

Challenges of and approaches to data collection across platforms and time: Conspiracy-related digital traces as examples of political contention

Annett Heft, Kilian Buehling, Xixuan Zhang, Juni Schindler & Miriam Milzner

To cite this article: Annett Heft, Kilian Buehling, Xixuan Zhang, Juni Schindler & Miriam Milzner (2024) Challenges of and approaches to data collection across platforms and time: Conspiracy-related digital traces as examples of political contention, *Journal of Information Technology & Politics*, 21:3, 323-339, DOI: [10.1080/19331681.2023.2250779](https://doi.org/10.1080/19331681.2023.2250779)

To link to this article: <https://doi.org/10.1080/19331681.2023.2250779>



© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 28 Aug 2023.



[Submit your article to this journal](#)



Article views: 2105



[View related articles](#)








[View Crossmark data](#)



Citing articles: 9 [View citing articles](#)

Challenges of and approaches to data collection across platforms and time: Conspiracy-related digital traces as examples of political contention

Annett Heft , Kilian Buehling , Xixuan Zhang , Juni Schindler , and Miriam Milzner 

ABSTRACT

Taking the example of conspiracy-related communication online as one form of contentious politics, this study examines the data collection challenges for multidimensional comparative research across platforms, time, and cultural embeddings. It compares the architectures and features relevant to data collection, access regimes, and use cultures for a set of digital platforms and communication venues. Differentiating between actor- and content-based strategies, this study discusses the potentials and limitations of these approaches, considering differences in platforms, temporal dynamics, and cultural embeddings as well as several layers of equivalence. The discussion highlights crucial insights into designing data collection strategies in multidimensional comparative studies.

KEYWORDS



Data collection strategies; cross-platform research; comparative research; political contention; conspiracy theories; social media; platform architectures; functional equivalence

Introduction

Nowadays, a significant part of political contention and mobilization is performed through digital communication and distributed in a hybrid and networked digital information ecology (Häussler, 2021). This ecology of communication and information circulation is constituted by a range of social networking platforms, such as Facebook or Twitter; messenger and microblogging sites, such as Telegram; image and discussion boards, such as 4chan and Reddit; and alternative and legacy media sites online. These venues all come with specific platform architectures, features, and afforded utilities for specific actor groups (Bossetta, 2019; Evans, Pearce, Vitak, & Treem, 2017), governance structures, and access regimes that fundamentally influence the *data collection* possibilities and limitations in such sites. Although decisions concerning these multifaceted platform characteristics can influence empirical analyses, they are rarely discussed at length (Mahl, von Nordheim, & Guenther, 2022).

In addition to the question of the ways in which platform peculiarities pose challenges to valid data collection, the matter becomes even more complex if we acknowledge the nature of digital information ecologies. Digital communication seldom remains contained within one specific platform, as

platforms and communication venues are mutually interrelated. First, technological features enable content to be easily spread within and across several platforms. This affordance of networked communication can contribute to the diffusion of topics and narratives from platforms providing spaces for fringe use cultures and *dark participation* (Quandt, 2018) to broader audiences, which might therefore influence societal discourses at large. Second, acts of political contention are not only intertwined across platforms through linking, sharing, and forwarding features, representing just one meaning of the “cross” in “cross-platform.” Political and challenger actors who contend with existing rules and procedures also often maintain accounts on several platforms. They strategically leverage platform-specific features to adapt their messages to distinct audiences and platform-specific use cultures (Ekman, 2022). Even without direct digital references between platforms, we can expect that users observe discourses across platforms and that debates on contentious issues are marked by mutual interference – again contributing to the whole of societal discussion. If we acknowledge this inherent and double cross-platform nature of digital communication, then

CONTACT Annett Heft  annett.heft@fu-berlin.de  Freie Universität Berlin and Weizenbaum Institute for the Networked Society, Garystr. 55, Berlin 14195, Germany

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

the question is how *equivalent data collection* from several platforms and communication venues can be facilitated to enable cross-platform and platform-comparative studies, as well as whether viable approaches to deal with vast platform differences are available. Comparative studies on these acts of contentious claims-making and mobilization across multiple platforms and communication venues have only recently become more frequent (e.g. Frischlich, Schatto-Eckrodt, & Völker, 2022; Yarchi, Baden, & Kligler-Vilenchik, 2020), which is not surprising given the task complexity involved.

The same applies to research on the spread of and mobilization through conspiracy theories. Studies researching conspiracy-related content online have often focused on single platforms (Gallagher, Davey, & Hart, 2020; Tuters & Hagen, 2020), context-specific events, or particular actor groups (Bevensee & Ross, 2018; Knight, 2008; Wilson, 2017). Conspiracy theories are central parts of contentious practices. As in any other type of digitally mediated political contention, understanding whether and how conspiracy theories appear on and potentially transcend the boundaries of specific platforms is crucial for our understanding of the formation of contentious public debate. Conspiracy theories are also deeply rooted in a particular time and culture (Barkun, 2013), adding these dimensions to comparative designs and thus increasing their complexity. Against this background, our study uses the phenomenon of conspiracy theories as an exemplary case for our discussion of the challenges of and approaches to data collection in multidimensional comparative studies on political contention. We ask the following questions:

What are the methodological and practical challenges of different platform architectures, governance and access regimes, and use cultures for data collection across platforms and time?

What approaches could facilitate equivalent data collection from multiple platforms while also considering temporal dynamics and cultural embeddings?

What are the implications of (partially unresolvable) limitations for valid data collection?

To address these questions, we briefly introduce the example of conspiracy-related communication

as one form of contentious politics. We then compare the architectures and features relevant to data collection, access regimes, and aspects that stand out for particular use cultures in the context of conspiracy theories for a set of platforms and digital communication venues. We highlight the challenges that researchers face when collecting data in a cross-platform and cross-time comparative design and juxtapose the data collection possibilities organized by the differentiation between actor- and content-based strategies. We discuss the potentials and limitations of these approaches, considering differences in platforms, temporal dynamics, and cultural embeddings, as well as several layers of equivalence. The discussion highlights crucial insights into designing data collection strategies in multidimensional comparative studies that extend beyond our example of conspiracy-related content to a wider range of digital political communication.

Comparative data collection: The example of conspiracy-related content

When social movements and political entrepreneurs strive to articulate their claims and mobilize political contention, social media platforms and a vast array of digital communication technologies provide a central infrastructure for the organization of contentious politics. One form of content that can contribute to political mobilization is the narration and distribution of conspiracy theories. Conspiracy theories have been defined as the “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” (Keeley, 1999, p. 116). The key characteristics of conspiracy theories are an intentionalism suspected behind events and actions, a dualism between a small group of conspirators and those who are affected, and the secrecy in which connected actions and processes occur (Baden & Sharon, 2021; Barkun, 2013; Butter & Knight, 2020; Mahl, Schäfer, & Zeng, 2023). Not every form of public conspiracism or communication linked to or labeled as conspiracy theory exhibits all defining characteristics. Particularly in public debate, ostracizing an actor group as a conspirator or an explanation as a conspiracy

can serve many political functions. Hyzen and den Bulck (2021) argue that such strategies are used to denigrate opposing actors and undermine prevailing institutions, values, and beliefs. Harris (2023, p. 6) sees their political utility in their “counter-official nature.” In this respect, conspiracy theories are central parts of contentious practices and are closely linked to populist communication (Bergmann, 2018). With the concept of conspiracy-related content, we refer to the full public communication *on* and *about* (alleged) conspiracy theories in various communication venues, including their narrations, counter-narrations, and debunking, as well as neutral observation forms.

