Universität Trier

Potentials of Digital Behavioural Trace Data: An Application from Radicalisation Research

Cumulative Dissertation
for obtaining the academic degree
Doctor of Philosophy (Dr. phil.)

Faculty I
University of Trier

Submitted by

Veronika Batzdorfer

Supervisors:

Prof. Dr. Michaela Brohm-Badry, University of Trier
Prof. Dr. Michael Bosnjak, University of Trier

January 05, 2022

Dissertationsort: Trier

A picnic. Picture a forest, a country road, a meadow. Cars drive off the country road into the meadow, a group of young people get out carrying bottles, baskets of food, transistor radios, and cameras. They light fires, pitch tents, turn on the music. In the morning they leave. The animals, birds, and insects that watched in horror through the long night creep out from their hiding places. And what do they see? Old spark plugs and old filters strewn around... Rags, burnt-out bulbs, and a monkey wrench left behind... And of course, the usual mess—apple cores, candy wrappers, charred remains of the campfire, cans, bottles, somebody's handkerchief, somebody's penknife, torn newspapers, coins, faded flowers picked in another meadow.

—Arkadi Strugatzki & Boris Strugatzki, Roadside Picnic

Acknowledgements

I would like to thank my family, Oliver and the bears to bear with me even when the roof was on fire. Further, I am grateful for the advice of my supervisors Prof. Michael Bosnjak and Prof. Michaela Brohm-Badry who patiently guided me throughout my research endeavours and supported me when it came to feasibility questions. Most notably, I am most grateful to my colleagues at ZPID starting with Holger Steinmetz, Martin Kerwer, André Bittermann and Judith Tinnes. Thank you for always lending a sympathetic ear to me even when the string of errors that was issued was longer than this page. I have learned so much from you all. I am grateful for having you as my companions, both for your integrity, character and expertise.

Abstract

Behavioural traces from interactions with digital technologies are diverse and abundant. Yet, their capacity for theory-driven research is still to be constituted. In the present cumulative dissertation project, I deliberate the caveats and potentials of digital behavioural trace data in behavioural and social science research. One use case is online radicalisation research. The three studies included, set out to discern the state-of-the-art of methods and constructs employed in radicalisation research, at the intersection of traditional methods and digital behavioural trace data. Firstly, I display, based on a systematic literature review of empirical work, the prevalence of digital behavioural trace data across different research strands and discern determinants and outcomes of radicalisation constructs. Secondly, I extract, based on this literature review, hypotheses and constructs and integrate them to a framework from network theory. This graph of hypotheses, in turn, makes the relative importance of theoretical considerations explicit. One implication of visualising the assumptions in the field is to systematise bottlenecks for the analysis of digital behavioural trace data and to provide the grounds for the genesis of new hypotheses. Thirdly, I provide a proof-of-concept for incorporating a theoretical framework from conspiracy theory research (as a specific form of radicalisation) and digital behavioural traces. I argue for marrying theoretical assumptions derived from temporal signals of posting behaviour and semantic meaning from textual content that rests on a framework from evolutionary psychology. In the light of these findings, I conclude by discussing important potential biases at different stages in the research cycle and practical implications.

Zusammenfassung

Verhaltensspuren, die sich aus der Interaktion mit digitalen Technologien ergeben, sind vielfältig und zahlreich. Ihre Eignung für die theoriegeleitete Forschung muss jedoch erst noch konstituiert werden. In dem vorliegenden kumulativen Dissertationsprojekt befasse ich mich mit den Grenzen und Möglichkeiten von digitalen Verhaltensspuren in der verhaltens- und sozialwissenschaftlichen Forschung. Ein Anwendungsfall ist die Online-Radikalisierungsforschung. Die drei einbezogenen Studien sollen den aktuellen Stand der Methoden und Konstrukte in der Radikalisierungsforschung an der Schnittstelle von traditionellen Methoden und digitalen Verhaltensspurdaten aufzeigen. Erstens zeige ich auf der Grundlage einer systematischen Literaturübersicht von empirischen Arbeiten die Prävalenz digitaler Verhaltensspuren in verschiedenen Forschungsbereichen auf und arbeite Determinanten und Ausprägungen von Radikalisierungskonstrukten heraus. Zweitens extrahiere ich auf der Grundlage dieser Literaturübersicht Hypothesen und Konstrukte und integriere sie in einen Rahmen aus der Netzwerktheorie. Dieser Graphen an Hypothesen macht wiederum die relative Bedeutung der theoretischen Überlegungen deutlich. Eine Implikation der Visualisierung der Annahmen im Feld ist die Systematisierung von Engpässen bei der Analyse digitaler Verhaltensspurdaten und die Schaffung von Grundlagen für die Entwicklung neuer Hypothesen. Drittens stelle ich ein Proof-of-Concept für die Einbeziehung eines theoretischen Rahmens aus der Verschwörungstheorieforschung (als spezifische Form der Radikalisierung) und digitalen Verhaltensspuren vor. Ich plädiere dafür, theoretische Annahmen, die aus zeitlichen Signalen des Posting-Verhaltens abgeleitet werden, mit semantischen Bedeutungen aus Textinhalten zu verbinden, die auf einem Theorierahmen aus der Evolutionspsychologie beruhen. Vor dem Hintergrund dieser Ergebnisse schließe ich mit einer Abwägung wichtiger potenzieller Verzerrungen in verschiedenen Phasen des Forschungszyklus und praktischer Implikationen.

Contents

Publications included in the dissertation

Study 1:

Batzdorfer, V., Steinmetz, H., & Bosnjak, M. (2020). Big Data in der Radikalisierungsforschung: Eine systematische Übersichtsarbeit. *Psychologische Rundschau*, *71*, 96-102. https://doi.org/10.1026/0033-3042/a000480

Study 2:

Batzdorfer, V., & Steinmetz, H. (2020). Reviewing radicalization research using a network approach. *Journal for Deradicalization*, (23), 45-95. https://journals.sfu.ca/jd/index.php/jd/article/view/361/235

Study 3:

Batzdorfer, V., Steinmetz, H., Biella, M., & Alizadeh, M. (2021). Conspiracy theories on Twitter: Emerging motifs and temporal dynamics during the COVID-19 pandemic. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-021-00298-6

List of Figures

List of Tables

Chapter 1

Introduction

Digital services are tightly interwoven with our daily routines. These services provide entertainment, education, or social networking. Much like we play sports with fitness wearables that combine sensor data (such as positional coordinates or acceleration rate) with context data (like the local weather), we engage with social media platforms in a morning routine or purchase products online and write product reviews. The large-scale adoption and pervasiveness of digital devices and Internet services throughout all areas of life shows that technology is not only an instrument per se, but it fundamentally changes and shapes social interactions (Ruths & Pfeffer, 2014). In this transforming role, platform and device machine learning algorithms exert their influence by recommending similar videos or products we might like based on our past behaviour (Heuer et al., 2021), determining the format and content type that can be posted by interface design (e.g., up to 280 characters) (Gligoric et al., 2018), functioning as information gatekeepers in what is displayed to the user, or influencing platform engagement by gratification metrics (e.g., liking). The rise of new services and their usage have resulted in new phenomena that did not exist a decade ago, such as polarisation processes (Barberá et al., 2015), influencing campaigns by automated accounts, or technologically augmented financial services that employ credit scores.

In the process of consuming new forms of information and services, individuals leave behind large amounts of traces from interactions with these devices and online platforms (Howison et al., 2011; Jungherr, 2018). Digital behavioural trace data can be defined as evidence of past activity in online information systems that are essentially pre-existing (Howison et al., 2011). They are *pre-existing* as the data obtained are usually generated in a process other than for research (Landers et al., 2016). Platforms such as Twitter were not designed for systematic research, but to maximise user engagement and revenues. For research purposes, researchers need to infer higher-order constructs of interests from the "raw" digital behavioural

trace data, or more precisely, what the platform provides as data types. Different trace types can be differentiated according to their informative gain.

Table 1.1: *Types of digital behavioural traces on social media*

	Raw data	Aggregation types
Content traces	tags, search queries, textual postings, profile description, video transcript	lexical markers (e.g., sentence length, vocabulary diversity) syntax (e.g., passive constructions) semantics and pragmatics (e.g., speech acts, topics, sentiment)
Interaction traces	up-voting, sharing links, quoting, rating a review, calling, marking as read	involvement rate (e.g., posting rate, turn taking frequency, pauses, temporal patterns of retweets)
Metadata	post edit, timestamps, location coordinates, web browsing history, picture dimensions	number of places visited, time difference between consecutive posting, reply, or post moderation

Traces can be differentiated based on *content*, *interaction* as well as *metadata* (see Table 1.1).

Essentially, technological innovations that provided momentum to this data access are the extraction of online data and restructuring and pre-processing large quantities of qualitative data for quantitative natural language processing. *Content* traces refer to user-generated textual data, as for instance, tags on the holiday pictures, textual postings such as tweets, or profile descriptions on social media platforms. Text can be transformed into numerical data, which then are aggregated into meaningful variables within texts, across users or across-time (for an overview on sentiment, see Algaba et al., 2020); the results of such analyses include the categorisation of content into predetermined categories (i.e., "lexicon-based") or exploratory discovered latent thematic structures (e.g., "topic modeling", Gerlach, Peixoto, & Altmann, 2018). Naturally occurring language, as one type of digital behavioural trace data offers intriguing insights into higher-order constructs such individual opinions, emotions, intentions (e.g., speech acts, Austin, 1975), cognitive processes (Humphreys & Wang, 2018) or mental health status (De Choudhury et al., 2014). Examples for insights derived from linguistic style analysis have been the prediction of age by emotion words (Nguyen, Smith, & Rose, 2011), gender predicted by the usage of function words (articles, pronouns or conjunctions) (Newman, Groom, & Handelman, 2008), or the prediction of mental health status (Seabrook et al., 2018), or emotional states by the frequency of positive and negative words (Beasley & Mason, 2015).

Interaction traces refers to user interactions with platform content that are not text-based, such as up-voting a posting, sharing cross-platform links, quoting another posting, or rating a product review. Here, the occurrence of an interaction may be taken as evidence for construct such as social relationship (e.g., the frequency of upvotes signaling popularity amongst peers or attention). However,

this greatly depends on the social context under which an interaction is executed. Further, with regards to user interactions traces, aspects that could be considered are time series decomposition into trend, seasonality, or error (Jebb et al., 2015). For instance, Dalal et al. (2014) claim that there are rhythmic mood cycles in individuals which would result in a sentiment series with seasonal fluctuations (see also Larsen & Kasimatis, 1990). Such rhythms can be well integrated into approaches such as self-regulation theory. Thereby, influences (external or internal) trigger changes and the individual counteracts them in order to restore homeostasis (see *set-point theory*, Lucas et al., 2004).

Metadata refers to data and by-products that arise from technologies that operate in the background (e.g., users requesting information from a server) (Howison et al., 2011). Examples are comment trees or timestamps of when review texts are submitted, location coordinates of the user, sequential web browsing history, information verifying purchases, or the dimensions of the picture posted. Such metadata, in turn, can be operationalised as context data for theoretical constructs such as lockdown measures compliance by means of individual mobility patterns over time. However, questions of construct validity arise based on the implicitness of these measures. Without measuring the meaning of these features by the user it remains unclear, validity problems pose (e.g., a friendship feature on social media does not necessary figure as an apt proxy for a friendship construct) (Jungherr, 2018).

A variety of raw digital behavioural trace data can be retrieved from social media platforms, and they are used by platform developers and companies to develop new applications or for marketing purposes. Given the abundance and accessibility of digital behavioural trace data, the question begs as to their standing in research. On the one hand, such data types and new methods of analysis are regarded as an "epistemological revolution" (Golder & Macy, 2014; Kitchin, 2014), offering seminal capabilities to overcome weaknesses of traditional methods (e.g., systematic error by *recall bias* or *observer-expectancy effects*) (Marres, 2015; Prior, 2009). Yet, on the other hand, they are accused at the same time of lacking theory and proper measurement theory (Jungherr, 2018) and shifting away from causality to pure discovery (Kitchin, 2014). Despite the value of digital behavioural traces, they have not been frequently discussed with respect to a proper measurement theory in fields of the behavioural or social sciences (Jungherr, 2018). Multiple aspects of dealing with such new data types and analysis approaches are still in development, starting with suitable research questions and potential use cases that offer a theoretical gain (Howison et al., 2011), reaching over to biases and their consequences (Malik, 2018; Ruths & Pfeffer, 2014; Sen et al., 2021; Tufekci, 2014), linking survey data with such data types (Beuthner et al., 2021), or sharing and archiving such data (Proferes et al., 2021; Weller & Kinder-Kurlanda, 2016).

This thesis, thus, provides both theoretical insights for researchers when working with digital behavioural trace data (as to systematising the radicalisation research domain and operationalising identified constructs), as well as practical implications for addressing potential biases at different stages in the research cycle.

1.1 Problem statement and research gaps

Departing from the great amount of available digital behavioural traces, the value of different trace types for theory-rooted research is, yet, to be established. One application area is radicalisation research (and as a specific form, conspiracy theory research). There are three reasons why the use of digital behavioural traces can be fruitful in this realm.

First, access to digital behavioural trace data allow unobtrusive insights into the actual behaviour of hard-to-reach individuals in their natural environment and with high granularity (Landers et al., 2016). Second, virtual communities represent a facilitating arena for broadcasting radical beliefs, as well as for connecting and recruiting vulnerable subjects. The diffusion of conspiracy theories exemplifies the discursive power, emotional contagion, and speed of user-generated content on social media platforms. It further stresses the relevance to understand causal underpinnings of potential determinants and outcomes, as these may have detrimental effects (van Mulukom et al., 2020). In particular, belief in conspiracy theories has been associated with greater resistance to scientific evidence such as vaccination and climate science (Lewandowsky, Oberauer, & Gignac, 2013). Third, most notably, there is a need to harness traces of digital behaviour and activities to determine their suitability and limitations for new application fields, much like has been done with measurement theories for traditional survey methods or experimental studies (Groves et al., 2011). Particularly in psychometrics, extensive theory has been developed, linking measures and entities of interest (e.g., depression, intelligence) via the specification of formal *measurement models* (e.g., Item Response Theory, for the evaluation of the validity and reliability of questionnaires). In this regard, aspects of validity are based on theoretical assumptions (e.g., the correlation among different measures) and well established in psychology and other fields (Groves et al., 2011). Such systematic efforts are still to be developed for the field of digital behavioural trace data. The precision of concepts is one necessary condition for the theoretical importance. Hence, defining outcomes and determinants of radicalization and different forms and yet also distinguishing it from non-radicalization are inherently important to advance the field.

Overall, the following research gaps can be derived:

(i) Given a wealth of research in the field of radicalisation, an integrated view to guide study designs—across approaches (survey, experimental approaches, as well as digital behavioural trace data analyses)—is lacking.

(ii) In radicalisation research exist a great variety of definitions of radicalisation (Schmid, 2013). These, in turn, are sometimes inconsistent and reflect a phenomenon that is difficult to demarcate (Kou et al., 2017). In attempting to understand the minimal sufficient determinants of radicalisation processes, approaches range from pathological manifestations to cognitive or trait-based explanations, but few approaches include an actionable definition for the online sphere (Klein, Clutton, & Dunn, 2019).

(iii) Longitudinal perspectives on the evolution of such extreme beliefs are rare in the field. Yet, dynamics of change over time and within-subject variability are an inherent scientific interest of behavioural, social and educational researchers which can be addressed with time series analysis. I argue that the possibility to unobtrusively observe people's actual behaviour in their natural environment bears potential for many research areas (Harlow & Oswald, 2016; Landers et al., 2016). The potential is defined by the characteristics of such data sources, that is, they are highly dense, longitudinal in format, and contain a series of outcomes of interest measured every day, every hour and allow researchers insights into dynamic processes, trends, and fluctuations across time. For example, voicing an opinion follows a temporal pattern like any other behaviour, involving abrupt spikes or long-term changes in its frequency and inertia that reflect system responses to external shocks, as interactions with the physical world (Jebb et al., 2015). However, so far there are insufficient applications that use these time series features in a theory-based way.

1.2 Main contributions

The following dissertation sets out to systematise and integrate research strands in radicalisation research, as an exemplary research field. It makes the following contributions:

1. **Operationalisable definition of radicalisation**
   Research lacks an operationalisable definition of radicalisation. It is, frequently, conceptualised either, too diffuse to be actionable in a social media context, or too specific for a given context, to be adopted for another environment. I provide for an actionable definition of radicalisation (with a focus on conspiracy theories) that is both sufficiently *generic* and scalable, as well as *specific* enough (see *Section 2.3*). It is generic, as radicalisation is operationalised in the context of digital behavioural trace data in a bottom-up manner and specific, as the approach combines a theory-

rooted notion from evolutionary psychology (with the five criteria: agency, threat, secrecy, coalition and pattern).

2. **Theoretical integration through network theory**
   I approach the research field by analysing the complex structure of all considered theoretical constructs and hypotheses formulated by scholars of the field and analyse differences in relevance and interconnections of constructs across methodological approaches (*Section 2.2*). This network-based approach paths the way for rigorous comparison of theories and allows to combine the constructs of several theories in a new way. I further ask for more clear sets of propositions postulated in theories and caution to consider issues such as endogenous selection bias.

3. **Construct validity of digital behavioural trace data**
   I provide practical guidance on the potentials and caveats of digital behavioural trace data with respect to different stages in a research life cycle (see also Figure 1.1 below). As I concluded in *Section 2.1* that there is a need of a "data theory" (Landers et al. 2016), considerations of potential biases are to be made at all stages, beginning from the: operationalisation, and sampling, reaching over to the analysis and validation stage. Certain biases can hardly be mitigated as, for instance, when sampling and data access are reliant on commercial enterprises and platform affordances or when training or validation (generalisability of machine learning models) are hampered by the general unavailability of ground truth. However, when aiming for advancing measures of radicalisation and drawing valid inferences with digital behavioural trace data this requires careful and transparent considerations of conceptual and underlying assumptions of how data were generated. The different stages of the research life cycle aim to encourage critical discussion on new research designs and their documentation.

1.3 Quality decision points in the research cycle

In the present section I take up the aspect of measurement theories for digital behavioural trace data and stress the need to characterise the ramifications throughout different stages of a research cycle (see also Sen et al., 2021). I subsequently focus on different stages (see Fig. 1.1), starting with **(i)** the conceptualisation phase of constructs of interest and their operationalisation, extending to **(ii)** sampling a target population, and **(iii)** analysing such data types and **(iv)** validating the steps. In each of the respective stages, the researcher has to make crucial decisions which impact subsequent stages and the possible inference from digital behavioural trace data possible. Worse, depending on the decisions, inference my be hampered or invalid at all.

Firstly, in the **conceptualising stage,** defining a concise theoretical construct requires substantive domain knowledge for it to be delineable against other con-

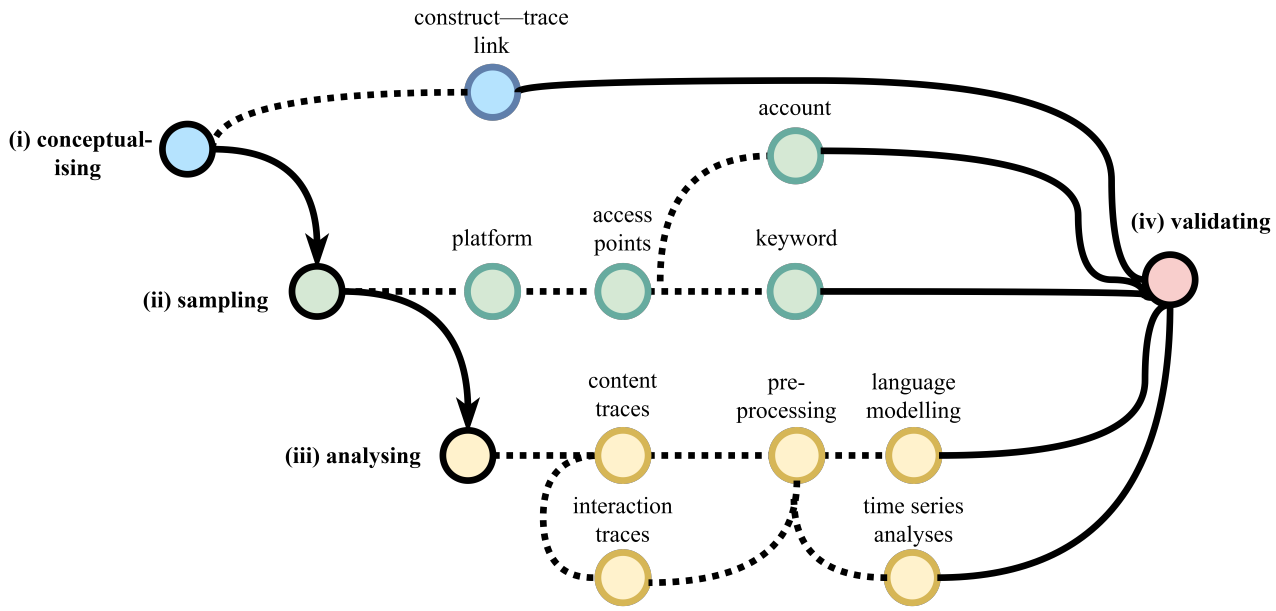

Figure 1.1: Decision points for digital behavioural trace data at different stages of the research cycle

structs. By extension, if interested in causal inference, explicitly stating the theoretical assumptions of how the data were generated makes potential threat of bias amenable to other researchers (Elwert & Winship, 2014). Further, operationalising the theoretical construct (*construct—trace link*) is largely influenced by the availability of data (due to the non-reactive nature of digital behavioural trace data). Construct validity can be compromised if the operationalisation is not specific enough for the construct of interest. In the realms of natural language processings, if a construct of interest is set out clearly, dictionaries (i.e., pre-determined lists of words) can be useful which are a top-down approach. However, if a construct cannot be clearly operationalised, more exploratory approaches (i.e., unsupervised machine learning) are more suitable which posit a bottom–up approach (Humphrey & Wang, 2018). Secondly, at the **sampling stage**, choosing a social media *platform* defines the type of digital behavioural trace data obtainable (e.g., video or picture-related). Beyond the type of obtainable data, platform affordances (e.g., platform design tailored to trending content) distort how users express themselves (Ruths & Pfeffer, 2014). This requires a careful consideration of the degree to which the realities of a platform can reflect the theoretical construct of interest. Further, *access points* to social media platforms (e.g., APIs) determines the amount of digital behavioural traces that can be obtained, as well as the extent of artifacts (e.g., bots, missing data). In order to sample a platform, a whitelist of user *accounts* can be used. This affects the generalisability and reusability of such data sets that are very specific to the use case. Alternatively, *keyword*/hashtag/geo-reference-based sampling may be employed. However, such approaches bear the potential of endogenous selection bias (conditioning on a variable that is caused by two other

variables) (Elwert & Winship, 2014). A combination of sampling strategies, alongside validating the obtained sample may mitigate some of the potential biases. Thirdly, at the **analysing stage**, data normalising choices, for both, *interaction* or *content* traces, affect the construct validity, as this may lead to the exclusion of users or data based on certain attributes. *Pre-processing* choices may relate to the aggregation method, which aims to reduce the dimensionality of text and temporal data and define noise and artifacts. Further, analytical choices at the level of *language modelling* or *time series analyses* may distort how the construct is estimated and lead to incorrectly specified models. For instance, regarding the automated analyses of natural language, model tuning and hyperparameter optimisation choices affect the extent and granularity of semantic meaning eventually considered. Fourthly, the **validating stage** connects back to all the previous stages. For instance, validating pre-processing decisions, such as filtering potential automated accounts, regarding the proportion of false positives and false negatives, gives a sense of the amount of bias present.

These stages of the research cycle lay the backbone of the present work, described in more detail in Chapter 2 with the empirical studies. The first study sets out to discern the prevalence of digital behavioural trace data and systematise determinants and outcomes of radicalisation in a systematic review of empirical studies. This study connects to the first three steps in the research cycle, as it aims to characterise the constructs, the sampling process and methodological approaches, across digital behavioural trace data studies, survey and experimental studies. The second study integrates the uncovered constructs to a framework of relations (i.e., in the form of a *graph* of constructs that connects each of them by the hypotheses in research articles). This graph connects to the first stage of the research cycle and makes the relative importance of theoretical considerations explicit and opens up perspectives for discerning bottle-necks in the analysis of digital behavioural trace data. The third study provides a proof-of-concept for incorporating a theoretical framework from conspiracy theory research (as a specific form of radicalisation) and digital behavioural traces. This work tries to address all four phases of the research cycle. Lastly, in the general discussion I reconsider implications regarding the decision points in the research cycle and future research avenues.

