Riding the Spider: A Network-Sampling Framework for Multi-Platform Data Collections

Philipp Kessling / Felix Victor Münch*

Research on the digital networked public sphere is not only hindered by challenges in data access but also by a lack of common standards for describing and implementing data collection independently of the form of access or technologies employed. These challenges are particularly pronounced in cross-platform research. In this article, we propose a network-sampling framework to conceptualize, implement, and document explorative data collections in a generalizable, readily operationalizable, and interoperable way. Building on the theoretically established components of the networked public sphere, the concept of multilayer networks, explorative network sampling, and legal and technical realities of cross-platform data access, we segment the data collection process into four modules: a Connector, a Parser, a Filter, and a Sampler. This framework enables researchers not only to describe their data collection in a precise and reproducible way but also to follow guidelines on for developing interoperable software implementations of these modules or to propose new modules themselves.

Key words: network sampling, cross-platform data collection, networked public sphere, multilayer networks, interoperability, replicability

1. Introduction

The study of digital, networked communication is challenged by a threefold, interconnected increase in complexity (Strippel et al. 2018). First, the media and communication landscape has diversified into a multitude of platforms and their associated affordances (see, e.g., Breiter and Hepp 2018). Second, the methods used to study these platforms have become increasingly complex, partly due to the adoption of approaches from the natural and computer sciences (Berry 2011; Lazer et al. 2020). Third, platforms have repeatedly modified their data access policies and interfaces, often restricting or entirely cutting off access to researchers. These changes pose significant obstacles to maintaining consistent and reliable data sources for longitudinal studies and cross-platform comparisons (Bruns 2019; Freelon 2021). Consequently, there is a pressing need for a formal framework to describe data collections that supports the study of larger structures of the public sphere, that is generalizable across studies, and remains reproducible over time.

Meanwhile, pressing societal issues underscore the need for a birds-eye view of digital publics. These include the fragmentation of democratic, liberal societies into smaller, less interconnected communities, which may contribute to polarization (Brüggemann and Meyer 2023; Esau et al. 2024) as well as the strategic use of mis- and disinformation (Rogers 2023; Quandt et al. 2019), which can erode trust in public institutions and established media (Frischlich and Humprecht 2021). Research that investigates the digital networked public and can capture both its short- and long-term structures, from individual to global scales,

* Philipp Kessling, M.A., Leibniz Institute for Media Research | Hans-Bredow-Institut (HBI), Warburgstraße 30b, 20354 Hamburg, Germany, p.kessling@leibniz-hbi.de, https://orcid.org/0000-0002-1739-446X;
Dr. Felix Victor Münch, GESIS Leibniz-Institut für Sozialwissenschaften, Knowledge Technologies for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Köln, Germany, Felix.Muench@gesis.org, https://orcid.org/0000-0001-8808-6790.

DOI: 10.5771/1615-634X-2026-1-52

is therefore crucial. While studies of short-term effects remain feasible despite the restrictive work-to-rule approach of major platforms toward researcher access (Rau et al., 2025), long-term and large-scale analyses—whether within single platforms (Guan et al. 2022; Münch et al. 2021; Bruns and Moon 2019; Bruns et al. 2017) or across multiple platforms—are severely slowed and rendered prohibitively costly, if not altogether impossible.

This challenge necessitates a shift toward reliable and reproducible data acquisition if research independent and free from conflicts of interest, outside the tightly controlled domains of major platforms—is to remain viable. Current efforts to address this issue, however, largely consist of uncoordinated (though necessary and commendable), single-use attempts to circumvent platform-imposed restrictions. Methods for collecting data at scale, particularly for empirically mapping digital publics both within and across platforms remain fragmented and difficult to reproduce (Wiedemann et al. 2023).

In this article, we propose and exemplify the necessity and feasibility of a generalizable, readily operationalizable framework to conceptualize, implement, and document large-scale data collections that are independent of individual platforms, yet compatible with empirically established theories of the networked public sphere. In the remainder of this article, we discuss theoretical models of networked publics (Bruns 2023; Friemel and Neuberger 2023) and examine their relationship to explorative, network-based sampling of trace data. In doing so, we aim to demonstrate an approach for gathering relevant datasets that

1. integrates a variety of data collection methods in a flexible manner, which is necessary due to the unreliability of platforms, and
2. openly and consistently documents the collection intent and process in a replicable way.

To achieve this, we propose dividing the data collection process into four distinct modules: a Connector, which retrieves research data from platform APIs, scrapers, databases, files, and other sources; a Parser, which extracts meaningful network edges from the retrieved data; a Filter, which selects or rejects edges and nodes based on researcher defined conditions; and a Sampler, which selects edges and nodes for continuation of the sampling process. This modular segmentation allows researchers to describe their data collection approach clearly and reproducibly, while also providing a framework for developing interoperable software implementations of these modules or proposing new ones.

In Section 2 we motivate a multilayer network perspective grounded in theoretical concepts of the (networked) public sphere(s). In Section 3, we introduce approaches to explorative network sampling and assess their suitability for research within this paradigm. Section 4 presents an example case, and in Section 5, we deduce a sampling framework for data collections.

2. Digital, Networked and Public Spheres

Discussing the public sphere today necessarily involves the *Internet*, where most public communication either occurs directly or is documented and reverberated, influencing society. This produces collectable trace data at varying levels of ephemerality, making communication more observable and therefore measurable—a development that contributed to the computational turn in the social sciences (Lazer et al., 2009). Even on early Internet message boards, and increasingly with the emergence of the Blogosphere and the early stages of Twitter in the mid- to late 2000s, previously silent users of media with limited audiences became what Bruns (2008) calls 'produsers'—media users who not only consume media as passive recipients but also produce their own content for audiences across the growing World Wide Web. Combined with the interconnected technical nature of the web, this led to a media landscape shifting from centralized, broadcast-oriented media toward networked, distributed, fine-grained contributions from anyone willing to participate. As a result, members of the general public and non-traditional content producers now share the arena with established media outlets, rendering old theoretical frameworks for understanding the public sphere outdated (Schmidt et al., 2017). In particular, the notion of a virtual stage dominated by normatively guided deliberation processes (Habermas, 1962) is challenged, prompting the need for new conceptual building blocks to describe networked public spheres. For instance, Bruns (2023) and Friemel and Neuberger (2023) define the basic entities in the digital public as follows:

- *Actors*: Humans and/or aggregations of humans in organizations, as well as automated computer systems that interact with humans or with other automated systems, such as chatbots or social bots[1].
- *Content*: Actor-generated content hosted on one or more platforms and shaped by the platforms' affordances.
- *Platforms*: As discussed in Section 4, platforms and their affordances often create barriers that prevent actors and content from leaving them. Nevertheless, substantial cross-platform diffusion occurs (see below), and different platforms play distinctive roles in the public sphere. Historically, the choice of platform for research was largely dictated by data availability, but this approach is far from ideal.