As is the case with political contention studies in general, research on conspiracy-related content online has often focused on single established social media platforms, such as Twitter (Graham, Bruns, Zhu, & Campbell, 2020; Mahl, Zeng, & Schäfer, 2021) and Facebook (Bruns, Harrington, & Hurcombe, 2020). Others examine platforms facilitating secure spaces for the coordination of contentious actions (Herasimenka, 2019) and the development of conspiracy narrations within specific communities, such as on 4chan, 8kun (Tuters & Hagen, 2020; Tuters, Jokubauskaitė, & Bach, 2018), Reddit, Gab, and Telegram (Busbridge, Moffitt, & Thorburn, 2020; Garry, Walther, Mohamed, & Mohammed, 2021; Zeng & Schäfer, 2021). In addition, studies often concentrate on singular events in a specific national context or particular actor groups (Bevensee & Ross, 2018; Knight, 2008; Wilson, 2017). This is no surprise, as conspiracy-related content is particularly difficult to detect, and the more so when automated text collection and analysis are involved. In principle, this is due to the inherent blurriness of the concept and the difficulty of distinguishing talk about possible conspiracies from, as Baden and Sharon (2021, p. 90) call it, “conspiracy theories proper.” Even when specific, *ex ante*-defined conspiracy theories are studied, their appearances can often be ambiguous, and the explicitness of related talk can vary depending on the discourse context, such as the platform’s communication styles and community norms (Baden & Sharon, 2021). Subcultural milieus, such as those found on 4chan and 8kun, demonstrate their belonging to a community by consciously using insider

abbreviations, floating signifiers, and slang that are difficult to detect and understand from the outside (Frischlich, Schatto-Eckrodt, & Völker, 2022; Nissenbaum & Shifman, 2017; Tuters & Hagen, 2020). Outside observations of conspiracist interpretations e.g. in traditional news media will likely rely on different denominators, depending on news outlets’ characteristics and journalistic styles (Bruns, Hurcombe, & Harrington, 2022). Expressions might also change over time as conspiracy narratives are adapted to latch on to and integrate current developments and crises. They might vary in terms and explicitness across countries, as conspirational thinking and narratives are differently embedded and accepted across time and different cultures (Barkun, 2013). What is more, the actors spreading conspiracy theories are likely to be less institutionalized, thus making it more difficult to detect and classify them and to collect relevant information across platforms and time.

Many of these characteristics apply to the actors in and the content of digitally mediated political contention in general. The discussion that follows thus addresses the challenges of and approaches to data collection in a more general sense. However, the case of conspiracy theories offers a prime example, as it combines general and specific challenges and helps illustrate solution strategies.

To systematize our discussion of data collection strategies, we broadly differentiate between two approaches (Heft & Buehling, 2022). The first is an *actor-based approach* in which scholars use known actors or accounts identified *a priori* as access points to gather their communication. In the case of conspiracy-related communication, these are often actors or sites that have been linked to conspirational or otherwise problematic content on blacklists, fact-checking sites, or prior research (e.g., on alternative media; Rooke, 2021), or ideological entrepreneurs, such as Jordan Peterson or Alex Jones (Hyzen & den Bulck, 2021). The second type is the *content-based approach*. Studies following this approach often focus on one or more *a priori* known conspiracy theories and use case-specific key terms and hashtags (e.g., #5GCoronavirus, #Pizzagate) (Graham, Bruns, Zhu, & Campbell, 2020; Leal, 2020) or more encompassing dictionary-based procedures, such as the computational dictionary for the study of

right-wing populist conspiracy discourse (RPC) by Puschmann, Karakurt, Amlinger, Gess, and Nachtwey (2022), to construct the data corpus of a study. The extent to which one of these strategies or a combination of both is viable for *comparative data collection* from the vast array of platforms and communication venues online depends not only on a study's aim but also on several platform characteristics, which are discussed in the following section.

Platform characteristics and their consequences for data collection

For comparative studies across platforms and communication venues, thorough insights into platforms' general architectures (Bossetta, 2019) and their ways of structuring content and enabling access through various features are paramount (Pearce et al., 2020), as these fundamentally shape data collection possibilities and limitations. The platform architecture defines the form of communication infrastructure that is established. The content- and actor-related characteristics of platforms and online media determine the units of analysis that are possible, how content can be organized and found, and how individual pieces of information are accessible (Table 1). While outlining platform specifics and how they enable or restrict data collection in general, we acknowledge that each platform is distinct and that architectures and access regimes can change considerably over time. Furthermore, we can only highlight some relevant aspects for *data collection* across the board, while specific sampling strategies (e.g., based on engagement measures) or possible levels of comparative data analysis (Rogers, 2019, Chapter 10) are beyond the scope of our discussion.

Discussion platforms

Discussion forums such as 4chan, 8kun, or Reddit offer a room-based interaction architecture in which communication is usually organized in a threaded, interconnected way (Frischlich, Schatto-Eckrodt, & Völker, 2022; Tuters, Jokubauskaitė, & Bach, 2018). They typically consist of thematically organized boards and specific threads (4chan, 8kun) or subreddits (Reddit).

Communication always starts with an initial submission or post, followed by comments ordered sequentially, which are kept together in a specific thread (Frischlich, Schatto-Eckrodt, & Völker, 2022; Prakasam & Huxtable-Thomas, 2021).

Discussion boards enable a content-based search within and across specific boards and subreddits, such as the subreddit r/conspiracy analyzed by Samory and Mitra (2018). When it comes to content-related characteristics, the full context of a particular post within these boards can usually be evaluated only if the preceding discussion is known. Therefore, one off-topic post could mislabel an entire thread and introduce significant noise to the dataset. Researchers must decide whether the whole board, specific threads, the full thread, or certain thread elements constitute the relevant unit of analysis. In terms of actor-related characteristics, 4chan users can start and contribute to discussions without registering, and upload text and images anonymously. On Reddit, no user account is required to access most content, which can be found through a general search or the selection of specific subreddits. Reddit also offers a high anonymity level for users, as registration requires no personal data and thus allows multiple and invalidated personas by the same individual (Prakasam & Huxtable-Thomas, 2021). Thus, while platform features allow open access to various content forms, linking this content to identifiable actors is prevented by the platforms' policies and features that afford high anonymity. Accordingly, actor-based approaches to data collection are not feasible within discussion boards. Content-based approaches, however, are viable if the content is accessed live or if an archive provides external access.

Networked social media

Social networking platforms, such as Facebook, and micro-blogging platforms, such as Twitter and Gab, offer a network-based interaction architecture in which communication is organized in an interconnected way. These platforms provide infrastructures through which users can upload and disseminate content via a recognizable identity (profile). Overall, the communication infrastructure is built on persistent identities, as usage always

Table 1. Key Facts.