1.4 Data

1.4.1 Social media as a data source

There are three main reasons to opt for social media platforms, as a data source for research, over other online services.

First, social media platforms, such as Twitter, are *openly accessible* for individuals. Users can equally contribute (upon registration) to discussions. The content of a broadcasted tweet, may include personal thoughts, expressions of emotions or opinions on recent events, media, behaviour, and politics. Particularly, the short format of tweets (up to 280 characters) or other low-barrier non-content interaction forms encourage users to update their Twitter accounts multiple times per day, and this activity over a longer time course enables research on dynamic changes and fluctuations with high granularity. Second, almost all conversations are in principle *observable* for researchers. Only a small fraction of tweets or accounts are set to private. This allows for the unobtrusive observation of naturally occurring behaviour. Third, the retrieved digital behavioural trace data are *rich* in content and interactions traces, as well as metadata (e.g., allow for historical observations) and hence particularly suitable for within- and between-subject designs. They further allow to distinguish individual users and their behaviour over time, unlike anonymous forums (e.g., *8kun*). Particularly Twitter has been extensively studied by researchers owing to its ease of data access and predefined categorisation of user activities (e.g., tweet text) (Jürgens & Jungherr, 2016).

Acknowledging, the choice of social media platform might induce potential biases within the research cycle. This is the case, as every social media platform has its own terms for defining concepts and providing functionalities. For instance, certain digital behavioural trace data cannot be queried, such as click-behavioral patterns of the users, "unfollowed" users, posts viewed but not interacted with, time spent on the platform, as would be possible with access to raw log data.

1.4.2 Data set construction

Access points. There are different access points to digital behavioural trace data, depending on the research design and device or platform structure. For instance, access points might range from requesting data from Twitter, YouTube, Instagram or Reddit, over to building own crawlers that scrape structured content from web pages where no public access points are available. Most popular are *Application Programming Interfaces* (APIs), an access point chosen for this thesis (Lomborg & Bechmann, 2014; Proferes et al., 2021). Such access points are for instance established for Youtube (which allows to search for content based on keyword queries and obtain the video, playlists and user activities such as upvoting, comments, favouriting) or Instagram (allowing to obtain comment trees relating to postings, friendship information of users or geolocation). They form a software interface of a website to connect with other software (Lomborg & Bechmann, 2014). These access points, published by a company, open up gateways for various parties to access defined data types, irrespective of the coding language. Most notably, the

provider defines which types of data can be accessed and interacted with. In this sense, APIs figure as a communication channel between programs rather than an explicit research data access.

The procedure for a researcher is as follows. A researcher who is interested in the data of an online platform, issues an requests for instance via a so called GET-function access (this usually refers to obtaining data). This data transfer process is mediated by the hypertext transfer protocol (HTTP) and the API, in turn, confirms whether access is granted to specific data in the database. Upon confirmation the receiver returns data (e.g., in *JSON* format) to the sender. These interaction opportunities allow companies on the one hand to either openly share data for other participants to deploy apps or other websites to interact with it or, on the other hand, to allow restricted participants access to more sensitive, curated data types and to monetise such proprietary data (Lahey, 2016; Lomborg & Bechmann, 2014). One major advantage for research approaches, is the interoperationability and possibility to automate the sampling, pre-processing and archiving procedure, which makes this data access efficient (Lomborg & Bechmann, 2014).

Caution needs to be exercised when using Twitter access points. As outlined in the second step of the research cycle (Figure 1.1) using an API endpoint might induce bias to the final sample retrieved. That is, access points for Twitter comprise the Streaming API and REST API. The Streaming API is oriented toward forward searches to capture in real time content on a post-level and author information that can be searched by keyword and hashtag searches. However, restrictions are implemented in the form of rate limitations and the amount of data that can be accessed. In the free version of the API access point, only 1 percent of tweets posted in real time can be obtained. This limitation to a percentage of the data stream poses problems to the sampling, as the premises under which precisely this 1 percent are selected are obscure. The REST API encompasses a search endpoint with which one can query Twitter specifically based on user-level that is user handles (i.e., self-assigned user-names). This allows to obtain also historical tweets for a user of up to 3,200 tweets, for the free access version. Furthermore, the number of requests that can be issued is restricted as well. This access point has been used in *Section 2.3.3* in order to model user behaviour relating to posting conspiracy beliefs over the course of time. I tried to mitigate endogenous selection bias by broadly identifying users based on keywords and subsequently sampling their historical tweet timeline and validating the selection rather than only focusing on hashtag or keyword-based identified tweets and their temporal patterns alone.

Pre-processing. One of the challenges when dealing with digital behavioural trace data is to *pre-process* raw data in order to obtain variables that concur with psychological constructs (see also Figure 1.1). In order to use text as data, free-text needs to be converted into a machine-readable format. As text data are

high-dimensional (i.e., the number of features is higher than the number of observations) (see Gentzkow et al., 2019), dimensionality needs to be reduced. There are multiple ways to achieve this. One relates to *tokenising* (that is reducing text data to meaningful entities such as single words, multi-word compounds, phrases that represent text features). However, tokenisation is not trivial for different languages (e.g., Chinese) and further decisions during the process and parameters chosen (e.g., the size of word units chosen) may lead to different measures. Raw data (particularly social media text data) often contain high rates of noise, such as abbreviations, misspelled words, *stopwords* (i.e., words with little semantic meaning such as conjunctions), informal language, non-ASCII characters, such as ampersands ("@") and punctuation. Reducing unstructured textual data to tokens, removing stopwords and noise, harmonises and narrows the textual data to diagnostic features (Gentzkow et al., 2019; Murphy, 2012). Choices of what constitutes noise depend very much at the down-stream tasks (e.g., clustering or prediction tasks) and eventually choices (such as choosing specific stopwords or augmenting to named entities) effect the results of the analyses (Mneimneh et al., 2021).

Another dimensionality reduction, usually, is reducing the dependencies of language (e.g., the sequencing of words) considered in text documents (e.g., tweets) (Gentzkow et al., 2019). For the sake of reducing language complexity, one paradigm is the *bag-of-words* vector representation. Within this word representation, one assumes that the frequency of words is indicative for its meaning (irrespective of the ordering of words) (Humphrey & Wang, 2018). Text data are then represented in an occurrence matrix—in which rows indicate the tokens and columns the documents, such as tweets. This representation allows to query documents by words and given similar column vectors, they can be considered to be related. With language modelling, assigning semantic meaning is challenging, as there is no one-to-one relations with a word or specific vocabulary and a specific meaning, due to the context sensitivity of language. With distributional semantics the assumption is that words that *co-occur* in similar contexts likely hold similar meaning. Hence, for the empirical study see *Section 2.3.3*, various pre-processing decisions, implications as well as choices at the stage of analysing have been deliberated.

The following chapter is concerned with addressing the analytic questions of systematising the field and providing a use-case for the theory-rooted application of digital behavioural trace data.

Chapter 2

Empirical Studies

This chapter lays out the three studies conducted to discern potentials of digital behavioural trace data for the application area of radicalisation. The first study reported is a translation of the article that appeared in German: Batzdorfer, V., Steinmetz, H. Bosnjak, M. (2020). Big Data in der Radikalisierungsforschung. Eine systematische Übersichtsarbeit [Big Data in Radicalization Research. A Systematic Review]. *Psychologische Rundschau*, *71*(2), pp. 96-102. Besides myself, Dr. Holger Steinmetz and Prof. Dr. Michael Bosnjak contributed to the creation of the first article. Dr. Steinmetz and Prof. Dr. Bosnjak acted as supervisors of my work. Regarding the second article (*Section 2.2*), besides myself, Dr. Holger Steinmetz contributed to the creation of the article, who consulted and supervised my work. Concerning the third article (*Section 2.3*), besides myself, Dr. Holger Steinmetz, Dr. Marco Biella and Dr. Meysam Alizadeh contributed to the creation of the article. Holger Steinmetz primarily supervised and consulted regarding time series analyses, whereas Marco Biella and Meysam Alizadeh consulted during the drafting of the article.

2.1 Study 1: Big data in radicalisation research. A systematic review

2.1.1 Abstract

The present study provides an overview of (i) the goals, data sources, and methods of digital behavioural trace data analysis chosen in radicalisation research, as well as exemplifies some of the results of these studies, and (ii) analyses the similarities and differences with traditional studies such as questionnaires or experimental studies. This systematic overview is based on 63 studies, of which, however, only a small proportion ($k = 18$) used digital behavioral trace data, while the majority consist of traditional approaches ($k = 52$). The results show that digital behavi-

oural trace data studies were largely aimed at identifying individuals with radical attitudes and predicting the development of radical views. Overall, behavioral trace data open up previously untapped potential for the analysis of personality profiles and the investigation of dynamic social interactions of those susceptible to extremist recruitment.

2.1.2 Introduction

The research of digital traces of behavior, e.g. postings on social media sites, click-behavior on websites or networking data of persons, has gained momentum in recent years. Such data offer an understanding of phenomena as they occur in real time in their natural environment (Landers et al., 2016). Thus, the collection of behavioral trace data enables the direct observation of behavior (e.g. acceptance of group norms, postings, or networking with people) and its determinants (e.g. personality traits) in a social context (e.g. in social networks on social media platforms) - with a low risk of bias, which is often present in traditional methods such as questionnaire data (Marres, 2015). Examples are the collection of data from online-forums, instant messaging and social networks such as Facebook or Twitter (Kosinski et al., 2016).

For research on radicalisation, access to digital behavioral trace data provides not only insights into the behavior of hard-to-reach individuals in situ (e.g., people with extremist attitudes), but also the observation of precisely those social environments in which radicalisation takes place and by which it is promoted (Ebner, 2019). For example, the online-milieu around platforms such as Gab, 4chan, 8chan, or Discord has been identified as a significant site of radicalisation processes after the Christchurch assassination or the leak of the right-wing extremist forum 'Iron March' (Munn, 2019). The fact that these milieus are difficult to regulate, are only partially visible from the outside and operate in the guise of anonymity, seems to promote escalating dynamics and raises questions about the conditions under which extremism-promoting beliefs, attitudes and dispositions arise (Munn, 2019; Pelzer, 2018).

Against the background of the value of digital behavioral trace data, the question therefore arises as to the relative importance of research with behavioral trace data. In particular, this study has the following purposes: Firstly, a systematic overview sheds light on the research goals, data sources, and methodological approaches that are the focus of current research with behavioral digital trace data and the results of this research. Secondly, similarities and differences of such studies with "traditional" approaches (questionnaire studies and experimental studies) are highlighted, in order to illustrate how the different research approaches complement each other.

Background

While research into radicalisation tendencies and their determinants has a decades-long tradition, the relevance of this research has increased in recent years (Schuurman, 2018). Using trace data in this area is fruitful not only because of the unobtrusive collection of data (as opposed to questionnaires), but also because it offers the possibility of analyzing radicalisation processes on social media platforms and thus at the very place where they take place. In this context, Ebner (2019) speaks of such platforms as "radicalisation machines" (p. 10), which enable radicalisation processes to the present extent.

In addition to the data that can be extracted by social media platforms, another source of digital behavioral trace data are open-source data (e.g. consisting of data sources such as PIRUS or ECDB). These sources provide information such as media reports, event data, and material from extremists, government documents, trial records and press releases from the American-speaking world. These sources provide anonymised background information on individuals who have links to extremist organizations or who have themselves demonstrated ideologically motivated criminal activities. These background characteristics can be demographic or biographical features, or information on mental health, ideological background, and time period of radicalisation, group dynamics or recruitment mechanisms. This possibility of viewing offline-characteristics is much more limited for social media approaches. Only political attitudes of the users (cf. Fernandez, Asif & Alani, 2018) as well as geographical localization and possibly related general sociographic data (cf. Mitts, 2019) can be extracted or inferred from statements.

As described at the beginning, there is a lack of a systematic overview of the use of such data sources and research designs, their questions and the comparison with traditional designs. In the following, the systematic of the literature search will be described, followed by a presentation of descriptive characteristics of all identified studies. Finally, a systematic analysis of the behavioral trace data studies will be undertaken.

2.1.3 Method

The search for relevant research was based on the PRISMA guidelines (Moher et al., 2009), which divide the search process into the steps 'identification of publications', 'screening', 'proficiency testing' and 'inclusion'. Selection criteria were **(i)** the application of research designs with possibilities for quantitative analysis (digital behavioral trace data, self-reports or experiments); **(ii)** focus on the following forms of radicalisation: political extremism (e.g. right or left extremism), religious fundamentalism (e.g., Islamism), nationalist/separatist extremism, 'single-issue' extremism (e.g. environmental protection or abortion) or ideologically independent

extremism; **(iii)** research on radicalisation determinants at the *micro-level* (e.g. psychological predispositions), *meso-level* (exposure to radical social environments) or *macro-level* (structural conditions, such as housing segregation or unemployment rates).

The selection of studies included those that focused on violent manifestations of radicalisation as well as its determinants. These were, for example, violent convictions and attitudes of persons or the willingness to use violence. In contrast, studies that investigated broader attitudes or dispositions (e.g., right-wing authoritarianism or social dominance) were excluded. The search was carried out for the period 2005-2019—especially since, beginning with the second wave of terrorism research and the emergence of new methods, the phenomenon of radicalisation increasingly came into the focus of research (Pape, 2009). The search was conducted using five databases and six other resources (e.g., *PubPsych*). Finally, only studies that focused on populations in the USA and Europe were included.

The information extracted from the articles falls into four categories: **(i)** survey mode (digital behavioral trace data, self-reports, experiments), **(ii)** analysed behavioural determinants (psychological dispositions, demographic characteristics, exposure to radical contexts, emergence of radical framework conditions), **(iii)** results of radicalisation processes (violent behavior, readiness for violent behavior, attitudes towards extremism, type of extremism), **(iv)** population (e.g. geographical context, sample size, age distribution). The initial screening of the publications was carried out by three independent coders. Selected full texts were checked for suitability by the first author.

2.1.4 Results

Study Description

Of the 6,602 studies resulting from the database search, only 63 met the inclusion criteria. This is due to a very high proportion of qualitative or purely conceptual papers. As expected, the majority of the studies were studies with traditional designs—i.e., based on self-reports ($k = 38$) and experiments ($k = 14$). A small part was related to the collection of trace data ($k = 18$). This group could in turn be differentiated into studies that collected behavioral trace data on social media ($k = 8$) and those that were based on open-source secondary data ($k = 10$).

Regardless of the design, about 27 percent of all studies ($k = 17$) dealt with Islamist fundamentalist extremism. The remainder of the studies focused on ideologically independent extremism ($k = 12$), right-wing extremism ($k = 6$), left-wing extremism ($k = 1$), nationalist/separatist forms ($k = 1$) and mixed forms ($k = 13$). A total of 101 samples were examined - about 30 percent of these ($k = 31$) consisted of adults from the general population, while the rest were students/pupils

($k = 29$), Muslim sub-populations ($k = 9$) or other special sub-populations such as offenders ($k = 10$) or activists ($k = 6$).

Aims, data sources and methods of behavioral analysis

Table 2.1 summarises the results of the overview of studies based on digital behavioral trace data. In terms of interest, these studies can be categorized into the following groups: (a) analysis of the role of experiences of discrimination and deprivation in the process of radicalisation ($k = 3$), (b) identification of radicals and prediction of their development ($k = 6$), (c) characterization of individuals with regard to psychological predispositions $k = 3$, or (d) comparison of different groups (e.g. of 'lone wolves', gangs, converts, or by type of offence) ($k = 6$). As far as the data source is concerned, about half of the trace data studies used open-source data ($k = 10; 55.6\%$) and the other half used social media data ($k = 8; 44.4\%$). The latter referred exclusively to the platform 'Twitter'. As can be seen in Table 2.1, open-source-based studies mostly analysed information on criminally convicted persons.

In terms of content, they mainly referred to Islamist radicalisation and its determinants. Such determinants were, for example, marital status or the existence of intact relationships, mental health, trauma, or post-traumatic stress disorders (LaFree et al., 2018). Exemplary for the characterization of personal psychological predispositions is the work of Jasko et al. (2017) who, based on the PIRUS (Profiles of Individual Radicalisation in the United States) data, used a sample of almost 1,500 political extremists. The most important outcome variable was whether the illegal act committed was violent (e.g. bombing) or non-violent (e.g. illegal protest). It was found that individuals more often used violence to pursue their ideological goals when they had experienced failure situations at work and when they had problems in social relationships. These results provide evidence of the connection between the motivation to feel significant and the use of political violence.

Another example of an open-source-based study is the study by Pyrooz et al. (2017), which used a combination of the PIRUS database and the National Longitudinal Survey of Youth (NLSY97) database to compare two types of groups - criminal but non-political gangs ('street gangs') and extremist groups. The aim of the study was to identify differences between the groups in terms of length of membership and demographic, family and socioeconomic characteristics. In addition, the authors investigated whether members of extremist groups had a history as gang members. The result was that only six percent of extremist persons had previously been in a street gang. With regard to group membership, only marital status (gang members were less often married and less often parents), ethnicity (whites were more likely to be in extremist groups, non-white minorities more likely to be

Table 2.1: *Prevalence of determinants and outcomes of radicalisation*

	Islamism		Polit. & Islam.		Non-ideol.	RWE/+
Determinants	SM	OS	SM	OS	OS	OS
Psychological Dispositions						
Psychopathology/ mental health	0	0	0	1	2	0
Personality profiles	2	0	1	1	1	0
Trauma, injustice and alienation	1	1	0	0	0	0
Personal status and rewards	1	0	0	0	0	0
Disinhibition of moral inhibitions	0	0	0	0	0	0
Self-control	0	0	0	0	1	0
(Institutional) trust	1	0	1	0	0	0
Risk taking	1	0	0	0	0	1
Intolerance of ambiguity	0	0	0	0	0	0
Consumption (drugs, alcohol)	0	0	0	0	2	0
Demographic Characteristics						
Work history	0	0	0	3	1	0
Educational background	1	1	0	1	1	1
Family status	0	0	0	2	0	1
Military experience	0	0	0	3	0	0
Social relations (intimate, peers, family)	0	0	0	1	1	1
School or work success	0	0	0	2	1	0
Crime-related Background Characteristics						
Criminal record	0	0	0	1	1	1
Parental violence and abuse	0	0	0	1	1	0
Further Individual Features						
Gender	2	1	0	3	0	0
Age	2	1	0	2	0	1
Religion	0	0	0	1	1	0
Ethnicity	0	0	0	2	1	0
Socioeconomic status	0	0	0	0	0	0
Political orientation	0	0	0	0	0	0
Exposure to Radical Contexts						
Social network: radical peers/ family members	2	0	1	1	1	2
Gang affiliation	0	0	0	1	1	1
Emergence of Radical Environments						
Housing segregation	0	0	0	0	0	1
Sociodemographic (share of poverty, unemployment, religion	1	0	0	0	0	1
Media	2	0	0	0	1	1
(Foreign) politics	3	0	0	0	0	1
Group-related Grievances						
Relative deprivation (marginalization)	1	1	0	0	0	0

Note. Included are 18 publications. SM = social media, OS = open source, Polit. & Islam. = political extremism in combination with Islamism, non-ideol. = non-ideological extremism, RWE/+ = right-wing extremism in combinatin with single-issue or Islamism

in street gangs) were more predictable. The role of gang membership depended on the religious community in question: while people with a Christian background were far more likely to belong to a street gang, the opposite was true for members of all other religious communities. Finally, members of extremist groups show an ostensibly higher level of education than members of street gangs.

In contrast to the open-source-based studies, studies focused on social media either analyzed postings using 'Text mining' or applied networking approaches to investigate social relationships between people. The studies based on the postings pursued the goal of classifying postings, e.g. in terms of the extent to which they reflected the perception of discrimination (Lara-Cabrera, Gonzalez-Pardo & Camacho, 2019), signaled support for extremist groups (Fernandez, Asif & Alani, 2018), or showed signs of incipient radicalisation. The latter was operationalised,

for example, through the first use of ideological rhetoric or the dissemination of fundamentalist content from known accounts, by the individual (Rowe & Saif, 2016).

To categorize the postings, linguistic features of the statements were used. These were stylistic features (e.g. the omission of sentence parts and the capitalization of whole text parts as markers for introversion and frustration), content-related terms (e.g. hashtags, ideological or political terms such as the naming of war zones) and terms which, although not related to content, nevertheless prove to be predictive (e.g. emotion words such as 'ugly' or 'nasty' which reflect affective processes) (cf. Alizadeh et al., 2019).

In addition to the text analysis of the postings, some studies aimed to analyze the networks of individuals, e.g. what role the networking density of participating individuals plays (Reganti et al., 2017), or what predictions metadata (e.g. existence of an account suspension or geographical data of individuals) provide for radicalisation. However, these were then only occasionally combined with "open-source data" such as regional election results or unemployment rates in order to estimate the spread of political attitudes or structural disadvantage in the immediate vicinity of the users (cf. Bail, Merhout & Ding, 2018; Mitts, 2019). For example, Mitts (2019) examined whether membership of an extremist group was influenced by experiencing anti-Muslim hostilities. In the study, postings from jihadist Twitter accounts were extracted and then classified according to various dimensions of ISIS-sympathy and persons were assigned to geographical locations. It was shown that people who were located in regions where anti-Muslim parties are strongly represented were more likely to show signs of radicalisation than others in less hostile locations (Mitts, 2019). It must be noted, however, that although regional unemployment and the occurrence of terrorist attacks have been statistically controlled, this is only weak evidence of the assumed effect.

Similarities and differences to traditional studies

While trace data studies provide unique results due to these special data sources and forms of analysis, a comparison with 'traditional' studies (experimental studies or studies based on self-reports) also shows some overlaps. Experimental studies ($k = 4$) and questionnaire studies ($k = 9$) focused on the impact of experiences of discrimination and deprivation. For example, Bäck et al. (2018) investigated in their laboratory experiment the effect of social exclusion on the acceptance of the political attitudes of a radical group. The basis of the experiment was the 'cyberball paradigm' in which participants play an online-game with (allegedly) other people. In the study with 71 students, half of the people in the exclusion condition suddenly stopped being involved in the game. When the persons received

a message from a fictitious member of a radical left-wing group after the end of the game, it became apparent that those persons who were particularly sensitive to rejection had an increased tendency to adapt their attitudes to those of the radical group.

Furthermore, similarities between traditional studies and trace data studies focusing on the influence of peer groups on the imitation or reinforcement of extreme political attitudes or behavior were found. Dahl (2017) used social network analysis to investigate how peers affect the attitudes and values (including advocacy of political violence) of young people in Sweden and whether these attitudes and norms influence their choice of friends. It was found that peers influence attitudes towards migrants, but the same effect does not apply to general political (universalistic) value orientations. In contrast, a universalistic peer network showed a reducing effect on support for political violence.

The most obvious difference between trace data studies and traditional studies is the form of data collection. Here, behavioral trace data have the enormous advantage of extracting behavioral data not affected by self-perception and desirability tendencies and this also in a far larger number of cases than in traditional studies. In contrast, trace data are less helpful when it comes to measuring psychological characteristics such as personality traits, where aspects such as reliability or validity are often unclear or, in the worst case, insufficient. And even if, for example, linguistic features of a text prove to be of predictive use, it is often unclear which construct was actually measured here. In this context, traditional questionnaires are irreplaceable despite their weaknesses. For future research, forms of triangulation would be helpful, in which both behavioral data on trace data are collected, enriched by traditional measurement with questionnaires. Similarly, field or natural experiments in combination with both data sources should make it possible to investigate the impact of interventions or naturally occurring events (e.g. changes in legislation) on radicalisation processes.

2.1.5 Discussion

Considering the importance of digital trace data - especially extracted from social media platforms and open-source sources - this overview of the field of radicalisation research shows that not only is the number of studies on this topic limited (cf. Schuurman, 2018), but also the range of analysed platforms: Although social media platforms essentially represent the social spaces in which radicalisation processes take place (Ebner, 2019), the results show that only a few studies analyse social media data. The sole focus on Twitter in this context is already criticised by Parekh et al. (2018). Lesser-known platforms such as *4chan* have so far been insufficiently considered in terms of their relevance and reach for the radicalisation

process (Schmid & Forest, 2018). In view of the intensive linkage and interaction of social networks (cf. Johnson et al., 2019), a holistic view across platforms is lacking, as is an answer to the question of whether determinants and conducive environments that have been analysed on one platform can be generalised to others. This is of relevance, especially since predominantly verbal behavior is observable on Twitter, while other platforms are more strongly characterised by visual elements (e.g. so-called 'memes' - i.e., rapidly spreading images with pointed verbal expressions) (Munn, 2019). Other platforms, such as the 'Iron March Forum', are strongly characterized by anonymity, irony and acronyms and cannot be quantified with classical text mining approaches. The latter illustrate new challenges in the evaluation and transferability of previous theoretical assumptions to these milieus.

