CONVERGENCE

Convergence: The International Journal of Research into New Media Technologies
2022, Vol. 28(4) 940–961

DOI: 10.1177/13548565221091809
journals.sagepub.com/home/con


Measuring the diffusion of conspiracy theories in digital information ecologies

Annett Heft and Kilian Buehling
Weizenbaum Institute for the Networked Society and the Free University of Berlin, Berlin, Germany

Abstract

Digital platforms and media are fertile breeding grounds for disinformation and conspirational views. They provide a variety of communication venues for a mixed set of actors and foster the diffusion of content between actor groups, across platforms and media, and across languages and geographical spaces. Understanding those diffusion processes requires approaches to measure the prevalence and spread of communicative acts within and across digital platforms. Given the increasing access to digital data, computational methods provide new possibilities to capture this spread and do justice to the interrelated nature and hybridity of online communication. Against this background, the paper focuses on the spread of conspiracy theories in digital information ecologies. It provides a review of recent methodological approaches to measuring conspiracy-related content online regarding the (a) *prevalence* and (b) *diffusion* of conspiracy theories. To that end, the paper differentiates between social network analysis approaches and computational techniques of automated text classification. It further discusses how far these and related computational approaches could facilitate studying the diffusion of conspiracy theories across different actor types, languages, topics and platforms. In doing so, it takes the specific nature of online communication and challenges in the field of conspiracy-related content into account.

Keywords

Diffusion, conspiracy theories, digital information ecologies, digital platforms, social network analysis, automated text classification, comparative research

Introduction

Digital platforms and media are fertile breeding grounds for disinformation and conspirational views (Bergmann, 2020; Lazer et al., 2018). These communication venues provide new

Corresponding author:
Annett Heft, Weizenbaum Institute for the Networked Society and the Free University of Berlin, Institute for Media and Communication Studies, Garystr 55, Berlin, 14195, Germany.
Email: annett.heft@fu-berlin.de

opportunities and lowered access barriers for all kinds of actors to become public, gain visibility and spread their messages (Benkler, 2006; Bruns, 2008). In particular, actors who challenge democratic procedures and established knowledge can communicate their messages and opinions unmediated, unbound by filtering gatekeeper structures and journalistic quality assurance (Bruns, 2008; Neuberger, 2009). Digital platforms, in their broadest sense, have been defined as 'digital infrastructures that enable two or more groups to interact' (Srnicek, 2017: 24). From a sociological and communicative perspective, platforms enable and structure 'the curation of social relations and social behavior' on the internet (Dolata, 2019: 195, translation by authors). These interaction spaces facilitate, for example, radical actors or specific groups such as vaccination opponents to connect with each other and foster and reinforce the visibility and spread of each other's messages. Digital technologies also allow users believing conspirational views 'to participate in crowd-sourcing 'evidence'' (Zeng and Schäfer, 2021: 4) for their narratives.

What's more, digital communication processes are not limited to single platforms and communication venues in digital space or local places. Instead, a wide range of different platforms and media with diverse technological affordances is available. They offer specific possibilities for different actor types with different styles and languages to establish connections, even across different geographical spaces. Platforms include social media, such as blogs, microblogs and social networking sites, as well as search engines and other aggregators. In addition to platforms, websites are central communication venues for traditional mass media and other actors. Altogether, these communication venues form digital information ecologies (Häussler, 2021) marked by hybridity of the fora and agents of communication, bringing actors from traditional communication roles together with a new diversity of speakers, commenters and automated acts of communication. The digital technologies of platforms and media enable the diffusion and 'spillover' of topics and narratives between different actor groups – between social media, semi-public communication spaces, hyper-partisan alternative media outlets and traditional journalistic mass media (Pfetsch et al., 2013). The 'success' of such communication in terms of salience and impact is also determined by the diffusion processes from niche to broader publics. Understanding those diffusion processes requires approaches to measuring the prevalence and spread of communicative acts within and across platforms in mutually interconnected digital information ecologies.

Taking this ecology perspective as our starting point, we focus our review and discussion on how to measure diffusion within such an ecology to the question of the prevalence and diffusion of conspiracy theories (CTs). A CT, according to Keeley (1999: 116), is 'a proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons—the conspirators—acting in secret'. Definitions of CTs regularly refer to intentionalism (everything has been planned), the dualism between a small group of evil conspirators and innocent victims, and the secrecy in which connected actions and processes take place as central characteristics (Baden and Sharon, 2021; Barkun, 2013: 3–4; Butter and Knight, 2020a: 1; Mahl et al., 2022). The term CT is sometimes used interchangeably with the term conspiracy narrative. Yet, the concept of conspiracy narrative focuses more on the narrative plot, narrative style and structure of any conspiracy, which juxtaposes an official narrative with a secret one (Seidler, 2016: 40–41).

In addition, research on CTs is often closely interlinked with research on disinformation. This frequently leads to blurred conceptual boundaries (Mahl et al., 2022). Disinformation describes the act of intentionally spreading false or inaccurate content for strategic purposes (Fallis, 2015; Zimmermann and Kohring, 2018). CTs, in contrast, describe a specific type of content that can comprise true and untrue allegations that are difficult to differentiate from reasonable theorizing about suspected conspiracies (Baden and Sharon, 2021). As Butter and Knight (2020a) highlight, the content is believed by its disseminators in many circumstances.

Given the conceptual complexity and narrative style of CTs, they are difficult to operationalize empirically. Many studies use discourse-analytical approaches and manual content analyses to realize in-depth text-based investigations of CTs (for recent overviews on the research field, see Butter and Knight, 2020b; Mahl et al., 2022). Instead of scrutinizing methods used for capturing CTs overall, we focus on the spread of CTs in the hybrid environment of digital online communication. Given the increasing access to digital data, computational methods provide new possibilities to capture this spread in more encompassing ways and do justice to the interrelated nature and hybridity of online communication. Overall, the procedures of automated approaches have become increasingly precise. Yet, those methods have their own prerequisites and trade-offs, and tracing the spread of CTs across multiple platforms, publics and transnational communication spaces remains particularly challenging methodologically (Giglietto et al., 2020). This 'cross'-perspective, however, is at the center of our interest.

Against this background, the aim of our paper is twofold. First, we provide a review of recent methodological approaches to measuring conspiracy-related content online regarding the (a) *prevalence* and (b) *diffusion* of CTs. To that end, we systematize how the diffusion of CTs can be conceptualized based on literature focused on CTs and related concepts and which methods have been used in this respect in communication-related research. Second, we discuss how far approaches used in CT-related studies and additional computational methods not yet applied to conspiracy-related content could facilitate studying the diffusion of CTs across different actor types, languages, topics, and platforms. We differentiate between social network analysis approaches and computational techniques of automated text classification and discuss the potential and limitations of these methods with respect to the hybrid and interrelated nature of online communication and the specific challenges of conspiracy-related content. As a result, we hope to contribute our identification and systematization of trade-offs to the discussion of methods, which could provide orientation for further studies.

The systematization and discussion provided in this paper are based on two literature reviews[1] examining recent research from two perspectives: First, we searched the EBSCO Communication Source database and communication-related fields in the Web of Science (SSCI) database for studies with a) CTs as the main subject, combined with b) a focus on digital media, social media or online communication. The abstracts of all articles found with this approach have been checked for whether they deal with our approaches to diffusion as conceptualized below. If so, they were considered in detail. Our second search was not confined to conspiracy-related studies but assessed studies dealing with (a) diffusion in particular and (b) text classification, topic modeling, or network analysis methods in (c) the context of digital media, social media, or online communication. From this second set of literature, we reviewed studies addressing cross-platform and cross-language applications of those methods. We acknowledge that our approach focuses on communication-related research, including disciplines such as communication, sociology, political science, psychology or linguistics, and will thus not cover the full multidisciplinarity of the field (Mahl et al., 2022).[2] In particular, recent developments in computer science (e.g. details on supervised learning) are beyond the scope of this paper. The combination of EBSCO's field-specific Communication Source database and the multidisciplinary Web of Science should, however, provide an extensive article coverage even though the possibility of missing journals that might be covered in other databases exclusively remains. We synthesized the literature in textual and graphical form with respect to the cross-actor, cross-language, cross-topic, and cross-platform nature of diffusion processes addressed by our second research question.

Analyzing the prevalence and diffusion of conspiracy theories online – overview and conceptualization

When approaching the prevalence and diffusion of CTs in digital texts, two questions must be answered in the first place, namely how to actually 'find' a CT and how to conceptualize diffusion in digital communication.