Further, Friemel and Neuberger (2023) theorize link types between actors and content as *production, reception*, or *curation*. Depending on a platform's affordances, these types can be mapped to platform-specific connections. The affordances of a *single* platform already produce a complex interactive system with multiple modalities, interaction types, and entity classes (boyd and Ellison 2007). For example, Facebook supports different types of actors, such as user profiles and organizational pages, as well as various social ties, including 'friendships,' 'follows,' and 'likes.' However, given the observable shift from a few dominant social media platforms (Facebook, Twitter, Instagram) toward a more fragmented landscape—including platforms such as TikTok, Telegram, Bluesky, Mastodon, Threads, and Truth Social—single-platform research can no longer sustain the empirical foundations needed to advance our understanding of the public sphere. This underscores the necessity for a data collection and sampling framework that provides a generalizable, cross-platform, and researcher-definable model of connections between actors and content within their respective public spherules. From communication processes generating the above links, higher order constructs emerge:

Personal Publics: Enabled by connections within and across platforms, communication on social media in general and former Twitter in particular *"is happening in networked, distributed conversations: single tweets forming the basic units […] are bundled […] in the constant stream of information within a personal timeline, filtered via social connections made explicit"* (Schmidt 2014).

Issue Publics: These often materialize as *hashtag publics*. More broadly, they can be understood as transitive intersections of multiple personal publics centered around a specific issue or hashtag (Bruns and Burgess 2011).

Network of Public "Spherules": Personal and issue publics often give rise to longer-lasting figurations (Hasebrink and Hepp 2017; Hepp and Hasebrink 2014), that is, repeating and, over the mid- to long-term, stable meso-scale patterns. These can be regarded as small public spherules in their own right, such as right-wing counter-publics on Telegram or journalistic networks on Twitter.

[1] We use this term broadly to acknowledge that automation exists in social networks, whether benign or malicious.

Combined with the need to account for the temporality of all these constructs, the (digital) public must be understood as a complex system characterized by emergent and dynamic phenomena (Waldherr 2017). These phenomena within networked publics arise from technical and socio-technical interactions, fundamentally based on the diffusion of a single content piece or digital resource over one or multiple steps. Due to the networked public sphere's centralization and structuring by large platforms, many digital resources can be considered to be members of these platforms. For objects residing on a platform, interactions between them can be captured as trace data. However, empirically investigating effects and phenomena in the networked public sphere requires a concrete operationalization of the linkage types under study. Accordingly, Friemel and Neuberger's categorization of *connections* (see above) can be interpreted as abstract roles to which specific linkage types within and across platforms can be mapped (2023). Useful trace data can include, for example:

1. Account-to-account interactions within a given platform (*reception*),
2. Account-to-content links, such as authorship indicators (*production*), and
3. In-platform links or hyperlinks that may direct to content or accounts on other platforms (*curation*).

To model a subset of the networked public sphere, various network models can be applied to facilitate analysis. Within a single platform, when considering only one type of linkage, a simple directed or undirected network may suffice. Extending the model to multiple platforms, and assuming that large groups of nodes represent objects of a single type—such as users of on a single platform—enables the networked public sphere to be represented as a layered network, with one node type per layer (see Kivelä et al., 2014, for an in-depth discussion of different formalizations). Allowing both multiple platforms and different linkage and node types within a single layer results in a network that is both multilayer and multiplex. In some cases, additional actor-to-account matching may be desirable to meet the requirements of multilayer network modeling.

3. Explorative Network Sampling and Web Crawling

To gather empirical insights within these models, networked data is required. Random sampling of content or actors typically produces networks that are too sparse to support inferences about global structures, making network sampling methods necessary. Network sampling involves drawing a sample based on network-like relations and can be applied either for network down-sampling (reducing the size and order of a network) or in an exploratory manner. Exploration usually begins at specific nodes, with their neighborhoods mapped by following edges (Hu and Lau 2013).

The networked public sphere model described above is largely unknown empirically. Random node samples, the network analogue of conventional quantitative social science sampling, are insufficient for revealing structural patterns, as they tend to be too sparse. Instead, subsets of this network *can* be explored from a limited set of digital resources by following, for example, links from content objects to other content or actors. For example, one might examine the timelines or profile pages of a given actor and the content embedded within them. Each content object contains outgoing links to both actors and other content objects—for instance, mentions of other actors or hyperlinks to additional content. By following these outgoing links, the neighborhood of the initial nodes can be discovered. Jost et al. (2023) apply this approach to map actor structures on Telegram.

The most widely known, large-scale samples of digital resource-networks with URL-encoded relations are referred to as crawls, which have long underpinned Internet search engines. Today, several openly accessible crawls exist, such as CommonCrawl ("Common Crawl - Overview" n.d.) which provides regular crawls of the public Internet, with a single crawl encompassing multiple terabytes of website data. Historically, crawlers have been implemented for Internet research, for example, IssueCrawler was used to map pre-platformized versions of the Internet, such as the Blogosphere (Rogers 1996; Rogers 2010).

Although the basic technical approach is similar, the collection strategies used by search engines differ from those required for researching digital platforms and public networked spheres. For instance, search engines regularly revisit digital resources (Wolf et al. 2002) but often omit social media posts. Moreover, websites frequently optimize their content to be discoverable and highly indexed by search engines. In contrast, studying the structure of networked public spheres requires uncovering the internal structures of platforms to effectively navigate the barriers—"fences"—built by these platforms.

The literature documents a wide range of algorithms for exploratory network sampling beyond the exhaustive link-following typically employed by crawlers. Although many strategies are available, each algorithm introduces distinct biases and affords different properties in the resulting sampled network (Leskovec and Faloutsos 2006). For example, classical social network sampling procedures such as snowball sampling, can be adapted for use in URL-networks; however, depending on the breadth of the search (i.e., number of links followed per step) these methods may either become highly dependent on the initial seed set or, with greater breadth, remain overly local (Goodman 1961).

Ricaud et al. (2020) present a generalization of many network sampling algorithms that enables configurable, probability-based sampling incorporating both edge and node metadata. This approach is particularly well-suited for research on digital platforms and, by extension, the networked public sphere. Depending on the research objective, alternative methods, such as breadth first, depth-first tree extraction, random-walk sampling, or their derivatives, may be applied.

However, the theoretically optimal algorithm is not necessarily viable in practice. In the context of sampling networks from platforms, the choice of strategy is constrained by economic viability: some approaches are more costly than others in terms of API calls or the volume of scraping traffic required. Consequently, strategies that yield informative networks with the fewest possible requests are often preferable (Coscia and Rossi 2018). Taken together, selecting an appropriate algorithm and carefully parameterizing it allows for fine-grained control of how far and broadly the amplifying process traverses the overall, yet unknown, network.