While questionnaire studies are often criticized for the risk of bias due to measurement errors and desirability tendencies, digital behavioral trace data analysis also face problems: While demographic characteristics can easily be extracted, the extraction of context data (e.g. number of retweets, number of friends) and user-generated content (e.g. text content, likes of other users' statements, self-reported individual differences) must be done with respect to the target construct, taking into account the context in which the behavioral traces were created when interpreting them (cf. Landers et al., 2016). In order for digital behavioral trace data analyses to acquire theoretical relevance, it is essential to integrate them into a '*data* or *measurement theory*' that conceptualises behaviour as a product of the interaction between person and situation (Landers et al., 2016).

Finally, digital behavioral trace data analyses offer an understanding of radicalisation, which is caused by determinants that partly stem from the biographical course of development (e.g. experienced deprivation). While this is a clear causal focus, existing studies are based almost exclusively on cross-sectional designs. With the newly emerging possibilities offered by digital behavioral trace data, the focus should be on the integration of traditional approaches and new technologies in order to map the process character. As an example, approaches such as online field experiments on the dissemination of emotional states in social networks, as already implemented by Kramer et al. (2014), could provide new insights into the milieu and have heuristic significance and explanatory value.

2.1.6 Implications

This study unveiled various research gaps. On the one hand, I will argue for the benefits of introducing longitudinal perspectives in this field in *Section 2.3*. I take this aspect up when setting out the theoretical gain of time series analyses to understand within-subject and between-subject dynamics, based on digital behavioural trace data. On the other hand, I argue to strengthen the aspect of a measurement

theory and acknowledging data-generating aspects throughout different stages in the research cycle (see Figure 1.1). Further, aspects of construct validity, sampling biases, decisions in the analysis stage and validation are taken up in *Section 2.3*.

2.2 Study 2: Reviewing radicalisation research using a network approach

The following section corresponds to a slightly abridged version of the initially submitted yet published paper "Reviewing radicalisation research using a network approach" (Batzdorfer & Steinmetz, 2020).

2.2.1 Abstract

This study provides a an innovative approach to systematic review. We apply a network approach for analysing the most prevalent constructs and related hypotheses in the literature. Network analysis is particularly useful in this context because, it allows the visualisation of the structure of constructs and hypotheses proposed in the field as well as the identification of crucial concepts. The review reveals differences across empirical approaches and closes with a discussion of over- and underresearched constructs, their generalisability across research approaches, and potentials for future research. We conclude by recommending a stronger integration of constructs and perspectives as well as a more rigid consideration of causal inference.

2.2.2 Introduction

In an effort to understand the causes of violent extremism, alongside how it develops and persists, a plethora of research was produced (Horgan, 2008). Notwithstanding the intense interest in the issue of radicalisation, the field still lacks a coherent understanding of the structures and cognitive and emotional processes by which some individuals come to adopt extremist ideologies and engage in ideologically motivated violence (Borum, 2011; Sageman, 2014; Wolfowicz, Litmanovitz, Weisburd, & Hasisi, 2019). Recent research has begun to investigate causal mechanisms (e.g., the role of criminogenic constructs such as low self-control or social control, see Opp, 2019). Extant research on radicalisation has been characterized by a lack of applied empirical methods or a focus on selective populations (e.g., mainly focusing on radical Islamists, see Klausen, Campion, Needle, Nguyen, & Libretti, 2016), and a narrow focus on the choice of dependent variables (e.g., only studying successfully committed violent acts) (cf. LaFree, Jensen, James, & Safer-Lichtenstein, 2018).

Because studies on political radicalisation are extremely diverse, an overview of the various scientific perspectives, constructs, hypotheses, and analytical approaches would lay the groundwork for cumulating knowledge and enable the creation of guidelines for future research. In recent years, a number of review papers have been published (Desmarais, Simons-Rudolph, Brugh, Schilling, & Hoggan, 2017; Hassan, et al., 2018; McGilloway, Gosh, & Bhui, 2015; Pelzer, 2018; Vergani, Iqbal, Ilbahar, & Barton, 2018) that shed light on the current state-of-the-art. Some reviews have a broad focus, covering different radicalisation risks, protective constructs or correlates (Christmann, 2012; Lösel, King, Bender, & Jugl, 2018; Wolfowicz et al., 2019), while a smaller number focus on a specific selection of constructs, such as social cohesion (Grossman & Tahiri, 2015). Likewise, some systematic reviews attempted to evaluate the psychometric properties of existing measurement instruments, such as Scarcella and colleagues' (2016) investigation of risk assessment tools, which mainly focused on self-reports of attitudes toward terrorism, extremism, or radicalisation. While having tremendously increased the knowledge in the field, limitations of these reviews include their focus on specific data sources and research approaches (e.g., self- report research), whereas an overall integrative overview is missing.

Of the aforementioned reviews, the meta-analysis by Wolfowicz et al. (2019) is the most comprehensive approach to date. The authors quantitatively summarized effect sizes of 57 studies referring to 60 individual level protective and risk factors for radical attitudes, intentions, and behaviours. The study resulted in a rank-order of effect sizes. The present study seeks to build on this meta-analysis. Whereas Wolfowicz et al. (2019) provided solid evidence about the strengths of relationships, our study approaches the field by analyzing the complex structure of all considered theoretical constructs and hypotheses formulated by scholars of the field. This is achieved by applying a network approach (Van de Wijngaert, Bouwman, & Contractor, 2014; Wasserman & Faust, 1994), that allows us to visually represent the whole field, with its constructs represented by nodes and its hypotheses represented as directed edges connecting the nodes. The network analysis also enables us to identify central constructs and hypotheses, to compare the network of constructs and hypotheses across research approaches and, thus, to identify facilitators and limitations for testing certain hypotheses. Most importantly, the network analysis provides a basis for future research as it can help to identify crucial constructs to generate causal models and to make decisions about necessary control variables. By doing so, our paper contributes to the growing literature on causal modelling (e.g., Pearl, 2009; Shrier & Platt, 2008). A second goal of the study is to compare the network structure and, thus, analyse differences in relevance and interconnections of constructs across methodological approaches (e.g., survey research, experimental research and social media research). As our study

focuses on hypotheses and theoretical perspectives in the field with an emphasis on their structure, we provide an additional unique perspective on the field that fruitfully adds to the quantitative results provided by the meta-analysis by Wolfowicz et al. (2019). By doing so, our study shows the unique value and, thus, the synergetic potential of both quantitative meta-analyses and network approaches.

Radicalisation research: Determinants and research approaches

Recently, research on political radicalisation has become of tremendous interest for scientists and politicians as well as the general public. Especially crimes and terror attacks in cities like New York, Brussels, Christchurch, El Paso, or Paris, and an increased polarization of political discourse and ostentatious displays of emotional outrage on social media channels have led researchers to increase their efforts in the investigation of potential determinants of radicalisation processes.

Despite the intense interest in the issue of radicalisation, establishing a generic approach to examining the phenomenon has been hindered by the heterogeneous and ambiguous conceptualization of "radicalisation" in relation to concepts like "terrorism," and "extremism" (Schmid, 2013). Pathways into violent extremism are multilevel and involve factors spanning macro-, meso-, and micro- levels of analysis, combining intra- and interindividual dynamics and societal processes, while some factors are consistently reported across different contexts and across various ideological and political hues.

While the main focus of this research is the development of violence-promoting attitudes and beliefs or behaviors, existing studies diverge in their focus on potential determinants or chosen research approaches. Research on radicalisation is motivated by the interest in the causal processes leading to extremism, not only to understand social and cognitive processes leading to society-endangering perspectives, but also as a means to develop potential interventions.

To organise determinants, it is helpful to rely on multilevel theory (see Franc & Pavlovic, 2018; Schmid, 2013). From this perspective, determinants located on the *micro-level* reflect psychological constructs such as factors that comprise moral and cognitive propensities (e.g., authoritarianism), personality constructs (e.g., low self-esteem), demographic characteristics, experiences that increase the propensity to form extremist attitudes (traumatic events, military experiences), or political or religious affiliations. Determinants on the *meso-level* relate to the milieu of the radicalising person and, in particular, concern the processes and characteristics of the social groups or the influence of significant others. This social environment acts as a socialization background and serves as the surroundings for normative influences, the transfer of critical information, as well as emotional support and reinforcement of beliefs and attitudes. Finally, *macro-level* determinants are char-

acteristics or events on the regional or societal level, for instance, globalisation and modernisation (leading to alienation from values of society or loss of credibility of government and state structures) and foreign policy interventions (perceived as foreign occupation). Additionally, objective markers of inequality (e.g., national poverty) can exacerbate the subjective perception of deprivation and injustices.

Beyond the differences in their focus on a variety of constructs, studies have applied different research approaches to test hypotheses. Most research mainly applied *survey approaches*, to measure psychological constructs, such as personality traits, perceptions of deprivation, group threat, or uncertainty (Doosje, Loseman, & Van den Bos, 2013). Others measured psychological health (e.g., Bhui et al., 2019) or the prevalence of radical attitudes in the general population (Loza, 2011). In contrast, *experimental approaches* attempted to manipulate experiences of discrimination and deprivation and investigated their impact on radicalisation-prone attitudes or behavior (e.g., Dechesne, 2009), or analyzed the influence of media consumption on extremist attitudes (e.g., Frischlich, et al., 2015). The studies, focusing on *digital trace data*, gathered data from either social media platforms (e.g., postings on Facebook, or Twitter) or open sources (e.g., databases like PIRUS or ECDB, which contain coded information on individual background characteristics, based on media reports or government documents). This type of studies investigated radicalisation processes as a result of discrimination and deprivation experiences (e.g., Mitts, 2019) or attempted to identify users with radical attitudes (e.g., Egan et al., 2016). Others compared the demographic or psychological profile of different groups (e.g., of "lone wolves", gangs, converts, or types of offenses; e.g., Kerodal, Freilich, & Chermak, 2016; LaFree et al., 2018).

The use of network theory for the integration of research

For decades, there has been an ongoing discussion on how to synthesize the literature to integrate the diverse perspectives, analytical approaches, and conclusions. While the most original form of a narrative review has been, and still is, an important source of orientation for a field, its subjective character has led to criticisms with regard to the selection biases when searching for and collecting articles or the subjective biases of the reviewer when interpreting and integrating the research (e.g., Tranfield, Denyer, & Palminder, 2003). As a result of these criticisms, a strong focus on systematic reviews emerged, especially in medicine and related fields that focus on evidence-based decision processes (Pawson, 2006; Sacket, Rosenberg, Gray, Haynes, & Richardson, 1996; Tranfield, et al., 2003). Likewise, to quantitatively summarize research results and to investigate the heterogeneity in the field with regard to the results, meta-analyses have become widespread (Cooper, 2017).

Finally, there are approaches to systematically compare theoretical frameworks used in a field (Opp & Wippler, 1990).

In contrast to the established approaches, the application of network theory and related analytical procedures, as a means to summarize the perspectives, hypotheses, and constructs held in a scientific field, is new (Van De Wijngaert et al., 2014; McGlashan, Johnstone, Creighton, de la Haye, & Allender, 2016). Networks are used in a number of different fields and for analyzing different phenomena, ranging from, social groups and dynamics (e.g., Borgatti, Mehra, Brass, & Labianca, 2009; for social capital, see Burt, 2000), *communication structures* (Bavelas, 1950), *construct definitions and measurement* (e.g., application to psychopathological constructs, see Borsboom & Cramer, 2013), to *causal inference* (e.g., directed acyclic graphs, see Elwert, 2013; Pearl, 2009). As explained later in detail, the gist of these different applications is that agents or entities (e.g., persons, symptoms, or constructs) can be described with regard to their structural relationships to other agents or entities. These structural relationships can represent interpersonal relationships, logical connections, or causal effects, and the overall system can be described by a *graph* that represents the structure of *nodes* (e.g., persons, variables) and *edges* as their connections (relationships, causal relations). In recent decades, network theory has been associated with a host of analytical procedures to derive and analyze properties of the whole graph (i.e., on the graph-level of analysis) and to identify important nodes by their location in the network (i.e., on the node-level of analysis).

Van De Wijngaert et al. (2014) emphasized the merits of applying network theory for the purpose of integrating research in a field. From this perspective, a research field focusing on some phenomenon can be represented as a graph which consists of nodes, representing constructs (e.g., radical attitude or personality traits) and the edges representing the hypotheses held in the field. Whereas overall network theory allows edges to be either undirected or directed, an edge in the present network-based review is always directed and represents a causal hypothesis formulated in the field. Figure 2.1 represents an example. In the figure, a directed edge linking personality and extremist attitudes would represent the hypothesis of one or several studies that some personality trait has a causal effect on radical attitudes. Furthermore, the different number of posed hypotheses can be visualized by the degree of thickness of edges referring to the node. Differences in the prevalence of certain constructs under consideration can be illustrated by the size of the nodes. In this example, Figure 2.1 indicates that the field was dominated by hypotheses focusing on the role of extremist attitude for extremist behavior whereas the examined papers seldom hypothesized the role of demographics.

Beyond the intuitive appeal of representing an entire field in one graph, a wide array of network analytical methods can be applied to quantitatively characterize

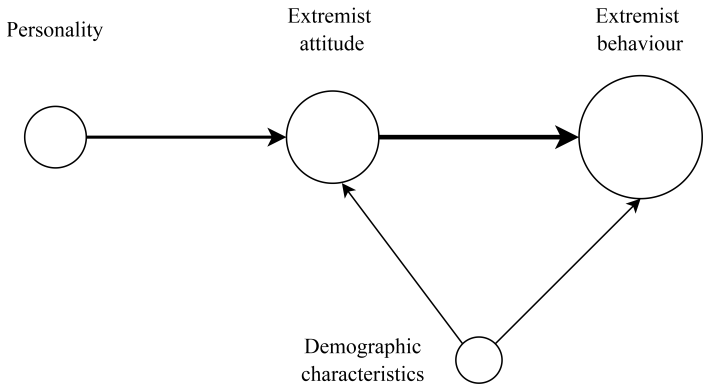

Figure 2.1. Exemplary network structure

the domain and to identify central constructs. Finally, the structure of the graph can be used to inform the field about potential opportunities to generate causal models (Elwert, 2013), including mediating processes (MacKinnon, Fairchild, & Fritz, 2007) or to reduce the danger of confounding bias (Vanderweele, 2019). For instance, from the network in Figure 2.1, one could conclude that extremist attitudes mediate the effect of personality on radical behavior (cf. Ajzen, 2005) or that demographic characteristics—due to their joint effect on extremist attitudes and behavior—confound the relationship between both. An important implication of the approach is that parts of the network may stem from exclusive sets of studies, in which some studies focused solely on one relationship, but not on others.

Finally, the network approach provides a basis to decide whether sampling specific subpopulations with a specific profile or values of some variable (e.g., focusing on only individuals already radicalised) is appropriate in order to avoid endogenous selection bias (Elwert & Winship, 2014). In this regard, Elwert and Winship suggest caution when selecting subsamples on the basis of some dependent variable.

The present study represents an attempt to use network theory to integrate the extant research on radicalisation to form a global network structure that illustrates the current state of thinking as well as the dominant and less dominant constructs and hypotheses. By creating different networks for the diversely used research approaches (i.e., survey research, experimental research, and research using online trace data), network analysis allows us to identify approach-specific constructs and perspectives in radicalisation research.

2.2.3 Method

Our inclusion/exclusion criteria and search strategy drew on Wolfowicz et al. (2019) who used the two-pyramid model (McCauley & Moskalenko, 2017). That is, in a similar vein, we distinguished cognitive and behavioral radicalisation and considered radical attitudes, intentions and behaviors as useful determinants and out-

comes in the radicalisation process. Likewise, the choice of relevant databases was informed but not limited by those of former meta-analyses. We applied the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting standards to describe the search flow and screening results and guarantee transparency. To identify papers, we searched in five databases and search engines (e.g., PubPsych, Medline, PsycINFO, SSRN, ISI Science, ACM Digital Library, JSTOR, The Campbell Collaboration Library, NCJRS) together with handsearching (e.g., Voxpol Network of Excellence, International Centre for Counter-Terrorism [ICCT] or Perspectives on Terrorism). We focused on the literature spanning a 15-year publication range (2004 – 2019), reflecting the point at which the concept of "radicalisation" started to appear more frequently in literature (Neumann & Kleinmann, 2013).

We included studies which had applied (i) empirical research approaches or analysis that formulated explicit hypotheses (digital behavioral trace data, self-reports or experiments); (ii) focused on the following forms of radicalisation: political extremism (e.g., right-wing or left-wing extremism), religious fundamentalism (e.g., Islamism), nationalist/separatist extremism, "single-issue" extremism (e.g., environmental protection or abortion), or ideologically independent extremism; (iii) focused on populations in the U.S. and the European region, in order to guarantee comparability by similar cultures and economic prerequisites (cf. Zhirkov, Verkuyten, & Weesie, 2014). The selection of studies included those that had focused on violent manifestations of radicalisation (e.g., violence-promoting beliefs, attitudes, intentions, or behaviors) as well as its determinants. This differentiation of behavior and beliefs connects to the issue that the latter is much vaguer which in turn extends the scope of possible measures targeting beliefs (cf. Wolfowicz et al., 2019). In contrast, we excluded studies that solely investigated broader attitudes or dispositions (e.g., right-wing authoritarianism or social dominance) without direct connection to radicalisation. Due to the comparable search terms, databases, and inclusion criteria, we ended up with a sample with a large overlap especially with the meta-analysis by Wolfowicz et al. (2019) providing the opportunity to integrate their results with the results of the network approach.

Network measures

We calculated various forms of centrality measures to analyze properties of the nodes (i.e., the analyzed constructs) in the network. Overall, the centrality concept reflects the importance of a node in the network, resulting from its location and structure of relationships to other nodes. Applied to our context, a high-centrality construct would reflect the prominence and importance of a certain radicalisation

construct. The centrality measures we take into consideration are *in-degree centrality*, *out-degree centrality*, *closeness centrality*, and *betweenness centrality*.

In-degree centrality ($D^{+}(v)$). This measure reflects the number of directed edges the target node receives. Applied to our context, a construct with high in-degree centrality is often conceptualized as a dependent variable.

Out-degree centrality ($D^{-}(v)$). This measure reflects the number of directed edges originating from the target node. Applied to our context, a construct with high out-degree centrality is often hypothesized as a causal determinant of other constructs.

Closeness centrality (C_v). Closeness centrality is the most intuitive measure on the importance of a target node and is defined as the reciprocal of the sum of paths by which the node is connected to all other paths. At an extreme, a node may be directly related to all other nodes, thus, resulting in a closeness centrality value of 1. The more other nodes the target node has to pass to reach another node, the lower the closeness centrality and the lower the numerical value. In our context, a construct exhibiting a strong closeness centrality is one that is the main focus of all the research examined here as illustrated by the fact that many hypotheses directly address this construct.

Betweenness centrality ($B(v)$). Betweenness centrality reflects the "broker" or bridging" function of a node connecting otherwise disconnected partitions of a network. In particular, a high betweenness centrality occurs when the target node is located within many indirect connections between other nodes. This concept has become popular in Burt's (2000) structural holes theory that describes the conditions of high-power individuals in complex networks. Applied to our context, target constructs with a high betweenness centrality are powerful bridge builders between distant constructs and may reflect either *mediators* (i.e., variables, transmitting an effect from the cause to the outcome), *confounders* (i.e., variables that affect two target variables and create a spurious relationship), or *colliders* (i.e., variables that are mutually influenced by two variables) (Elwert, 2013). Hence, identifying those constructs provides a fruitful basis for guiding future research with regard to clarifying the potential causal role of the respective construct.

Network density. On the level of the network, we analyzed the density of the network, which reflects the density or scarcity of hypotheses in the field. A dense network is a network in which the number of edges is close to the maximum. A network with small number of ties is called scarce. The density of a network is calculated by dividing the number of edges in the network by the number of edges possible, in case the network is a completely linked network. It ranges around values between 0 and 1 in the binary number system. The 0 value demonstrates that there are no ties between constructs. Applied to the area of systematic reviews, a dense research field implies lack of parsimony (Van De Wijngaert et al., 2014),

that is, a proliferation of constructs without integration into an overall framework with common pathways and mediating processes.

Analytical procedure

Coding of articles. We coded the articles according to four categories of information: (i) the analyzed constructs, that is, constructs on the micro-level (i.e., individual-related constructs), meso-level (i.e., group and relationship-related constructs), and macro-level (i.e., societal constructs), (ii) information about the hypotheses, and (iii) the chosen research approach (i.e., survey approaches, experimental research, and digital trace data approaches).

To organize the constructs and to analyze the constructs and hypotheses with a network model, we aggregated constructs to higher-level constructs. Table 2.2 depicts the constructs extracted from the studies and the higher-order constructs.

Analyses. After data extraction, the hypotheses were transformed into a "node and edge list," which contained the pair of the respective independent and dependent variables implied in the hypothesis and the unique ID of the respective studies to enable referring the study to additional attribute information (e.g., the applied research approaches). The order of the pairing is meaningful, as it indicated which construct was hypothesized as an independent variable and which was hypothesized as a dependent variable. After creating the node and edge list, we calculated the network measures (e.g., betweenness centrality). The network statistics were calculated using the *igraph* package in the software R (R Core Team, 2018). The edge and node list was imported in the open-source network visualization software Gephi (https://gephi.org/).

2.2.4 Results

Descriptive Results

The data extraction led to a total of 57 articles containing 777 constructs which—when aggregated to 25 higher-order constructs (see Table 2.2)—resulted in 244 hypotheses containing a unique combination of independent and dependent constructs. Table 2.3 shows the number of studies and the number of constructs considered in the three research approaches.

Overall, the majority of studies ($k = 27$) applied a survey approach and used self-report questionnaires to measure target constructs whereas 14 studies conducted experiments and 16 gathered trace data. Survey studies predominantly measured demographic variables ($k = 15$) or social status (k = 16) as these variables are easily measured via self- report and reflected research that aimed at targeting at-risk individuals on the basis of these surface-level indicators. Likewise, studies

Table 2.2: *Coding of categories extracted from hypotheses and their respective definitions*

Construct	Higher-order construct
Individual-related constructs (micro-level)	
Non-violent behavior (e.g., protest, support for non-violent organisations)	Activism
Criminal activity before radicalisation (conviction, violence against property or people)	Criminal history
Potential trauma, triggering events, abused childhood	Critical events
Gender, age, marital status, ethnicity, citizenship	Demographics
Stable individual traits (personality, intelligence, self-control, coping skills, need for order, extroversion, risk seeking, authoritarianism)	Dispositions
Genetic factors	Genetics
Search for purpose in life, significance, uncertainty avoidance	Meaningfulness
Military training and serving military services	Military experience
Psychological disorder or chronic impairment of wellbeing or social functioning (mortality salience, psychosis proneness, depression)	Psychological health
Ideology, support for instrumental violence (voice grievances, desire to hurt others, opposition to equality, persuasiveness of radical content)	Radical attitudes
Violent (attempted) offense (e.g., bombing) or unusual behavior (e.g., travel abroad, lifestyle changes, risky behavior), delinquency	Radical behavior
Religious membership (e.g., Christianity)	Religious affiliation
Attitudes toward duties and morality (e.g., self-sacrifice for a higher cause)	Religious beliefs
Religion-related behaviors (e.g., prayer frequency, conversion, mosque attendance)	Religious practices
Education, income, employment, status seeking	Social status
Emotional responses and sensitivity (e.g., situational hatred, frustration, affective valence)	State
Drug or other substance consumption or addiction	Substance abuse
Group and relationship-related constructs (meso-level)	
Commitment and loyalty, or development of close group relationships (ingroup identification, gang member, social support)	Cohesion
Shared beliefs and attitudes, biases in evaluation of events or people (ingroup superiority, symbolic threat, collective relative deprivation)	Group processes
Connectedness to family and intimate relationships and social control	Significant others
Rejection or exclusion by the group or individual representatives of a group (target of prejudices, socially isolated)	Social exclusion
Peer pressure, recruiting or influence of information sources/narratives (propaganda consumption, epistemic authority figures, peer immersion, lexical homophily)	Social influence
Societal constructs (macro-level)	
Dual (ethnic) identity, alienation or distance to people and mainstream society (perceived identity incompatibility)	Integration
Population-level estimates of disadvantage: economic (GDP, poverty rate) or sociopolitical (political participation, share of foreign-born residents, hate crimes)	Objective inequality
Individual perceptions of deprivation: economic (income dissatisfaction) or sociopolitical (legal cynicism, anti-government beliefs, unfair treatment by police, religious suppression)	Subjective inequality

Note. Examples for categories extracted (left column) are nonexhaustive

Table 2.3: *Number of studies across higher-order constructs and research approaches*

Construct	Research approaches		
	Survey approaches	Experimental approaches	Trace data approaches
Individual-related constructs (micro-level)			
Activism	1	3	3
Criminal history	4	0	4
Critical events	5	2	5
Demographics	15	1	6
Dispositions	13	8	5
Genetics	1	0	0
Meaningfulness	7	4	1
Military experience	0	0	5
Psychological health	6	3	4
Radical attitudes	27	11	12
Radical behavior	8	3	8
Religious affiliation	7	0	1
Religious beliefs	8	3	2
Religious practices	5	0	3
Social status	16	0	7
State	1	8	1
Substance abuse	2	0	2
Group and relationship-related constructs (meso-level)			
Cohesion	6	4	9
Group processes	10	2	2
Significant others	3	0	5
Social exclusion	11	2	6
Social influence	9	2	6
Societal constructs (macro-level)			
Integration	14	0	0
Objective inequality	9	0	3
Subjective inequality	18	3	5
Total number of studies	27	14	16

with a survey approach often measured radical attitudes and intentions ($k = 27$) or dispositions ($k = 13$), as these constructs, due to their subjective nature, are suitable for measurement by self-reports. Constructs belonging to the "integration" category ($k = 14$) were exclusively investigated by self-reports and referred, for instance, to dual identity and perceived identity incompatibility (see Simon, Reichert, & Grabow, 2013).

