing science by citation mapping. *Journal of the American Society for Information Science 50*(9), 799–813. https://doi.org/10.1002/(SICI)1097-4571(1999)50:9<799::AID-ASI9>3.0.CO;2-G

Sober E (ed.) (2006) *Conceptual Issues in Evolutionary Biology*, 3rd Edn. Cambridge, Mass: MIT Press.

Starke C and Lünich M (2020) Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy 2*, e16. https://doi.org/10.1017/dap.2020.19

Steyvers M, **Smyth P**, **Rosen-Zvi M and Griffiths T** (2004) Probabilistic author-topic models for information discovery. In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*. New York, NY: ACM Press. p. 306. https://doi.org/10.1145/1014052.1014087

Tang J, **Zhang J**, **Yao L**, **Li J**, **Zhang L and Su Z** (2008) ArnetMiner: Extraction and mining of academic social networks. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*. Las Vegas, Nevada, USA: ACM Press. p. 990. https://doi.org/10.1145/1401890.1402008

Varone F, **Ingold K**, **Jourdain C and Schneider V** (2017) Studying policy advocacy through social network analysis. *European Political Science 16*(3), 322–336. https://doi.org/10.1057/eps.2016.16

Victor JN, **Montgomery AH and Lubell M** (2017) *The Oxford Handbook of Political Networks*. Oxford: Oxford University Press.

Vries E d, **Schoonvelde M and Schumacher G** (2018) No longer lost in translation: Evidence that Google translate works for comparative bag-of-words text applications. *Political Analysis 26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Ward MD, **Stovel K and Sacks A** (2011) Network analysis and political science. *Annual Review of Political Science 14*(1), 245–264. https://doi.org/10.1146/annurev.polisci.12.040907.115949

Wasserman S and Faust K (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Yang L, **Cao X**, **Jin D**, **Wang X and Meng D** (2014) A unified semi-supervised community detection framework using latent space graph regularization. *IEEE Transactions on Cybernetics 45*(11), 2585–2598.

Yang S, **Keller F and Zheng L** (2016) *Social Network Analysis: Methods and Examples (1er édition)*. Los Angeles: Sage Publications.

Ye F, **Chen C and Zheng Z** (2018) Deep autoencoder-like nonnegative matrix factorization for community detection. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York, NY: ACM Press. pp. 1393–1402.

Zhang H, **Qiu B**, **Giles CL**, **Foley HC, and Yen J** (2007) An LDA-based community structure discovery approach for large-scale social networks. In *2007 IEEE Intelligence and Security Informatics*. New York, NY: IEEE. pp. 200–207.

Zhou D, **Ji X**, **Zha H and Giles CL** (2006) Topic evolution and social interactions: How authors effect research. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. New York, NY: ACM Press. pp. 248–257.

Cite this article: Malaterre C and Lareau F (2024). Inferring social networks from unstructured text data: A proof of concept detection of hidden communities of interest. *Data & Policy*, 6: e5. doi:10.1017/dap.2023.48