4. Cross-Platform Data Modeling and Acquisition in Practice

Research on interplatform phenomena faces substantial challenges in data logistics and the operationalization of units of analysis, as differences in platform affordances hinder the identification of functional equivalences. As Heft et al. (2024) note, *"[…] 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."*

At a technical level of abstraction, however, the Internet can be understood as a collection of interlinked digital resources, with platforms representing large aggregations of such resources, particularly when considering the definition of Very Large Online Platforms

(VLOPS)[2]. All of these resources are identified by a Universal Resource Identifier (*URI*), which on the Web typically takes the form of a Universal Resource Locator (*URL*) (Berners-Lee 1994). URLs enable Hyperlinks (i.e., (hyper)text containing embedded addresses), and can be regarded as the *connective tissue* of the Web, linking resources to one another and data exchange via APIs (Nielsen et al., 1996).

Hyperlinks create network structures between resources that were early on crawled and analyzed by researchers and search engines using network-analytic methods and metrics—such as Pagerank (Brin et al. 1998)—to understand the structure of the Web and rank websites by relevance. The scale of web crawling required for global analyses and rankings is immense; for example, the openly accessible CommonCrawl hosts multiple terabytes of data for a single crawl ("Common Crawl – Overview" n.d.).

Nevertheless, even such efforts fail to capture a rapidly growing segment of the Internet dominated by URLs primarily intended for use via APIs. Unlike the 'traditional' web of websites, APIs are interfaces that enable the fine-grained control over access and thus lend themselves to commodifying that access. The commodification of user generated content on *platforms*, and the resulting shift toward data connections between websites via APIs, was central to early discussions of platformization (Helmond 2015).

These walled gardens expand, as platforms like Instagram or Twitter/X increasingly restrict public data access, for example, by requiring logins for previously public content like tweet replies or Instagram stories. Yet even these environments remain collections of digital resources that largely possess URLs, despite platforms' efforts to channel users toward proprietary apps—thereby inserting additional layers of machines, behavioral analysis, and access control between users and content. This persistence of URLs reflects platforms' own interests in linking to content and steering users from the open Web into their controlled ecosystems.

A technical approach to addressing this power imbalance lies in diversifying data access risks by developing and supporting multiple ways of collecting platform data. Beyond official APIs, mechanisms such as data access requests and web scraping can provide researchers with at least a resource-constrained access. Applying (network) sampling strategies to the URL-based networks described above further mitigates the need for complete data, given the inherently skewed distributions of social media data (Barabasi and Albert 1999). Consequently, this approach enables the interchangeable use of diverse data sources, including APIs, web-scraping, data donations, tracking data, data repositories, and DSA-mandated data access (Ohme et al. 2023). Suitable sampling strategies also allow these sources to be used while reducing the likelihood of triggering automated rate limits and similar restrictions.

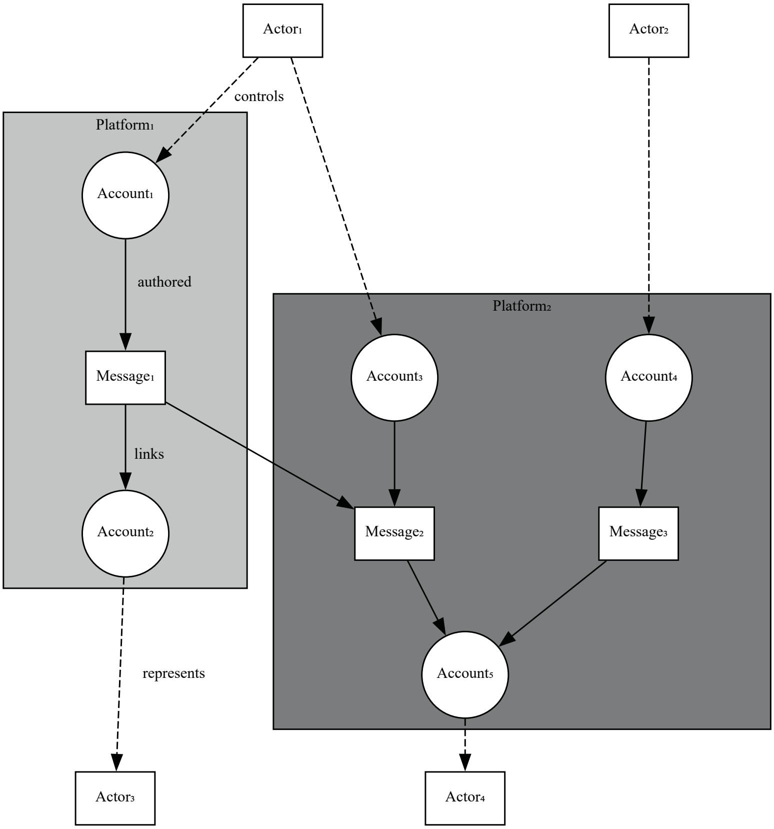
At the same time, this approach also introduces numerous additional degrees of freedom that threaten research reproducibility. It aggravates a foundational problem that hampers the integration of heterogeneous data sources: the imprecise description of data collection processes in much of the existing literature. In many cases—without intending to "name and blame," though examples are readily found—such descriptions are limited to broad statements about which data were collected from which platform and over which time frame, while crucial details such as the exact timing of collection are omitted. When sampling methods such as "snowball" or "random" sampling are mentioned, the specific variant, implementation, and parameterization are often left unspecified. Given this widespread lack of precision, and the substantial downstream effects that even minor changes in data collec-

[2] **V**ery **L**arge **O**nline **P**latform**s** are a legal category for enterprises introduced by the European Union's Digital Services Act (DSA).

tion strategies can produce (Olteanu et al. 2019; Sen et al. 2021), we argue that standards for specifying social media and web data collections are both lacking but urgently needed to improve repeatability and reproducibility. More formal specifications of data collection processes would significantly support researchers in replicating and validating results, as well as in repeating data collections within longitudinal or multi-phase research projects.

Taking an abstract view, the platform-specific connections discussed above can be represented as actor-message-platform networks that rely on URL-based pointers for at least the inter-platform connection$_1$. Consider Figure 1, which illustrates a minimal actor-message network with two edge types spanning two platforms. Beginning the exploration with two known actors—actor$_1$ and actor$_2$—we retrieve the messages authored by these actors via their platform-specific accounts: account$_1$ and account$_3$ for actor$_1$ and account$_4$ for actor$_2$. Actor$_1$ has posted messages on two platforms, platform$_1$ and platform$_2$, whereas actor$_2$ has posted only on platform$_2$. Based on the message texts, we identify two additional actors—actor$_3$ and actor$_4$—who are interacted with through these messages.