Constructs considered in the category of experimental approaches were either experimentally manipulated (e.g., the experience of social exclusion, see Pretus et al., 2018) or measured as an outcome or covariate. Analogously, the constructs considered most frequently were dispositions ($k = 8$) and radical attitudes and intentions ($k = 11$) (e.g., perceived persuasiveness of radical content or the advocacy of violence for political goals), or emotional states (e.g., situational hatred or frustration, $k = 8$).

The studies that had collected digital trace data from social media and open sources ($k = 16$) focused on the role of cohesion in groups ($k = 9$), for instance,

established in open sources through extremist group membership or movement-related tattoos (see Kerodal et al., 2016). Similarly, radical behavior figured prominently in open sources ($k = 8$), distinguishing pre-attack behavior, lifestyle changes, and types of crimes (spontaneous vs. planned, offenses against property vs. civilians) (see, e.g., Corner & Gill, 2015; Sweeney & Perliger, 2018). On behalf of social media records, constructs reflecting radical attitudes comprised positive statements about ISIS ideology or expressed threats against others (see Mitts, 2019).

Figure 2.2 shows the network of constructs and hypotheses illustrating the radicalisation field. Overall, the research field reflects a substantially dense network (density = .407), implying a vast number of hypotheses and a lack of a parsimonious structure. Table 2.4 reports the associated network measures. Whereas the centrality measures reflect the number of hypotheses linking two constructs, their weighted forms consider the number of studies which had tested a referring hypothesis. In particular, the weighted in-degree centrality reflects the number of hypotheses expressing an effect on the respective construct weighted by the number of studies which had tested such a hypothesis.

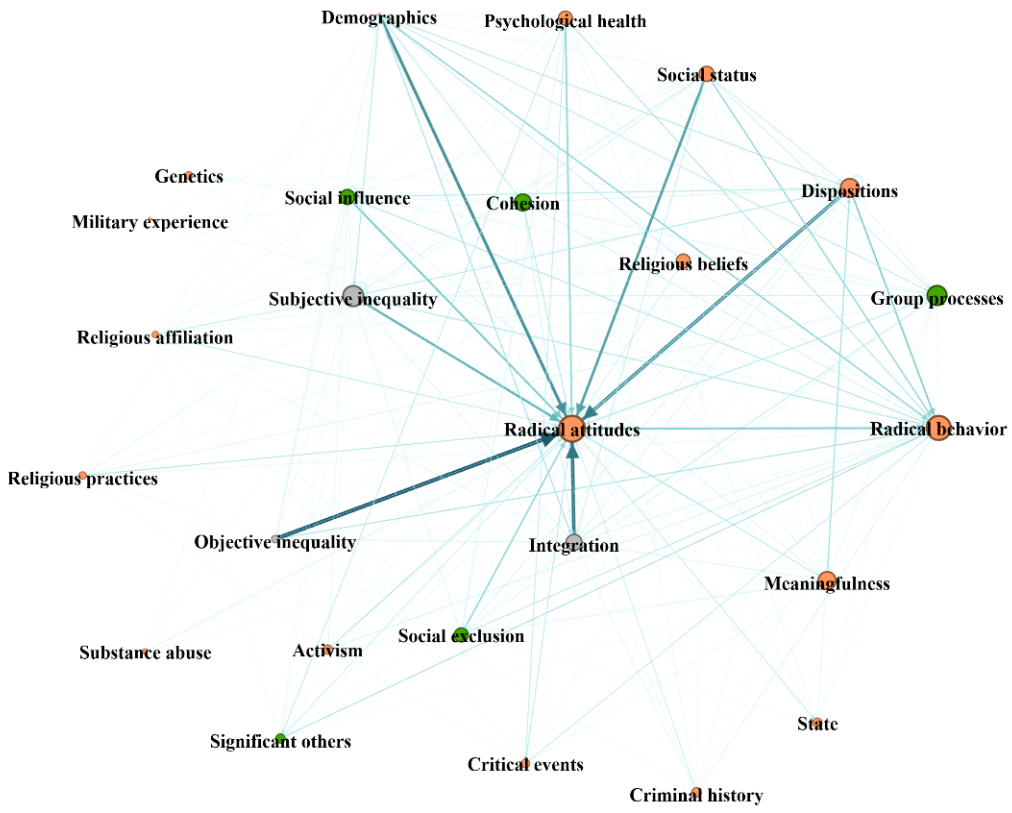

Figure 2.2. Network of hypotheses. Nodes represent constructs in hypotheses (node color: orange = micro-level construct, green = meso-level construct, gray = macro-level construct; width of edges is scaled to the occurrence frequency; node size is scaled to the respective node's in-degree centrality)

Table 2.4: *Network metrics based on constructs of self-reports, experimental, and trace data hypotheses*

Construct	Closeness centrality ($C(v)$)	In-degree centrality ($D^+(v)$)	Out-degree centrality ($D^-(v)$)	Weighted in-degree centrality	Weighted out-degree centrality	Betweenness centrality ($B(v)$)
Radical attitude	0.697	25	14	605	110	101.98
Radical behavior	0.418	22	1	275	3	0
Subjective inequality	0.697	18	14	98	116	24.28
Group processes	0.657	17	11	97	58	22.05
Dispositions	0.742	16	16	111	174	31.06
Meaningfulness	0.697	15	13	68	51	22.61
Cohesion	0.622	14	11	56	59	20.47
Integration	0.622	13	11	78	108	11.79
Social influence	0.657	12	13	59	96	15.88
Social status	0.852	12	19	24	146	44.95
Psychological health	0.742	11	15	23	107	62.38
Social exclusion	0.639	11	11	26	74	5.41
Religious beliefs	0.575	11	8	39	26	15.54
State	0.489	6	5	25	7	0.34
Activism	0.590	6	9	11	38	11.78
Significant others	0.548	6	6	22	32	0.16
Criminal history	0.469	6	4	10	24	0.22
Critical events	0.548	5	6	10	34	0.75
Objective inequality	0.605	4	10	10	119	6.29
Religious practices	0.500	4	5	10	30	0.34
Genetics	0.500	4	5	10	11	0.11
Religious affiliation	0.719	3	14	10	54	2.51
Substance abuse	0.434	2	2	10	9	0
Military experience	0.460	1	3	10	7	0.13
Demographics	0.800	0	18	10	195	0

In-degree centrality and out-degree centrality. As can be seen in Table 2.4, the construct considered most frequently was the presence of radical attitudes, which was considered as a central outcome of 25 antecedents and a determinant of 14 constructs. The most frequently considered determinants of radical attitudes, were objective inequality, subjective inequality, demographics, integration, social exclusion, social status, and dispositions. With regard to the overall number of expected incoming and outgoing effects, most relevant constructs were subjective inequality ($(D^+(v)) = 18$, $(D^+(v)) = 14$), group processes ($(D^+(v)) = 17$, $(D^+(v)) = 11$), dispositions ($(D^+(v)) = 16$, $(D^+(v)) = 16$), meaningfulness ($(D^+(v)) = 15$, $(D^+(v)) = 13$). These constructs were assumed to be effective for other constructs as well as hypothesized as important outcomes

Closeness centrality. With regard to the overall importance in the network (i.e., closeness centrality), social status ($C_v = .852$) and demographic characteristics ($C_v = .800$) were most central constructs, followed by dispositions ($C_v = .742$) and psychological health ($C_v = .742$): These constructs were directly related to a vast number of other constructs, indicating their role as central background variables to important outcomes.

Betweenness centrality. As aforementioned, constructs with a high betweenness centrality connect constructs in the field. These connections either represent a mediating structure (e.g., the target construct is hypothesized as a mediating mechanism between to other constructs), a *colliding* structure (i.e., the target con-

struct is expected to have an incoming effect by two other constructs), or the *confounder* structure (i.e., the target construct is supposed to act as a common cause of two other constructs). Whereas betweenness centrality represents the importance of a construct as a bridge builder, the weighted in-degree centrality and weighted out-degree centrality provides an impression about the assumed role of a certain construct. It should be noted, that a certain causal function of a construct is only conceivable with regard to a considered pair of constructs and that the following considerations represent a general evaluation of this function.

As Table 2.4 shows, radical attitude has the highest value of betweenness centrality ($B(v) = 101.98$); both its high degree of in-degree centrality ($B(v) = 25$) as well as its high level of out-degree centrality ($B(v) = 14$) indicates that it represents the core hypothesized mediator in this field as it received a substantial number of effects and in turn emitted a substantial number (mainly towards radical behavior). The weighted forms of both centrality measures emphasize that this seems to be the focal perspective in the literature. Similarly, the betweenness centrality of psychological health was ($B(v) = 62.38$) and the latter had an almost equal number of in-degree and out-degree centrality thus signalling its potential as a mediator of certain pairs of variables and a collider or confounder of others. As stated earlier, the causal role of a construct always depends on the pair of target constructs. In our case, studies most frequently hypothesized it as a common cause—and, thus, confounder—of the relationship between radical attitude and radical behavior. One example is found in the study by Ellis, Bixby, Miller, and Sideridis (2016) in which anxiety and depression predicted sympathies for violent protest and terrorism, as well as delinquency. Social status ($B(v) = 44.95$) functioned most frequently in a similar way as a confounder of the relationship between radical attitude and behavior (cf. Baier, Manzoni, Bergmann, 2016, investigating the effect of school achievement on right-wing attitudes and behavior). Likewise, dispositions ($B(v) = 31.06$) implied a confounder function in some studies (see Baier et al.'s, (2016) analysis of the effect of risk-seeking on left-wing attitudes and behaviour).

Analysis of approach-specific networks

Beyond the overall integration of studies in the field of radicalisation, our paper strives to investigate differences across the applied research approaches. Table 2.5 shows the differences between the research approaches with regard to the number of studies which had measured a respective construct as well as the in-degree centrality and out-degree centrality. Further, we characterized each construct according to whether the differences in both centrality measures reflect a predominant perspective of the construct as a rather independent variable (i.e., determinant) or

dependent variable (i.e., consequence) or both. We classified the role as independent versus dependent when the ratio between both exceeded 1.5.

As Table 2.5 shows that there are some differences between the approaches. *First*, and not surprisingly, all approaches focused on radical attitudes to a comparable degree. In contrast, the focus on the radical behavior itself was highest in trace data research, probably due to the focus of open source studies on coded behavioral data such as Profiles of Individual Radicalization in the United States (PIRUS). *Second*, and according to our expectations, dispositions were most frequently investigated in survey studies and experimental studies, probably due to the ease of measuring respective constructs with questionnaires. The same result and interpretation holds for meaningfulness, but interestingly not for other constructs that indicate some kind of reflection or subjective assessment (e.g., psychological health, religious beliefs) which were investigated comparably often in the three approaches. A substantial contrast is the number of survey studies focusing on integration (50 %) and subjective inequality (67 %).

With regard to the presumed causal role of the constructs, most constructs were regarded as determinants as well as consequences of other constructs. The percentage of these "mixed roles," however, varied across the approaches: Whereas 14 of the 25 constructs were hypothesized as independent as well as dependent, this was only the case for five constructs in experimental research and six constructs in trace data research. It should be noted that these results do not imply a state ambiguity or arbitrariness, but rather reflect a potential role of several constructs as more or less explicit mediating variables.

Table 2.5: *Number of studies and network statistics across research approaches*

	Survey research				Experimental research				Trace data research			
Construct	Nr. of studies (%)	In-degree centrality ($D^+(v)$)	Out-degree centrality ($D^-(v)$)	Prominent causal role	Nr. of studies (%)	In-degree centrality ($D^+(v)$)	Out-degree centrality ($D^-(v)$)	Prominent causal role	Nr. of studies (%)	In-degree centrality ($D^+(v)$)	Out-degree centrality ($D^-(v)$)	Prominent causal role
Radical attitude	27 (1.0)	22	10	Endogenous	11 (.79)	12	3	Endogenous	12 (.75)	15	7	Endogenous
Group processes	10 (.37)	17	10	Mixed	2 (.14)	1	3	Exogenous	2 (.13)	0	2	Exogenous
Subjective inequality	18 (.67)	16	14	Mixed	3 (.21)	3	0	Endogenous	5 (.31)	1	3	Exogenous
Military experience	0				0				5 (.31)	1	3	Exogenous
Radical behavior	8 (.30)	14	1	Endogenous	3 (.21)	5	0	Endogenous	8 (.50)	18	1	Endogenous
Meaningfulness	7 (.26)	14	11	Mixed	4 (.29)	3	5	Mixed	1 (.06)	1	0	Mixed
Dispositions	13 (.48)	13	15	Mixed	8 (.57)	7	6	Mixed	5 (.31)	1	3	Exogenous
Psychological health	6 (.22)	10	3	Endogenous	3 (.21)	1	5	Exogenous	4 (.25)	1	10	Exogenous
Religious beliefs	8 (.30)	10	7	Mixed	3 (.21)	2	2	Mixed	2 (.13)	0	1	Mixed
Social status	16 (.59)	10	18	Exogenous	0				7 (.44)	2	4	Exogenous
Social exclusion	11 (.41)	9	10	Mixed	2 (.14)	0	7	Exogenous	6 (.38)	2	2	Mixed
Integration	14 (.52)	13	11	Mixed	0				0			
Social influence	9 (.33)	10	12	Mixed	2 (.14)	0	3	Exogenous	6 (.38)	6	1	Endogenous
Cohesion	6 (.22)	8	11	Mixed	4 (.29)	7	1	Endogenous	9 (.56)	6	3	Endogenous
Activism	1 (.04)	5	0	Endogenous	3 (.21)	2	2	Mixed	3 (.19)	0	8	Exogenous
Criminal history	4 (.15)	5	3	Mixed	0				4 (.25)	1	1	Mixed
Significant others	3 (.11)	5	6	Mixed	0				5 (.31)	1	2	Exogenous
Genetics	1 (.04)	4	5	Exogenous								
Critical events	5 (.19)	3	4	Mixed	2 (.14)	0	4	Exogenous	5 (.31)	2	2	Mixed
Objective inequality	9 (.33)	3	9	Exogenous	0				3 (.19)	1	2	Exogenous
Religious affiliation	7 (.26)	3	13	Exogenous	0				1 (.06)	0	1	Exogenous
Religious practices	5 (.19)	3	5	Mixed	0				3 (.19)	1	2	Exogenous
Substance abuse	2 (.07)	1	1	Mixed	0				2 (.13)	1	1	Mixed
Demographics	15 (.56)	0	18	Exogenous	1 (.07)	0	4	Exogenous	6 (.38)	0	3	Exogenous
State	1 (.04)	0	1	Exogenous	8 (.57)	6	4	Mixed	1 (.06)	1	0	Mixed

Note. Number in parentheses are proportions of studies within the respective approach

2.2.5 Discussion

This systematic review intended to illustrate, summarize, and integrate the research focusing on determinants and outcomes of radicalisation constructs. To this end, we applied an innovative network approach to graphically represent radicalisation research and to statistically analyze the role, prevalence, and centrality of the constructs and hypotheses. Moreover, we investigated how the perspectives and focused constructs vary across research approach.

The most striking result was the quantity of constructs investigated over the years and even our aggregation procedures still resulted in 25 higher-order constructs located on the individual level, group level, or societal level. The results from the network analysis further revealed a substantially dense structure, indicating a lack of parsimony of the field (see also Wolfowicz et al., 2019).

One part of the explanation may lie in the historic development of the research on radicalisation, starting with the focus on surface-level demographic constructs (e.g. age or gender) and psychological health in order to identify radical individuals (cf. Stern, 2016). Further research efforts moved to disentangle the specificity problem (cf. Sageman, 2014) namely, why only some individuals out of the population confronted with the same determinants (e.g., discrimination experiences), in fact radicalise. This in turn may reflect a variety of further determinants considered in research to address particularly the lack of specificity for attitudinal extremism (cf. Slootman & Tillie, 2006). However, when partitioning the network according to the publication year of the study and comparing post- hoc the two resultant subnetworks (2014-2019 and 2005-2013) both density values did not yield substantive differences, which might speak against the historic explanation of the lack of parsimony.

A further explanation may be potential differences in the predictors of the different extremism ideologies (e.g., right-wing extremism vs. religious extremism), which might account for the heterogeneity of determinants and thereby network patterns. The apparent fragmentation additionally increased as other research approaches such as experimental research and trace data research developed and added contributions to the literature. As an example, scholars have traditionally assumed that "social influence" is a major determinant of radicalisation. While historically, social influence rather referred to the influence of peers or traditional media, technical developments of other media sources (e.g., the Internet and social media) were integrated in the overall concept of social influence (see Taylor et al., 2015), which represented the assumption that the development of radical attitudes is a direct consequence of contacts with extremist social media content. Apart from the increased broadness of the overall social influence concept, the review by Odag, Leiser, and Boehnke (2019) raised doubts on this assumption as the

literature lacks sufficient investigations that could explain the link between media effect and constructs of radicalisation.

While it is beyond the scope of this systematic review to recommend any particular framework, one basic approach to understand an individual's broader motivation-set would be to organize constructs in the multilevel framework on which our coding was based (cf. Schmid, 2013). Consequently, as a next principle, organizing constructs on a continuum ranging from distal or broad (demographic, personality, societal), over proximal or more radicalisation focused (e.g., group processes, cohesion, experiences) to radical attitudes and behavior, reflects the interplay of circumstances, beliefs, attitudes and behavior (cf. the reasoned-action approach, Fishbein & Ajzen, 2010). This is as well reflected in the general meaning framework by Kruglanski and colleagues (2014) in which the individual's quest for significance is a major motivational driver for violent extremism. Especially the need for restoration of a sense of purpose and meaning in interaction with societal processes, alongside group dynamics through which the individual comes to share violent ideology and narratives might lead to different degrees of radicalisation (ranging from passive support to self- sacrifice).

Evaluation of the results. Coercing study-specific constructs to higher-order constructs faces a trade-off between parsimony and precision. In particular, reducing the number of the myriads of "bloated specifics" (Cattell, 1978) into organized, and integrated higher-order constructs achieves parsimony of constructs, as it enables to identify generic principles inherent in radicalisation research, across extremism types. The approach presented is an economic representation of an etiological network, linking causes and effects and allows to clarify and represent domain knowledge inferred from hypotheses.

One example for a broad construct in our network are dispositions. Decomposing dispositions into their lower-level constructs revealed the prominence of constructs like authoritarianism or low self-control (impulsivity and risk-seeking). For instance, studies showed that authoritarian individuals tend to hold antidemocratic social attitudes, are rigidly attached to traditional values, uncritically accept authorities and are intolerant toward opposing views. Authoritarianism was frequently hypothesized to predict psychological uncertainty or willingness to engage in extreme means (Rieger, Frischlich, & Bente, 2017). The results of our network analysis can be integrated with prior research. In their meta- analysis, Wolfowicz et al. (2019), identified risk and protective factors for different outcomes of radicalization and presented a rank-order of these factors according to their effect sizes, in which authoritarianism had a relatively large effect (Wolfowicz et al., 2019). Similarly, when contextualizing the high closeness centrality of dispositions and thus importance in the network, self-control emerged as an important construct. The role of low self-control for radical behaviour was also found to have a relatively

large effect in the meta- analysis by Wolfowicz et al. (2019). These factors have also been investigated by Pauwels and Svensson (2017) who found an interaction between the degree of extremist beliefs and self-control in reducing the propensity for radical behaviour. Finally, constructs like integration, demographics, or peers and religion emerged as prominent foci of prior research. Our review found that the integration construct (with an out-degree centrality, ($D^+(v) = 11$) figured in the network as antecedent for radical attitudes, cohesion, as well as group processes (see Coid et al., 2016; Ellis et al., 2016; Simon, Reichert, & Grabow, 2013). Again, our findings can be contextualized by those found by Wolfowicz et al. (2019) and their critical discussion on the role of low integration as a risk factor for radicalization, for which they found modest effects for radical intentions and behaviour. Furthermore, higher-order constructs such as demographics (out-degree centrality, ($D^-(v) = 18$) were frequently hypothesized. Similarly, Wolfowicz et al. (2019) found these to be among the most commonly examined factors, albeit displaying small and sometimes non-significant effect sizes on radical attitudes and behaviors. In contrast, their analysis found that radical peers was important risk factors for radical attitudes and behaviours. But this also connects to the central point of the network that multiple constructs reaching from individual to social levels play into the connection of radical attitudes and radical behaviors which in turn have been most prevalent in the network. In this regard, Wolfowicz and colleagues (2019) argued there are both arguments for and against a risk effect of religious beliefs and practices in the radicalization process. They showed on the one hand small effects on the radical attitude whereas on the other hand the importance of the identification with the group was shown to be more important (Wolfowicz et al., 2019).

By forming higher-order dispositional constructs, we illustrate that adversarial personality traits (low self-control), traits implying an identity-weakness (low self-esteem), opportunities for engagement (salient injustice narratives that imply dissatisfaction with the "system" and blames on the outgroup and threats) and anxiety-related traits (uncertainty- aversion, need for structure) may prompt an engagement in radical groups or radical attitudes (see also McGregor, Hayes, & Prentice, 2015).

With regard to the comparison of the research approaches, our results demonstrated the dominance of survey research and a comparably lower number of trace data studies. However, the sole focus on Twitter in this context has been criticized by Parekh et al. (2018). Lesser known platforms (such as *4chan*) have yet to be sufficiently considered in terms of their relevance and reach for the radicalization process (Schmid & Forest, 2018). In view of the intensive linkage and interaction of social networks (cf. Johnson et al., 2019), a holistic view across platforms is lacking, as is an answer to the question of whether determinants and conducive

framework conditions that have been analyzed on one particular platform can be generalized to others. This is of relevance, especially since mainly verbal behavior can be observed on Twitter, while other platforms are more strongly characterized by visual elements (e.g., so-called "memes", i.e., quickly spreading images with verbal expressions) (Munn, 2019). Other platforms, such as the "*4chan*", are strongly characterized by anonymity, irony, and acronyms and cannot be quantified with classical text mining approaches. The latter illustrates new challenges in the evaluation and transferability of previous theoretical assumptions to these milieus. While questionnaire studies are often criticized for the risk of bias due to measurement errors and desirability trends, digital behavioral trace data analysis also faces measurement problems: While demographic characteristics can easily be extracted, the extraction of contextual data (e.g., number of retweets, number of friends) and user-generated content (e.g., text content, "likes" of other users' statements, self-reported individual differences) must be done with respect to the target construct, taking into account the context in which the behavioral trajectories were created when interpreting them (see Landers et al., 2016).