Detecting conspiracy theories and conceptualizing prevalence

With respect to the first question, the literature suggests two main approaches that both begin with known conspiracy topics or conspirational actors identified a priori instead of detecting unknown actors or CTs. In the first, *actor-based* approach, known actor or account characteristics serve as the starting point to identify conspiracy-related communication, using accounts and media with specific characteristics regarding trustworthiness, conspicuous features and prior behavior. For example, as starting points of their design and data collection, scholars use actors and sites that have been reported on blacklists as problematic for disinformation and conspirational content or conspiracy-related accounts, groups or sites that have been classified in other research (Boberg et al., 2020; Giglietto et al., 2020), sometimes even equating the actors with the content they produce and circulate. Thus, conspiracy-related character is the gateway to further analysis and not its result, which sometimes introduces conceptual blurriness and a mixing of disinformation and conspirational content (ISD Digital Research Unit, 2020a; 2020b). Other examples include analyzing specific r/conspiracy-focused subreddits on Reddit (Klein et al., 2019; Samory and Mitra, 2018a) or particular threads on the Stormfront website (Wilson, 2017) or 8kun (Zeng and Schäfer, 2021).

The second and more prevalent *topic-based* approach is that studies focus on one or more known CTs and use (lists of) case-specific key terms and hashtags (e.g. #5GCoronavirus, #Pizzagate) to capture related communication (Ahmed et al., 2020; Graham et al., 2020; Leal, 2020; Wood, 2018). Such approaches are viable if the overall CT is known before the analysis. They can be applied in cross-topical and cross-language studies, as discussed in more detail below. However, each communicative venue has its own characteristics, with platforms generally more open to dictionary approaches while websites and online fora require additional prior steps of corpus generation. Recent studies show promising combinations of the actor- and keyword-based approaches (Garry et al., 2021; Mahl et al., 2021; Zeng and Schäfer, 2021). These approaches allow for capturing the prevalence – that is the absolute presence and importance of conspiracy-related content – in a defined (static) time span and, subject to a comparative study design, an assessment of their relative salience (degree of their relative importance).

While both actor- and keyword-based approaches are limited to the study of pre-defined actors and topics, a third strand of research aims to manually or automatically detect CTs without prior actor- or case-specific knowledge. We expand on such approaches below.

Conceptualizing diffusion

Concerning the second question, the conceptualization of diffusion, we follow literature that understands diffusion as 'a social (communication) process through which new ideas, technologies, products, or processes spread among the members of a particular social system via specific communication channels over time' (Kreps, 2017: 1; see also Rogers, 1983: 5). Regarding CTs and digital communication, we deal with idea and information diffusion and differentiate between two concepts and operationalizations: reference-based diffusion and content-based diffusion.

Reference-based diffusion. Reference-based diffusion understands diffusion as the spread of the exact same item or material online, which is enabled by technical affordances. The reconstruction of such diffusion processes is based on the references, that is, identifiers such as hyperlinks, which establish a direct relationship between items. Their analysis can be static, using aggregate analysis and one-time snapshots not actually capturing the *diffusion process* as such, but actors who have agency within those processes. Studies measuring *diffusion dynamics* focus on the diffusion process through time-series analysis or the measurement of cascade dynamics.

Content-based diffusion. Another approach is, how Buhl et al. (2018) call it, a content-based approach. Here, the crucial criterion is that 'the same event' is the subject matter (Buhl et al., 2018: 85), without the necessity of direct links. Diffusion is understood as the result of independent decisions by agents observed in retrospect and interpreted as the result of an expected diffusion process. Grounded on news event diffusion research (Rogers, 2000), the authors argue that 'the reconstruction of news diffusion processes in general and within the online news ecosystem in particular does not aim at direct relationships among the population under study, which may be inscribed into text or software traces, but more generally at process patterns, which emerge from the timing of both dependent and independent adoption decisions by individual elements […]' (Buhl et al., 2018: 85). *Similarity* in the prevalence and content patterns of a CT is thus defined as an indicator and snapshot of the underlying dynamic diffusion process. Such similarity measurement mandatorily requires comparative designs that – across actors and platforms and both the static (on time-point) and dynamic (multiple time-points) – capture the prevalence and spread of certain events.

In a similar way, other research differentiates between *explicit* connections (such as hyperlinks) and *implicit* links between content (understanding shared quotes or similar keywords as shared content) (Kim et al., 2013). Taking this differentiation as part of our analytical framework, we discuss the extent to which computational methods can support research on the diffusion of CTs in hybrid information ecologies below. Regarding content-based diffusion, we first discuss several content classification approaches that have been used in the research field and general studies of communication science. Concerning reference-based diffusion, we then discuss social network analysis approaches employed within and beyond the research on CTs.

Text classification approaches to content-based diffusion

The content-based approach conceptualizes content-related similarity as an outcome of a diffusion process and measures and compares patterns in the emergence of content-related features. Two central questions have to be addressed in this respect: (1) What is the main research question and thus starting point of a study, and which conceptualization (and measurement) serves which aims? (2) Which dimensions should be used to conceptualize and measure similarity, and how well are these measurements suited to capture CTs in a comparative perspective (see *Introduction*)? On the conceptual level, CTs can be specified as a *topic* or issue as such. Measurements include a simple keyword approach or various topic modeling procedures if operationalized as topics and subtopics. Yet, each CT entails a particular narrative plot, a constellation of actors and their arguments that ties in with more specific concepts, such as *narratives* and *frames.* On the design level, the question is whether the main aim is to analyze one specific conspiracy based on an existing text corpus that the researcher wants to sub-classify in more specific units (case-specific corpora), or whether the task is to detect multiple CTs in a more general corpus. We have already seen that many conspiracy-related studies follow the former approach and discuss the related methods first. Concerning the task of finding (different) CTs in non-case-specific corpora, we refer to classification approaches based on

linguistics at the end of this section. We also discuss the static versus dynamic perspectives of these approaches.

Classifying similarity based on topics

Dictionary-based approach. One frequently used approach in the research field is using (lists of) case-specific key terms such as 'Gates' (Gerts et al., 2021) or 'coronavirus' (Shahsavari et al., 2020), or specific hashtags such as #Pizzagate (Leal, 2020) or #Chemtrails (Mahl et al., 2021) to gather material on the (potential) prevalence of a particular CT. After steps of data cleaning and aggregation, such material can provide an instant snapshot on the overall prevalence of a CT as a topic.

In general, such a dictionary approach can be applied to all searchable text corpora. Dictionaries can be created for each CT of interest, although the difficulty of finding terms that unambiguously represent a CT varies from one case to another (Mahl et al., 2021). The approach is, in general, adaptable to different platforms, communication venues, and actor groups. Yet, each digital environment will require specific adaptations (for instance, with respect to Boolean operators) and not all identifiers are applicable across platforms (e.g. hashtags). In addition, such dictionaries can be developed across languages. Given that searchable corpora are available or can be created, the dictionary approach is a viable option for a priori known CTs, but clearly limited to them. For detecting CTs without already knowing the narrative, scholars experiment with automated text classification techniques that rely on linguistic features (see *Using semantics to detect conspiracy theories*), but likewise resort to manual classification (Baden and Sharon, 2021) due to the complexity of the task.

With respect to the theoretical construct CT, the approach is admittedly broad and unprecise. The simple dictionary approach neither differentiates between the actors involved in CTs nor the detailed narrative patterns. Thus, many studies treat this step as a part of the data gathering procedure rather than data analysis and combine it with subsequent analysis procedures.