Figure 1: Example of an actor-message network in which actors interact with other actors via platform-specific accounts. Messages mediate these interactions and may link to messages on other platforms



Further, the simplified actor-message network illustrates how a network-based sampling of the networked public sphere can be conducted. Given a seed set—the actors near the top of Figure 1 as inputs to the process—it is necessary to specify accounts on a per platform basis. Using a rule set that defines how to interpret message content and metadata, these instructions are executed for each observed message associated with each account. In this way, relationships between messages and additional entities within the networked public sphere are identified—applied to our example, this yields two further accounts for which the same process can be repeated.

To illustrate our argument, we outline a research scenario that highlights both the challenges and potential solutions. For the sake of conciseness, we focus on possible linkages between two platforms; however, the process described here can be extended to additional platforms. We use Telegram and YouTube as example cases, as both are highly relevant platforms for political communication but have not yet been examined as extensively in terms of their structural characteristics as platforms such as Twitter.

Telegram's publicly viewable entities include channels, which can function either as one-to-many broadcasting venues or many-to-many forums with potentially thousands of participants. Replies to individual posts, reactions, and other interactive features can be enabled by the channel's owners or moderators[3]. From this, two relevant and distinct entity types emerge: channels and messages. Channels are controlled by one or more users, and when multiple moderators are active, each message is signed with the corresponding username. The message's text can contain a variety of marked-up references, such as hashtags or hyperlinks. Information—in the form of messages—can be easily amplified within Telegram, as channels can repost messages from other channels. This reposting process replicates the original message within the new channel's context while providing a backlink to the original message in the source channel.

Within YouTube, entities exhibit similar interactions, with channels able to post videos, though they lack the ability to repost content. Between the two platforms, directed connections exist from Telegram to YouTube, as hyperlinks embedded in Telegram messages point to YouTube resources. These connections can reference either a specific video or a channel (see Figure 2 (a)).

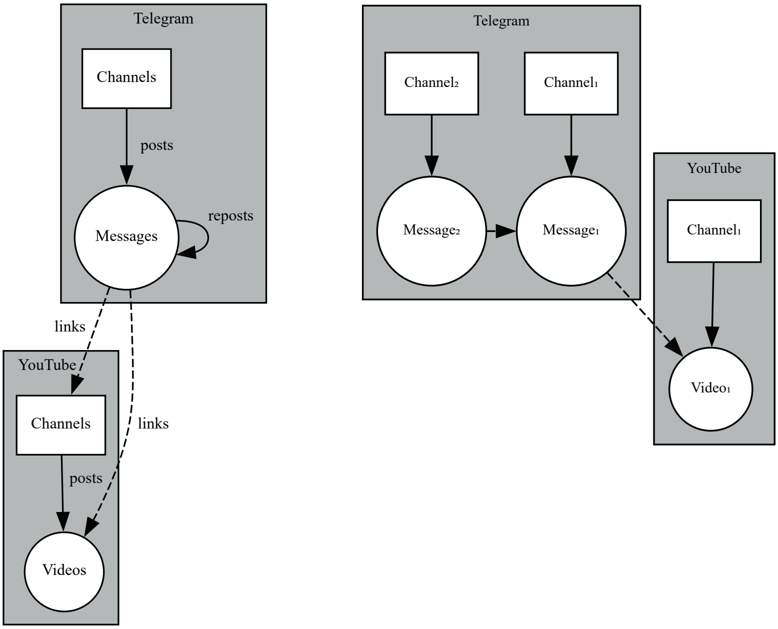
For linkages tracing the diffusion of information within Telegram, we use resharing connections captured in the trace data. For example, as shown in Figure 2 (b): a Telegram channel$_1$ posts a message$_1$ which is subsequently amplified by channel$_2$. Channel$_2$ thereby generates a message$_2$ which links back to message$_1$. Message$_1$ also contains a reference to a YouTube video$_1$ hosted on YouTube channel$_1$. In this way, Telegram channel$_1$ amplifies information from YouTube channel$_1$ and is, in turn, amplified by channel$_2$. Zooming out from this minimal example, since channels have more than one message and introducing more channels into our considerations, reveals a latent network that connects actors, accounts or channels on specific platforms and content objects which allow for a near infinite amount of connections.

[3] For more information on Telegram channels, see https://core.telegram.org/api/channel.

Figure 2: Structural diagrams of possible scenarios of information diffusion within and between the two platforms, Telegram and YouTube, allow for a virtually infinite number of connections

(a) Schematic representation of possible interactions between object types on both Telegram and YouTube.

(b) Example of linkages occurring within and between the two platforms.



5. The Network Sampling Framework

Integrating networked public sphere theory (Section 2), explorative network sampling approaches (Section 3), and the practical realities of cross-platform data acquisition (Section 4), the proposed framework represents *actors* as nodes connected through their messages—following Friemel and Neuberger (2023)—via productive, receptive, or curative links. Nodes are necessarily URL-addressable objects, such as profile pages or a blog roll associated with an actor, and are connected to other nodes that also represent actors. Connections are established through features of communication outputs, including feeds, messages, postings, videos, and other metadata extracted from the respective platforms. The interpretation of these objects is flexible and can be adapted to the researcher's specific use case; for example, one could extract links from a message by identifying mentioned accounts or by querying the platform's followers API endpoint. Accordingly, the framework is designed to be modular and adaptable to different combinations of platforms and their intrinsic connection and object types.

Given this model, the realities of cross-platform data acquisition and the complex yet sparse nature of social networks require a network generation process in which edges and nodes are selected exploratorily, sampling the neighborhoods of the original seed nodes while remaining as agnostic as possible about the data source. This process can be generalized into the following sequence of steps, executed by four modules (see Table 1 for an overview of the generalized steps):

1. A sampling round (or hop) begins by identifying the seed nodes. For the first hop these nodes must be provided by the researcher. The seed nodes are assigned to a network layer—for example, accounts on BlueSky or channels on YouTube.
2. Each layer is associated with a *connector* that retrieves the desired information for its nodes.
3. Retrieved information is stored in a database and then processed by the *parser* for the layer. Parsing generates edges, which may connect nodes within the same layer or across different layers.
4. Edges are filtered according to researcher-defined criteria, such as including only messages containing a specific keyword or written in a particular language.
5. A new set of seed nodes is generated by evaluating the sampled network using a *sampler* assigned to each layer.