Platform/ Media	Genre	Design			Content		Actor		Collection
		general	content-chunking	interaction	unit of analysis	anonymity	access	unit of analysis	
4chan	discussion platform	general	content-chunking	room-based	submission	yes	open	-	live, externally archived
Reddit	discussion platform	thematic boards and thematic subreddits and threads	variable-sized, interconnected	room-based	comment submission	yes	open	user profile	live, externally archived
Telegram	instant messenger	private chats, groups and channels	variable-sized, self-contained, interconnected	hybrid	comment public message private message	yes no no no	open open open closed	user profile no	selectable stored history
Gab	micro-blogging	profiles, posts, and groups	variable-sized, self-contained, interconnected	network- based	post	no	open	user profile	open
Twitter	micro-blogging	profiles and tweets	fixed-sized, self-contained, interconnected	network- based	comment tweet	no no	open closed	user profile	selectable stored history
Facebook	social network	profiles, posts, private and closed groups	variable-sized, self-contained, interconnected	network- based	comment post	no no	closed selectable	user profile	selectable stored history
YouTube	format-oriented publishing	video content and user profiles	variable-sized, self-contained	broadcast- based	comment video	no no	open open	user profile	open stored history
Online Media	format-oriented publishing	articles, thematically organized	variable-sized, self-contained	broadcast- based	comment article	no yes/no	selectable (paywall)	whole site/ editor	stored history, externally archived

requires some form of registration and self-presentation via usernames or descriptions, resulting in appearances as identifiable personas (Frischlich, Schatto-Eckrodt, & Völker, 2022; Jasser, McSwiney, Pertwee, & Zannettou, 2023).

At the content level, the main analysis units comprise original posts (Facebook, Gab) or tweets (Twitter), which are self-contained in distributing their meanings. Several forms of comments inscribed in forward, reply, or retweet functions are also possible. These communication forms can but must not be interconnected, and they do not necessarily follow a sequential logic but can also take place simultaneously. Depending on the requirements of persistent personalization, which rather enhances pseudonymity in some instances (Frischlich, Schatto-Eckrodt, & Völker, 2022), several features enable data collection based on the actors and their content as the analysis units. For example, Bruns, Harrington, and Hurcombe (2020, p. 15) use the search query “(covid,corona,virus,epidemi,pandemi) AND (5 g)” to collect Facebook posts related to the dissemination of COVID-19/5 G conspiracy theories while classifying actors spreading this content as part of their data analyses. However, these platforms also afford users ways to limit the findability and accessibility of content and user information through choices in privacy settings (Frischlich, Schatto-Eckrodt, & Völker, 2022; Jasser, McSwiney, Pertwee, & Zannettou, 2023). As a result of this, for example the study by Bruns, Harrington, and Hurcombe (2020) was limited to Facebook public spaces while closed groups or private profiles can’t be collected.

Overall, data collection on Facebook and Twitter primarily relies on application programming interfaces (APIs), which are open to researchers upon request. While Twitter at the time of writing supports a full-archive search, the Facebook API “CrowdTangle” limits data access to public pages and groups. As for Gab, scholars either scraped the platform (Fair & Wesslen, 2019) or used APIs (Jasser, McSwiney, Pertwee, & Zannettou, 2023; Zannettou et al., 2018) to enable large data collections.

Publishing-oriented platforms and online media

Format-oriented publishing platforms, such as YouTube, or the vast array of online alternative and legacy news media offer a broadcast-style

interaction architecture in which articles (online media) or videos (YouTube) stand on their own, not requiring direct relations to prior content.

At the content level, the main units of analysis can consequently be described as self-contained. However, users can refer to articles or videos through comment sections. On YouTube, content can generally be found by searching for specific terms or actors who can be identified by their registered user accounts. For example, Allington and Joshi (2020) use an actor-based approach to data collection, starting from the account of David Icke, an actor frequently described as a conspiracy theorist, and collecting his videos and related comments from his account. As for online media websites, while individual articles published on a website may or may not be linked to a specific author, research regularly assumes that the output of media websites (articles) represents the medium as a whole. Some format-oriented publishing platforms, such as YouTube, allow for both content- and actor-based approaches to data collection, while in the case of online media websites, a hybrid approach is necessary, as the actor (the analysis unit, i.e., the particular medium) must be defined in advance to create or assess a searchable corpus of website content.

Overall, format-oriented publishing platforms offer content that is regularly openly accessible, although commercial online media may sometimes restrict access through paywalls or membership-based access regimes. The feasibility of data collection is then highly dependent on the availability of archived content. For instance, while YouTube offers a search function similar to that of the front end, querying a variety of alternative and legacy media requires the availability of comprehensive archives, each with its own data quality and reach limitations (Blatchford, 2020). Approaching this challenge with a hybrid strategy of combining actor- and content-based data collection reveals the limitations of each website in terms of the availability of permanently archived content, native search functions, or APIs (Freelon, 2018).

Hybrid platforms

Telegram’s communication infrastructure, composed of channels and chat groups, renders it

a hybrid in our framework. Chat groups within Telegram afford room-based threaded discussions between uniquely identifiable users. Channels, on the other hand, show characteristics of broadcast-oriented platforms in which posts can be attributed to the channel's administrators and sometimes also provide a discussion option for readers.

Content in public groups and channels is generally openly accessible but can only be found if the name of the channel or group is known or an access link is available. Telegram user profiles neither entail information about the user's other posts in channels or chat groups nor require information about other chat groups in which the user is active (except if one's own account is a member of the same group as the focal user is).

An actor-based collection strategy for Telegram would require abstraction from individual users and call for the identification of whole chat groups and channels as actors. Schulze et al. (2022) have identified three exemplary channels related to QAnon that were used to collect all publicly available posts from these channels. In the case of chat groups, the resulting disregard for a multitude of speakers would need to be justified. Lacking an adequate search function, content-based approaches require a similar hybrid strategy as do media platforms. That is, a full sample of previously discovered actors (channels or chat groups) needs to be collected, which can then be queried for their content in a second step.

Time-related challenges

Assessing temporal dynamics is crucial for obtaining reliable data and valid results. Previous research has shown that access to digital trace data diminishes over time (Buehling, 2023; Schatto-Eckrodt, 2022; Walker, 2017), reducing data quality and possibly impairing subsequent results. The general evolution of issues impedes the content-based detection of specific content over time. This effect is amplified in research on contentious political communication, such as the propagation of conspiracy theories, which are altered and reposed throughout their life cycles. Furthermore, the actors involved in such communication might leave the public arena for a variety of reasons (Sillaber, Chimiak-Opoka, & Breu, 2013). In the

following, the temporal complications requiring consideration are grouped at the platform, individual user, and issue context levels, which are mutually intertwined. These challenges might occur either because of temporal changes at the time of content creation or the time lag between content creation and data collection.

Platform and medium level

At the platform level, data content changes and deterioration are determined by factors rooted in platform governance and architecture. Every social media platform differs in content quality standards, moderation practices, and enforcement capabilities, which are themselves context and time specific. A sophisticated framework of acceptable and unwanted user behaviors has been implemented on most platforms over time (Gorwa, 2019), although differences may arise from various platform-internal and -external factors. Platforms' codes of conduct can be enforced through content de-amplification, content deletion, and temporary/permanent user bans (Gorwa, 2019).

While content moderation is constantly evolving, most changes are enforced retroactively. Therefore, the time of data collection significantly influences subsequent results. Previously salient content and influential actors become invisible after content moderation, unless they have been previously archived. Consequently, in cross-platform studies, platforms might appear to differ in their historical contents when the only real difference is the retroactive enforcement of varying codes of conduct.