Topic modeling. Another frequently used approach for both further classifying case-specific corpora of conspiracy-related content and non-case-specific classifications of CTs within broader text corpora is topic modeling (Boberg et al., 2020; Gerts et al., 2021; Sha et al., 2020; Smith and Graham, 2019). Topic modeling is an inductive method used for exploring, categorizing, and comparing the content composition of large corpora of digital texts that relies on a text-mining algorithm to discover latent topics based on the bag-of-words approach. Estimates about the topical content of documents are made by identifying frequently co-occurring terms (Roberts et al., 2016). Topic models can be applied to case-specific, pre-classified corpora (such as those specified by keyword approaches) or non-case-specific text corpora (for a general review of the concepts and applications of automated content analysis see Grimmer and Stewart, 2013). For example, Gerts et al. (2021) apply topic models on pre-classified data to identify 'subtopics' and their evolution over time. In their study, the 'Gates theory' used to describe the CT that Bill and Melinda Gates have funded, patented or otherwise economically benefited from COVID-19 was sub-classified into two topics. One focused more specifically on the relationship of the coronavirus outbreak and the pandemic with the Gates Foundation, and the other combined several CTs about Bill Gates, COVID-19, vaccines, Soros, etc. (Gerts et al., 2021). Faddoul et al. similarly use topic modeling to discern the main topics from a broader corpus of conspiracy-related content (not pre-classified into one specific CT, but where content has been classified as conspirational a priori) and analyze their salience over time (Faddoul et al., 2020). They identify extremely broad topics, namely 'alternative science and history', 'prophecies and online cults' and 'political conspiracies and QAnon' (Faddoul

et al., 2020). Used to sub-classify the content of six purposively selected public anti-vaccination Facebook pages (with potentially conspirational content), Smith and Graham (2019) utilize latent Dirichlet allocation (LDA) topic modeling for their text classification. The study results in topics that discern specific CTs (e.g. 'Zika Virus and Gates Foundation' or 'Chemtrails and Agriscience'), but likewise classifies texts referring to 'Media, censorship, and 'cover up'' and even broader areas, such as 'Activism' (Smith and Graham, 2019: 1322).

Topic models can also be used for measuring the dynamic development within CTs over time. For example, a study by Gerts et al. suggests dynamic topic modeling as a method that allows for analyzing temporal changes in topics (more specifically, changes in within-topic word importance) over time (Gerts et al., 2021). The authors highlight that this method allowed them to 'identify overlaps between theories' (Gerts et al., 2021: 12) based on the relevance of words distinguishing one topic from another. Another approach that allows for examining how topic proportions change over time and how topic proportions differ across different actors and platforms is structural topic modeling (STM). STM allows for using metadata (such as timestamps or actor characteristics) as covariates in the model (Benoit et al., 2018; Roberts et al., 2016).

These examples highlight the potential and challenges of topic modeling with respect to capturing CTs in particular, the method in general, and its application across different corpora and platforms. In topic modeling, the meaning of 'topic' is assessed empirically. What 'topic' means is neither conceptualized theoretically nor clearly linked to a theoretical conceptualization (Maier et al., 2018a: 95). Scholars argue the empirically resulting topics can theoretically be interpreted as issues (Maier et al., 2018b: 6), understood as contentious or controversial topics (Dearing and Rogers, 1996). Some even interpret the results of topic models through the lens of frames, understood as interpretative frameworks that define problems, diagnose causes, make moral judgments, and suggest remedies (Entman, 1993). Yet, this interpretation is more contested (Walter and Ophir, 2019). As the examples show, the results of conspiracy-focused studies identified types or *thematic areas* of CTs in Faddoul et al. (2020) but no *specific* CTs, while the research by Gerts et al. identified subtopics within one CT (Gerts et al., 2021: 5). In accordance with discussion concerning the method in general, interpreting the theoretical meaning of topic modeling results is highly dependent on topic characteristics, context, and topical granularity, so that 'the burden of making sense of the results is still on the researcher (Jacobi et al., 2016: 103). While the statistical granularity of a topic model (e.g. number of topics) can be controlled via hyperparameters, the interpretive granularity depends on the topic itself. The results of topic models will likely represent different levels of interpretive granularity, even within the same model (see the discussion in Shadrova, 2021). The actual conceptual distinction between topics and sub-topics, and thus the question of granularity, cannot be easily solved by the method. Comparative research based on sub-corpora modeled with individual topic models poses significant challenges for the comparison of the model outputs. Since each model is fitted inductively to the material, models may necessitate different numbers of topics to reach optimal statistical distinctiveness and thus different levels of statistical granularity. The different interpretive granularities of topics further hinder direct comparability across cases, platforms and media. That is, the topics in the sub-corpora may represent conceptually distinct constructs (for a broader discussion, see Lind et al., 2021). On the other hand, combining corpora from different platforms poses challenges concerning different text styles, formats, etc. These characteristics need to be considered regarding their influence on results.

Conspiracy-related research on transnational and global diffusion can consider approaches allowing for cross-lingual classifications of conspirational content. The problem of cross-lingual text classification is intensively addressed in current research (Chan et al., 2020; Lind et al., 2019, 2021; Reber, 2019). Scholars provide guidance for constructing multilingual term lists for a

keyword-driven approach to collecting conspirational content across countries and for multilingual dictionary construction (Maier et al., 2021). For topic modeling, machine translation is one of the strategies to consolidate data across languages (for an overview, see Chan et al., 2020). Yet, as Chan et al. (2020) argue, this method is not reproducible over time because the underlying algorithms deployed by commercial machine translation providers might be subject to improvements. As such, the authors propose a technique that enables reproducible cross-lingual topic extraction across time by using contextualized word embeddings. One additional advantage highlighted by the authors is that systematic language differences are filtered out and not clustered in specific topics, which has been a common problem of comparative research. Approaches like the ones discussed in this section could foster the cross-lingual classification of content in research on conspiracy-related texts.

Classifying similarity based on narratives

The issue remains that the bag-of-words approach underlying topic modeling does not accurately represent the interrelated narrative construct of a CT. CTs are conceptualized as narratives, and an approach specifically designed to capture the key narrative pattern of CTs has been proposed by Samory and Mitra (2018b). First, the authors extracted topics from a dataset comprising all submissions and comments in the r/conspiracy subreddit on Reddit (from 2008 to 2017) with a specific topic modeling procedure (Samory and Mitra, 2018b: 6–8). Second, the conspiratorial agents, the actions they perform, and their targets (aims) are understood as key conceptual elements of a CT and computationally detected as 'agent-action-target triplets in conspiratorial statements'. Those triplets constitute 'narrative-motifs' (Samory and Mitra, 2018b: 1), defined as 'recurring patterns of conspiratorial agents, actions, and targets' (p. 2). Empirically, the agent-action-target-construct is operationalized as subject-verb-object triplets and measured via word embeddings and clustering procedures. Based on the overarching narrative-motif for triplets within a cluster (Samory and Mitra, 2018b: 6–10), the authors report the salience of CTs, the narrative construction or framing within CTs, and their overlap.

Another approach to capturing the narratives underlying CTs has been developed by Shahsavari et al. (2020). The authors apply machine-learning methods to two corpora of social media posts and journalistic news reports to automatically detect COVID-19 conspiracies and analyze the interplay and flow of conspiracy narratives. Based on the sentences in the corpus, syntax relationships (e.g. between nouns and verbs or subjects and verbs) are extracted using natural language processing and aggregated into contextual groups by utilizing the contextual word embeddings provided by Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2019).

These conspiracy-focused research examples are comparable with methods-related research that aims to provide a better computational operationalization of the framing concept, which shares some similarities with the concept of conspiracy narratives. One such approach has been proposed by Walter and Ophir (2019), who suggest operationalizing frames as communities in a network of topics. The so-called Analysis of Topic Model Networks process combines topic modeling, network analysis, and community detection. Topics resulting from the LDA procedure are interpreted as frame elements, which are mapped as networks based on the relationship between the frame elements and grouped through community detection into dense clusters interpreted as 'frame packages' (Walter and Ophir, 2019: 248). The experiences from such approaches could be valuable for further developing text classifications in the context of CTs.

All these approaches can be applied to static and dynamic study designs and are suited for comparative analyses. They can be used for cross-conspiracy comparative research and enable a more fine-grained operationalization of CTs. Like the classification of the topic, they are primarily

concerned with the conspirational content and less with the actors (sources or speakers) who push a particular narrative. The higher precision with respect to capturing the construct of CTs comes with more complex analytical procedures, which will likely further complicate cross-platform and -language comparative studies. The question of combining and comparing material from different text sources and the task of cross-language comparative analyses on the semantic layer remain particularly challenging.