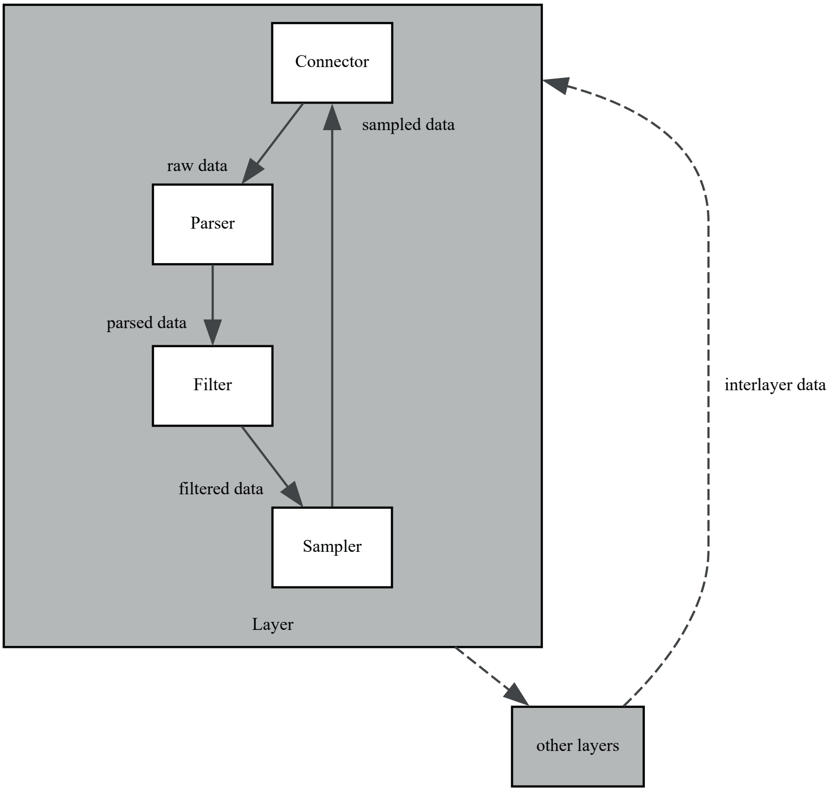
This process is repeated until a stopping condition is met, such as achieving a predetermined sample size for each layer.

Table 1: At the top-level, the sampling framework consists of the following modules

Module	Attributes	Description
Connector	Data connection specification	Describes a program or template for retrieving data for a resource: It enables the collection of information on both actors and content, for example, by scraping public Telegram channels, accessing BlueSky's API, or loading local data files.
Parser	Edge Rules Node Rules	The data returned by the connector and stored in the network layer is processed using a specified rule set, generating nodes and edges from the raw data.
Filter	Edge Rules Node Rules	Rejects specific edges or nodes prior to sampling based on defined conditions. Filtering can be applied to the network's topology, metadata of actors or content, or the content itself—for example, by language.
Sampler	Sampling Algorithm Specification	For each inter- and intralayer—or for the network as a whole —an algorithm is specified to select nodes and edges for sampling. Examples include snowball, rank-degree, forest-fire, or probability-based approaches, depending on metadata or structural properties of the network (cf. Spikyball sampling).

Seeds are the set of nodes from which a network exploration process begins. For each layer, the desired node identifiers are provided to initiate explorative sampling. The format of these identifiers depends on the type of *connector* used; for example, exploring a follower network or account timelines requires account handles or IDs, whereas other connectors may accept content IDs or hashtags. Accordingly, each entry in the database must be mapped to an account handle or ID, as specified in the retrieval configuration (cf. Listing 1).

Figure 3: Exemplary sampling process for a single layer



Listing 1: Example of a multi-layer seed definition containing accounts from two German broadcast programs

```
seeds:
  tiktok:
    - 'DW News': dwnews
    - 'ZDF Heute': zdfheute
  twitter:
    - 'DW News': dwnews
    - 'ZDF Heute':ZDFheute
  facebook:
    - 'DW News': deutschewellenews
    - 'ZDF Heute':ZDFheute
  instagram:
    - 'DW News': dwnews
    - 'ZDF Heute': zdfheute
```

5.1 Connector

We refer to logical data sources as connectors, which encapsulate API endpoints, web-scraping routines, or direct access to databases or files. The input to a connector is always a node identifier, and its output is processed by the *parser* to generate edges, which are then added to the corresponding network layer along node metadata.

Introducing this abstraction as a common, unified interface enhances the stability of data collection processes. Since APIs evolve over time, the *abstraction* allows the underlying implementation to adapt without affecting the interface. The same principle applies to web scraping, where code often breaks due to website changes. In some cases, fallback implementations may be desirable—for example, switching from an API to scraping if needed. Similarly, databases or data files can be queried directly, facilitating the integration of existing datasets into ongoing retrieval tasks.

Formulating relevant connections is inherently platform-dependent. For example, in a database of public speakers, account information may exist for four platforms: Facebook, Instagram, Twitter/X, and TikTok. A connector to the TikTok Research API could access the endpoint that reports which other channels a given channels follows, and return both account information (such as subscriber- and like-count) and connection information—in this case, a list of channels followed by the given channel. Another connector could retrieve the network in the opposite direction, reporting followers, while a third could collect the account's post timeline. Similarly, other connectors can wrap API endpoints from additional platforms, databases or files[4].

Table 2: Input and output listing of the connector module

Input	**Output**
Identifier: A valid identifier for the node to be retrieved, such as a username or numeric ID, depending on the necessities of the implementation. **Table specifications**: A named set of process specifications, where each specification defines a process that accepts the identifier and returns a list of **records**. These processes can be scripts, programs, or other executable routines. **Common attributes**: Additional parameters applied to all processes listed in the table specifications. Examples include a function to retrieve an OAuth token or commonly used parameters such as timeouts, page sizes, or hard limits.	A keyed data collection, where the keys correspond to the table names, and values contain the retrieved records.

[4] As an example, consider an implementation that wraps around a Python package for the TikTok Research API https://github.com/Leibniz-HBI/spiderexpress-tiktok.

Listing 2: Example of a connector defined for the TikTok Research API. It accesses the follower endpoint, retrieving both channel information and account connections, and returns these data for further processing

```
layer:
  tiktok:
    connector:
      infos:
        type: request
        endpoint: https://open.tiktokapis.com/v2/research/user/info/
        request_body_template: '{"user_name": "$node_name"}'
        method: POST
      connections:
        type: request
          endpoint: https://open.tiktokapis.com/v2/research/user/fol-
lowing/
        request_body_template: '{"user_name": "$node_name"}'
        method: POST
```

5.2 Parser

As noted above, URLs themselves do not carry intrinsic meaning—hyperlinks have no inherent semantics. However, because the networked public sphere is highly centralized from a platformization perspective, platform-specific rules can be useful. For example, posts on X.com follow a predictable URL schema: https://x.com/$username/status/$post_id. Similar rules can be defined for other platforms, as their technical systems generally produce predictable hyperlink structures.