Researcher access to social media data is also limited by the platforms' terms of service, materialized in the API access granted. The access options and information granularity enabled are platform specific and time dependent themselves (van der Vlist, Helmond, Burkhardt, & Seitz, 2022). Changes in API governance can subsequently impair intertemporal within-platform data comparability (Ho, 2020) and complicate cross-platform comparability. Similarly, researcher access to online news content is limited by access to and the completeness of databases. Because of the challenging endeavor of archiving online news, existing commercial databases such as Factiva yield

inconsistent or incomplete records (Blatchford, 2020). Collecting online news data is challenging, requiring human intervention and the use of database combinations to obtain a more complete representation of the content studied (Blatchford, 2020).

Platforms also differ in inscribing content persistence in their technical architectures. While content persistence is afforded on most social media platforms, content creation and data collection on 4chan are shaped by ephemerality. The platforms' automated deletion of threads that decline in activity or become too old implies that posts can only be streamed with a native API. Studies relying on retrospective data collection, for example, to trace the genealogy of conspiracy theories back to 4chan (De Zeeuw, Hagen, Peeters, & Jokubauskaite, 2020), are only possible via third-party archives. Researchers reluctant to use unverifiable third-party archives are impelled to set up a custom archival pipeline and forego historical data (Tuters, Jokubauskaitė, & Bach, 2018).

Issue context

When studied over time, the words identifying a topic under research, or the predefined composition of communicators, can change partly independent of platform and individual user actions. While this is true for most issues in online communication, this dynamic becomes particularly clear in the collection of conspiracy-related content. Discourse about a specific conspiracy might change its focus, evolve, and absorb adjacent topics in the course of the legitimization strategies brought forward by its proponents. For example, the QAnon conspiracy theory relies on historic anti-Semitic narratives and more recent conspiracy theories, such as Pizzagate, while constantly being updated by so-called Q Drops (Garry, Walther, Mohamed, & Mohammed, 2021).

Scholars applying an actor-based data collection strategy might encounter dynamics in actor compositions, caused by either a natural dynamic of visibility in the discourse (e.g., Rogers, 2020) or by the actors' choices to leave the public arena. If not considered, this could lead to data undersampling in periods when predefined actors' communication was not prevalent, although there was a discourse about the topic. A naturally evolving turnover of

dominant speakers becomes even more drastic when actors delete their profiles on a platform after withdrawing from a debate or retiring from their careers, as this might imply a retroactive deletion of all of their past communication (Bachl, 2018).

Individual user level

At the individual user level, platform and time dependence also manifests in social media posts' content and ephemerality. Previous studies have shown that voluntary post deletion is highly prevalent, especially if the post involves a topic deemed as socially undesirable, such as bullying, profane language, and intoxicant use (Almuhimedi, Wilson, Liu, Sadeh, & Acquisti, 2013). User-driven content moderation in the realm of a personal social media page (Gagrčin, 2022) can also result in content ephemerality.

Walker (2017) not only shows that data quality is inversely correlated to the time lag between social media content creation and data collection but that ephemerality in political communication also depends on the contentiousness of the topics discussed. As an explanation for this, Bastos (2021) proposes the possible regret of posting subprime-quality content. Neubaum and Weeks (2023) recognize message ephemerality as an affordance, allowing individuals to voice political opinions they perceive as possibly harmful to themselves if archived forever. In this respect, ephemerality can be used in a political strategy of deliberate provocation and polarization through contentious, manipulated, or illegal content (Münch, 2021). Buehling (2023) shows that message ephemerality resulting from post deletions in conspirational chats potentially biases computational content and social network analysis results. Individual user behavior on specific platforms also changes as a result of platform governance, as the use of certain words might trigger content deletion or account bans. The same topic might undergo a change in characteristic terms as users adapt to and evade content moderation efforts by using deliberate misspellings or dog whistles (Moran, Grasso, & Koltai, 2022).

Users might further adapt their posting behaviors to the structural content ephemerality built into platforms' architectures. On 4chan, for

example, users circumvent automated message deletion by self-archival via specific post structures, such as *general posts* (Tuters, Jokubauskaitė, & Bach, 2018), which need to be considered in data collection.

Equivalent data collection across platforms and time

Cross-platform studies not only better align with the double interrelated nature of digital communication; they also enable the assessment of platform architecture influences on communication and mobilization patterns itself (Matassi & Boczkowski, 2023; Pearce et al., 2020). However, comparability between and across platforms is a severe issue, as the same objects might not be available or might have different meanings across platforms (Rogers, 2017a, 2019). In addition, the time-related and cultural characteristics described above lead to a multidimensional comparative setting (see Figure 1), which considerably impedes data collection.

Comparative research more generally approaches this challenge with the concept of *functional equivalence* (Kolb, 2002; Wirth & Kolb, 2004) – the objects or units do not have to be equal across several system contexts, but “the functionality of the research objects within the different system contexts must be equivalent” (Wirth & Kolb, 2004, p. 88) that is, provide a “common basis of the comparison” (Kolb, 2002, p. 4). Scholars distinguish between construct, item, and method equivalence. Construct equivalence refers to the theoretical construct of interest and whether

it can be considered equivalent across several systems, such as platforms or cultures. Item equivalence considers whether single items, such as data collection search terms, lead to equivalent content across contexts and how this can be ensured. Method equivalence refers to the entire research process, namely, to an equivalent selection of analysis units (sample, e.g., actor and content units), application of the research instruments (e.g., codebooks and dictionaries), and procedures at the administrative level (Kolb, 2002; Wirth & Kolb, 2004). Using the example of conspiracy-related research and the framework of actor- and content-based approaches (see Section 2), in the following, we discuss data collection strategies to enhance *equivalent data collection* for cross-platform and platform-comparative studies, as well as for time- and culture-sensitive studies.

Actor-based strategies

Whether an actor-based approach relying on a priori defined actors as units and starting points of data collection across platforms is possible depends fundamentally on platform features and user choice. It requires platforms providing persistent identification mechanisms, user choices enabling open access, and communication units being clearly attributable to an identifiable actor. This content attributability to individual actors can also differ in preciseness. In platforms such as YouTube and Facebook, individual author information is available per communication unit (e.g., a video or post). In online media, units such as articles are regularly attributed to the medium,

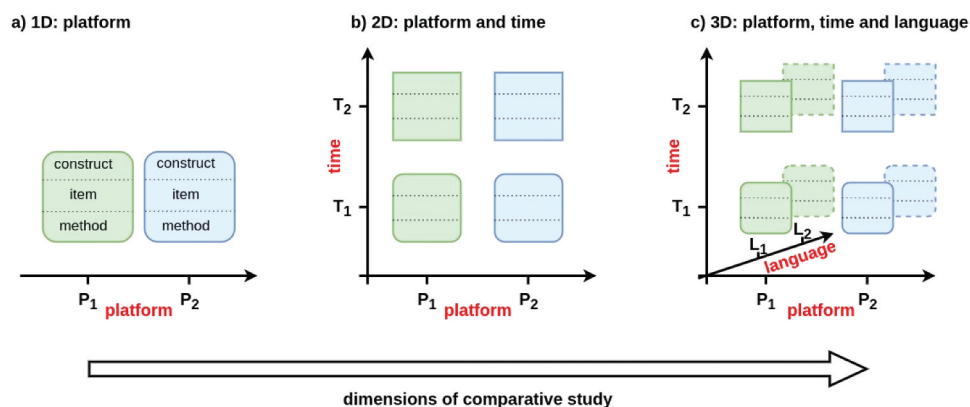


Figure 1. Dimensions comparative study.

even if each article can be written by different authors.