Using semantics to detect conspiracy theories

Regarding how to detect a CT in the first place, researchers have started to look at the semantic characteristics of CTs and use them for automated recognition. Studies have shown that people who express conspiratorial beliefs are more likely to use terms that can be related to, for example, defiance and distrust or a conspirational worldview (Klein et al., 2019). A study by Gerts et al. (2021) showed that tweets containing COVID-19 conspiracies rated higher on negative, distinct emotions such as fear, anger and disgust than tweets on other topics. These computational approaches build on ideas initially developed in manual mixed methods analyses that showed rumor-related content contained specific phrases and linguistic patterns linked to expressions of uncertainty (Starbird et al., 2016). To conceptualize expressed uncertainty, Starbird and colleagues differentiate between 'shielding' and 'milling' expressions (2016: 362). The first type attributes responsibility for information to an external source and questions its accuracy with expressions such as 'possibly' or 'unconfirmed'. The second type represents uncertainty expressed through speculative questions and statements of incredulity, hope, or fear. Yet, such constructs are only partially represented by individual terms or sentiments and difficult to isolate with automated procedures. Further, 'off-the-shelf' dictionaries for sentiment classification are highly context specific and require revalidation (Chan et al., 2021). While linguistic and semantic features provide a starting point to search for CTs in large datasets and across platforms without a priori knowledge of a specific case, improvement of computational approaches is needed to advance their performance and in particular to narrow the gap between constructs and their empirical representations.

Social network analysis approaches to reference-based diffusion

Hyperlinks are a 'technological capability that enables one specific website (or webpage) to link with another' (Park, 2003: 49), providing for a basic structural element of the internet. By using hyperlinks, actors can exchange information – establishing diffusion roads through digital communication networks. Social media platforms and online communication more broadly provide various technological affordances for referencing actors and material through hyperlinking – integrated through share, forward, and retweet applications, etc., or embedded in message texts. Reference networks established in this way provide researchers with various possibilities for measuring the diffusion of CTs using network analysis procedures.

When researching the spread of CTs in online networks based on shared references, the approaches chosen by scholars can be roughly classified into three categories. First is a cross-section or one-point-in-time approach analyzing a snapshot of social media network interactions concerning one or more CTs. This approach is often used to infer the influence of specific actors (e.g. social bots or influential users) on diffusion dynamics. A second approach is using several network snapshots to gain knowledge about network and link sharing behavior over time. Another approach to actor-sensitive analysis of reference-based conspiracy spread is the modeling of the diffusion process

itself. The main focus of this third approach is to analyze factors influencing the cascading reference sharing behavior associated with viral social network posts.

Static social network analysis providing diffusion snapshots

The one-point-in-time approach and time-dependent network analyses are mostly executed by defining one or more pre-defined starting points and collecting all references connected to them. These starting points can be specific to actors' websites (Garry et al., 2021; Giglietto et al., 2020) or news articles (Shao et al., 2017) that are assumed to spread disinformation or CTs. To generate dictionaries of such actors and websites, authors use fact-checking platforms (e.g. snopes.com, politifact.com, factcheck.org) that flag online articles and their providers if the content is questionable or they show questionable editorial behavior associated with journalistic misconduct. Other starting points are hashtags linked to agreed-upon CT content (Gruzd and Mai, 2020) or co-occuring keywords (Graham et al., 2020). Usually, the hyperlinks referring to those starting points form the reference-based corpus of the following network analyses.

The aim of such analyses is to measure the influence of real and (semi-)automated actors or 'inauthentic coordinated behavior' (a term introduced by Facebook that tries to incorporate all non-organic individual behavior on social platforms (Gleicher, 2018; Graham et al., 2020)) on the spread of CTs or false information. Coordinated inauthentic behavior in research on disinformation spread is determined via several stylized facts about between-actor behavior within network clusters. These include the repeated co-occurrence of tweet sequences, co-sharing of images, and co-sharing of Twitter handles (Pacheco et al., 2021), as well as the near-simultaneous sharing of hyperlinks and retweets (Giglietto et al., 2020; Graham et al., 2020; Pacheco et al., 2021) or coordinated fast deletion of tweets (Elmas et al., 2019). Those behaviors are complemented with scores determined by bot-like behavior of individual accounts and provided by tools such as Botometer (Yang et al., 2020) or BotSlayer (Hui et al., 2019). Bots and other coordinated networks, again, are theorized to be employed for different purposes, ranging from the spread of CTs and political disinformation (Ferrara, 2020; Giglietto et al., 2020; Graham et al., 2020; Shao et al., 2017) or simulating mass support for a political matter (Elmas et al., 2019), to commercial interests (Elmas et al., 2019; Graham et al., 2020; Pacheco et al., 2021) or news bots (Ferrara, 2020; Giglietto et al., 2020). Meanwhile, the choice of indicators used for detecting clusters of coordinated inauthentic behavior strongly impacts the results of such analyses. For example, Pacheco et al. (2020) utilize a different measure for inauthentic user behavior and come to diametrically different results than Starbird and Wilson (2020) after investigating the same issue, a challenge also investigated by Rauchfleisch and Kaiser (2020a).

Overall, the main aim of such analyses is not mapping the diffusion process itself but the characterization of actors involved in spreading CTs or disinformation. The actor and reference-driven approach to CT diffusion in social networks tries first to identify whether a CT is promoted by references included in a publicly shared message. Hyperlinks are the foundation of the reference-based approach, even though they differ in scope and function on different platforms. One has to distinguish between platform-internal and platform-external hyperlinks. Platform-internal references are exemplified by public replies, comments, embeddings, or forwarded (or retweeted) messages within a platform. Platform-external hyperlinks include the embedding of sources from outside the specific platform via URLs or other interfaces. Using hyperlinks pointing to news websites or blogs, regardless of their journalistic reputation or dubiousness, is a robust approach because the inclusion of hyperlinks is usable on nearly all social media platforms. References to parts of the internet that are not under the realm of social media platforms qualify as cross-media

analysis. A comparative study of the use of such hyperlinks across social media platforms would have to consider platform-specific affordances, for example, the general difficulty for regular users to share or embed hyperlinks on Instagram and Snapchat; the limited, different use of hashtags on Facebook and Telegram; or functional distinctions in message replies on Twitter, YouTube, and Gab. Hyperlinks to news sites and blogs can be used as a unique identifier of whether a CT is promoted if the referenced webpage is flagged as promoting a CT. This is already done by numerous fact-checking websites that label wrongful information. Research conducted on CTs should qualitatively assess the contents of the blacklists produced by fact-checking services. Furthermore, fact-checking sites are mostly language- or country-specific, which is why cross-country comparisons will have to consider different fact-checking providers or rely on the limited diffusion of foreign-language content in a given country. The analysis of platform-to-platform hyperlinks, as is used in Wilson and Starbird (2020) and Ahmed et al. (2020) to identify which clusters of users leverage YouTube as a source to gain credibility on Twitter, are a feasible way to arrive at actor-based cross-platform analyses for platforms where individual posts can be referenced via hyperlinks (excluding, for example, Telegram).

Another starting point is a combination of hashtags that refer to specific CTs and the collection of all related references. This approach is not limited to single CTs if a dictionary of unambiguous hashtags for different theories can be established, which also allows for analysis of their salience and interactions of CT spreaders (Mahl et al., 2021). The use of hashtags, again, differs from one social media platform to another (Highfield, 2018; Potnis and Tahamtan, 2021) and is limited to social media platforms (thus excluding legacy media websites and blogs), adding complexity to their comparison. Cross-country comparisons can be conducted by including different country-specific hashtags or keywords in the dictionary.

The use of classifiers differentiating between human and (semi-)automated actors can shed light on the actor composition, and possibly sharing intentions, that lie behind salient CTs in social networks. Staying with (semi-)automated actors, a cross-country analysis within single platforms is necessary as bot clusters from different regions in the world seem to be active in major hashtag trends (Graham et al., 2020; Gruzd and Mai, 2020).

Tracing diffusion via dynamic social network analysis

Another approach for the in-depth study of CT diffusion via network actors is the analysis of social network evolution derived from a time series of social network snapshots, aided by qualitative analysis of influential actors and processes. With the prerequisite of a clearly identifiable starting point for the CT or rumor, such as the Pizzagate conspiracy (Leal, 2020), the car attack at the 'Unite The Right' rally at Charlottesville, Virginia being a false flag action (Krafft and Donovan, 2020), and the connection between COVID-19 and 5G towers (Bruns et al., 2020), the role of actors and network dynamics in CT diffusion can be mapped out thoroughly.

Using Facebook data, Bruns et al. (2020) analyzed the salience of the 5G theory over time and space in terms of country-specific pages and accounts propagating the theory. An additional qualitative evaluation of the data allowed for identifying different phases of the propagation of the CT and related key actors, such as local news or influential celebrity accounts. Krafft and Donovan (2020) and Leal (2020) apply a social network evolution approach to case studies of a completed CT lifecycle, that is a theory tracked from its origin to its definite rebuttal. By analyzing via trace-ethnography the role of specific network actors, as well as the importance of different hyperlink references to other social media platforms or legacy media websites to strengthen the conspiracy

argument (Leal, 2020) and gain credibility, a deep understanding of the importance of the network structure and function of cross-network entanglements can be extracted.