Leveraging this predictability, relevant identifiers of posts or accounts can be extracted from URLs that reference these resources. A minimal syntax for extracting the necessary identifiers from hyperlinks can be derived from the following steps: select a field from the datum's content or metadata, access the specified field, optionally apply a regular expression with a capture group, and assign the resulting edge the appropriate edge type. Using the identifier of the processed resource and the extracted reference(s), edges can then be formed between the corresponding nodes.

Similarly, node metadata can be extracted from trace data. For example, a *connector* (cf. Section 5.1) that retrieves both account information and the account's social connections provides two types of data. The account information typically contains fields that can be directly mapped within the framework. Social connections can be extracted and mapped in a comparable manner, allowing them to be represented as edges in the network.

Table 3: Input and output listing of the parser module

Inputs	Output
Data: A keyed dictionary in which each key corresponds to a list of records. **Edge rule set**: A list of edge extraction specifications. Each specification defines which objects and fields to access for both the source and the target of an edge, as well as any additional metadata to extract. Every specification must	A keyed data collection in which the keys correspond to table names and the values contain the retrieved records.

Inputs	Output
specify an edge type and the layers in which the source and target nodes reside. **Node rule set**: A list of node extraction specifications, where each specification maps a field from the data object to the node metadata within the network.	

Listing 3: Example of parser rules that process the data gathered by the connector described in the previous example, extracting both edge and node information for inclusion in the network

```
parser:
  edges:
    source: $node_name
    target:
      type: follow
      field: connections.data.[].name
      regex: null
    type: repost
  node:
    display_name:
      field: infos.display_name
      regex: null
    subscriber_count:
      field: infos.suscriber_count
      regex: null
```

5.3 Filter

Depending on the use case, it may be necessary to reject portions of the network parsed in the previous step. For example, certain actors could be excluded for data protection or privacy reasons, such as by setting a minimum subscriber or follower count for inclusion. It may also be useful to filter posts containing specific keywords to maintain topical relevance, or to select content in a particular language to focus on a single "language sphere."

Table 4: Inputs and outputs of the filter module

Inputs	Outputs
Network: Node and edge data as produced by the **parser**. **Edge/node rule set**: Expressions evaluated using the network as an input.	The network with edges and nodes removed according to the evaluation of the specified rule sets.

5.4 Sampler

This module processes the *filtered* network output and determines the seed set for the next sampling hop. It can also use the previous sampler state as input when employing a stateful strategy, for example, to avoid revisiting nodes or edges that have already been sampled.

As discussed in Section 3, the choice of sampling strategy affects exploration patterns, and different strategies are suited to different use cases (see Hu and Lau 2013 for an overview). Examples include Snowball sampling (Goodman 1961), Forest Fire (Leskovec and Faloutsos 2006), Rank-degree sampling, which ranks nodes based on known degrees in the original network (Voudigari et al. 2016; Münch et al. 2021), and Spikyball, a generalization of multiple strategies that leverages various platform metadata for probability-based sampling (Ricaud, Aspert, and Miz 2020).

Table 5: Input and output listing of the sampler module

Inputs	Outputs
Network: Node and edge data from the relevant layer(s). **Sampler specification**: Defines the strategy to use and, optionally, any parameters required by the strategy. **Sampler state:** Optional state information, used by stateful strategies to avoid revisiting nodes or to manage multi-hop walks.	A new seed set, as well as the sampled network, if the sampling algorithm is stateful, the updated sampler state.

5.5 Limitations

The proposed framework assumes a multilayer network structure, which requires matching nodes across layers. Consequently, it operates at the actor-level, where each actor may control multiple accounts across multiple platforms. This approach is feasible when using a known set of actors with a pre-established mapping to their accounts. However, when combined with explorative network sampling, it becomes necessary to (automatically) match accounts both within a single platform and across platforms, assigning them to individual actors. Research methods for this task are still in early stages, such as mining and clustering accounts based on behavioral characteristics (Bruns et al. 2025).

A further limitation is that using URL-based network sampling of digital resources inherently creates a strong dependence on the network structure of the underlying spherules. If certain digital resources are not connected to the portion of the hidden network where the sampling begins, they cannot be discovered, meaning some remain entirely undetected. Nevertheless, content that reaches a sufficient level of virality is likely to be captured, and for most relevant networks, the majority of nodes belong to a so-called giant component.

Another limitation of focusing on URLs is that information diffusion can bypass these unique identifiers. For example, users often share screenshots of posts or repost content, breaking the chain of observable links. This limitation adds to the inherent multimodality of contemporary social media and web platforms, which cannot be captured by URL-based networks alone. However, it is possible to enrich the discovered networks post-hoc with semantic interlinking, such as tracing paraphrased texts or tracking the diffusion of memes.

On the technical side, the proposed approach—and web scraping in general—can be undermined if platforms stop assigning URLs to content or limit their usefulness. Without accessible URLs, profiles and content objects cannot be reached from the open Internet. For example, Meta allows WhatsApp channels to have an invite link (a URL), but the channel contents are not publicly accessible. Nevertheless, not exposing content or user profiles via URLs is a significant drawback for platforms, as it prevents important content from being referenced externally, whether in media publications or by search engines.

Lastly, the current diversification of platforms, particularly the Fediverse, presents unique challenges. Content can flow between platforms without relying on URLs to reference other posts. However, this content is not uniformly distributed across all instances, as some instances can block others. This situation complicates the assessment of platform affordances, since content can exist independently of any single platform, and individual platforms may implement widely varying policies and technical approaches.

6. Conclusions and Outlook

In this article, we proposed a networked-based, exploratory sampling framework for analyzing networked public spheres. Building on theoretical foundations from media and communication studies, as well as legal and technical considerations, we developed a process that enables researchers to capture and sample data from multiple sources in a well-documented, interoperable, and repeatable manner.

Our main contribution lies enhancing documentability, (inter-)operationalizability, and repeatability. By structuring the data collection process into the four proposed modules (Connector, Parser, Filter, and Sampler, see Table 1) and their associated rulesets, researchers can not only describe their data collection precisely and reproducibly but also obtain guidance for building their own module implementations. These implementations can interface seamlessly with others, reducing dependence on a single data source. We developed this framework alongside an experimental example implementation as a Python command-line tool, which is publicly available and actively maintained[5].

The proposed framework can be further extended to address the limitations discussed in Section 5.5. For example, a new module could be added to automatically match accounts to actors based on topological, metadata, or behavioral features. Likewise, additional modules and extensions are both possible and desirable to enhance the framework's functionality and adaptability.