Comparative research across platforms with an actor-based strategy is thus only viable for non-anonymous platforms at the individual or aggregate level (e.g., a full medium and a full chat). In terms of data access, the platforms would need to enable a search per actor, and for online media, the sites would have to provide archived content, or other archives would have to be available for search and data collection. Regarding the example of conspiracy-related communication, an actor-based strategy for data collection would mean that group-based platforms, such as 4chan, with their limited actor identifiability, are excluded; this is difficult to justify, though, given these platforms' relevance for this type of content.

Regarding construct equivalence, the challenge then is to select actors who – across platforms, time, and cultural contexts – function to represent comparable conspiracy-related actors. For the individual-level analysis in distinct cultural contexts, one strategy can be to select the same single actors who are active with accounts on several platforms in the same time span, also facilitating comparability at the method level. This selection will generally be most viable for actors with a higher degree of institutionalization, such as party actors or politicians; this is because they deliberately enact their public voices through several but identifiable accounts, often also directly referencing their various platform-specific online appearances in their communication and cross-posting their content to maximize reach and impact (Bossetta & Schmøkel, 2023). In the field of conspiracy-related communication, however, ideological entrepreneurs (Hyzen & den Bulck, 2021) also tend to self-brand in a recognizable way using their clear names or brand pseudonyms, and they cross-mention their appearances on several platforms or directly link to their various accounts across platforms.

However, in the context of political contention, we must consider that actors in this field are likely less institutionalized, more heterogeneous, and more difficult to detect across platforms, time, and cultures. A second general option is to resort to actor types and pursue a design in which the actors chosen represent the same characteristics and systemic functions

across platforms, time, or cultures. This could be, for example, actors who share a comparable functional role (e.g., as hyperpartisan media actors or online influencers), a comparable position in a shared field (e.g., based on activity or engagement metrics, such as mentions; or based on network metrics), and other meta-data-based similarities (McNerney et al., 2022).

With respect to item and method equivalence, functionally comparable data collection would profit from actor content that is persistent and accessible historically, enabling comparisons across actors and time. Our overview has shown that content persistence is a considerable challenge that can go as far as full actor accounts vanish as a result of platform governance or individual user decisions. However, the actor-based strategy could be particularly fruitful in carving out different communication styles and forms of conspiracy-related content at different communication venues.

Content-based strategies

For content-based strategies, we focus on approaches that use case-specific keywords, denominators (e.g. hashtags), or comprehensive dictionaries for data collection. Keywords are words with purposive meaning that act “as the key to a cipher or code” (Rogers, 2017b, p. 82, following the New Oxford American Dictionary). Dictionaries consist of keyword sets that can be accompanied by a set of rules (van Atteveldt, Welbers, & van der Velden, 2019). As Rogers (2017b, p. 83) highlights, keywords can be “parts of programmes, anti-programmes or efforts at neutrality,” which should be considered when designing queries and dictionaries. For our example of conspiracy-related content, this means that keywords need to equally capture content contributing to the narration of conspiracy theories (programs), content that challenges these narrations or contributes to counter-narration and debunking (anti-programs), and content that neutrally relates to both.

These content-based strategies must always be adjusted to account for platforms' architectures and use cultures. At the conceptual level, the question is which terms are key to, for example,

a specific conspiracy-related discourse and the extent to which these keywords differ across platforms and time and require adjustments for equivalent data collection. While keyword and dictionary construction is fundamental, the influences of keyword selection and validation are far less acknowledged and lack standardization (Mahl, von Nordheim, & Guenther, 2022). Studies on conspiracy theories often use a small or event-specific set of keywords, often without explicit validation (Bruns, Hurcombe, & Harrington, 2022; Starbird, 2017; Zeng & Schäfer, 2021).

Aiming for a broader collection of content across platforms and time, studies need to acknowledge the different communication styles and use cultures on various platforms and the potential changes in relevant terms across time and language areas. This is particularly relevant in the context of conspiracy-related content, which has been shown to evolve across time, depending on recurring events and the platform-specific and cultural embeddings of contentious narratives.

Recent approaches put more emphasis on dictionary creation, expansion, and validation, such as the dictionary coherence, augmentation, validation, and analysis (CAVA) approach developed by van Atteveldt and Chan (2022); this approach allows researchers to construct data-driven dictionaries by determining words semantically similar to preselected keywords as measured through word vector representations (Bojanowski, Grave, Joulin, & Mikolov, 2017). The CAVA approach also offers means of validating a dictionary (construct equivalence). An example of a validated dictionary for conspiracy-related research is the RPC-Lex (Puschmann, Karakurt, Amlinger, Gess, & Nachtwey, 2022), consisting of 10,829 unique keywords for the study of right-wing populist conspiracy (RPC) in German-language texts.

To adapt dictionary development for cross-platform research, we propose that dictionary validation and expansion should be based on *platform-specific corpora*. This accounts for platform-specific use cultures and potential differences in keywords representing the same construct, albeit with likely differences in the share of programs, anti-programs, and neutrality and differences because of platform-specific styles. This can be achieved through a workflow

starting with a theoretically defined seed dictionary, which is computationally expanded on the basis of relevant keywords extracted from platform-specific text samples. The equivalence of cross-platform data collection at the construct level can then be pursued by combining the platform-based expanded and validated dictionaries into a single dictionary to be used across all data corpora.

If the research aim involves understanding variations across multiple dimensions (e.g., in the narration of conspiracy theories that have been highlighted for their time dependency and cultural differences), the approach can also account for semantic changes in concepts over time and national or cultural contexts. This can be ensured by conducting dictionary development and expansion based on keywords derived from time-, language-, and platform-specific text samples.

Validating single keywords and their translations in different cultural contexts is crucial to ensure item equivalence (Lind, Eberl, Heidenreich, & Boomgaarden, 2019). Rather than relying solely on keyword translation, the data-driven keyword expansion approach proposed here can capture new and comparatively equivalent keywords that additionally reveal language patterns and nuances in different cultural contexts.

Finally, method equivalence entails decisions to be made on the best possible equivalence of analysis units, starting from the question of the extent to which a submission, post, video, or article is functionally equivalent. While a process of dictionary development as described above should foster equivalence in the research instrument, whether it can be applied in the same way depends on platform-specific data collection possibilities, such as differences in search functions per platform (e.g., enabling an unlimited dictionary or limiting the number of keywords), and the database accessible for data collection. Depending on successful context-specific adaptation, content-based strategies could be particularly viable for multi-dimensional comparative studies across platforms, time, and cultural contexts.

Conclusion

Taking the example of conspiracy-related communication online as one form of contentious politics,

this study examined the methodological and practical challenges of equivalent data collection for multidimensional comparative studies across different platforms, time, and cultural embeddings. Interest in cross-platform research has grown recently for its capacity to enrich the theoretical understanding of how platform-specific characteristics and their appropriation influence political communication (Bossetta, 2019; Matassi & Boczkowski, 2023) and its more encompassing approach to a networked digital information ecology (Zannettou et al., 2018). Theoretically, the concepts of digital architecture (Bossetta, 2019) and affordances (Evans, Pearce, Vitak, & Treem, 2017) provide frameworks to facilitate comparative endeavors and help ascertain the platform features and functionalities that can be considered functionally equivalent.

Comparatively less attention has focused on the practical problems of *data access and data collection* online from a cross-platform perspective as, for example, inscribed in platforms' access regimes and dependent on the availability and structures of archived data (but see Burgess & Matamoros-Fernández, 2016; Pearce et al., 2020; Rogers, 2017a).