2.2.6 Implications

Whereas traditional behavioral sciences have emphasized the role of measurement models or theories that connect data with supposed theoretically important entities, this is seldom the case in social media research. Hence it is crucial that researchers formulate such models and explicate theoretical links (i.e., causally or logically) between measured data and referring constructs. One further route can be to seek multiple indicators for the same construct under investigation, as some indicators might be more closely related to each other than taken in isolation.

Finally, digital behavioral trace data analyses offer an approach to understand radicalization, which is caused by determinants that partly stem from the biographical course of development (e.g., experienced deprivation). While this is a clear causal focus, existing studies are based almost exclusively on cross-sectional approaches. With the newly emerging possibilities offered by digital behavioral trace data, the focus should be on the integration of traditional approaches and new technologies to map the process character. As an example, approaches such as online field experiments on the dissemination of emotional states in social networks, as already implemented by Kramer, Guillory, and Hancock, (2014), could provide new insights into the milieu and have heuristic significance and explanatory value.

The main strength of applying a network theoretical approach is that the network summarizes the more or less explicit causal hypotheses in the field and the resulting role of the constructs within the causal structure. As the network analysis indicated, some constructs were uniformly hypothesized as mediators (e.g.,

radical attitudes) whereas most constructs were most often expected to be causes as well as outcomes, implying their potential role as *confounders* (i.e., variables affecting two or more other target constructs) or *colliders* (i.e., variables which are outcomes of two or more target constructs). While the experimental research reviewed in this paper has the immense strength of enhancing causal interpretability due to the randomization of the hypothetical construct, survey research and studies relying on trace data are naturally much more plagued by biases resulting from the observational data. While this state of affairs has resulted in a resignation and problematic jargon, avoiding causal concepts and using rather imprecise "relationship" rhetoric (cf. Pearl & MacKenzie, 2018), our study provides a basis for improving statistical models in order to reduce causal biases (see also Antonakis et al., 2010) by the following means:

First, considering potential *confounders* of a targeted relationship provides a basis for controlling for relevant variables. The list of higher-order constructs and those constructs contained in the primary studies (see Table 2.2 and Table 2.5) provide a checklist of constructs which could be considered as potential confounders for a particular relationship (as practical examples, see the studies by Shrier, & Platt, 2008; or Vahratian, Siega-Riz, Savitz, & Zhang, 2005; or the theoretical basis in Vanderweele, 2019).

Second, *colliders* are less known to the field but represent an equally valid threat to causal inference (Elwert, 2013; Pearl, 2009; Rohrer, 2018), especially when it comes to the question of which variable a researcher should control and which should s/he not control. In this regard, controlling for colliders will introduce a bias in the estimate of the effect. As a simple rule and with reference to the graph in Figure 2, we recommend not to control for a variable that likely receives an arrow from the hypothetical exogenous variable as this will either represent a collider or a mediator (Pearl, Glymour, & Jewell, 2016; Rohrer, 2018). An alternative form of collider bias is endogenous selection bias, which emerges when a subgroup is drawn on the basis of a dependent variable (Elwert & Winship, 2014). For instance, focusing on a subsample of persons with a radical attitude may induce a bias on potential effects of a model with radical attitude as a mediator or outcome. Again, as a simple rule, we would recommend not to select a subsample based on a variable that is a dependent variable in the considered model. As before, the network analysis and the list of constructs may provide a basis for deciding which relevant variables the considered model may contain.

Limitations of the present study. While we stress the contributions of our study, we see three aspects that could cause some scepticism. First, we focused on the networks of proposed hypotheses instead of actual results, which probably would have resulted in a sparser network. However, this approach perfectly represents our main goal—to summarize the theoretical perspectives in the field.

Although estimating a network with empirical effect sizes is attractive, such an approach would have run into difficulties as the relationships between constructs substantially vary in the number of studies on which they are based (Cheung & Chan, 2005) resulting in ambiguity about the relevant sample size necessary for statistical tests. While this problem has been solved in confirmatory approaches to meta-analytical structural equation models (i.e., a multivariate extension of meta-analysis, see Viswesvaran & Ones, 1995), it is still an open problem in exploratory approaches (such as networks or causal search algorithms, see Glymour, 2004). At the same time, our results and their discussion may guide the selection and incorporation of central constructs into a future meta-analytical model.

Second, our comparison of the research approaches was qualitative and subjective. As the network structures were not nested, application of inferential statistics was not possible, resulting in perhaps spurious differences. Third, and related to this issue is the fact that research approaches did not only vary in the constructs but also in the populations that provided the data. Studies substantially differed with regard to whether they were based on a clear conceptualization of a population at all (vs. using ad-hoc samples) or whether they applied some systematic sampling process (vs. selecting a sub-group of individuals based on some characteristic). Analogous to our plea for using integrative theoretical frameworks more, we would recommend to more clearly conceptualize a referent population and to at least attempt to approach ideal forms of sampling in contrast to selecting individuals either ad-hoc or based on some characteristics. Our discussion on potential endogenous selection biases provided a theoretical basis based on a graph to consider the circumstances where this is appropriate versus problematic.

In the present systematic review, we applied an innovative network theoretical approach to synthesize the hypotheses in a research field. By these means, our analyses provide a snapshot of the collective thoughts on determinants and outcomes within the radicalization context of a whole community of researchers. As the contribution intended, we hope to have delivered some basis on what the community focuses on, its hypotheses and assumptions, as well as differences and similarities between the various approaches. The results give an impression about a field developed by integrating vastly different perspectives, constructs, and assumptions, and they clearly indicate that the time is rife for their integration.

2.3 Study 3: Conspiracy theories on Twitter: Emerging motifs and temporal dynamics during the COVID-19 pandemic

2.3.1 Abstract

The COVID-19 pandemic resulted in an upsurge in the spread of diverse conspiracy theories (*CTs*) with real-life impact. However, the dynamics of user engagement remain under-researched. In the present study, we leverage Twitter data across 11 months in 2020 from the timelines of 109 CT posters and a comparison group (*non-CT* group) of equal size. Within this approach, we used word embeddings to distinguish non-CT content from CT-related content as well as analysed which element of CT content emerged in the pandemic. Subsequently, we applied time series analyses on the aggregate and individual level to investigate whether there is a difference between CT posters and non-CT posters in non-CT tweets as well as the temporal dynamics of CT tweets. In this regard, we provide a description of the aggregate and individual series, conducted a STL decomposition in trends, seasons, and errors, as well as an autocorrelation analysis, and applied generalized additive mixed models to analyse nonlinear trends and their differences across users. The narrative motifs, characterised by word embeddings, address pandemic-specific motifs alongside broader motifs and can be related to several psychological needs (epistemic, existential, or social). Overall, the comparison of the CT group and non-CT group showed a substantially higher level of overall COVID-19-related tweets in the non-CT group and higher level of random fluctuations. Focussing on conspiracy tweets, we found a slight positive trend but, more importantly, an increase in users in 2020. Moreover, the aggregate series of CT content revealed two breaks in 2020 and a significant albeit weak positive trend since June. On the individual level, the series showed strong differences in temporal dynamics and a high degree of randomness and day-specific sensitivity. The results stress the importance of Twitter as a means of communication during the pandemic and illustrate that these beliefs travel very fast and are quickly endorsed.

2.3.2 Introduction

Humans are prone to search for causal explanations of events driven by the need to learn and adapt. Among the myriad of event types, the interpretation of social and political events is especially important as these may lead to exploitation or other threats for the individual or group. As an extreme form of interpreting events, conspiracy theories (CTs), that is, sets of beliefs about the existence of a hidden and powerful coalition of people or organisations with malevolent agendas, have become

a prominent research field (Douglas, Sutton, & Cichocka, 2017). This is partially due to the assumption that CTs may prompt a radicalisation process in which individuals develop beliefs immune to falsification (van Prooijen & Van Vugt, 2018). Additionally, as CTs often trigger the need to defend against perceived threats, they may elicit behaviour either detrimental to the individual (e.g., isolation) or the social environment (e.g., deviant behaviour).

Research shows that crisis situations and dramatic events (e.g., natural disasters) or terror events cause a high level of uncertainty and, thus, foster the emergence of conspiracy ideation (Lin, Margolin, & Wen 2017; Samory & Mitra, 2018). Such events are usually complex while their causes and remedies are unknown, as media coverage is most often contradictory and incomplete. In order to rationalise such phenomena and decrease personal uncertainty and lack of control, rumours and conspiratorial ideation might provide coping strategies for collective sensemaking. One instantiation of such a situation is the COVID-19 pandemic that started at the beginning of 2020. Not only has the uncertainty about the spread of the disease affected collectives and individuals but also the resultant public health interventions (e.g., the range of non-pharmaceutical measures being implemented by governments around the globe with a direct impact on social, economic lives and individual behaviours and wellbeing) (Hale et al., 2020). These public health interventions profoundly impacted sensemaking (e.g., distrust of authorities) and behavioural responses (e.g., decreased willingness for vaccination or increase in deviant behaviour) (Bertin, Nera, & Delouvée, 2020; Freeman, et al., 2020; Šrol, Mikušková, & Cavojova, 2020).

In recent years, social media platforms have not only become a viable means for individuals to inform themselves but also a platform to disseminate conspiracy ideation (van Mulukom et al., 2020). As such, these platforms are not only the relevant environment where CTs evolve but also a viable data source for research. This latter aspect of platforms leads us to the question how to gather text that is indicative of CTs beyond the simple focus on predetermined search terms, which make compiling an exhaustive list of synonyms and related concepts a challenging task. More importantly, in contrast to past research utilising keywords or hashtag-based identified samples of CT users, focusing on derogatory language or taking keywords alone as a sufficient indicator for a CT user, we adopt an iterative procedure that has a theoretical foundation in evolutionary psychology (i.e., not every remark about Bill Gates represents a CT). As a remedy, with the approach from distributional semantics we are able to delineate tweets which co-occur with each other and hence hold related semantic meaning and which serve to expand our initial scope. Most notably, we do not presume that every posting by a CT user is in fact a conspiracy tweet. This allows for differentiated individual human

behaviour as degrees of engagement with a concept that is derived from theory, alongside the variability of postings over time.

Likewise, the focus on social media platforms allows an in vitro view on the temporal dynamics of content creation and communication by means of an intensive longitudinal perspective. As any other behaviour, expressing CTs is a temporal process with probable nonlinear dynamics involving slow trend changes as well as abrupt chaotic spikes. Investigating these dynamics can provide insights on the psychological underpinnings and their rational and strategic versus affective and impulsive characteristics. Likewise, phenomena such as inertia and long-term trend changes can give insights into possible radicalisation processes in which people, when considering CTs, create a positive feedback loop, resulting in respective behaviour for a period of time or even in a durable fashion.

The present paper aims at exploiting these two merits of social media platforms. First, by using word embeddings, we investigate CTs utilising a data-based approach (i.e., *vector semantics*) that assigns meaning to a word by the distribution of words around it, combined with paradigmatic examples. With this natural language processing approach, we explore the context around COVID-19 discourse from a semantic perspective, in a time span when the conspiracy beliefs and narratives have emerged and spread. Second, to analyse the temporal dynamics, we apply a time series perspective (Box-Steffensmeier et al., 2014) and investigate, in an unobtrusive way, the temporal characteristics of user behaviour on social media as collective responses alongside individual ones. In this regard, we provide a description of the series of tweets both aggregated across individuals as well as individual series, conducted an STL decomposition in trends, seasons and errors, as well as an autocorrelation analysis (Hyndman & Athanasopoulos, 2018). We further attend to applied generalized additive mixed models to analyse nonlinear trends and their differences across users (Simpson, 2018). Moreover, we conducted a structural break analysis of the series of CT tweets in 2020 that could provide hints on the responsiveness to external events (Zeileis & Kleiber, 2005).

In particular, the paper adopts an exploratory perspective and aims to answer the following questions: **(1)** Which CT motifs emerged in the pandemic, and which terms are indicative of these motifs? **(2)** Do the CT group and non-CT group differ in the temporal dynamics of their posting behaviour of overall COVID-19-related content—that is, are there differences in the nonlinear trends, within-week rhythms of posting (i.e., *seasonality*), and degree of autocorrelation indicating inertia vs. randomness of tweets? **(3)** What are the temporal dynamics (e.g., trends, seasonality, autocorrelation) of CT tweets, and **(4)** are there inter-individual differences between users in these dynamics? The difference between processes on an aggregate versus individual level is a dominant issue in the social sciences. In this regard, scholars have repeatedly stressed not to trivially generalise results from

one level to the other, both, with regard to social systems in general (Klein & Kozlowski, 2000), and culture or social media, in particular (Kern et al., 2016).

To answer these questions, we present results of a social media analysis of $N = 218$ Twitter users (among them $n = 109$ CT posters) who have tweeted content with CT content over a period of approx. 11 months (from January until November 2020). This group is contrasted with $n = 109$ Twitter users who have not posted messages containing CTs (i.e., non-CT posters). Our study offers two major contributions, that is, firstly, providing a proof-of-concept to differentiate conspiracy language and to characterise it by linguistic similar indicators and psychological needs. Secondly, we assess how time series methods can enrich a theory-rooted view on dynamic user engagement. More specifically, we deliver an important contribution to the data science community, which rests on a substantive theoretical basis on which we build our automated NLP pipeline and time series analyses. The theoretical concepts—in our case these are concepts stemming from evolutionary psychology—aim to characterise forms of individual engagement with conspiracy content (regarding content types, as well as differentiating CT opinions from non-CT content). We deem such a theoretical foundation as fruitful for three reasons: First, distinguishing the variability of individuals in voicing conspiracy content and some of the underlying motivations against aggregated system dynamics allows us to analyse individual behaviour in a social context. Second, considering our temporal focus, we gain knowledge about trends (and their variability across users) that provide information about a possible radicalisation as well as temporal characteristics of the posting behaviour (i.e., whether it is systematic vs. impulsive) or structural breaks (as system responses to shocks that may hint at coping behaviour or persistent maladaptations) which can be taken into consideration when developing interventions (Hyndman & Athanasopoulos, 2018). Third, differentiating CT content from non-CT content with an approach from distributional semantics is scalable. In the next section, we provide the theoretical background on conspiracy theories that provides the basis of our word embedding approach.

Background

Conspiracy theories (CTs). A plethora of definitions regarding conspiracy theories exist that are at times contradictory and reflect a phenomenon that is hard to actionably delineate (Kou et al., 2017). Likewise, as understanding the minimal sufficient determinants for radicalisation processes, frameworks span pathological manifestations, cognitive or trait explanations, yet few approaches adopt an actionable definition (Klein, Clutton, & Dunn, 2019). We depart from a view on CTs that are defined as the belief that hidden coalitions of powerful individuals follow an agenda that intends or causes harm to society, the particular in-group of the

individual, or the individual specifically. While mistrust, criticisms, and specific claims are often erroneously regarded as CTs, van Prooijen and Van Vugt (2018) pointed out five criteria that define a CT which are adopted for this study. The first criterion is the perception of a *pattern* that leads individuals to connect events or specific observations to an integrated whole. Second, individuals assume an underlying *agency*, that is, they attribute intentionality of actions. This propensity results from the overall tendency to form social knowledge that strives to understand and predict human decisions and their behaviour. Third, people assume the joint acting of *coalitions*—in the vast majority of a more powerful group compared to one's own group. Fourth, the person thinks the plans of this group present a *threat* to the person or in-group, and fifth, either the group or its plans are *secret*, which makes it difficult to find clear evidence for the convictions and falsify them.

While research and especially the public discussion tends to view CTs as irrational, an expression of a pathological mind (Oliver, & Wood, 2014), or an extremist political attitude, van Prooijen and Van Vugt (2018) emphasise the evolutionary roots of CTs as a functional adaptation to persisting actual threats by hidden coalitions or at least side-products of specific functional adaptations, such as the tendency for pattern recognition or harm detection sensitivity. They note, however, that while being functional for the vast history of humans, this "hyperactive agency-detection system" (p. 773) has lost its usefulness in modern society, and CTs are now the result of this innate sensitivity being confronted by apparent cues, ubiquitous in the internet and social media era.

Adopting a more psychological perspective, Douglas et al. (2017) claim that CTs serve the fulfilment of three basic needs–an *epistemic need* to understand the world and the causes and consequences of relevant events, an *existential need* to avoid harm, achieve security, control the environment, and a *social need* to preserve a positive social identity. Especially the latter helps to understand that CTs often evolve, caused by the perception of intergroup conflicts, discrimination, or relative deprivation (e.g., Crocker et al., 1999; Stempel, Hargrove, & Stempel, 2007). Douglas et al. (2017) stress that although CTs aim to fulfil these needs, they fail to do so. Specifically, epistemic needs are unfulfilled as the individual creates CTs immune to falsification, unrealistically complex, irrational and unfounded. Likewise, the need to gain control and reduce uncertainty are unmet as the individual increases his/her perception of being the victim of powerful others. Empirically this has been shown to lead to reduced activities that actually would increase control (e.g., political engagement, see Jolley & Douglas, 2014). Likewise, CTs result in ongoing resentment and distrust against other groups or institutions that, in the long term, excel immediate feelings of superiority of the in-group. Empirically, research on CTs has evolved in a variety of fields, such as psychology, political science, sociology, medicine, or anthropology. Topics have been likewise diverse,

ranging from overall theoretical discussions (Clarke, 2002; Douglas et al., 2019), anti-science CTs, often discussed with the example of anti-climate change CTs (Douglas & Sutton, 2015; Hornsey, & Fielding, 2017; Lewandowsky, Oberauer, & Gignac, 2013) or anti-vaccination CTs (Guidry et al., 2015; Hornsey et al., 2018; Jolley & Douglas, 2017), and the role of demographic predictors (Goertzel, 1994), political predictors, such as political orientation (van Prooijen et al., 2015), or psychological predictors (Barron et al., 2014; Douglas & Sutton, 2011; Swami et al., 2011; Swami, et al., 2014).

An additional reason why investigating conspiracy beliefs is crucial in the COVID-19 context is related to the link between these beliefs and the rejection of scientific knowledge. Specifically, conspiracist ideation has been linked to greater opposition to scientific advancements such as vaccinations and climate science (Lewandowsky, Oberauer, & Gignac, 2013). Moreover, conspiracy content has been found on many online platforms (Kata, 2010). These types of content allude to for-profit collusion between vaccination promoters and pharmaceutical companies or cover-ups hiding the vaccine's side effects while it promotes "rebel doctors" who break away from medical establishments, refusing to support scientifically supported policies. Moreover, they relate conspiracy theories to the COVID-19 situation as they can be easily spread over social media, such as Twitter, contributing to the dangerous impact of these media on vaccination hesitancy (Chadwick et al., 2021). Given that time series approaches are scarce in the literature on conspiracy theories, we introduce central concepts of times series analyses in the following section.

Understanding the dynamics of CTs. Behaviour on social media is of scientific and practical interest because of the low barriers to post content and, thus, the high chances to be able to analyse impulsive actions due to emotional processes or reactions to stimuli (e.g., news events). Such social media communication may affect public risk perceptions during other crisis events, like the Zika virus in the United States in 2016 (Chan et al., 2018). Research found that CTs identified in social media posts differed from rumors in their temporal pattern, that is, CTs peaked multiple times in a period whereas rumours showed single peaks and recession patterns (Starbird, 2017). Further, the long-term elaboration and reinvention of CTs was shown in work by Nied et al. (2017) indicating that respective groups on Twitter comprise individuals with diverse ideologies and beliefs. This sets the stage for fruitfully investigating the particularities vs. generality of online posting behaviour across time.

While fields such as economy or ecology have a tradition in investigating dynamic processes, only recently the behavioural sciences adopted such approaches. Among these, time series analysis has been applied considerably seldom (Jebb et al., 2015). This is disadvantageous, as beyond their methodological capabilities,

time series concepts have theoretical benefits due to the possibilities for conceptualizing temporal dynamics. This is the case for the occurrence of *linear* or *nonlinear trends*, *autocorrelation* (Hamaker et al., 2018), *systematic fluctuations* (periodicity and seasonality, see Almagor & Ehrlich, 1990), and *structural breaks* in mean level, trend, or variance (Caporale & Grier, 2005) that show the behaviour of the system in response to sudden and emerging external or internal events. These concepts all share the fundamental purpose that we learn something about the underlying dynamics of psychological entities (e.g., beliefs and attitudes), emotional processes and their rhythms, regulations (or their failure), and long-term (mis)adaptations versus learning in the form of ongoing disequilibria.

Aggregate versus individual dynamics. With regard to the analyses of the temporal characteristics of the posting behaviour, our paper considers these characteristics on the aggregate level (i.e., the sum of postings of the overall groups) as well as on the individual level that focuses on individual users and their differences. By doing so, our approach adopts a multilevel perspective that is ubiquitous in the social sciences. Research involving hierarchical systems conceives individual entities (e.g., individuals) nested within higher-order entities (e.g., work teams, organisations, or other collectives). Inherent in this perspective is the emphasis that characteristics of the various levels are ontologically different, with the most prominent concept being "emergent systems" stressing that higher order entities may have a unique ontological status that cannot be deduced by its components. In the case of posting behaviour, an aggregate perspective, inspecting a part of the collective may be fruitful as an instantiation of collective sensemaking. Beyond these ontological issues, scientific approaches across disciplines have always been subject to the difference between *nomothetic* versus *idiosyncratic* perspectives and potential generalisations of scientific results versus particularities. In the most extreme example, single case designs have emerged but remained limited in their popularity (Edgington, 1987). Other scholars, in contrast, have proposed that both perspectives should not be viewed as contradictions; rather, studies investigating both perspectives should be conducted. In our paper we follow the latter perspective and analyse the aggregate series in addition to the individual series and their differences.

Research questions

As aforementioned, our study intends to apply word embeddings to identify terms signifying CTs and their semantic relationships and to analyse how CT tweets unfold over the first year of COVID-19 in 2020. By these means, we learn the temporal characteristics of tweeting CTs, their trends, dynamic profile, extent of external sensitivity, and systematic versus impulsive (or random) parts. To this

end, we compare different groups of people and series of different content (CTs vs. overall COVID-19-related content). We emphasize that the comparison does not aim to explain differences between the groups beyond the characteristics of their posting behaviour.

In particular, we investigate the dynamics of CTs from two perspectives. Namely, we differentiate the aggregate level and the individual level and potential differences in their ontological status, processes, and temporal dynamics. Table 2.6 summarises the research questions.

Table 2.6: *Levels of research questions*

	CT group	Non-CT group
Aggregate Level (averaged tweets)	**RQ1**: Identify semantically similar expressions of CTs	
	RQ2: Comparison of the posting behaviour on overall coronavirus-related (i.e., non-CT) content	
	RQ3: Identify temporal characteristics of CT tweets	
Individual Level	**RQ4**: Individual differences in temporal characteristics	

Note. CT = conspiracy theory; RQ = research question

On the aggregate level, we focussed on the temporal characteristics of the overall posting behaviour and investigated the total proportion of tweets posted across each day between January and November of 2020. The *CT group* consists of individuals who posted conspiracy-related content. To identify differences versus commonalities with Twitter users who do not post CTs and their posting behaviour, we identified a *non-CT group*. The first research question (**RQ1**) relates to characterizing CT tweets from the CT group not only in terms of what information they post but also how users formulate their posts of coronavirus-related content. For this purpose, we use word embeddings to identify semantically similar term vectors underlying the concepts.

The second research question (**RQ2**) focussed on the comparison of the two groups. To have a common ground, we directed our attention towards tweets with COVID-19-related content, containing information on infection rates, social distancing measures, recommendations to wear a mask, etc. We explicitly excluded CT tweets (see the Methods section on word embeddings).

With the third research question (**RQ3**), we analysed the temporal characteristics of the CT-related tweets. To this end, we focussed on two relevant variables: The overall number of active users posting CTs each day and the mean proportion of CT tweets of all tweets across the days. Of particular interest was the analysis of a potential trend in the number of users and proportion of tweets. Further, the autocorrelations examined for the proportion of CT tweets provide information

on potential inertia and, thus, the degree of recovery of the aggregate to mean levels. Finally, the exploration of structural breaks can give rise to interpretations of external events causing these changes.