A question that remains open—and that will inform both substantive research and the further development of this framework—is under which circumstances content posted on one platform crosses over to other platforms or sites, becoming linked from external sources. This process is naturally influenced by the affordances of both the source and target platforms: for example, YouTube videos are frequently linked from other platforms, reflecting YouTube's role as a content search engine rather than solely a social network. In contrast, linking Instagram posts or videos has only recently begun to gain traction. Additionally, cultural and language-specific practices shape platform affordances, resulting in differences in how platforms are used across national and linguistic contexts.

Finally, the sampling algorithms and their suitability for specific research tasks must be carefully evaluated. For instance, some research questions may require more localized data, where smaller nodes are prioritized before bridging communities through well-connected, larger nodes, necessitating different algorithms than those used to study the macrostructure of platform figurations. Another promising avenue for multidisciplinary research is the development of sampling strategies for multilayer networks, which consider not only individual layers or inter-layer connections but also the multilayer network as a whole from a global perspective.

With these questions in mind, we hope this contribution will support empirical research on digital publics and their interconnectedness, enhancing scalability, scope, repeatability, and longitudinal observability. The distribution and diffusion of content across multiple

[5] https://github.com/Leibniz-HBI/spiderexpress.

platforms is a research area of growing importance today and is likely to become even more critical in the near future.

References

Barabasi, Albert-László, and Reka Albert. 1999. Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Berners-Lee, Tim. 1994. "Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as Used in the World-Wide Web." Request for {{Comments}} RFC 1630. Internet Engineering Task Force. https://doi.org/10.17487/RFC1630

Berry, David. 2011. The computational turn: Thinking about the digital humanities. *CULTURE MACHINE*, *12*.

boyd, danah m., and Nicole B. Ellison. 2007. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication* 13 (1): 210–30. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Bossetta, Michael. 2019. *The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election* (No. arXiv:1904.07333). arXiv. https://doi.org/10.48550/arXiv.1904.07333

Breiter, Andreas, and Andreas Hepp. 2018. "The Complexity of Datafication: Putting Digital Traces in Context." In *Communicative Figurations: Transforming Communications in Times of Deep Mediatization*, edited by Andreas Hepp, Andreas Breiter, and Uwe Hasebrink, 387–405. Transforming Communications – Studies in Cross-Media Research. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65584-0_16

Brin, Sergey, Rajeev Motwani, Lawrence Page, and Terry Winograd. 1998. "What Can You Do with a Web in Your Pocket?" *IEEE Data Eng. Bull.* 21 (2): 37–47.

Brüggemann, Michael, and Hendrik Meyer. 2023. "When Debates Break Apart: Discursive Polarization as a Multi-Dimensional Divergence Emerging in and Through Communication." *Communication Theory* 33 (2–3): 132–42. https://doi.org/10.1093/ct/qtad012

Bruns, Axel. 2008. "Life Beyond the Public Sphere: Towards a Networked Model for Political Deliberation." *Information Polity* 13: 71–85.

Bruns, Axel. 2019. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Bruns, Axel. 2023. "From 'the' Public Sphere to a Network of Publics: Towards an Empirically Founded Model of Contemporary Public Communication Spaces." *Communication Theory* 33 (2–3): 70–81. https://doi.org/10.1093/ct/qtad007

Bruns, Axel, and Jean Burgess. 2011. "The Use of Twitter Hashtags in the Formation of Ad Hoc Publics." In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, edited by A. Bruns and P. De Wilde, 1–9. United Kingdom: The European Consortium for Political Research (ECPR).

Bruns, Axel, Kateryna Kasianenko, Vish Padinjaredath Suresh, Ehsan Dehghan, and Laura Vodden. 2025. Untangling the Furball: A Practice Mapping Approach to the Analysis of Multimodal Interactions in Social Networks. *Social Media + Society*, *11*(2). https://doi.org/10.1177/20563051251331748

Bruns, Axel, and Brenda Moon. 2019. "One Day in the Life of a National Twittersphere." *Nordicom Review* 40 (s1): 11–30. https://doi.org/10.2478/nor-2019-0011

Bruns, Axel, Brenda Moon, Felix Victor Münch, and Troy Sadkowsky. 2017. "The Australian Twittersphere in 2016: Mapping the Follower/Followee Network." *Social Media + Society, 3*(4). https://doi.org/10.1177/2056305117748162.

"Common Crawl – Overview." n.d. https://commoncrawl.org/overview. Accessed January 28, 2026.

Coscia, Michele, and Luca Rossi. 2018. "Benchmarking API Costs of Network Sampling Strategies." In *2018 IEEE International Conference on Big Data (Big Data)*, 663–72. https://doi.org/10.1109/BigData.2018.8622486

Esau, Katharina, Tariq Choucair, Samantha Vilkins, Sebastian F. K. Svegaard, Axel Bruns, Kate S. O'Connor-Farfan, and Carly Lubicz-Zaorski. 2024. "Destructive Polarization in Digital Communication Contexts: A Critical Review and Conceptual Framework." *Information, Communication & Society* 0 (0): 1–22. https://doi.org/10.1080/1369118X.2024.2413127

Freelon, Deen. 2021. “The Post-API Age Reconsidered: Web Science in the ’20s and Beyond.” In *13th ACM Web Science Conference 2021*, 3–3. Virtual Event United Kingdom: ACM. https://doi.org/10.1145/3447535.3466177.

Friemel, Thomas N, and Christoph Neuberger. 2023. “The Public Sphere as a Dynamic Network.” *Communication Theory* 33 (2–3): 92–101. https://doi.org/10.1093/ct/qtad003

Frischlich, Lena, & Edda Humprecht (2021). *Trust, Democratic Resilience, and the Infodemic*. https://doi.org/10.5167/UZH-202660

Goodman, Leo A. 1961. “Snowball Sampling.” *The Annals of Mathematical Statistics* 32 (1): 148–70. https://doi.org/10.1214/aoms/1177705148

Guan, Lu, Xiao Fan Liu, Wujiu Sun, Hai Liang, and Jonathan Zhu. 2022. “Census of Twitter Users: Scraping and Describing the National Network of South Korea.” *PLOS ONE* 17 (November): e0277549. https://doi.org/10.1371/journal.pone.0277549

Habermas, Jürgen. 1962. *Strukturwandel der Öffentlichkeit – Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. 1990th ed. Suhrkamp.

Hasebrink, Uwe, and Andreas Hepp. 2017. “How to Research Cross-Media Practices? Investigating Media Repertoires and Media Ensembles.” *Convergence, 23*(4): 362–77. https://doi.org/10.1177/1354856517700384

Heft, Annett, Kilian Buehling, Xixuan Zhang, Juni Schindler, and 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. https://doi.org/10.1080/19331681.2023.2250779

Helmond, Anne. 2015. “The Platformization of the Web: Making Web Data Platform Ready.” *Social Media + Society* 1 (2): 205630511560308. https://doi.org/10.1177/2056305115603080

Hepp, Andreas, and Uwe Hasebrink. 2014. Kommunikative Figurationen – ein Ansatz zur Analyse der Transformation mediatisierter Gesellschaften und Kulturen. In *Von der Gutenberg-Galaxis zur Google-Galaxis: Alte und neue Grenzvermessungen nach 50 Jahren DGPuK* (pp. 343–360). UVK Verlagsgesellschaft.