Although it is shown that this approach is useful for both cross-country and cross-network analysis, the resource intensity of the qualitative assessment requires a reduction in the complexity of the data. This is achieved by limiting the analysis to a clearly confined case or reducing the number of social networks analyzed. The insights gained with this qualitative and ethnographic network evolution approach, on the other hand, are very instructive for our understanding of diffusion processes on its own. In addition, the actor- and topic-related insights can foster subsequent large-scale quantitative studies just as much as follow-up qualitative analyses such as those by Bruns et al. (2020) contribute to a deeper understanding of automatically classified data. To the best of our knowledge, such actor-based, large-scale studies of network evolution and link formation in CT spread, either within or between networks, utilizing stochastic actor-oriented models (Snijders, 2001) or temporal exponential random graph models (Leifeld et al., 2018), have yet to be conducted.

Measuring diffusion processes as cascades

To analyze diffusion dynamics, another approach focuses on cascades inferred by single CT-related references within networks. Vosoughi et al. (2018) define cascades as unbroken retweet chains with a single identifiable origin. Cascades differ, for example, according to their depth (i.e. the number of intermediating users that contribute to the indirect spread of a post) or size (i.e. the total number of users involved in a cascade). Bessi et al. (2015) and Del Vicario et al. (2016) investigate the diffusion and interaction patterns of conspirational and scientific posts on Twitter, relating the different cascade patterns to the respective topic and the isolation of the actors' sub-clusters. On the basis of actor and cascade characteristics, Vosoughi et al. (2018) estimate that false news spreads faster than accurate news, irrespective of bot involvement. Friggeri et al. (2014) analyze visual misinformation and rumor content, showing a larger cascade depth for misinformation.

Cascades of social network posts can also be inferred by self-exciting point process models or Hawkes processes (Kim et al., 2020; Kobayashi and Lambiotte, 2016; Zhao et al., 2015). While better-known point process models, such as the Poisson process, model inter-event arrival times according to a common distribution and are, therefore, 'memoryless' (Rizoiu et al., 2017: 4), self-exciting point processes are able to capture whether the arrival of one event increases the probability of the next event. Using maximum likelihood estimation on real-world cascade data of retweets (Murayama et al., 2021; Rizoiu et al., 2017) or social media posts containing references flagged as disinformation or CTs (Papakyriakopoulos et al., 2020; Zannettou et al., 2017) enables researchers to predict cascades of CT posts.

Taking other actor characteristics into account, such as the follower count of a tweeting user on Twitter (Kong et al., 2021), allows for cross-actor analysis, while cross-topic analysis depends on the construction of a sample that includes cascades concerned with different topics, and is therefore the most useful approach for comparative analysis. Since cascades of retweets are independent of the language, cross-language analysis is not affected by this model, while a comparative analysis of cascade dynamics in a dataset comprised of cascades in different languages remains a possibility. Cross-platform and -media analyses are possible via multivariate Hawkes processes (Papakyriakopoulos et al., 2020; Zannettou et al., 2017) that model the increased probability of a post on a certain platform, depending on a changing amount of posts on a different platform. Paudel et al. (2021) show a possible framework of using Hawkes process modeling to bridge the gap between automated content analysis approaches and social network analysis approaches. Their

model measures the cross-platform diffusion of CTs based on the automated topic classification of social media posts and their time-dependent appearance on different platforms.

Discussion and conclusion

Taking the hybrid and interconnected nature of digital online communication seriously, our paper was interested in the extent to which recent computational methods provide possibilities to capture the prevalence and diffusion of CTs online and how far these approaches could facilitate studying these phenomena across different actor types, languages, topics, and platforms. In the following, we systematize our discussion of the potential and limitations of these methods with respect to the 'cross'-nature of online communication (see Figure 1). This can only be a tentative assessment of the comparative strengths and weaknesses of the methods, leaving many specific application questions aside. After that, we take another step back and reflect on more general trade-offs of the approaches with respect to the specific challenges of conspiracy-related content.

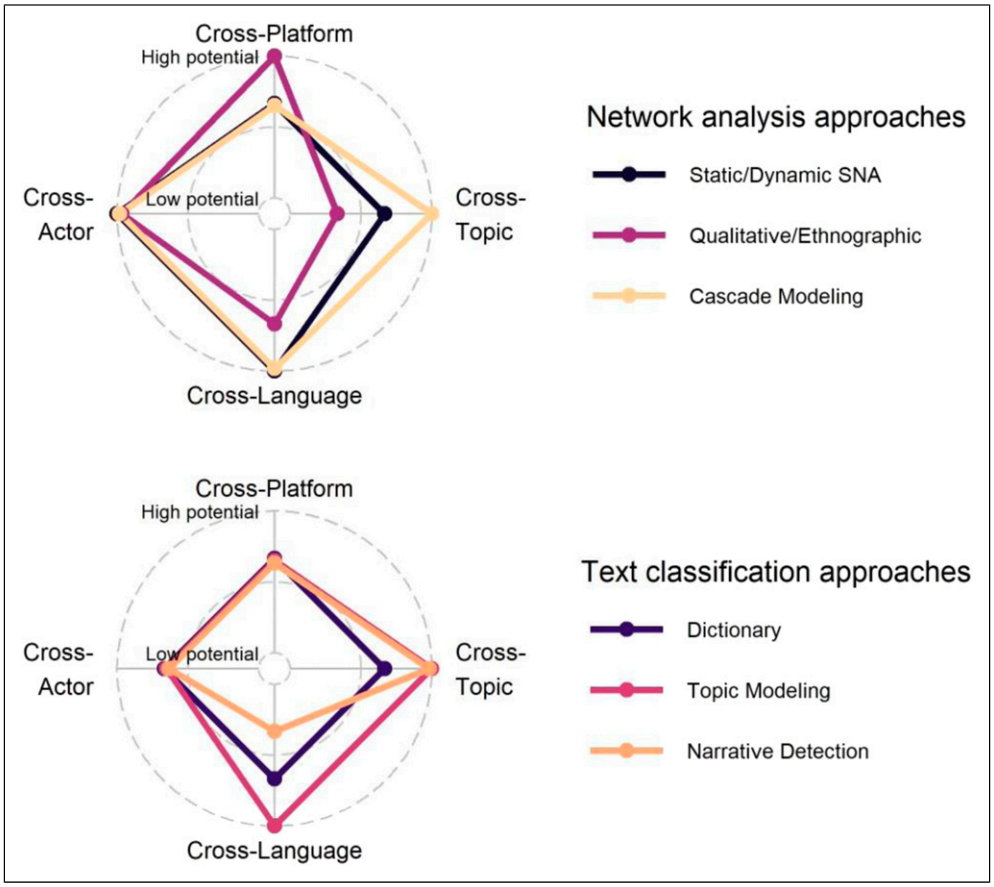

Figure 1. Methods' suitability for cross-analyses.

Social network analysis approaches

Actors. With respect to the diversity of actors gaining unmediated voice online, the actor-focused social network and cascade analyses are suitable to capture the static and dynamic relationship between actor types, which can be defined via community detection approaches or – although more resource-intensive – by manual classification.

Languages. Network analysis approaches stand out by their inherent cross-language scope, given the dataset has not been pre-limited in the stage of its construction, for example through language-bound dictionaries or other language-specific selection criteria.

Topics. Comparative network analyses across several topics are complicated by the fact that dictionary-based databases likely differ in the depth and breadth of their representation of specific CTs. While structural network and actor metrics can be compared across several topic networks, sample-based influences are difficult to estimate. Ethnographic and qualitative procedures have the advantage of high case sensitivity, which likewise impedes cross-topic comparisons.

Platforms. Reference-based network analysis approaches are generally well-suited for cross-platform comparative studies, especially for static diffusion snapshots across platforms. Yet, differences in platform-internal reference usage according to specific technical affordances complicate these analyses, as do external CT sources. Careful corpora and dictionary construction is, therefore, a prerequisite. Reference-based dynamic modeling is thus primarily a viable platform-internal option, while cascade modeling across platforms can be applied based on both content- and reference-based indicators.