Finally, the fourth research question (**RQ4**) focussed on the individual level of analysis. Here, we aim at analysing the series of tweets for each user separately. Consequently, estimation of their trends across 2020 allows us to learn about differences between individuals—most notably differences in the functional form of trends (i.e., linear vs. nonlinear) and directions (i.e., upward trends vs. downward trends). In addition, differences in the autocorrelation coefficients indicate differences in inertia vs. fast recovery (e.g., after emotionally triggering news). In summary, the inter-individual approach provides an empirical basis for future research targeting predictors or explanatory factors of these differences.

2.3.3 Method

Data collection and preprocessing

Information retrieval. In order to identify CT users, we manually searched for matching keywords in the advanced search of the web version of Twitter and then retrieved the matching tweets and tweet handles. This offers the opportunity to model individual trajectories over time, capture the occurrence of the target keywords at different points in the year, and not be compromised by algorithmic sampling biases that favour the most recent, trending postings. We collected a list of thematic keywords, potentially indicative for conspiracy theories, based on research articles (Shahsavari et al., 2020; Jiang et al., 2020), and third-party sources (Brennen et al., 2020). The selection was based primarily on the occurrence of thematic topics that were flagged as misinformation by the "EUvsDISNFO"-database in 2020 (EUvsDISINFO, 2020). The search queries comprised the bespoken keywords, as well as a reference to the broader COVID-19 context (e.g., 'pandemic'). Using each of these keywords, we queried the Twitter user interface for a seven-month period and retrieved users with matching tweets for each month. By sampling per month 10 random users that matched the queries resulting in 420 sampled and annotated tweets, we are able to capture users over the year of 2020, in contrast to sampling with the Streaming API or Search API, which are primarily used for forward searches. Thereby, we could address the potential bias of only sampling users that have been highly active in a recent short time interval of the particular sampling time. Subsequently, we employed two independent coders who annotated user's tweets (as containing potential CT content). For the annotation, only original tweet content was considered, to avoid confounding by third party opinions (e.g., retweets). We retained, however, links to external sources or sharing of picture material. The underlying coding scheme was based on a catalogue of five

criteria (comprising: agency, secrecy, coalition, threat, pattern) of which at least three criteria needed to be met in a tweet to be indicative for expressing a CT, along biographical information.

Cohen's kappa was assessed on the binary decision of including or excluding an account. In particular, coders were trained on Twitter-specific platform affordances (posting types and conventions, non-standard abbreviations and symbols) and conspiracy-specific characteristics (e.g., examples for deceptive intentions or coalitions). A sample of 10 random users was coded as a pre-test. Subsequently, we turned to the whole dataset of 420 matching tweets and a first round of coding was conducted. We achieved a Cohen's Kappa of $k = .60$, which indicates an agreement of 79.76% (Landis & Koch, 1977). After the initial coding, in a second round, disagreement between raters on ironic or allusive tweets (e.g., "swamp" referring to the deep state coalition) was resolved by consulting the respective tweet history and profile. This process of resolving disagreement led to the inclusion of $N = 203$ Twitter accounts. We acknowledge that potential conspiracy accounts might have been excluded due to factors such as extremely short tweets, incomprehensible abbreviations, lack of sentence structure and the use of hashtags only. To establish a non-CT group, we sought to identify Twitter users exhibiting a tweet behaviour focusing on coronavirus-related content but non-CT-related. For this purpose, we conducted a keyword search with Twitter's Search API (e.g., corona OR covid OR pandemic) to identify common users.

Querying Twitter. In the next step, we used the Twitter REST API to harvest the available timelines (e.g., all tweets, retweets, or replies) of the selected accounts of both groups. This process referred to the public user timelines (i.e., the tweet history of the user) of each identified account and was conducted on November 8, 2020. For each of the timelines we were able to retrieve a maximum of 3,200 tweets, resulting in individual time series of unequal lengths. The unequal lengths imposed no limitations for the aggregate level analyses, as we calculated the average percentage of tweets among all tweets.

User preprocessing. After retrieving the timelines, we applied the following criteria for inclusion of potential CT accounts as well as for the non-CT accounts. First, to remove dormant or entirely inactive accounts, we included only accounts that had posted status updates across a three-month period. Second, the accounts had to be owned by English-speaking users. In particular, we excluded accounts using English words at a rate of less than 80%, computed at the tweet level. Third, we used the R-package *tweetbotornot* (Kearney, 2020) to exclude, with a probability of 80% and higher, accounts that were created by a bot. The functionality of the package takes into account features on the user level (e.g., profile information, account creation date) and tweet level (rate of status updates, or word complexity) (Kearney, 2020). In order to assess the extent of bias when classifying bots (false

negatives or false positives), we manually annotated a random sample of initial CT accounts (40 out of 132) as well as non-CT accounts (40 out of 520). We based the classification on the user profile, the degree of human creativity and specificity of content, follower and friend count, extent of duplicates, and degree of automation (see Chu et al., 2012). We calculated the intercoder-reliability (Rauchfleisch & Kaiser, 2020) for the CT accounts ($k = 0.6$) and non-CT accounts ($k = 0.5$). More false positives were found for the latter type, that is, human Twitter users were classified as a bot by TweetBotorNot. Eventually, after the filtering, this resulted in 109 accounts for the CT group and 333 accounts for the non-CT group, for the latter we drew a random sub-sample of 109 accounts. Pertaining to the face-validity of the non-CT group, we drew a random sample of non-CT accounts ($n = 40$) from 109 accounts and annotated a random tweet, each of them by two coders, following the five-criteria scheme. Similarly, as with the CT group we calculated agreement for the binary categorisation of the tweet as conspiracy or not. This resulted in an agreement rate of 97.5 percent, with one tweet being flagged as conspiratorial.

Tweet preprocessing. We applied three steps of text pre-processing. First, tweets were converted to lowercase type, and then all links, HTML tags, ampersands, mentions, hashtag symbols, stopwords and non-ASCII characters were converted or removed. Second, we converted expressions of emphasis from tweets (e.g., elongations such as 'heyyyy') to normal text. Third, we then tokenised the text using sets of up to two terms (i.e., two-grams). This procedure resulted in $N = 142,559$ tweets with $2,963,424$ tokens for the CT posters as well as $N = 95,394$ tweets with $2,558,504$ tokens for the non-CT posters.

Distributional semantic model

Text documents from social media pose a substantive challenge when inferring latent information such as that which is needed for discriminating between conspiracy and non-conspiracy content in RQ1. *Word embeddings*, which range under the family of distributional semantic models, offer a state-of-the art approach for representing words in vector space to understand, at a word level, semantic meaning, but also to extract document similarity from them (here: tweets). More specifically, terms are represented with real numbers as a vector in continuous n-dimensional vector space, and the distance between the vectors denotes semantic similarity of the underlying construct. Word vectors exploit a spatial analogy, so that similar words have similar spatial relationships (Chollet & Allaire, 2017).

Global vectors for word representation. We use word embeddings for RQ1 to characterise emerging motifs with the respective related terms. More specifically, *Global Vectors for Word Representation* (GloVe) algorithm was used to discover latent vector representations in unannotated textual data (Pennington,

Socher, & Manning, 2014). With this method, word embeddings can be inferred from word co-occurrence matrices. GloVe is an unsupervised method for learning word representations based on log-bilinear regression that captures both global and local statistics of the term co-occurrence information (Pennington, Socher, & Manning, 2014). This method of incorporating global statistics of word co-occurrences performs well even with small corpora (Pennington, Socher, & Manning, 2014). Pennington and colleagues (Pennington, Socher, & Manning, 2014) showed, in experiments regarding the word analogy task, that a 100-dimensional GloVe model outperforms HPCA vectors or vLBL. The authors of GloVe, argue for their approach by setting out that both count-based matrix factorisation methods and predictive neural network methods suffer several disadvantages (Pennington, Socher, & Manning, 2014).

Methods regarding global matrix factorisation consider statistical information but they perform less optimally on internal evaluation tasks like the word analogy task that try to find semantically similar words (Pennington, Socher, & Manning, 2014). Neural network methods like the skip-gram architecture (which try to predict the context word from a target word) perform better on the analogy task, but conversely show shortcomings on the global statistics of the corpus (Pennington, Socher, & Manning, 2014). GloVe combines the best of both worlds as it allows us to consider the global context by the ratio of conditional probabilities to model the vector representations, as well as linear structures of vector spaces as likewise captured by *word2vec* (Pennington, Socher, & Manning, 2014).

Specifically, the GloVe algorithm uses co-occurrence probability ratios in the training phase of the word embeddings and accounts for rare co-occurrence word pairs (Pennington, Socher, & Manning, 2014). The weighted least-squares objective function J indirectly factorises the term–co-occurrence matrix (X), where w_i,w_j are word vectors and V denotes vocabulary size (Pennington, Socher, & Manning, 2014) (see equation 1). The objective of the GloVe training is to minimise the difference between the dot product of word and context word vectors ($W_i^T w_j$) and the logarithm of the word co-occurrence probability of the word embeddings ($log(X_{ij})$). In order to avoid that rare co-occurrences are overweighted, a cost/weighting function $f(X_{ij})$, is applied to the model (see equation 1). This function reduces the weight of co-occurrences appearing fewer times then the cutoff value x_{max}.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(W_i^T w_j + b_i + b_j - logX_{ij})^2 \quad (1)$$

Our workflow for representing the respective tweet corpora as word embeddings (for each the CT group and the non-CT group) is shown in Fig. 2.3.

Firstly, after pre-processing the raw tweets **(step 1)**, we build a vocabulary of tokens (bigrams) from the corpus. These can then be represented as a global

term-co-occurrence matrix (TCM) **(step 2)**— which takes into account the ratio of co-occurrence probabilities (by a pairwise context window). With GloVe, the cost function is directly optimised which allows for a more global context, as the dot product of two word vectors equals the number of times the terms co-occur. For the training, the GloVe algorithm uses the stochastic gradient descent algorithm to factorise the log of the TCM (Selivanov & Wang, 2018). The resultant GloVe weight matrix consists of 2 vector types: main vectors and context vectors which are summed up (Selivanov & Wang, 2018). Eventually, each token is represented as one real-valued vector of D-dimensions **(step 3)**. The embedding dimensions in turn specify the complexity of the model and space into which we try to "embed" the tokens. The semantic similarity between two vectors can then be queried by similarity measures like cosine similarity **(step 4)**.

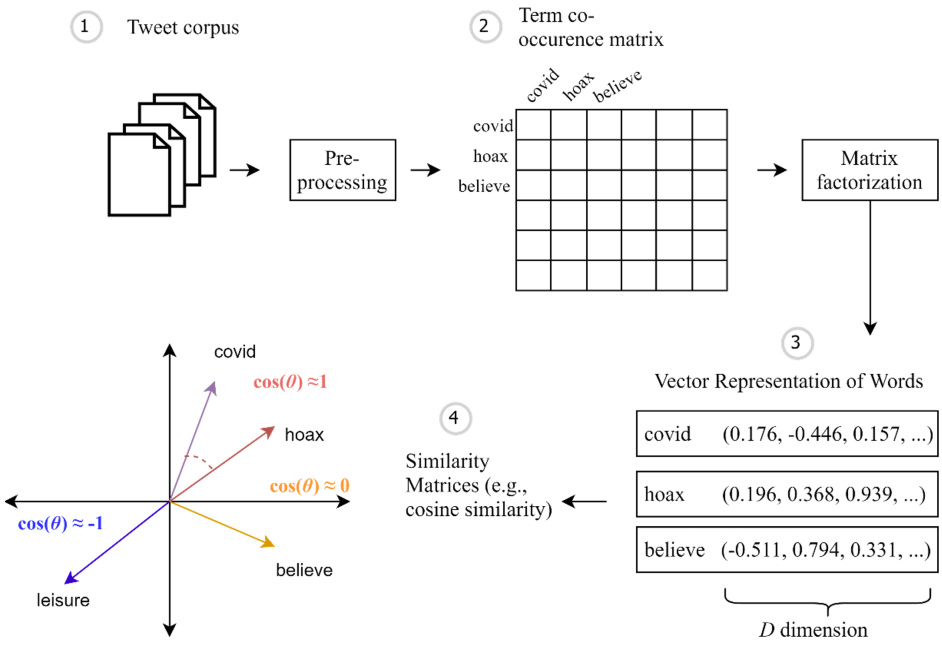

Figure 2.3. Framework for constructing *Global Vectors for Word Representation* (GloVe) models and measuring similarity

Within this framework we vectorise text by (i) constructing symmetric, window-based TCMs from the pre-processed tweet "documents", and (ii) fitting GloVe models to the TCM for CT posters and non-CT posters. In order to assess the GloVe model performance, we perform *intrinsic evaluation* (i.e., a word analogy task). This is a direct evaluation of the GloVe model performance—based on the hit-miss ratio of predicting a set of query terms and semantically related target words (Elekes et al., 2018). Here we used the BATS (Gladkova, Drozd, & Matsuoka, 2016) and Google Analogy data set (Mikolov et al., 2013). In our experimental set up we tested different settings for the hyperparameters. We tested the accuracy for different GloVe *dimensions* $(50, 100, 150, 200, 250)$ and *window sizes* $(3-12)$. As for the window size this denotes the context of a word that extends before and

after a target term. Words that appear further away in the context from a word are given less weight than words closer to the respective word (Pennington, Socher, & Manning, 2014). Thereafter, we adopted GloVe models with 100 *dimensions* and a *context window* of 8, which are simplest and showed best performance. We fixed the number of *iterations* to 20 and a *convergence threshold* of 0.001, so that training stopped if the maximum number of iterations is reached or the change in loss is lower than the convergence threshold. Furthermore, the number of co-occurrences within the weighting function $f(X_{ij})$ denoted by x_{max} was set to 10.

Concept mover's distance. In a next step tweets are discriminated against as conspiracy and non-conspiracy, ignoring user-related variables. This categorisation is guided by the semantic similarity of the word vectors with a custom CT lexicon, as well as a custom coronavirus lexicon. The general COVID-19 dictionary is based on the Yale Medicine vocabulary, hence it comprises overall categories that relate to: linguistic variation of coronaviruses (e.g., "SARS"), medical-response (e.g., "remdesivir"), prevention (e.g., "stay_home"), spread of the disease (e.g., "outbreak"), and transmission (e.g., "symptomatic") (Katella, 2020). The CT dictionary comprised a set of seed terms as identified in the original article by van Prooijen and Van Vugt (2018), as well as by the EUvsDISNFO-database, as well as Part-of-Speech-Tagging of tweets (e.g., noun phrases for coalitions) for each category: agency (e.g., "plandemic"), threat (e.g., "eugenics"), coalition (e.g., "capitalist"), pattern (e.g., "great_awakening"), and secrecy (e.g., "mole"). Hence, the selection of seed terms connects to the initial five-category system of manual annotation, by building on these premises. The initial vocabularies for both dictionaries were enhanced by retrieving the 20 semantically most similar terms by cosine similarity of the GloVe vectors, associated with these seed terms, which were then selected based on relevance by human judgement.

Calculating the semantic similarity is achieved with the *concept mover's distance* (CMD) algorithm (Stoltz & Taylor, 2019). As a development from the original word mover's distance function (Kusner et al., 2015), the CMD algorithm captures the semantic similarity between documents (i.e., the word embeddings of documents and averaged terms generated from the dictionary in vector space) even at instances when they do not share exact words (Stoltz & Taylor, 2019). One example for the general principle can be seen in Fig. 2.4 when the relative cost of moving components in tweets (T_1 or T_2) toward a target concept (T_{pseudo}) is shown. As overall the cost for the first tweet is relatively lower, T_1 can be taken to engage more with the concept.

The calculation with the CMD is based on the "Relaxed Word Mover's Distance" (Kusner et al., 2015) which tries to find the minimum cost to transform the embedded words of a specific document to words from other documents in the

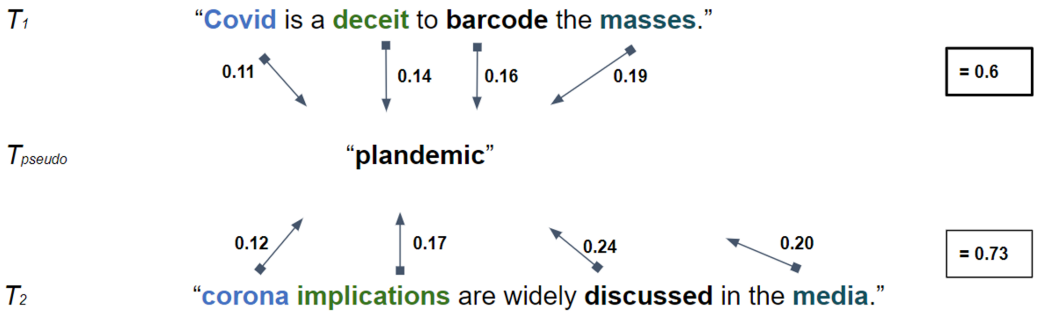

Figure 2.4. Illustration of the Concept Mover's Distance principle (with T_1 and T_2 representing fictitious tweets and T_{pseudo} a "pseudo"- document comprising only one term) (see also Kusner et al., 2015)

embedding space (Stoltz & Taylor, 2019). In this vein, the CMD algorithm allows us to determine the similarity of the words in a tweet document and "pseudo"-document which relates to theoretical concepts of interest and must not necessarily be of equal length. It returns a list of standardised distances which are inverted for the convenience of interpretation. With CMD, the cost (i.e., cosine similarity) of moving concepts in vector space is assigned as incoming and outgoing weights to a document (see Stoltz & Taylor, 2019). Hence, semantically similar concepts that appear closer in vector space (i.e., they require less "effort" to be moved) can be classified respectively based on these weights as CT (relating to the CT dictionary) and coronavirus-specific (based on the coronavirus dictionary). One of the major advantages of this technique is using the relational word meaning for assigning common groups, instead of relying on discrete, entirely a priori determined CT categories.

More specifically, for each of the five categories in the conspiracy dictionary, a *centroid* (the averaged concepts) in the word embedding is calculated. Next the distance from each "tweet document" of the CT posters to the centroid is calculated with the CMD (with large CMD values indicating large concept engagement). For interpreting the results of this procedure, we adopted a threshold of ≥ 0.8 for the closeness to conspiracy categories. To ensure, for the CT posters a genuine focus on coronavirus concepts, without containing potential CT content, juxtaposed pairs for these two semantic directions are constructed (Stoltz & Taylor, 2019). For this, the semantic direction was combined with the concept mover's distance. That is, a list of antonym pairs was generated based on the general COVID-19 dictionary and then a subsample of equal size from the CT dictionary was drawn that pose theoretical antonyms to the general COVID-19 terms. In this context general COVID-19 vectors are treated as "additions", whereas conspiracy vectors as "subtracts". This involves treating CT concepts (e.g., "scamdemic", "biowarfare" or "ankle_monitors") as antonyms to general coronavirus concepts (e.g., "vulnerable_people", "vaccine" or "stay_home"). In a next step, the difference between

the respective vectors was obtained and averaged. Eventually, we obtained one side of the continuum, which can then be quantified regarding its distance to the tweet documents with the concept mover's distance algorithm, with a threshold of ≥ 0.8.

We validated the CMD-based classification for the CT group by randomly sampling 100 tweets (50 each for the CT-classified tweets as well as the non-CT-classified tweets). We further annotated them and compared human and CMD-based classification. This resulted in a classification by CMD with a precision of 0.8, by which 10 tweets were found, mostly due to their brief length, to be non-CT content by human annotators (i.e., false positives) and 40 were true positives. Further, this resulted in a *recall* of 0.89 (with 5 tweets being classified as false negatives). This eventually results in a fair *F-measure* of 0.84.

Time Series Analyses

We used a variety of approaches to investigate the time course and temporal dynamics of tweets both on an aggregate level (RQ2 and RQ3) and on the individual level (RQ4). The percentage of CT-related content was calculated on a daily basis. Hence, if on a certain day a user has posted 20 tweets and half of them were CT-related, the relevant number is .50 for the respective day. This allows us to use the person as the reference system which enables estimating meaningful within-person trajectories (see RQ4). Hence, an increase of the involvement with CT across time would be visible in an increase in the proportion irrespective of how large the overall number of tweets was. With regards to the minimum number of tweets, this was zero due to the days on which there was no posting behaviour. The goal was to measure the CT-content per user per day—not the degree of CT-conviction (behind the posts) for which setting the value to zero would have been inappropriate.

First, we plotted the aggregate time series to facilitate illustration and visual exploration, thereby obtaining a first indication of linear or nonlinear trends and the occurrence of seasonal variations. Next, we calculated the autocorrelation function and partial autocorrelation function due to two substantive reasons–that is, to judge the level of inertia (positive autocorrelation) versus bouncing (negative autocorrelation) of the behaviour as well as to evaluate whether the estimated time series models require enlargement by an ARIMA (autoregressive integrated moving average) component (Box-Steffensmeier et al., 2014). Further, formal evaluation of seasonality was based on a decomposition of each series into trend, season, and error (Hyndman & Athanasopoulos, 2018).

Second, we formally tested for linear and nonlinear trends, season effects, and potential group differences by means of *generalized additive models* GAMs, (Dominici et al., 2002, Simpson, 2018; Wood, 2017). GAMs extend generalized linear mod-

els by estimating nonlinear relationships between variables by smooth terms that can fit any degree of wiggliness. A penalty parameter prevents overfitting, with the result that the estimated curve is not wigglier than necessary. In a time series model, smooth terms can be estimated for the trend as well as nonlinear seasonality effects. A comparison with a linear trend model by means of an analysis of variance allows us to statistically differentiate both models. We estimated the GAMs with cubic regression splines and, as the dependent variables of interest were proportions (i.e., of CT-related content in the total number of tweets), we used a beta error distribution with a log link. In the case of RQ2 that involved a comparison between both groups (i.e., the non-CT group and CT group), we estimated differences in nonlinear trends and season effects by means of factor-smooth interactions.

Third, for RQ3, we conducted a structural break analysis (Zeileis & Kleiber, 2005) to explore potential breaks in the level or trend of a respective series. This was done to re-evaluate the causes of a nonlinear overall trend tested in the prior GAM but also to gain insights into critical events that prompted a rise in the proportion of CT-related content.

Fourth, we used a combination of time series approaches and *generalized additive mixed models* (GAMM) to analyse RQ4. Time series approaches consisted of the estimation of the degree of autocorrelation for each individual series in the CT posters, and the GAMM aimed at testing the overall nonlinear trend and seasonality as fixed effects and inter-individual differences in levels, slopes, and nonlinear functional forms by means of random effects. The differences in the level, slope, and nonlinear trends were tested following recommendations from the overall literature on multi-level models (Aguinis, Gottfredson, & Culpepper, 2013) and growth curve analyses (Bliese & Ployhart, 2002). Hence, we built the model in three steps.

In the first step, we tested a random intercept model incorporating a nonlinear time trend, estimated with cubic spline basis functions with $k = 100$ and a weekday predictor, estimated with thin-plate basis functions and a $k = 7$ weekdays. The number of basis functions were investigated by using the *gam.check* function in the *mgcv* package (Harezlak, Ruppert, & Wand, 2018), which indicated that higher numbers were unnecessary. Adding a random intercept tested for significance differences in the starting point of the individuals' timelines. The second step added a random slope. This step still contained the same nonlinear (fixed) trend for all persons but allowed for different trend strengths (i.e., slopes). Technically, the random slope was represented by a smooth interaction between the individual and the trend variable. Finally, the third step, replaced the former random effects with a random smooth component, allowing for individual differences in the functional form of the trend including intercept and slope differences. Residuals of the final model were checked for signs of autocorrelation which were not indicated.

2.3.4 Results

Description. The dataset comprised 109 individuals for each group, respectively, who had posted $N = 595,751$ status updates in total. Regarding the temporal behaviour of users, the CT group posted more tweets on average daily basis over the year than the non-CT group whilst showing a higher proportion of CT-related content in comparison to COVID-19-related content (see Table 2.7).

Table 2.7: *Descriptive statistics of tweets by account for the CT group and non-CT group*

Tweet characteristics per user	CT-group			non-CT group		
	M(SD)	Min	Max	*M*(SD)	Min	Max
Number of daily tweets	11.70 (11.60)	0.27	70.00	2.56 (1.79)	0.05	9.13
Number of daily corona-related tweets over the year	0.59 (0.69)	0.01	3.62	0.46 (0.57)	0.01	3.85
Proportion of corona-related tweets over the year	.04 (.02)	0.003	0.11	.12 (.12)	0.005	0.59
Number of conspiracy-related tweets over the year	5.33 (6.07)	0.11	44.9	-	-	-
Proportion of conspiracy-related tweets over the year	0.35 (0.13)	0.06	0.65	-	-	-

Identifying Semantically Similar Expressions of CTs (RQ1). Concerning the first research question, the narrative themes contained in tweets from the CT posters were not restricted solely to motifs centring around the pandemic or lockdown measures but emerged in a variety of broader motifs (see also Samory & Mitra, 2018) including: (i) events (e.g., "9/11", the killing of George Floyd), (ii) elections (Democratic and Republican party politics), and (iii) domestic politics ("Hunter Biden scandal"), (iv) globalisation (e.g., "global communism"), (v) intelligence operations (e.g., military operations, bioweapons), (vi) media ("mainstream media"), or (vii) mystic rituals and paedophile rings (e.g., cabal, satan).