In addition, ensuring equivalent data collection in comparative studies becomes more challenging when multiple dimensions (e.g., platform, time, and language) are considered. In analogy to other computational problems arising with high-dimensional data (Hastie, Tibshirani, & Friedman, 2009), this phenomenon could be called the *curse of dimensionality of comparative data collection*. From our insights into the distinct architectures and use cultures of several platforms/communication venues and time-related data collection challenges, we derived a discussion of actor- and content-based strategies – exemplified with the case of collecting conspiracy-related content online – and how they can be adjusted to platform peculiarities, cultural embeddings, and temporal dynamics. To tackle the curse of dimensionality, our discussion highlights the following crucial points for designing comparative data collection in studies on political communication and contention:

Whether a study is interested in what Rogers (2017a) calls *medium research*, that is, the influence of platform architectures on practices and content, in the social phenomenon (e.g., the distribution of

conspiracy-related content), or in both is the guiding step for the theoretical and methodological design of a multidimensional comparative study. Then, the theoretical construct to be measured must be defined in consideration of the potential differences between platforms and communication venues. We have exemplified this with the construct of conspiracy-related content, which is based on definitions of conspiracy theories but considers programs, anti-programs, and neutral content.

Sensitivity toward distinct use cultures on different platforms at different times and in specific contexts (e.g., language areas) and what these factors might mean for the study object (Rogers, 2017a) needs to then be translated into the study design and data collection. This includes a discussion of (a) the important platforms for the research question at hand, (b) the unit of analysis that is most relevant to the research question, e.g. rather an actor- or content-based unit or a combination; and (c) the analysis units that are actually available across all dimensions.

Based on this, a data collection and query strategy can be designed. We have highlighted some potentials of actor- and content-based strategies, especially ways to adjust dictionary-based strategies to platforms, use cultures, and time. Approaches relying on a small set of keywords or specific hashtags need to consider that they can be deliberately used when referring to a certain topic but can also be deliberately avoided (Massanari, 2017). Computationally expanded dictionaries must also ensure that they can capture the concept fully and equivalently across platforms, time, and context, as suggested by the platform-, time-, and language-dependent expansion procedure in our example.

However, functional equivalence does not mean neutralizing all platform differences; it means finding the analysis units that best represent a certain function (see also Pearce et al., 2020). This includes the fact that different platforms might necessitate different sampling strategies, both with respect to the units of analysis (actors, tweets, etc.) and the overall sampling strategy (from the full archive, sub-domains, etc.). Overall, research needs to reflect on how limitations and differences might influence the collected data and impede comparisons or can be accounted for through the analytical strategy.

This extends to the challenges posed by digital media data's ephemerality and differences in data accessibility for platforms and websites, demanding individual researcher decisions and trade-offs concerning costs, time expenditure, and completeness. Furthermore, large-scale social media data collection is always restricted to publicly available data, leaving vast amounts of discussion and mobilization occurring in closed communication channels in the dark (Burgess & Matamoros-Fernández, 2016). These limitations can be partly circumvented through emerging data collection designs relying on data donations or data sharing by researchers, both of which bear their own legal and ethical pitfalls (Assenmacher et al., 2022; van Driel et al., 2022).

All these aspects call for careful pre-studies and validations of the data collection process and a thorough discussion of its limitations. The increasing multidimensionality and complexity of collecting digital communication data remain challenges, particularly in the area of political contention. However, we consider current efforts to establish social media archives, enable data donations and sharing, and develop computational methods, such as the proposed content-based approach for the computational expansion of dictionaries derived from platform-, time-, and language-specific corpora, as promising avenues to further facilitate comparative studies in a constantly changing and fluid field.

Acknowledgments

This study has benefited from discussions with numerous colleagues, particularly Christian Baden and those in the jointly organized workshop *Advancing Cross-Platform Research in Political Social Media Communication* at the Weizenbaum Institute for the Networked Society in Berlin in 2022.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by grants from the German Federal Ministry of Education and Research (grant numbers 13N16049 [in the context of the call for proposals Civil Security – Societies in Transition] and 16DII135 [in the

context of the Weizenbaum Institute]). Dominik Schindler acknowledges support from the EPSRC (PhD studentship through the Department of Mathematics at Imperial College London) and from the Weizenbaum Institute (Research Fellowship).

Notes on contributors

Annett Heft heads the research group Dynamics of Digital Mobilization at the Weizenbaum Institute for the Networked Society, Berlin, and is a 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 and right-wing communication infrastructures, transnational communication, as well as quantitative research methods and computational social science.

Kilian Buehling is a research associate at the Freie Universität Berlin and the Weizenbaum Institute for the Networked Society. He studied economics at Technische Universität Dresden, and his research interests are information diffusion, transnational communication processes, and network dynamics of anti-democratic and conspiracy theory groups.

Xixuan Zhang is a research associate and doctoral candidate at the Freie Universität Berlin. She studied media and political communication at Freie Universität Berlin and media informatics at Technische Universität Berlin. Her research interests are in the fields of digital activism, online discourse, and the networked public sphere. She also researches the application of computational methods, ranging from text mining to network analysis and machine learning.




Juni Schindler is a PhD candidate in applied mathematics at Imperial College London. Drawing from network science and machine learning, he has developed methods for the computational social sciences and analyzed conspiracy-related online communication as a research fellow at the Weizenbaum Institute Berlin. He obtained his MSc in applied mathematics from Imperial College and his MA degree in digital media from Goldsmiths, University of London.

Miriam Milzner is a research associate and doctoral candidate at Freie Universität Berlin and the Weizenbaum Institute for the Networked Society. Prior to her PhD, she received her MA degree in journalism and communication studies from Freie Universität Berlin. Her research interests include the logics of digital information infrastructures, the hybrid media sphere, and the dynamics of (coordinated) information manipulation.

ORCID

Annett Heft  <http://orcid.org/0000-0001-6637-795X>

Kilian Buehling  <http://orcid.org/0000-0002-5244-7547>

Xixuan Zhang  <http://orcid.org/0000-0002-4636-3881>
 Juni Schindler  <http://orcid.org/0000-0002-8728-9286>
 Miriam Milzner  <http://orcid.org/0009-0007-1705-0281>

Declaration of conflicting interests

The authors do not report any potential conflicts of interest.