Text classification approaches

Actors. In general, the text classification approaches described in this paper can all be applied to corpora from different actors; although specific communication styles complicate corpus construction (e.g. dictionary approaches) in the first place. The challenge with topic modeling or narrative detection is rather that the method does not classify the actors spreading the content. This needs to be included by design, meaning either as model covariates or by combinations of content- and actor-based approaches.

Languages. The advances in handling multi-language corpora (Chan et al., 2020; Lind et al., 2019, 2021; Reber, 2019) enable cross-language analysis, at least for dictionary and topic modeling approaches, notwithstanding the complexity of multilingual dictionary construction and language transformations in topic models. To date, narrative detection based on semantic features is, to our knowledge, of limited applicability in cross-language analyses as its compatibility with current algorithmic and machine translation techniques is untested.

Topics. Content-based text-classification procedures are particularly fruitful for cross-topical comparisons. Yet, classifications can be rather coarse depending on the breadth of the initial text corpus (Faddoul et al., 2020), impacting all following investigations. A more fine-grained classification, again, can only be achieved if the researcher incorporates more prior assumptions in corpus construction.

Platforms. Cross-platform analyses are complicated as of now. We cannot yet estimate under which conditions different platform corpora can be combined to arrive at one cross-platform model, or if a different model is needed for each communication platform due to unique affordances and styles.

General trade-offs

For choosing the methodology for a comparative assessment of the salience and diffusion of CTs online, researchers must weigh various general trade-offs associated with the methods discussed in this paper. We highlight three of them below. The first trade-off is related to the research objective, the second one to a priori knowledge, and the final – and likely most severe one – to the construct validity when using computational approaches for analyzing conspirational content.

The first trade-off emerges with respect to the main theoretical objective and the dimensions a study is interested in. Applying the methods requires certain preconditions to be fulfilled, and a chosen analytical and methodological approach can exclude subsequent research objectives. For a thorough understanding of one or more specific CTs, they need to be explicitly pre-defined as data collection and corpus construction are based on a complete dictionary of actors, key terms, hashtags or hyperlinks associated with a given CT. While many features can be derived from a given dataset, a decision on at least one of these basic features has to be made. Compared to the interest in specific CTs, detecting CTs in general is primarily the domain of linguistic and semantic classifications, which is in its early stages. The question whether static observations or the modeling of diffusion processes are at the core of a research question, is most central with respect to study design and the time span for data collection. Both network and text classification approaches allow modeling dynamics in several ways, their choice very much driven by the main comparative interest. Again, a trade-off emerges between investigating the actual diffusion of CTs on an actor level in a social network or on a level that abstracts from the individual and measures CT diffusion as the salience of particular topics or narratives. Research purely analyzing the actors involved in the spread of a pre-defined CT via social network analysis, static and dynamic alike, abstracts from the actual construction of a conspiracy narrative and its topical variety. The networks are constructed and investigated using hyperlinks embedded from external websites or platforms, or represent platform-internal links used for the distribution of CTs. When analyzing external hyperlinks, an in-depth understanding of the shared contents (beyond URL classification) is crucial to infer CT spreading behavior. Cascade construction for self-exciting point-process models requires a similar selection of suitable identifiers and results in a model explaining the actual diffusion of a CT rather than its contents or construction. Investigating the salience and structures of CTs through topic modeling and narrative detection, on the other hand, devotes almost no attention to the actors themselves. Topic modeling is used to assess CTs, though the interpretation of model results is still debated. Topic models trained on a pre-defined, CT-specific corpus can give insight into the issues discussed (Maier et al., 2018b: 6) across time, while inferring CTs from topic model classifications based on a broader corpus is more difficult. Narrative detection in CTs (Samory and Mitra, 2018b; Shahsavari et al., 2020) allows for a deeper understanding of CT construction while maintaining the possibility of cross-topical analysis. To bridge the trade-off between actor and content-focused methods, there have been advances to combine topic modeling and network analysis to measure CT (Rauchfleisch and Kaiser, 2020b) or issue salience (Maier et al., 2018b) on an actor level. Yet, theoretical interpretation of the results of these procedures remains an issue, as discussed below (construct validity).

The second trade-off is the amount of prior case-specific CT knowledge and assumptions included in the corpus construction, positively affecting granularity and negatively affecting the scope of the subsequent analysis. This cannot be resolved by combining different methods and remains a

major challenge for each research endeavor setting out for an in-depth understanding of CT diffusion dynamics.

The third severe trade-off we want to highlight is the one between a most accurate representation of a CT as a theoretically defined construct and the application of automated procedures for handling large amounts of data. Especially the dictionary and topic modeling procedures allow no differentiation between data that represent communication *about* a CT and the communication *of* a CT. While the narrative detection approach provides first steps for capturing narrative patterns, its applicability and validity need to be tested in further studies. Finally, none of the methods discussed are able to disentangle CTs from reasonable theorizing about suspected conspiracies (Baden and Sharon, 2021). This inaccuracy can be partly evaded by a careful sample selection of actors investigated, which in turn adds to the list of prior assumptions shaping the constructed corpus. The more the performance of those approaches can be improved, the more pronounced their inherent limitations might become. Thus, depending on the main aim and research question, each project has to carefully weigh whether and to what extent the described computational methods are suitable for its objectives, which trade-offs are acceptable, and which combinations of classical manual approaches and automated procedures are most promising.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by grants from the German Federal Ministry of Education and Research (grant number 16DII125 and 13N16049 [in the context of the call for proposals 'civil security - societies in transition']).

ORCID iDs

Annett Heft https://orcid.org/0000-0001-6637-795X
Kilian Buehling https://orcid.org/0000-0002-5244-7547

Supplemental Material

Supplemental material for this article is available online.

Notes

1. For the type of literature review, see Grant and Booth (2009).
2. The list of included scientific disciplines and search strings used as well as the full list of considered studies are documented in the Supplementary Material to this paper.

References

Ahmed W, Vidal-Alaball J, Downing J, et al. (2020) COVID-19 and the 5G conspiracy theory: social network analysis of twitter data. *Journal of Medical Internet Research* 22(5): 1–9. DOI: 10.2196/19458.

Baden C and Sharon T (2021) Blinded by the lies? Toward an integrated definition of conspiracy theories. *Communication Theory* 31(1): 82–106. DOI: 10.1093/ct/qtaa023.

Barkun M (2013) *A Culture of Conspiracy. Apocalyptic Visions in Contemporary America*. Comparative Studies in Religion and Society. Berkeley: University of California Press.

Benkler Y (2006) *The Wealth of Networks. How Social Production Transforms Markets and Freedom*. New Haven [Conn.]: Yale University Press.

Benoit K, Watanabe K, Wang H, et al. (2018) quanteda: an R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30): 1–4. DOI: 10.21105/joss.00774.

Bergmann E (2020) Populism and the politics of misinformation. *Safundi* 21(3): 251–265. DOI: 10.1080/17533171.2020.1783086.

Bessi A, Coletto M, Davidescu GA, et al. (2015) Science vs conspiracy: collective narratives in the age of misinformation. *Plos One* 10(2): 1–17. DOI: 10.1371/journal.pone.0118093.

Boberg S, Quandt T, Schatto-Eckrodt T, et al. (2020) Pandemic populism: facebook pages of alternative news media and the corona crisis - a computational content analysis. Muenster Online Research (MOR) Working Paper, 6 April. Available at: http://arxiv.org/abs/2004.02566.

Bruns A (2008) Life beyond the public sphere: towards a networked model for political deliberation. *Information Polity* 13(1–2): 71–85. DOI: 10.3233/IP-2008-0141.

Bruns A, Harrington S, and Hurcombe E (2020) 'Corona? 5G? or both?': the dynamics of COVID-19/5G conspiracy theories on Facebook. *Media International Australia* 177(1): 12–29. DOI: 10.1177/1329878X20946113.

Buhl F, Günther E, and Quandt T (2018) Observing the dynamics of the online news ecosystem. News diffusion processes among German news sites. *Journalism Studies* 19(1): 79–104. DOI: 10.1080/1461670X.2016.1168711.

Butter M and Knight P (2020a) General introduction. In: M Butter and P Knight (eds) *Routledge Handbook of Conspiracy Theories*. London: Routledge, 1–8. DOI: 10.4324/9780429452734.

Butter M and Knight P (2020b) *The Routledge Handbook of Conspiracy Theories*. Abingdon: Routledge.

Chan C-H, Zeng J, Wessler H, et al. (2020) Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures* 14(4): 285–305. DOI: 10.1080/19312458.2020.1812555.