As aforementioned, the tweet content can be interpreted as touching the psychological needs of the person (i.e., existential, epistemic, or social needs). In this vein, when exploring the GloVe vector of the CT posters for similar vectors denoted by the cosine similarity (i.e., "c"), we were able to identify different realisations of these needs. Specifically, we considered cosine similarity values higher than .40 as indicating sufficient similarity. Further, values in the range from-1 to 0 by which values approximating 0 indicate low semantic similarity and conversely approaching 1 indicates high semantic closeness. Values with opposite polarity show diverging meaning.

For instance, threats of existential needs could be related to common uses of keywords such as "vaccinations" ($c = .45$ with "depopulation agenda"). These were connected with motifs referring to coalitions like pharma ($c = .44$ with "big_pharma"). Likewise, the keyword "deep_state" was associated with "hardware_us" ($c = .40$).

The notion of harm and threat was further taken up by a pervasive disapproval of media outlets, which is framed as a source of disinformation and a vehicle played by third parties to control the population (e.g., "news" was associated with "fake_news", $c = .56$ or "dangerous_lunacy", $c = .43$). Further, events like "9/11" (e.g., related terms were "pacification_psyop", $c = .46$; or "majority_murders", $c = .44$) were framed as staged and spun in a hidden fashion by "government insiders". Other sources of threat were prominent individuals (like Bill Gates) who were depicted as being in a quest for domination and personal gain (e.g., "satanic", $c = .49$; "overseas_spying", $c = .40$). Emerging current social movements like the Black Lives Matter ("blm") movement were incorporated into a threat narrative (e.g., "blm_antifa", $c = .59$; "terrorists", $c = .45$; "marxist", $c = .47$).

Relating to epistemic needs, pragmatic markers indicate the individual attention, assessment of causes and commitment to stances (Humphreys & Wang, 2018)] (e.g., "uncover" is associated with "growing_totalitarian", c = .40 or "batresearch_program", $c = .40$). Further, the element of secret agency and operations is used (e.g., "op" is associated with "chemicals_manufactured", $c = .42$; "trees_changed", $c = .41$; "programming_people", $c = .40$). In this vein, clear goals are set, like ending child trafficking or ending the lockdown and uncovering the truth beneath the surface. Researching information is turned into a game to solve the secret plot (essentially finding proof for why the official account is not true) and in the realms of satisfying social needs by belonging to those who see through (e.g. "research" was associated with "pedo_city", $c = .48$; "proven_scam", $c = .45$; "public_surface", $c = .41$). Henceforward, rhetorical tropes and epistemic markers of truth propositions and questioning coincidences play a functional role in engaging with conspiratorial ideation and further, as they are unlikely to be banned or shadow-banned, as a marker of shared interpretive frames in a consistent manner.

Group comparison of posting behaviour on overall COVID-19-related tweets (RQ2). The second research question centred around whether tweet posting behaviour focusing on general coronavirus content differed between the CT group and the non-CT group.

Fig. 2.5 shows the proportion of coronavirus-related tweets to the overall number of tweets in each group. As clearly depicted in the figure, from January to March both groups tweeted to a similar extent. From March onward, the non-CT group showed a constantly higher proportion of tweets than the CT group. In addition, there was a substantial increase in the proportion of tweets at the beginning of March. Three probable events causing the increase are first, extensive media coverage of the events in the northern regions of Italy (at the time, this was the area most affected by COVID-19 besides Wuhan in the Hubei Province, China), second the announcement of a countrywide quarantine in Italy on March 9 and

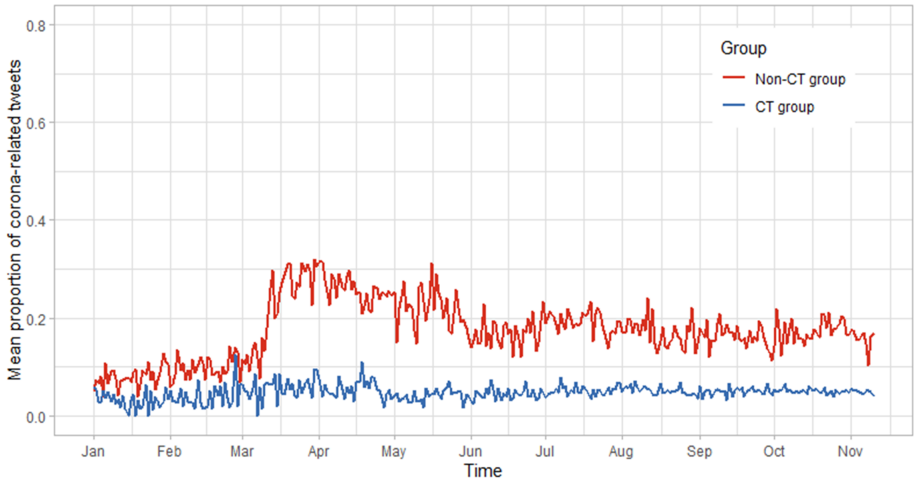

Figure 2.5. Mean proportion of COVID-19-related tweets by the CT group and the non-CT group

third subsequently the official declaration of the corona crisis as a pandemic by the WHO on March 11. Hence, particularly for the non-CT group, the "spread" of the coronavirus became prevalent in the following months (with associated terms such as "spread_covid19", $c = .66$; "stop_spread", $c = .57$; "prevent_spread", $c = .58$). Further, the declaration as a national state of emergency in the US resulted in public responses (for the non-CT group the GloVe model contained "emergency" associated terms like "health_emergency", $c = .54$ or "state_emergency", $c = .48$). As the subsequent analyses will show, the CT group in contrast focussed primarily on CT-related tweets.

With regard to temporal characteristics of the two series in Fig. 2.5, we estimated the autocorrelation and conducted an STL decomposition into the (nonlinear) trend, seasons, and remainder for both groups. The result was a substantially higher autocorrelation of the non-CT group ($r = .83$) than the CT group ($r = .11$), implying a stronger persistence in the posting behaviour from one day to the next. The STL decomposition suggested a weekly season effect with the tweet behaviour constantly high from Mondays to Thursdays and then rapidly dropping for the non-CT group. In contrast, no systematic pattern could be observed for the CT group. Based on the results from the estimation of the autocorrelation, we used the *auto.arima* function in R's *feasts* package to identify potential ARIMA models, and a model with no autocorrelated error structure was preferred.

In the last step, we analysed differences between both groups in the dynamics of the series using a GAM with a factor smooth interaction. We used cubic splines as the family of basis functions for the time trend and thin plate splines for the weekday.

Further, we set the number of basis functions to the highest possible number that led to a converging model. This was $k = 305$ for the time trend and $k = 7$ for

Table 2.8: *Results of the GAM investigating differences between the non-CT posters and CT posters in overall coronavirus-related tweets*

	Separate smooth model	
	B	*p*
Linear part of the model		
Intercept	-1.61***	<.001
Group: CT group	-1.53***	<.001
	EDF	*p*
Nonlinear part of the model		
Non-CT group: Season weekday	3.26***	<.001
CT group: Season weekday	1.00	.750
Non-CT group: Time trend	17.91***	<.001
CT group: Time trend	46.34***	<.001
R square	.92	
Deviance explained	.95	

Note. EDF = effective degrees of freedom (indicates the amount of wiggliness of a curve); EDF = 1 indicates a straight line; ***$p < .001$

the weekday season (i.e., the changing pattern across the week). We estimated two models. The first was a *separate-smooth model*, which results in the estimation of a season smooth and a time trend smooth separately in each group. The second model was a *difference smooth model* which—analogously to dummy interaction—estimates baseline smooths for the season and time trend (of the non-CT group) and difference smooths for the contrasting group (i.e., the CT group). Table 2.8 shows the results of the separate smooth model. Table 2.8 displays two types of coefficients. The coefficients in the linear segment are the regression intercept and the level difference between the CT group and the non-CT group. The EDF (effective degrees of freedom) in the nonlinear segment describe the nonlinear dynamics in the weekday effect and the overall time trend. The table shows that the non-CT group showed a significant weekday effect, while the CT posters did not. In addition, both groups showed a nonlinear trend which was significantly wigglier for the CT group ($p < .001$).

Identifying dynamics of CT-related tweets (RQ3). The third research question focussed on the time series of CT-related tweets and its characteristics. Fig. 2.6 shows the distribution of the number of CT posters across time (upper panel) and the proportion of tweets (lower panel) of this group. The figure shows that that number of individuals increased during 2020, revealing a horizontal spread of CT engagement (i.e., the number of involved persons) whereas the proportion showed an increase in the spring and a seemingly constant level for the rest of the year (we will re-consider this later regarding structural breaks). A preliminary GAM, not yet considering potential autocorrelation and seasonal-

ity, revealed a significant positive linear trend for the number of users ($B = .006$, $p < .001$) and proportions of tweets ($B = .0009, p < .001$) signifying an 8% increase from January to November. Furthermore, specifying a nonlinear trend in three GAMs resulted in a significantly better data fit than the linear variants in all three cases. It should be noted, however, that the number of users being represented on each day differed across the time, which cannot be reflected in the single-series GAM. This will be considered in the section on individual trends and their averages.

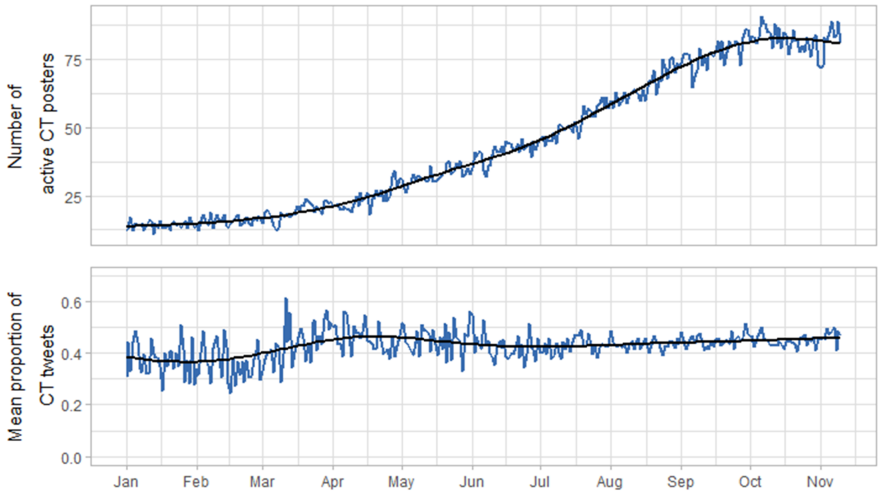

Figure 2.6. Distribution of CT posters (upper panel) and mean proportion of CT tweets (bottom panel)

The next two steps focussed on autocorrelation and seasonality. The autocorrelation function revealed a mean lag-1 autocorrelation of $r = .34$ for the tweets. Although this correlation was higher compared to the tweets with general COVID-19-related content ($r = .11$), this still indicated a lack of substantial autocorrelation (especially if compared to the $r = .83$ in the non-CT group). When estimating the GAM, we found that a simple model without an autoregressive and moving average component would best fit the data—therefore, we repeated the GAM by only including a weekly seasonal smooth that had been suggested by the STL decomposition. The results showed, however, no significant seasonality. Hence, the interpretation of the aforementioned seasonal pattern should be undertaken with caution. Overall, these analyses suggest that the posting behaviour of CT-related content contains a high degree of randomness and day-specific dynamics.

As the final analysis, we conducted a structural break analysis by investigating structural breaks in linear trends within segments. Despite the non-significant season effects, we based this analysis on the residuals of a former GAM with a nonlinear season estimate but no trend. A test focusing on the cumulated sum of standardised residuals (CUSUM fluctuation test) and the F-test ($F = 77.11, p < .001$)

indicated a significant deviation from the null hypothesis that all measures are reflections of the same data generating process. A subsequent analysis of variants with differing numbers of breakpoints showed that the Bayesian Information Criterion (BIC) suggested two breakpoints whereas the residual sum of squares indicated that all models with more than one breakpoints had equal fit. Fig. 2.7 shows the breakpoints and their confidence intervals (i.e., the grey area) for CT tweets of the CT group.

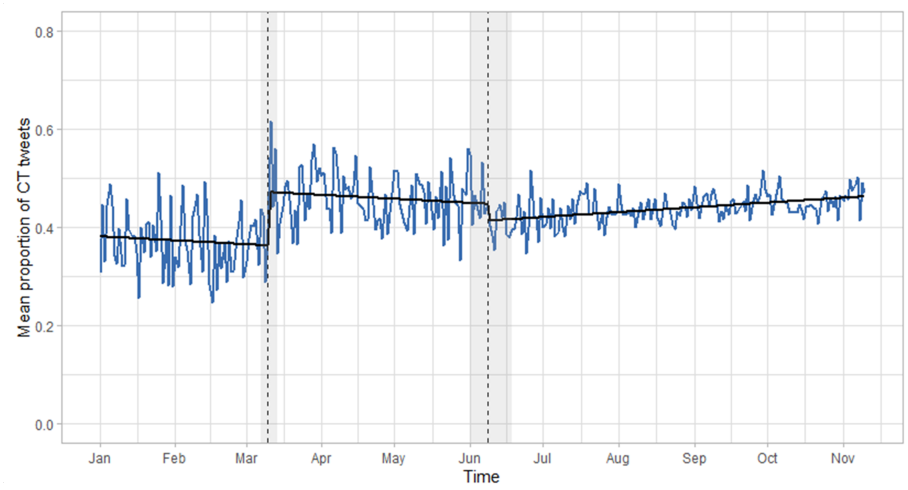

Figure 2.7. Structural break analysis of CT tweets for the CT group via the CUSUM and F-test. Gray areas indicate confidence intervals for two structural breaks on March 10, 2020 and June 8, 2020 (dashed lines)

The dates associated with the breakpoints were March 10 and June 8. Noteworthy events in this timespan are on the one hand the implementation of public health measures on March 8, 2020 which resulted in strict social distancing measures in highly affected European countries like France or Spain. This was followed by the WHO pandemic declaration. The second breakpoint falls into the time of the emerging "Black Lives Matter" Movement on June 3 and the George Floyd protests and tearing down of memorials in several other countries during the following days.

Furthermore, the division of the series in three segments resulted in non-significant trends in the first two segments (both $Bs = .0002, p = .46$ and $p = .24$, respectively) but a significant trend in the phase beginning on June 9 ($B = .0003, p < .001$). This effect, however, should not be overinterpreted due to the substantially higher power compared to the trend estimates in the first two segments.

Individual differences in temporal characteristics (RQ4). In addition to the analyses of the tweets on an aggregate level, we investigated the series of individual CT posters. As a consequence of the Twitter API regulations on the amount of available historical tweets of individual timelines, we had series

that varied in length, with an average length of 174 days (SD = 92), ranging from 37 to 315 days. Whereas the former analyses presented information about the overall dynamics of the aggregate posting behaviour, the following analysis focussed on inter-individual differences in the dynamics, including differences in the autocorrelation, level, slope, and functional linear and nonlinear trends.

With regard to inter-individual variations in the autocorrelation, we found substantial differences ranging from -.34 to .55 (M = .10, SD = .15). The first pattern was most often a result of switching between days on which a person tweeted CT content followed by one or several days of either not tweeting at all or tweeting only non-CT content. The positive autocorrelation consisted of consecutive series of days on which the person posted followed by several days of absence.

Table 2.9: *Results of the generalised additive mixed-effects model addressing inter-individual differences in time trends*

	Random intercept model		Random slope model		Random smooth model	
	EDF	p	EDF	p	EDF	p
Fixed effects						
Time trend	7.77***	<.001	7.77***	<.001	7.50***	<.001
Weekday	2.36	.083	2.34	.092	2.34	.095
Random effects						
Random intercept	102.80***	<.001	77.85***	<.001		
Random slope			71.42***	<.001		
Random smooth					266.14***	<.001
R square	.139		.156		.180	
Deviance explained	.009		.009		.010	
AIC	-423,743.1		-424,046.6		-424,468.2	

Note. EDF = effective degrees of freedom (indicates the amount of wiggliness of a curve); EDF = 1 indicates a straight line; ***p < .001; AIC = Akaike information criterion

As Table 2.9 reveals, the Akaike information criterion (AIC) was lowest for the random smooth model, indicating significant differences in the nonlinear dynamics between individuals. Furthermore, the explained variance was low for all models showing the large individual deviations, often spanning the range between zero tweets per day (and accordingly zero proportion of conspiracy content) up to 100 percent CT content. Finally, while the fixed effects for the time trend revealed a nonlinear average trajectory across time, there was no significant weekday effect.

To analyse the individual nonlinear trends and to judge the percentages of individual positive versus negative linear trends, we estimated specific single-person GAMs for the CT posters. To apply a comparison standard and not to overwhelm

the depiction, Fig. 2.8 shows the trends for those individuals for which at least 200 days of data were present. The figure shows the immense differences in level and nonlinear trends across time.

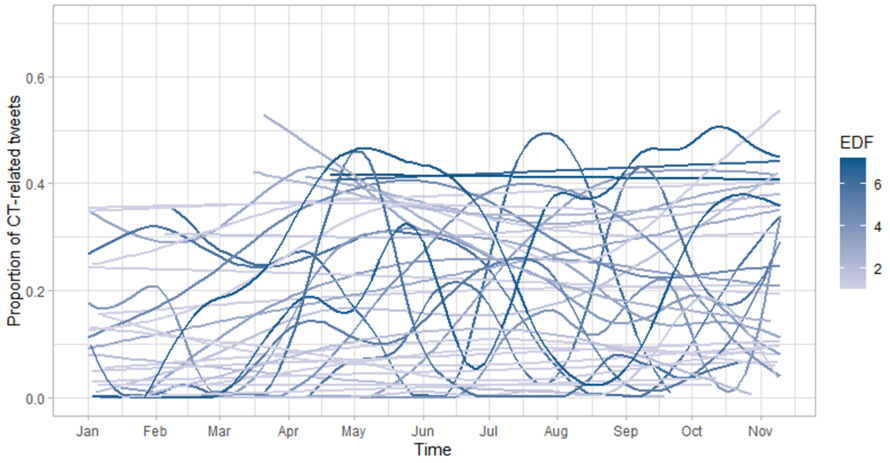

Figure 2.8. Proportion of CT-related tweets and effective degrees of freedom (EDF) for a subsample of individual CT posters (at least 200 days of posting behaviour)

To draw a conclusion about which percentage of the CT posters systematically increased or decreased the proportion of CT content, we estimated a GAM with a linear time trend while controlling for a nonlinear weekday effect. The regression coefficients had a mean of $B = -.001$ ($SD = .012$) with a min of $B = -0.06$ and a max of $B = .04$, which showed no overall trend but also inter-individual differences in the increase versus decrease of posted CT content.

2.3.5 Discussion

In the present article, we investigated the spread of CT tweets on Twitter throughout the first year of the COVID-19 pandemic in 2020. We used state-of-the-art text analytics (word embeddings) and time series analyses on Twitter timelines of 109 CT posters and non-CT posters, respectively, to investigate the content of CTs as well temporal characteristics of aggregate and individual series. Results showed that CT tweets fit well with claims of scholars emphasising the role of violation of existential needs in endorsing CTs (van Mulukom et al., 2020; Kay et al., 2009). In this regard, CTs can be interpreted as the individual's attempts to cope with an uncertain and dangerous situation and to attribute causes to external agents in order to gain control. While CTs have been shown to involve cognitive biases, they can be seen as evolved patterns to cope with existential threats and perceived powerlessness. This prepares the ground for user-generated content that refers to considerably few lucid coalitions (e.g., abstract references to the government, media, or concrete ones like Anthony Fauci) that are adapted to new events occurring.

The use of word embeddings to identify CTs and broaden conceptual knowledge

Our approach of using word embeddings, informed by a minimal set of theoretical constructs (agency, pattern, coalition, secrecy, threat), resulted in the identification of terms with related semantic meaning that further enrich our knowledge on conspiratorial worldviews and implicit language use. In finding CTs in which either the severity or existence of the pandemic is called into question (i.e., hoax) or that blame certain actors for causing the pandemic (i.e., Bill Gates, China, deep state), as a way of a collective sensemaking of events, our results align with those of van Mulukom et al. (2020). The latter of the two schemes exemplifies an integration with other pre-existing, conceptually unrelated CTs, for instance, relating to the "pizzagate conspiracy", anti-vaccination, "9/11 inside-job" or QAnon (see also Wood, Douglas, & Sutton, 2012). These strategies might eventually steer different prevention behaviours of the posters—that is—rejecting prevention measures altogether or only partially. In a similar vein as Samory and Mitra (2018) noted, albeit coalitions are easily discernible the other theoretical constructs (e.g., threat or pattern) are much more finely distinguished.

Analysing the temporal dynamics of CT tweets

Beyond the semantic analyses, the temporal analyses resulted in insights into the temporal dynamics of CT tweets on the average level and the individual level as well as differences between CT posters and non-CT posters and within the group of CT posters. First, we found substantial differences between non-CT posters and CT posters in the series of tweets focusing on motifs centering on coronavirus-related content. In particular, the series of these tweets of the CT posters had a remarkably lower level indicating that although having the same reactivity to coronavirus-related events (e.g., rising infection rates, governmental measures), CT posters tend to strongly respond with CT-related content. Hence, both groups differ on the abstraction level of their responses. This is most apparent when integrating the results of Fig. 2.5 and Fig. 2.6 to show that the CT group posted fewer tweets containing non-CT content compared to the non-CT posters.

As a second substantive result, the time series indicated a strong dynamic in the posting pattern of users in the CT group indicating a substantial impulsiveness of the posting behaviour. This was evidenced by a significantly stronger wiggliness of the overall series, the much lower autocorrelation, the lower level of weekly seasonality, and the lack of residual autocorrelation. The latter suggests that the behaviour of CT posters is an impulsive reaction to day-level events and not a stepwise and sustained distribution of CT content. This aspect has implications for the evaluation of the role of Twitter as a facilitator of an individual self-radicalisation.

The latter is also indicated by the negligible trend in the proportion of CT-related tweets across 2020. This result shows that the posting behaviour of the CT group as a collective does not indicate a disequilibrium or imbalance but can rather be represented as a stochastic process.

While these results concern aggregate level of analysis, analysing the individual level revealed a more complex and diverse picture. The analyses revealed substantial differences in the level of inertia—indicated by the strong differences in the individual autocorrelations—as well as the trends in terms of slope and functional form. In this regard, some individuals showed a linear upward trend and others a strong, dynamic reactivity. However, for those exhibiting a linear upward trend, this trend again was not substantial. The wiggliness of some series suggest that these individuals were more reactive to daily stimulations. In line with existing theory, this finding can be explained by the internally driven pattern of behaviour shown by a CT-prone person. This type of person is trying to make sense of the news he/she receives with the ultimate goal to fulfil his/her epistemic, existential and social needs by showing hyperactive pattern recognition, which turns into the maladaptive behaviour of endorsing conspiracy beliefs (Douglas, Sutton, & Cichocka, 2017). This erratic hyperactivity is striking if compared with the aggregate trend of the non-CT posters. This group showed greater inertia and more consistent engagement with mainstream content at the aggregate level, a behaviour pattern in stark contrast to the erratic reactivity of the CT posters.

2.3.6 Implications

Beyond providing the insights discussed before, our results may stimulate future research that addresses issues that are beyond the scope and possibilities of the present paper. First, as discussed above, our results point to a posting behaviour that can best be described as a stationary stochastic process, which again may be interpreted as a signal of calm in the ongoing discussions about social media, the spread of conspiracy convictions, and false narratives. It should be noted, however, that this interpretation only concerns the number of users sampled within this study and their behaviour but not a potential spread of conspiracy information and growth of social networks across future CT users. In this regard, one result in this study was the strong upwards trend of the number of users that—implied that the bulk of users sampled had emerged in the later part of 2020. Hence, we recommend investigating the potential divergence between personal radicalisation processes and a nonetheless possible spread of CT content.