References

- Allington, D., & Joshi, T. (2020). "What others dare not say": An antisemitic conspiracy fantasy and its YouTube audience. *Journal of Contemporary Antisemitism*, 3(1), 35–54. doi:10.26613/jca/3.1.42
- Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., & Acquisti, A. (2013). Tweets are forever: A large-scale quantitative analysis of deleted tweets. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 897–908. 10.1145/2441776.2441878
- Assenmacher, D., Weber, D., Preuss, M., Calero Valdez, A., Bradshaw, A. . . Grimme, C. (2022). Benchmarking crisis in social media analytics: A solution for the data-sharing problem. *Social Science Computer Review*, 40(6), 1496–1522. doi:10.1177/08944393211012268
- Bachl, M. (2018). An evaluation of retrospective Facebook content collection. In P. Müller, S. Geiss, C. Schemer, T. K. Naab, & C. Peter (Eds.), *Dynamische Prozesse der öffentlichen Kommunikation: Methodische Herausforderungen* (pp. 57–72). Herbert von Halem Verlag. doi:10.31219/osf.io/6txge
- Baden, C., & Sharon, T. (2021). BLINDED by the LIES? Toward an integrated definition of conspiracy theories. *Communication Theory*, 31(1), 82–106. doi:10.1093/ct/ctaa023
- Barkun, M. (2013). A culture of conspiracy: Apocalyptic visions in contemporary America. *Comparative studies in religion and society* (2nd ed., Vol. 15). Berkeley: University of California Press.
- Bastos, M. (2021). This account doesn't exist: Tweet decay and the politics of deletion in the Brexit debate. *American Behavioral Scientist*, 65(5), 757–773. doi:10.1177/0002764221989772
- Bergmann, E. (2018). *Conspiracy & populism: The politics of misinformation*. Palgrave Macmillan. doi:10.1007/978-3-319-90359-0
- Bevensee, E., & Ross, A. R. (2018). The alt-right and global information warfare. 2018 *IEEE International Conference on Big Data (Big Data)*, 4393–4402. 10.1109/BigData.2018.8622270
- Blatchford, A. (2020). Searching for online news content: The challenges and decisions. *Communication Research & Practice*, 6(2), 143–156. doi:10.1080/22041451.2019.1676864
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. doi:10.1162/tac1_a_00051
- Bossetta, M. (2019). *The digital architectures of social media: Platforms and participation in contemporary politics* (Doctoral Thesis). University of Copenhagen, Faculty of Social Sciences, Copenhagen.
- Bossetta, M., & Schmøkel, R. (2023). Cross-platform emotions and audience engagement in social media political campaigning: Comparing candidates' Facebook and Instagram images in the 2020 US election. *Political Communication*, 40(1), 48–68. doi:10.1080/10584609.2022.2128949
- Bruns, A., Harrington, S., & 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
- Bruns, A., Hurcombe, E., & Harrington, S. (2022). Covering conspiracy: Approaches to reporting the COVID/5G conspiracy theory. *Digital Journalism*, 10(6), 930–951. doi:10.1080/21670811.2021.1968921
- Buehling, K. (2023). Message deletion on Telegram: Affected data types and implications for computational analysis. *Communication Methods and Measures*, 1–23. doi:10.1080/19312458.2023.2183188
- Burgess, J., & Matamoros-Fernández, A. (2016). Mapping sociocultural controversies across digital media platforms: One week of #gamergate on Twitter, YouTube, and Tumblr. *Communication Research & Practice*, 2(1), 79–96. doi:10.1080/22041451.2016.1155338
- Busbridge, R., Moffitt, B., & Thorburn, J. (2020). Cultural Marxism: Far-right conspiracy theory in Australia's culture wars. *Social Identities*, 26(6), 722–738. doi:10.1080/13504630.2020.1787822
- Butter, M., & Knight, P. (2020). General Introduction. In M. Butter & P. Knight (Eds.), *Routledge handbook of conspiracy theories* (pp. 1–8). Routledge. doi:10.4324/9780429452734-0
- De Zeeuw, D., Hagen, S., Peeters, S., & Jokubauskaite, E. (2020). Tracing normification: A cross-platform analysis of the QAnon conspiracy theory. *First Monday*, 25(11). doi:10.5210/fm.v25i11.10643
- Ekman, M. (2022). The great replacement: Strategic mainstreaming of far-right conspiracy claims. *Convergence: The International Journal of Research into New Media Technologies*, 28(4), 1127–1143. doi:10.1177/13548565221091983
- Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, 22(1), 35–52. doi:10.1111/jcc4.12180
- Fair, G., & Wesslen, R. (2019). Shouting into the void: A database of the alternative social media platform Gab. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, 608–610. 10.1609/icwsm.v13i01.3258
- Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, 35(4), 665–668. doi:10.1093/oxfordhb/9780199760107.013.0018

- Frischlich, L., Schatto-Eckrodt, T., & Völker, J. (2022). *Rückzug in die Schatten? Die Verlagerung digitaler Foren zwischen Fringe Communities und "Dark Social" und ihre Implikationen für die Extremismusprävention* (CoRE-NRW Kurzgutachten, 4). Bonn: Bonn International Centre for Conflict Studies (BICC) gGmbH.
- Gagrčin, E. (2022). *Your social ties, your personal public sphere, your responsibility: How users construe a sense of personal responsibility for intervention against uncivil comments on Facebook*. New Media & Society. doi:10.1177/14614448221117499
- Gallagher, A., Davey, J., & Hart, M. (2020). *Genesis of a conspiracy theory: Key trends in qanon activity since 2017* (ISD Reports). Institute for Strategic Dialogue.
- Garry, A., Walther, S., Mohamed, R., & Mohammed, A. (2021). Qanon conspiracy theory: Examining its evolution and mechanisms of radicalization. *Journal for Deradicalization*, 26, 152–216.
- Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, 22(6), 854–871. doi:10.1080/1369118X.2019.1573914
- Graham, T., Bruns, A., Zhu, G., & Campbell, R. (2020). *Like a virus: The coordinated spread of coronavirus disinformation*. Canberra: The Australia Institute.
- Harris, K. R. (2023). Conspiracy theories, populism, and epistemic autonomy. *Journal of the American Philosophical Association*, 9(1), 21–36. doi:10.1017/apa.2021.44
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. doi:10.1007/978-0-387-84858-7
- 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
- Heft, A., & Buehling, K. (2022). Measuring the diffusion of conspiracy theories in digital information ecologies. *Convergence: The International Journal of Research into New Media Technologies*, 28(4), 940–961. doi:10.1177/13548565221091809
- Herasimenka, A. (2019). *Political organisation, leadership and communication in authoritarian settings: Digital activism in Belarus and Russia* (PhD thesis). University of Westminster, London. doi:10.34737/qy763
- Ho, J. C.-T. (2020). How biased is the sample? Reverse engineering the ranking algorithm of Facebook's graph application programming interface. *Big Data & Society*, 7(1), 205395172090587. doi:10.1177/2053951720905874
- Hyzen, A., & den Bulck, H. V. (2021). Conspiracies, ideological entrepreneurs, and digital Popular culture. *Media and Communication*, 9(3), 179–188. doi:10.17645/MAC.V9I3.4092
- Jasser, G., McSwiney, J., Pertwee, E., & Zannettou, S. (2023). 'Welcome to #GabFam': Far-right virtual community on Gab. *New Media & Society*, 25(7), 1728–1745. doi:10.1177/14614448211024546
- Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, 96(3), 109–126. doi:10.2307/2564659
- Knight, P. (2008). Outrageous conspiracy theories: Popular and official responses to 9/11 in Germany and the United States. *New German Critique*, 103(1), 165–193. doi:10.1215/0094033X-2007-024
- Kolb, S. (2002, May 24). *The functional equivalence concept as an approach to allow a Broad level of comparability in general media statistic*. Dortmund, Germany: ENTIRE-conference.
- Leal, H. (2020). Networked disinformation and the lifecycle of online conspiracy theories. In M. Butter & P. Knight (Eds.), *Routledge handbook of conspiracy theories* (pp. 497–511). Routledge. doi:10.4324/9780429452734-4_9
- Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. (2019). When the Journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, 13(21), 4000–4020.
- Mahl, D., Schäfer, M. S., & Zeng, J. (2023). Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research. *New Media & Society*, 25(7), 1781–1801. doi:10.1177/14614448221075759
- Mahl, D., von Nordheim, G., & Guenther, L. (2022). Noise pollution: A multi-step approach to assessing the consequences of (not) validating search terms on automated content analyses. *Digital Journalism*, 11(2), 298–320. doi:10.1080/21670811.2022.2114920
- Mahl, D., Zeng, J., & Schäfer, M. S. (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), 7(2). doi:10.1177/20563051211017482
- Massanari, A. (2017). #gamergate and the fapping: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3), 329–346. doi:10.1177/1461444815608807
- Matassi, M., & Boczkowski, P. J. (2023). *To know is to compare. Studying social media across nations, media, and platforms*. The MIT Press.
- McNerney, H. W., Spann, B., Mead, E. L., Kready, J., Marcoux, T., & Agarwal, N. (2022). Assessing the influence and reach of digital activity amongst far-right actors: A comparative evaluation of mainstream and 'free speech' social media platforms. *For(e)Dialogue*, 4(1). doi:10.21428/e3990ae6.60c47409
- Moran, R. E., Grasso, I., & Koltai, K. (2022). Folk theories of Avoiding content moderation: How vaccine-opposed influencers amplify vaccine opposition on Instagram. *Social Media + Society*, 8(4), 8(4). doi:10.1177/20563051221144252
- Münch, F. V. (2021). Es war einmal ein Tweet. Wie transparent soll Wahlkampf sein? In V. Hofmann & M. C. Kettemann (Eds.), *Plattformregulierung im Superwahljahr 2021: Ergebnisse rechtswissenschaftlicher, sozialwissenschaftlicher und datenwissenschaftlicher Studien zu Parteien und*