Chan C, Bajjalieh J, Auvil L, et al. (2021) Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: a large-scale p-hacking experiment. *Computational Communication Research* 3(1): 1–27. DOI: 10.5117/CCR2021.1.001.

Dearing JW and Rogers EM (1996) *Agenda-Setting*. Thousand Oaks/London/New Delhi: Sage.

Del Vicario M, Bessi A, Zollo F, et al. (2016) The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America* 113(3): 554–559. DOI: 10.1073/pnas.1517441113.

Devlin J, Chang MW, Lee K, et al. (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171–4186.

Dolata U (2019) Plattform-Regulierung. Koordination von Märkten und Kuratierung von Sozialität im Internet. *Berliner Journal für Soziologie* 69(3–4): 179–206. DOI: 10.1007/s11609-020-00403-9.

Elmas T, Overdorf R, Özkalay AF, et al. (2019) Ephemeral astroturfing attacks: the case of fake twitter trends. arXiv preprint. Available at: https://arxiv.org/abs/1910.07783.

Entman RM (1993) Framing: toward clarification of a fractured paradigm. *Journal of Communication* 43(4): 51–58. DOI: 10.1111/j.1460-2466.1993.tb01304.x.

Faddoul M, Chaslot G, and Farid H (2020) A longitudinal analysis of YouTube's promotion of conspiracy videos. arXiv preprint. Available at: http://arxiv.org/abs/2003.03318.

Fallis D (2015) The concept of disinformation. In: M Khosrow-Pour (ed) *Encyclopedia of Information Science and Technology*. 3rd edition. PA: IGI Global, 4720–4727. DOI: 10.4018/978-1-4666-5888-2.ch463.

Ferrara E (2020) What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday* 25(6): 1–25. DOI: 10.5210/fm.v25i6.10633.

Friggeri A, Adamic LA, Eckles D, et al. (2014) Rumor cascades. *Proceedings of the International AAAI Conference on Web and Social Media* 8(1): 101–110. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14559 (accessed 17 January 2022).

Garry A, Walther S, Mohamed R, et al. (2021) QAnon conspiracy theory : examining its evolution and mechanisms of radicalization. *Journal for Deradicalization* Spring(26): 152–216.

Gerts D, Shelley CD, Parikh N, et al. (2021) "Thought I'd share first" and other conspiracy theory tweets from the COVID-19 infodemic: exploratory study. *JMIR Public Health and Surveillance* 7(4): e26527. DOI: 10.2196/26527.

Giglietto F, Righetti N, Rossi L, et al. (2020) It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society* 23(6): 867–891. DOI: 10.1080/1369118X.2020.1739732.

Gleicher N (2018) *Coordinated Inauthentic Behavior Explained*. Available at: https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/ (accessed 22 July 2021).

Graham T, Bruns A, Zhu G, et al. (2020) *Like a Virus: The Coordinated Spread of Coronavirus Disinformation*. Canberra: Centre for Responsible Technology, The Australia Institute. Available at: https://eprints.qut.edu.au/202960/.

Grant MJ and Booth A (2009) A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal* 26(2): 91–108. DOI: 10.1111/j.1471-1842.2009.00848.x.

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. DOI: 10.1093/PAN/MPS028.

Gruzd A and Mai P (2020) Going viral: how a single tweet spawned a COVID-19 conspiracy theory on Twitter. *Big Data & Society* 7(2): 1–9. DOI: 10.1177/2053951720938405.

Häussler T (2021) Civil society, the media and the internet: changing roles and challenging authorities in digital political communication ecologies. *Information, Communication & Society* 24(9): 1265–1282. DOI: 10.1080/1369118X.2019.1697338.

Highfield T (2018) Emoji hashtags// hashtag emoji: of platforms, visual affect, and discursive flexibility. *First Monday* 23(9). DOI: 10.5210/FM.V23I9.9398.

Hui P-M, Yang K-C, Torres-Lugo C, et al. (2019) BotSlayer: real-time detection of bot amplification on Twitter. *Journal of Open Source Software* 4(42): 1706. DOI: 10.21105/joss.01706.

ISD Digital Research Unit (2020a) *Covid-19 Disinformation Briefing No. 1*. Available at: https://www.isdglobal.org/wp-content/uploads/2020/03/COVID-19-Briefing-Institute-for-Strategic-Dialogue-27th-March-2020.pdf.

ISD Digital Research Unit (2020b) *Far-right Exploitation of Covid-19*. Available at: https://www.isdglobal.org/wp-content/uploads/2020/05/20200513-ISDG-Weekly-Briefing-3b.pdf.

Jacobi C, Van Atteveldt W, and Welbers K (2016) Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism* 4(1): 89–106. DOI: 10.1080/21670811.2015.1093271.

Keeley BL (1999) Of conspiracy theories. *The Journal of Philosophy* 96(3): 109–126. DOI: 10.2307/2564659.

Kim M, Newth D, and Christen P (2013) Modeling dynamics of diffusion across heterogeneous social networks: news diffusion in social media. *Entropy* 15(10): 4215–4242. DOI: 10.3390/e15104215.

Kim M, Paini D, and Jurdak R (2020) Real-world diffusion dynamics based on point process approaches: a review. *Artificial Intelligence Review* 53(1): 321–350. DOI: 10.1007/s10462-018-9656-9.

Klein C, Clutton P, and Dunn AG (2019) Pathways to conspiracy: the social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *Plos One* 14(11): e0225098. DOI: 10.1371/journal.pone.0225098.

Kobayashi R and Lambiotte R (2016) TiDeH: time-dependent Hawkes process for predicting retweet dynamics. In: Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM), Cologne, 2016, pp. 191–200. AAAI Press.

Kong Q, Ram R, and Rizoiu MA (2021) Evently: modeling and analyzing reshare cascades with hawkes processes. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, New York, 2021, pp. 1097–1100. Association for Computing Machinery. DOI: 10.1145/3437963.3441708.

Krafft P and Donovan J (2020) Disinformation by design: the use of evidence collages and platform filtering in a media manipulation campaign. *Political Communication* 37(2): 194–214. DOI: 10.1080/10584609.2019.1686094.

Kreps GL (2017) Diffusion theory in integrative approaches. In: DL Cloud (ed) *Oxford Research Encyclopedia of Communication*. Oxford: Oxford University Press. DOI: 10.1093/acrefore/9780190228613.013.251.

Lazer DMJ, Baum MA, Benkler Y, et al. (2018) The science of fake news. *Science* 359(6380): 1094–1096. DOI: 10.1126/science.aao2998.

Leal H (2020) Networked disinformation and the lifecycle of online conspiracy theories. In: M Butter and P Knight (eds) *Routledge Handbook of Conspiracy Theories*. London: Rouledge, 497–511. DOI: 10.4324/9780429452734.

Leifeld P, Cranmer SJ, and Desmarais BA (2018) Temporal exponential random graph models with btergm: estimation and bootstrap confidence intervals. *Journal of Statistical Software* 83(1): 1–36. DOI: 10.18637/JSS.V083.I06.

Lind F, Eberl J-M, Heidenreich T, et al. (2019) When the journey is as important as the goal: a roadmap to multilingual dictionary construction. *International Journal of Communication* 13(21): 4000–4020. Available at: https://ijoc.org/index.php/ijoc/article/view/10578.

Lind F, Eberl J-M, Eisele O, et al. (2021) Building the bridge: topic modeling for comparative research. In: Paper presented at the 71st Annual ICA Conference, Virtual Conference, 2021.

Mahl D, Zeng J, and Schäfer MS (2021) From "nasa lies" to "reptilian eyes": mapping communication about 10 conspiracy theories, their communities, and main propagators on Twitter. *Social Media + Society* 7(2): 1–12. DOI: 10.1177/20563051211017482.

Mahl D, Schäfer MS, and Zeng J (2022) Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research. *New Media & Society*. OnlineFirst. DOI: 10.1177/14614448221075759.

Maier D, Waldherr A, Miltner P, et al. (2018a) Applying LDA topic modeling in communication research: toward a valid and reliable methodology. *Communication Methods and Measures* 12(2–3): 93–118. DOI: 10.1080/19312458.2018.1430754.

Maier D, Waldherr A, Miltner P, et al. (2018b) Exploring issues in a networked public sphere. Combining hyperlink network analysis and topic modeling. *Social Science Computer Review* 36(1): 3–20. DOI: 10.1177/0894439317690337.