Hu, Pili, and Wing Cheong Lau. 2013. “A Survey and Taxonomy of Graph Sampling.” arXiv. https://doi.org/10.48550/arXiv.1308.5865

Jost, Pablo, Annett Heft, Kilian Buehling, Maximilian Zehring, Heidi Schulze, Hendrik Bitzmann, and Emese Domahidi. 2023. “Mapping a Dark Space: Challenges in Sampling and Classifying Non-Institutionalized Actors on Telegram.” *M&K Medien & Kommunikationswissenschaft* 71(3–4): 212–29. https://doi.org/10.5771/1615-634X-2023-3-4-212

Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. 2014. “Multilayer Networks.” *Journal of Complex Networks*, 2(3): 203–71. https://doi.org/10.1093/comnet/cnu016

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall van Alstyne. 2009. *Computational Social Science. Science, 323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lazer, David, Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, et al. 2020. “Computational Social Science: Obstacles and Opportunities.” *Science* 369 (6507): 1060–62. https://doi.org/10.1126/science.aaz8170

Leskovec, Jure, and Christos Faloutsos. 2006. “Sampling from Large Graphs.” In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 631–36. Philadelphia PA USA: ACM. https://doi.org/10.1145/1150402.1150479

Münch, Felix Victor, Ben Thies, Cornelius Puschmann, and Axel Bruns. 2021. “Walking Through Twitter: Sampling a Language-Based Follow Network of Influential Twitter Accounts.” *Social Media + Society, 7*(1): 2056305120984475. https://doi.org/10.1177/2056305120984475

Nielsen, Hendrik, Roy T. Fielding, and Tim Berners-Lee (1996). Hypertext Transfer Protocol – HTTP/1.0 (Request for Comments No. RFC 1945). Internet Engineering Task Force. https://doi.org/10.17487/RFC1945

Ohme, Jakob, Theo Araujo, Laura Boeschoten, Deen Freelon, Nilam Ram, Byron B. Reeves, and Thomas N. Robinson. 2023. “Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking.” *Communication Methods and Measures, 0* (0): 1–18. https://doi.org/10.1080/19312458.2023.2181319

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." *Frontiers in Big Data* 2.

Pearce, Wareen, Suay M. Özkula, Amanda K. Greene, Lauren Teeling, Jennifer S. Bansard, Janna Joceli Omena, and Elaine Teixeira Rabello. 2020. Visual Cross-Platform Analysis: Digital Methods to Research Social Media Images. *Information, Communication & Society*, *23*(2), 161–180. https://doi.org/10.1080/1369118X.2018.1486871

Quandt, Thorsten, Lena Frischlich, Svenja Boberg, and Tim Schatto-Eckrodt. 2019. Fake News. In *The International Encyclopedia of Journalism Studies* (pp. 1–6). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118841570.iejs0128

Rau, Jan Philipp, Philipp Kessling, Gregor Wiedemann, and Felix Victor Münch. 2025. "Research Data Access in the Context of Art. 40 DSA for the German Federal Election: A Mixed Experience at Best." Frankfurt, Leizig. https://doi.org/10.58079/13ZE0

Ricaud, Benjamin, Nicolas Aspert, and Volodymyr Miz. 2020. "Spikyball Sampling: Exploring Large Networks via an Inhomogeneous Filtered Diffusion." arXiv. https://doi.org/10.48550/arXiv.2010.11786.

Rogers, Richard 1996. "The Future of Science and Technology Studies on the Web." *EASST Review* 15, 25–27.

Rogers, Richard. 2010. "Mapping Public Web Space with the Issuecrawler." In *Digital Cognitive Technologies: Epistemology and Knowledge Society*, edited by Claire Brossard and Bernard Rebers. London, England: Wiley.

Rogers, Richard. 2023. "'Serious Queries' and 'Editorial Epistemologies'." In *The Propagation of Misinformation in Social Media: A Cross-platform Analysis*. Amsterdam University Press. https://doi.org/10.5117/9789463720762

Schmidt, Jan-Hinrik. 2014. "Twitter and the Rise of Personal Publics." In *Twitter and Society*, 3–14. New York, Washington, D.C., Bern.

Schmidt, Jan-Hinrik, Lisa Merten, Uwe Hasebrink, Isabelle Petrich, and Amelie Rolfs (2017). Zur Relevanz von Online-Intermediären für die Meinungsbildung. Arbeitspapiere des Hans-Bredow-Instituts, 40, 107 S. https://doi.org/10.21241/SSOAR.71784

Sen, Indira, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. "A Total Error Framework for Digital Traces of Human Behavior on Online Platforms." *Public Opinion Quarterly* 85 (S1): 399–422. https://doi.org/10.1093/poq/nfab018

Strippel, Christian, Annekatrin Bock, Christian Katzenbach, Merja Mahrt, Lisa Merten, Christian Nuernbergk, Christian Pentzold, Cornelius Puschmann, and Annie Waldherr. 2018. "Die Zukunft der Kommunikationswissenschaft ist schon da, sie ist nur ungleich verteilt: Eine Kollektivreplik auf Beiträge im „Forum" (Publizistik, Heft 3 und 4, 2016)." *Publizistik* 63, 11–27 (Januar). https://doi.org/10.1007/s11616-017-0398-5.

Voudigari, Elli, Nikos Salamanos, Theodore Papageorgiou, and Emmanuel J. Yannakoudakis. 2016. "Rank Degree: An Efficient Algorithm for Graph Sampling." *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 120–29. https://doi.org/10.1109/ASONAM.2016.7752223

Waldherr, Annie. 2017. "Öffentlichkeit als komplexes System. Theoretischer Entwurf und methodische Konsequenzen." *M&K Medien & Kommunikationswissenschaft,* 65(3): 534–49. https://doi.org/10.5771/1615-634X-2017-3-534

Wiedemann, Gregor, Felix Victor Münch, Jan Philipp Rau, Phillip Kessling, and Jan-Hinrik Schmidt. 2023. "Concept and Challenges of a Social Media Observatory as a DIY Research Infrastructure." *Publizistik*, 201–223 August. https://doi.org/10.1007/s11616-023-00807-6.

Wolf, J. L., Squillante, M. S., Yu, P. S., Sethuraman, J., & Ozsen, L. (2002, May 7). *Optimal Crawling Strategies for Web Search Engines*. WWW2002.