- Plattformen im Bundestagswahlkampf* (Vol. 2.0.0, pp. 82–89). Hamburg: SSOAR - GESIS Leibniz Institute for the Social Sciences.
- Neubaum, G., & Weeks, B. (2023). Computer-mediated political expression: A conceptual framework of technological affordances and individual tradeoffs. *Journal of Information Technology & Politics*, 20(1), 19–33. doi:10.1080/19331681.2022.2028694
- Nissenbaum, A., & Shifman, L. (2017). Internet memes as contested cultural capital: The case of 4chan's/b/board. *New Media & Society*, 19(4), 483–501. doi:10.1177/1461444815609313
- Pearce, W., Özkula, S. M., Greene, A. K., Teeling, L., Bansard, J. S., Omena, J. J., & Rabello, E. T. (2020). Visual cross-platform analysis: Digital methods to research social media images. *Information, Communication & Society*, 23(2), 161–180. doi:10.1080/1369118X.2018.1486871
- Prakasam, N., & Huxtable-Thomas, L. (2021). Reddit: Affordances as an enabler for shifting loyalties. *Information Systems Frontiers*, 23(3), 723–751. doi:10.1007/s10796-020-10002-x
- Puschmann, C., Karakurt, H., Amlinger, C., Gess, N., & Nachtwey, O. (2022). RPC-Lex: A dictionary to measure German right-wing populist conspiracy discourse online. *Convergence: The International Journal of Research into New Media Technologies*, 28(4), 1144–1171. doi:10.1177/13548565221109440
- Quandt, T. (2018). Dark participation. *Media and Communication*, 6(4), 36–48. doi:10.17645/mac.v6i4.1519
- Rogers, R. (2017a). Digital methods for cross-platform analysis. In J. Burgess, T. Poell, & A. E. Marwick (Eds.), *The SAGE handbook of social media* (pp. 91–110). SAGE.
- Rogers, R. (2017b). Foundations of digital methods: Query design. In M. T. Schäfer & K. van Es (Eds.), *The datafied society. Studying culture through data* (pp. 75–94). Amsterdam University Press. doi:10.1515/9789048531011-008
- Rogers, R. (2019). *Doing digital methods* (1st ed.). Los Angeles: SAGE Publications.
- Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3), 213–229. doi:10.1177/0267323120922066
- Rooke, M. (2021). Alternative media framing of COVID-19 risks. *Current Sociology*, 69(4), 584–602. doi:10.1177/00113921211006115
- Samory, M., & Mitra, T. (2018). The government spies using our webcams. *ACM on Human-Computer Interaction*, 2 (CSCW), 1–24. doi:10.1145/3274421
- Schatto-Eckrodt, T. (2022). Hidden biases – the effects of unavailable content on Twitter on sampling quality. In J. Jünger, U. Gochermann, C. Peter, & M. Bachl (Eds.), *Grenzen, Probleme und Lösungen bei der Stichprobenziehung* (pp. 178–195). Köln: Herbert von Halem Verlag.
- Schulze, H., Hohner, J., Greipl, S., Girgnhuber, M., Desta, I., & Rieger, D. (2022). Far-right conspiracy groups on fringe platforms: A longitudinal analysis of radicalization dynamics on Telegram. *Convergence: The International Journal of Research into New Media Technologies*, 28(4), 1103–1126. doi:10.1177/13548565221104977
- Sillaber, C., Chimiak-Opoka, J., & Breu, R. (2013). Understanding and modeling usage decline in social networking services. In L. Uden, F. Herrera, J. B. Pérez, & J. M. C. Rodríguez (Eds.), *7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing* (Vol. 172, pp. 377–388). Springer Berlin Heidelberg. doi:10.1007/978-3-642-30867-3_34
- Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. *Proceedings of the 11th International Conference on Web and Social Media*, Montreal (pp. 230–239). doi:10.1609/icwsm.v11i1.14878
- Tuters, M., & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, 22(12), 2218–2237. doi:10.1177/1461444819888746
- Tuters, M., Jokubauskaitė, E., & Bach, D. (2018). Post-truth protest: How 4chan cooked up the pizzagate bullshit. *M/C Journal*, 21(3). doi:10.5204/mcj.1422
- van Atteveldt, W., & Chan, C.-H. (2022). CAVA: An open source R toolkit for dictionary coherence. *Adaptation, Validation, and Analysis*. R. <https://github.com/vanatteveldt/CAVA>
- van Atteveldt, W., Welbers, K., & van der Velden, M. (2019). Studying political decision making with automatic text analysis. In *Oxford research encyclopedia of politics*. Oxford University Press. doi:10.1093/acrefore/9780190228637.013.957.
- van der Vlist, F. N., Helmond, A., Burkhardt, M., & Seitz, T. (2022). API governance: The case of Facebook's evolution. *Social Media + Society*, 8(2), 8(2). doi:10.1177/20563051221086228
- van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures*, 16(4), 266–282. doi:10.1080/19312458.2022.2109608
- Walker, S. (2017). *The Complexity of Collecting Digital and Social Media Data in Ephemeral Contexts* [Thesis]. <https://digital.lib.washington.edu/443/researchworks/handle/1773/40612>
- Wilson, A. F. (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

- Wirth, W., & Kolb, S. (2004). Designs and methods of comparative political communication research. In F. Esser & B. Pfetsch (Eds.), *Comparing political communication. Theories, cases, and challenges* (pp. 87–111). Cambridge: Cambridge University Press.
- Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2020). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38(1–2), 98–139. doi:[10.1080/10584609.2020.1785067](https://doi.org/10.1080/10584609.2020.1785067)
- Zannettou, S., Cauleld, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Suarez-Tangil, G. (2018). On the origins of memes by means of fringe web communities. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 188–202. [10.1145/3278532.3278550](https://doi.org/10.1145/3278532.3278550)
- Zeng, J., & Schäfer, M. S. (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](https://doi.org/10.1080/21670811.2021.1938165)