Maier D, Baden C, Stoltenberg D, et al. (2021) Machine translation Vs. multilingual dictionaries: assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures*. Online first. DOI: 10.1080/19312458.2021.1955845.

Murayama T, Wakamiya S, Aramaki E, et al. (2021) Modeling the spread of fake news on Twitter. *Plos One* 16(4): 1–16. DOI: 10.1371/journal.pone.0250419.

Neuberger C (2009) Internet, Journalismus und Öffentlichkeit. Analyse des Medienumbruchs. In: C Neuberger, C Nuernbergk, and M Rischke (eds) *Journalismus Im Internet. Profession - Partizipation - Technisierung*. Wiesbaden: VS Verlag für Sozialwissenschaften, 19–105.

Pacheco D, Flammini A, and Menczer F (2020) Unveiling coordinated groups behind white helmets disinformation. In: Companion Proceedings of the Web Conference 2020, New York, 2020, pp. 611–616. Association for Computing Machinery. DOI: 10.1145/3366424.3385775.

Pacheco D, Hui P-M, Torres-Lugo C, et al. (2021) Uncovering coordinated networks on social media: methods and case studies. In: Proceedings of the AAAI International Conference on Web and Social Media (ICWSM), 2021. arXiv preprint. Available at: http://arxiv.org/abs/2001.05658.

Papakyriakopoulos O, Serrano JCM, and Hegelich S (2020) Political communication on social media: a tale of hyperactive users and bias in recommender systems. *Online Social Networks and Media* 15: 100058. DOI: 10.1016/j.osnem.2019.100058.

Park HW (2003) Hyperlink network analysis : a new method for the study of social structure on the web. *Connections* 25(1): 49–61.

Paudel P, Blackburn J, De Cristofaro E, et al. (2021) Soros, child sacrifices, and 5G: understanding the spread of conspiracy theories on web communities. arXiv preprint. Available at: https://arxiv.org/abs/2111.02187v1 (accessed 24 January 2022).

Pfetsch B, Adam S, and Bennett WL (2013) The critical linkage between online and offline media: an approach to researching the conditions of issue spill-over. *Javnost - The Public* 20(3): 9–22. DOI: 10.1080/13183222.2013.11009118.

Potnis D and Tahamtan I (2021) Hashtags for gatekeeping of information on social media. *Journal of the Association for Information Science and Technology*. Online first. DOI: 10.1002/asi.24467.

Rauchfleisch A and Kaiser J (2020a) The false positive problem of automatic bot detection in social science research. *Plos One* 15(10): e0241045. DOI: 10.1371/JOURNAL.PONE.0241045.

Rauchfleisch A and Kaiser J (2020b) The German far-right on youtube: an analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media* 64(3): 373–396. DOI: 10.1080/08838151.2020.1799690.

Reber U (2019) Overcoming language barriers: assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures* 13(2): 102–125. DOI: 10.1080/19312458.2018.1555798.

Rizoiu M-A, Lee Y, Mishra S, et al. (2017) A tutorial on hawkes processes for events in social media. arXiv preprint. Available at: http://arxiv.org/abs/1708.06401.

Roberts ME, Stewart BM, and Airoldi EM (2016) A model of text for experimentation in the social sciences. *Journal of the American Statistical Association* 111(515): 988–1003. DOI: 10.1080/01621459.2016.1141684.

Rogers EM (1983) *The Diffusion of Innovations*. 3rd edition. New York: The Free Press.

Rogers EM (2000) Reflections on news event diffusion research. *Journalism & Mass Communication Quaterly* 77(3): 561–576.

Samory M and Mitra T (2018a) Conspiracies online: user discussions in a conspiracy community following dramatic events. In: Proceedings of the 12th International AAAI Conference on Web and Social Media, 2018, pp. 340–349. AAAI Press.

Samory M and Mitra T (2018b) 'The government spies using our webcams'. In: ACM on Human-Computer Interaction, 2018, pp. 1–24. DOI: 10.1145/3274421.

Seidler JD (2016) *Die Verschwörung Der Massenmedien*. Bielefeld: transcript Verlag. DOI: 10.14361/9783839434062.

Sha H, Hasan M Al, Mohler G, et al. (2020) Dynamic topic modeling of the COVID-19 Twitter narrative among US governors and cabinet executives. arXiv preprint. Available at: https://arxiv.org/abs/2004.11692.

Shadrova A (2021) Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining and Digital Humanities*. Centre pour la Communication Scientifique Directe (CCSD). DOI: 10.46298/JDMDH.7595.

Shahsavari S, Holur P, Wang T, et al. (2020) Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science* 3: 279–317. DOI: 10.1007/s42001-020-00086-5.

Shao C, Ciampaglia GL, Varol O, et al. (2017) The spread of low-credibility content by social bots. *Nature Communications* 9. arXiv preprint: 4787. DOI: 10.1038/s41467-018-06930-7.

Smith N and Graham T (2019) Mapping the anti-vaccination movement on Facebook. *Information Communication and Society* 22(9): 1310–1327. DOI: 10.1080/1369118X.2017.1418406.

Snijders TAB (2001) The statistical evaluation of social network dynamics. *Sociological Methodology* 31(1): 361–395. DOI: 10.1111/0081-1750.00099.

Srnicek N (2017) *Platform Capitalism*. Cambridge: Polity Press (e-book).

Starbird K and Wilson T (2020) Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review* 1(1): 1–11. DOI: 10.37016/mr-2020-002.

Starbird K, Spiro E, Edwards I, et al. (2016) Could this be true? I think so! Expressed uncertainty in online rumoring. In: Proceedings of the Conference on Human Factors in Computing Systems, New York, 2016, pp. 360–371. Association for Computing Machinery. DOI: 10.1145/2858036.2858551.

Vosoughi S, Roy D, and Aral S (2018) The spread of true and false news online. *Science* 359(6380): 1146–1151. DOI: 10.1126/science.aap9559.

Walter D and Ophir Y (2019) News frame analysis: an inductive mixed-method computational approach. *Communication Methods and Measures* 13(4): 248–266. DOI: 10.1080/19312458.2019.1639145.

Wilson AF (2017) The bitter end: apocalypse and conspiracy in white nationalist responses to the Islamic State attacks in Paris. *Patterns of Prejudice* 51(5): 412–431. DOI: 10.1080/0031322X.2017.1398963.

Wood MJ (2018) Propagating and debunking conspiracy theories on Twitter during the 2015-2016 Zika virus outbreak. *Cyberpsychology, Behavior, and Social Networking* 21(8): 485–490. DOI: 10.1089/cyber.2017.0669.

Yang KC, Varol O, Hui PM, et al. (2020) Scalable and generalizable social bot detection through data selection. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020, pp. 1096–1103. DOI: 10.1609/aaai.v34i01.5460.

Zannettou S, Caulfield T, De Cristofaro E, et al. (2017) The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In: Proceedings of the ACM SIGCOMM Internet Measurement Conference, London, 2017, pp. 405–418. Association for Computing Machinery. DOI: 10.1145/3131365.3131390.

Zeng J and Schäfer MS (2021) Conceptualizing 'dark platforms'. Covid-19-related conspiracy theories on 8kun and gab. *Digital Journalism* 9(9): 1321–1343. DOI: 10.1080/21670811.2021.1938165.

Zhao Q, Erdogdu MA, He HY, et al. (2015) SEISMIC: a self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sidney, 2015, pp. 1513–1522. Association for Computing Machinery. DOI: 10.1145/2783258.2783401.

Zimmermann F and Kohring M (2018) "Fake news" als aktuelle desinformation. Systematische Bestimmung eines heterogenen Begriffs. *M&K Medien & Kommunikationswissenschaft* 66(4): 526–541. DOI: 10.5771/1615-634X-2018-4-526.

Author biographies

Annett Heft is head of the research group *Digitalization and the Transnational Public Sphere* at the Weizenbaum Institute for the Networked Society, Berlin, and senior researcher at the Institute for Media and Communication Studies, Freie Universität Berlin. Her main research fields are the comparative study of political communication in Europe, with an emphasis on digital public spheres, right-wing communication infrastructures, transnational communication as well as quantitative research methods and computational social science.

Kilian Buehling is a doctoral researcher in the research group *Digitalization and the Transnational Public Sphere* at the Weizenbaum Institute for the Networked Society, Berlin, and the Institute for Media and Communication Studies, Freie Universität Berlin. His previous work contains research in information science, quantitative innovation economics and scientometrics.