Second, the remarkable inter-individual differences point to individual or contextual determinants of these differences. In this regard, our study lacked the data to further investigate such determinants, most probably by integrating social me-

dia data with individual level data from, for instance, surveys. Survey data have a long tradition in the social sciences and allow researchers to measure relevant constructs (e.g., personality traits, political attitudes, demographic information) in a reliable and valid manner. Methodologically, such questions can both be approached by using modern multilevel models (e.g., with such person factors predicting features of the individual series of tweets (e.g., trend, wiggliness, inertia) as well as typological or cluster-based approaches targeting the identification of groups of individuals with a similar radicalisation process. Furthermore, the validity and robustness of discriminating conspiracy content from non-conspiracy content should be further corroborated by comparing our results to baselines.

Limitations of the study

The present study is confronted by some limitations which, although not critical for the main results of the study, should be taken into consideration for future research. First, Twitter's API rate limit led to timelines that differed substantially in the time span of retrieved content. As a consequence, an individual's time span showed a moderate albeit significant correlation ($r = -.25, p < .001$), indicating that longer series were wigglier than in cases where the contingent of tweets were spread more evenly across the time span. Without representing a limitation per se, we note that one of the influences on the series' dynamics may not be psychological but technically based.

Second, our results may be specific to the COVID-19 pandemic and not generalisable to other forms of societal events and their interpretation. Likewise, we recommend a careful interpretation of our study for CT processes beyond posting on Twitter, as these users may represent a sample that is not representative of the general population (Ruths & Pfeffer, 2014; Wojcik & Hughes, 2019). Hence, future research should investigate longitudinal processes of other platforms for measuring CTs as well as the key differences between the Twitter population and other populations. Likewise, extreme CT-prone persons may be banned from Twitter or adapt their behaviour so as not avoid being banned—thus, indicating an example of proxy population bias (Ruths & Pfeffer, 2014).

Third, another restriction on sample representativeness are implications of tweet deletion and account suspensions. A potential result of users deleting their tweets or accounts, setting accounts to private, or becoming suspended due to a violation of Twitter's Terms of Service might lead to the underrepresentation of misinformation content in a data sample (see Maddock, Starbird, & Mason, 2015). Thus, it needs to be acknowledged that the true rate of conspiracy content might be higher than stated and users with high trends in their postings might be missing. One step towards ameliorating this problem is concentrating further

on sampling user timelines with the REST API and adding a real-time component by refreshing the dataset over time and comparing changes which are due to deletion for this time period. This could be a viable approach for conducting historical tweet analyses, given that anonymity and ethical principles for users are considered (Maddock, Starbird, & Mason, 2015). Finally, a strength of our study is that it is founded on scholarly definitions of CT properties that are general and scalable and, hence, offers several implications for further research. Specifically, our analysis pipeline proved to be suitable for matching theoretical expectations in terms of both user's group and individual behaviours. Future research could take up, on the one hand, on contextualised word piece embeddings that mitigate issues of word sense disambiguation and provide bidirectional contexts (e.g., with *Bidirectional Encoder Representations from Transformers*) and, on the other hand, temporal word embeddings that allow for modelling language evolution, for instance, with probabilistic state space models where the word and context embeddings evolve with time. Such outlooks on word evolution may provide information to perspectives of how users adapt language as an indicator of increasing radicalisation. Such an approach could provide further information on the change of content meaning over time and provide fine-grained insights into emotional responses which evoke responses at short time intervals (e.g., minute scale). This type of semantic and temporal approach can add valuable information to theoretical assumptions on feelings of anxiety and lack of control, which have been the focus of survey studies (see Šrol, Mikušková, & Cavojova, 2021) to date. Taken further, when considering temporal dependencies of emotions on social media, dynamic modelling techniques for studying within-person processes can be fruitful. Eventually, using a case-control design holds the potential of inquiring causal questions, as to comparing the impact of certain events, user characteristics or social factors on user behaviours.

Chapter 3

General discussion

3.1 Contributions revisited

In the present section, I summarise the findings of this thesis by revisiting the contributions. While the findings referring to radicalisation processes do not automatically extend to every source of social media or thematic domain, the results on quality aspects are generic and have generalisable implications. In particular, I argue to cast a broad view from all phases in the research cycle on digital behavioural trace data (see Figure 1.1.) to make analyses transparent and generalisable. In particular, operationalising constructs according to the theoretical underpinnings is a challenge (in terms of capturing all relevant aspects of the construct), as automated textual analysis is a relatively new method for measuring psychological constructs. Notably, requirements for digital behavioral trace data differ depending on the *inference goal*. If researchers have research goals that extend descriptive analyses (e.g., classifying radical content) and, rather, aim for inference on the population level, data-generating processes that might have caused divergence between the sample frame and the target population are important (Groves et al., 2011). Different measurement outcomes of the same underlying construct may result depending on how the measurement process is shaped that is—which environmental influences are considered, which statistical procedures are used to model noise, which proxies are used in the application of the background theory. These points open up further development areas that the main contributions build on:

1. **Operationalisable definition of radicalisation**
   In the dissertation I addressed the short-comings measuring radical beliefs and behaviours (as identified in *Section 2.1*)—in particular with regard to the lack of a dynamic perspective. I provide this dynamic perspective and show that one can learn something from the dynamics of underlying psychological entities and their rhythms and regularities. I constitute, particularly, the value of interaction traces

and their temporal patterns for theory-rooted insights into long-term changes and inertia, as system responses to shocks and external events or abrupt, short-term spikes and trends (*Section 2.3.4*). Moreover, I constitute (*Section 2.1*) that the radicalisation outcome is multifaceted. This translates into perceiving conspiracy beliefs (as a radicalisation outcome) as involving multiple facets, such as a "hypersensitive agency detection system" (Van Prooijen and Van Vugt, 2018, p. 773) (*Section 2.3.2*). As no canonical conceptualisation of conspiracy beliefs exists, I adapt a theoretical framework from evolutionary psychology that views conspiracy beliefs in terms of functional adaptions and assumes a combination of five generic criteria (agency, threat, secrecy, pattern and coalition) to measure the construct. Further, I differentiate various psychological needs (i.e., epistemic, existential and social) as potential determinants. I quantify the *relative engagement* of online users with conspiracy beliefs by an iterative semi-supervised machine learning approach from information retrieval literature that allows to incorporate expert knowledge on various levels and vector models, to model semantic meaning by its context and by extension, to also differentiate from non-conspiracy content (*Section 2.3*). This work, thus, provides both theoretical insights for researchers on conspiracy beliefs, as well as practical implications for practitioners on automating and scaling up the differentiation, as well as drawing valid inferences from the data obtained.

2. **Theoretical integration through network theory**
   A second contribution of this dissertation is the integration of theories, hypotheses and constructs established in the field by network theory. This approach has two values. First, it presents a summary and integration of the theoretical foci held by researchers and hence, quantifies specifically research agendas. Second, it allows to set the stage for a more detailed causal thinking by viewing the networks of hypotheses as a network of causal claims. By doing so, developments in causal theory (e.g., graph theory, Pearl, 2009) can be applied and conclusions about the diverse role of the involved variables can be made (see *Section 2.2*). In particular, as the network analysis indicated, some constructs were uniformly hypothesised as having the role of mediators (e.g., psychological health). Most constructs were expected to be causes as well as outcomes, implying their potential role as *confounders* (i.e., a variable which figures as a common cause for two or more other target constructs) or *colliders* (i.e., a variable which is caused by two or more target constructs). Theorising about the presumed causal role of the variables in the network can be fruitfully used to advance research (e.g., by generating appropriate designs or deciding upon control variables). For this purpose, I curated a list of higher-order constructs and those constructs contained in the primary studies (see Table 2.2 and Table 2.5) that can serve together with the network statistics as a checklist to consider them as potential confounders for a particular relationship between target constructs.

3. **Construct validity of digital behavioural trace data**
   Within the scope of this dissertation, I note the advantage of digital behavioural trace data to extract data irrespective of desirability tendencies. The unobtrusive way of measuring constructs lends claims to their ecological validity. In *Section 2.1*, I further conclude that digital behavioural trace data pose challenges when it comes to measuring psychological dispositions, where aspects such as reliability or validity are often unclear. Even more so, one of the biggest hindrances to analyse validity is the problem to acquire auxiliary variables that could serve as validation criteria (e.g., demographic variables, political affiliation, or attitudes) (see also *Section 2.2*). Although certain user characteristics can be indirectly inferred such as gender or geolocation these, however, are limited in their scope and subject to uncertainty—when not enough digital behavioural trace data are available (Sen et al., 2021). Additionally, I argue in *Section 2.3* for multiple instances within the research cycle when inferences might be jeopardised by selection bias, platform specific norms and affordances, as well as distortions to measurement quality induced by pre-processing and modeling data. I address some of the challenges and introduce substantive documentation in *Section 2.3*. More importantly, I propose to consider the following *caveats* to validity which I exemplify in the:

   Sampling phase. I demonstrated how to **(i)** circumvent the limitations of API-based sampling by sampling individuals who were active over a substantive time and at different points in time (to avoid recruiting only the most active users or users with a very short activity span). Further, I recommend addressing the ephemeral nature of such data **(ii)** by sampling individual timelines in real-time and acknowledging deletions, moderations or revisions to user-generated content (which can inform potential confounding structures). Additionally, I illustrated how to mitigate selection bias **(iii)** by sampling on broad topics (rather than specific conspiracy theories) and to differentiate only later, based on a semi-supervised approach, conspiracy theory beliefs, which enhances the external validity of the approach (as being generalisable to other people). I moreover, argued to iteratively filter accounts **(iv)** (e.g., based on activity thresholds or automation) and qualitatively annotate user-generated content, based on theory-derived criteria to establish their internal validity and suitability for the construct. In a similar vein, I established the face-validity of my measure of conspiracy beliefs by validating the extent to which non-CT users post conspiracy content based on the inter-rater reliability of human-coding and classification.

   Analysing phase. I further illustrated the need to evaluate modeling decisions **(v)**. One instance is the intrinsic evaluation of the word embeddings performance, regarding the accuracy of the hyper-parameter settings chosen. Another instance is the validation of the accuracy of the concept movers distance classification by human ratings. However, I further call for testing the theoretical assumptions of conspiracy theories measured here, not in isolation, but against other alternative theories to account for anomalies. Further, assuring quality of the data **(vi)**, such

as accounting for ontological differences in different levels and modeling noise in time series analysis, mitigates biases at this stage.

Overall, my aim has been to provide a specific view on the requirements of employing different digital behavioural trace data types in radicalisation—with a focus on the content, scope, justification and limits of such approaches and further casting a bird's eye on the caveats in a research cycle that generalise over the research domain. The three contributions touched upon several aspects particularly, for Twitter-based studies that should be accounted for in future research designs. However, further research is needed when it comes to addressing issues of *external validity* that relate to how well findings from this Twitter study generalise to other (out-of-domain) settings, as well as, to other populations. Comparing, however, across platforms is difficult, as the traces of behaviour are not standardised. The obtained data can greatly vary according to the platform considered, as for instance, some online services primarily focus on user-generated content that is either text-based, pictures or video-based. Furthermore, data sources are shaped by different practices and formats, and due to proprietary rights, the obtained data are inevitably selective and subject to temporal change (Proferes et al., 2021; Tufekci, 2014). In an effort to address these hindrances to inference, openly accessible and findable research data sets would be a major leap forward. It would be a worthwhile research avenue to assess the extent to which different platform affordances affect the inference and how errors are propagated and potentially amplified, within the purview of the research cycle. Identifying further bottlenecks in the analysis of digital behavioural trace data across platforms and for different analysis strands should inform the development of rigorous documentation and establish transparent evaluation practices. Light-house projects in psychology, for instance, focused on developing an adaptable template for specifying and communicating the main elements of a study before conducting research (i.e., preregistration for psychology, see Bosnjak et al., 2021). Developing reference framework in a similar vein for digital behavioural trace data, even for exploratory work, would benefit questions of replicability and quality.

3.2 Implications

The presented studies point to further research avenues that relate to *data linkage*, as well as to *data reuse and sharing*. I have touched on the shortcoming of digital behavioural trace data of missing auxiliary variables and their shortcomings regarding theory-driven (causal) analyses. One way of overcoming the limitations is to enrich these data and cross-validate measurements for psychological beliefs, by relating established measurement instruments of survey methodology

to unobtrusive measures. Linking individual digital behavioural trace data, via the individual geo-coordinates, with *secondary data* of the environment—such as socio-demographic characteristics (e.g., crime rate, unemployment rate), environmental pollution or infrastructure—offers the objective analysis of contextual influences (as main and moderator effects of the environment). Another interesting future research endeavour could be the combination of digital behavioural trace data with *sensor data* (such as generated from smartphones or app usage) which offers new forms of granularity of human behaviour (e.g., individual level or aggregate level characteristics for a specific geographic region) (Beuthner et al., 2021). By extension, one could model, for instance, the influence of social behaviour measured via Bluetooth and GPS on changes in individual well-being. Further developments in psychology in the field of *Ecological Momentary Assessment* (see Lutz et al., 2018) offer fine-grained longitudinal insights into within-subject developments by being able to combine sensor data and questionnaires and by extension, linking online behaviors with offline outcomes.

Those benefits, however, can only be achieved by combining the work and expertise of different researchers, as the sampling of different data types requires different expertise. The opportunity to re-use data generated through research is vital, as it should encourage new research and address validity issues. The lack of gold standard data sets forms a particular challenge (not only in radicalisation research). Annotated data sets are the center of many machine learning developments and evaluation. Relating to various natural language processing tasks models may rely on faulty heuristics and learn spurious cues, present in the data. Models trained, for instance, on Wikipedia data may include latent variables that might be spuriously associated with the target construct. So that a model learns more about general Wikipedia article than about the construct of interest (e.g., due to an overrepresentation of certain linguistic patterns). Future research should explicitly account for such confounding variables by making use of causal graphs to explicitly encode causal assumptions (Pearl, Glymour, & Jewell, 2016).

When sharing digital behavioural trace data with the research community various challenges arise, that differ from traditional study designs. These relate on the one hand, to the circumstance that the data might contain privacy-sensitive information (e.g., geolocation information or names). On the other hand, contents from social media platforms are subject to proprietary regulations and often cannot be shared. Frequently, for Twitter data sets, *re-hydrateable* data sets are shared (containing only Tweet-IDs, which can serve to ”reconstruct” the original data set, by once more querying the respective API). Yet, data have an ephemeral structure and can be deleted, edited or moderated in the community. Hence, these information may be missing in the final obtained data set. This caveat profoundly impacts replicability of analyses. Further, cooperation with platform companies such as

Twitter to share their data unobstructedly or exploiting synthetic social media data to generate benchmark data for the public would be valuable for the research community.

References

Aguinis, H., Gottfredson, R.K., & Culpepper, S.A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*(6) 1490–1528. https://doi.org/10.1177/0149206313478188

Ajzen, I. (2005). Attitudes, personality and behavior. Milton-Keynes, England: Open University Press.

Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys*, *34*(3), 512-547.

Alizadeh, M., Weber, I., Cioffi-Revilla, C., Fortunato, S., & Macy, M. (2019). Psychology and morality of political extremists: evidence from Twitter language analysis of alt-right and Antifa. *EPJ Data Science*, *8*(1), 17.

Almagor, M., & Ehrlich, S. (1990). Personality correlates and cyclicity in positive and negative affect. *Psychological Reports*, *66*(3_suppl), 1159-1169. https://doi.org/10.2466/pr0.1990.66.3c.1159

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21, 1086-1120. https://doi.org/10.1016/j.leaqua.2010.10.010

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Bail, C. A., Merhout, F., & Ding, P. (2018). Using Internet search data to examine the relationship between anti-Muslim and pro-ISIS sentiment in US counties. *Science Advances*, *4*(6). https://doi.org/10.1126/sciadv.aao5948

Bäck, E. A., Bäck, H., Altermark, N., & Knapton, H. (2018). The quest for significance: Attitude adaption to a radical group following social exclusion. *International Journal of Developmental Science*, 1-12.

Baier, D., Manzoni, P., & Bergmann, M. C. (2016). Einflussfaktoren des politischen Extremismus im Jugendalter. Rechtsextremismus, Linksextremismus und islamischer Extremismus im Vergleich. *Monatsschrift für Kriminologie und Strafrechtsreform*, *99*(5), 171-198.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological Science*, *26*(10), 1531-1542.

Barron, D., Morgan, K., Towell, T., Altemeyer, B., & Swami, V. (2014). Associations between schizotypy and belief in conspiracist ideation. *Personality and Individual Differences*, *70*, 156-159. https://doi.org/10.1016/j.paid.2014.06.040

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, *22*(6), 725-730.

Beasley, A., & Mason, W. (2015, June). Emotional states vs. emotional words in social media. In *Proceedings of the ACM web science conference* (pp. 1-10).

Bertin, P., Nera, K., & Delouvée, S. (2020). Conspiracy beliefs, rejection of vaccination, and support for hydroxychloroquine: A conceptual replication-extension in the

COVID-19 pandemic context. *Frontiers in Psychology*, 11, 2471.

Beuthner, C., Breuer, J., & Jünger, S. (2021). Data Linking-Linking survey data with geospatial, social media, and sensor data. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS Survey Guidelines). https://doi.org/10.15465/gesis-sg_en_039

Bhui, K., Otis, M., Silva, M. J., Halvorsrud, K., Freestone, M., & Jones, E. (2019). Extremism and common mental illness: Cross-sectional community survey of White British and Pakistani men and women living in England. *The British Journal of Psychiatry*, 1-8.

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods*, *5*(4), 362-387.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892-895. https://doi.org/10.1126/science.1165821

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borum, R. (2011). Radicalization into violent extremism I: A review of social science theories. *Journal of strategic security*, *4*(4), 7-36.

Bosnjak, M., Fiebach, C., Mellor, D. T., Mueller, S., O'Connor, D. B., Oswald, F. L., & Sokol-Chang, R. (2021). A template for preregistration of quantitative research in psychology: Report of the Joint Psychological Societies Preregistration Task Force. *American Psychologist*. https://doi.org/10.31234/osf.io/d7m5r

Box-Steffensmeier, J. M., Freeman, J. R., Hitt, M. P., & Pevehouse, J. C. (2014). Time series analysis for the social sciences. Cambridge University Press.

Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of COVID-19 misinformation. Reuters Institute, 7(3), 1.

Burt, R. (2000). The network structure of social capital. *Research in Organizational Behavior*, 22, 345-423.

Caporale, T., & Grier, K. (2005). How smart is my dummy? Time series tests for the influence of politics. *Political Analysis*, *13*(1), 77-94.

Cattell, R. B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*, Plenum, New York.

Chadwick, A., Kaiser, J., Vaccari, C., Freeman, D., Lambe, S., Loe, B. S., ... & Yu, L. M. (2021). Online social endorsement and Covid-19 vaccine hesitancy in the United Kingdom. *Social Media+ Society*, *7*(2), https://doi.org/10.1177/20563051211008817

Chan, M. P. S., Winneg, K., Hawkins, L., Farhadloo, M., Jamieson, K. H., & Albarracín, D. (2018). Legacy and social media respectively influence risk perceptions and protective behaviors during emerging health threats: A multi-wave analysis of communications on Zika virus cases. *Social Science & Medicine*, 212, 50-59.

Chollet, F., & Allaire, J. J. (2017). Deep Learning with R. Manning Publications, Manning Early Access Program.

Christmann, K. (2012). *Preventing religious radicalisation and violent extremism: A systematic review of the research evidence*. Research Report. Youth Justice Board.

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*(1), 40-64.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. *IEEE Transactions on dependable and secure computing*, *9*(6), 811-824.

Coid, J. W., Bhui, K., MacManus, D., Kallis, C., Bebbington, P., & Ullrich, S. (2016). Extremism, religion and psychiatric morbidity in a population-based sample of young men. *The British Journal of Psychiatry*, *209*(6), 491-497.

Clarke, S. (2002). Conspiracy theories and conspiracy theorizing. *Philosophy of the Social Sciences*, *32*(2), 131-150.

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (Fifth ed.). Los Angeles, CA, USA: Sage.

Corner, E., & Gill, P. (2015). A false dichotomy? Mental illness and lone-actor terrorism. *Law and Human Behavior*, *39*(1), 23-34. `http://dx.doi.org/10.1037/lhb0000102`

Crocker, J., Luhtanen, R., Broadnax, S., & Blaine, B. E. (1999). Belief in US government conspiracies against Blacks among Black and White college students: Powerlessness or system blame?. *Personality and Social Psychology Bulletin, 25*(8), 941-953.

Dahl, V. (2017). Reducing adolescents' approval of political violence: The social influence of universalistic and immigrant-friendly peers. *Zeitschrift für Psychologie*, *225*(4), 302-312.

Dalal, R. S., Bhave, D. P., Fiset, J. (2014). Within-person variability in job performance: A theoretical review and research agenda. *Journal of Management*, *40*(5), 1396-1436.

De Choudhury, M. D., & De, S. (2014, May). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*. `https://ojs.aaai.org/index.php/ICWSM/article/view/14526`

Dechesne, M. (2009). Explorations in the experimental social psychology of terrorism: The struggle-violence link and its predictors. *Revue Internationale de Psychologie Sociale*, *22*(3-4), 87-102.

Desmarais, S. L., Simons-Rudolph, J., Brugh, C. S., Schilling, E., & Hoggan, C. (2017). The state of scientific knowledge regarding factors associated with terrorism. *Journal of Threat Assessment and Management*, *4*(4), 180.

Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, *156*(3), 193-203. `https://doi.org/10.1093/aje/kwf062`

Doosje, B., Loseman, A., & Van den Bos, K. (2013). Determinants of radicalization of Islamic youth in the Netherlands: Personal uncertainty, perceived injustice, and perceived group threat. *Journal of Social Issues*, *69*(3), 586-604.

Douglas, K. M., & Sutton, R. M. (2011). Does it take one to know one? Endorsement of conspiracy theories is influenced by personal willingness to conspire. *British Journal of Social Psychology*, *50*(3), 544-552.

Douglas, K. M., & Sutton, R. M. (2015). Climate change: Why the conspiracy theories are dangerous. *Bulletin of the Atomic Scientists*, *71*(2), 98-106.

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current directions in psychological science*, *26*(6), 538-542.

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3-35. https://doi.org/10.1111/pops.12568

Ebner, J. (2019). *Radikalisierungsmaschinen - Wie Extremisten die neuen Technologien nutzen und uns manipulieren*. Suhrkamp Verlag.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. Journal of Counseling Psychology, *34*(4), 437.

Egan, V., Cole, J., Cole, B., Alison, L., Alison, E., Waring, S., & Elntib, S. (2016). Can you identify violent extremists using a screening checklist and open-source intelligence alone? *Journal of Threat Assessment and Management*, *3*(1), 21-36.

Elekes, Á., Englhardt, A., Schäler, M., & Böhm, K. (2018, October). Resources to examine the quality of word embedding models trained on n-gram data. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 423-432).

Ellis, B. H., Bixby, C., Miller, A., & Sideridis, G. (2016). Understanding pathways to and away from violent radicalization among resettled Somali refugees. Boston: Boston Children's Hospital.

Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research*. (pp. 245-273). Dordrecht Heidelberg New York London: Springer.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*, 31-53. https://doi.org/10.1146/annurev-soc-071913-043455

EUvsDISINFO.: DISINFO DATABASE". https://euvsdisinfo.eu/disinformation-cases/

Fernandez, M., Asif, M., & Alani, H. (2018). Understanding the Roots of Radicalisation on Twitter. Paper presented at the *Proceedings of the 10th ACM Conference on Web Science*, Amsterdam, Netherlands. http://dx.doi.org/doi:10.1145/3201064.3201082

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York: Taylor & Francis.

Franc, R., & Pavlović, T. (2018). Systematic review of quantitative studies on inequality and radicalisation. *DARE* (725349).

Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., ... & Lambe, S. (2020). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological medicine*, 1-13.

Frischlich, L., Rieger, D., Hein, M., & Bente, G. (2015). Dying the right-way? Interest in and perceived persuasiveness of parochial extremist propaganda increases after mortality salience. *Frontiers in Psychology*, 6, 1222.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-74. https://doi.org/10.1257/jel.20181020

Gerlach, M., Peixoto, T. P., Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, *4*(7), eaaq1360.

Gladkova, A., Drozd, A., & Matsuoka, S. (2016, June). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop* (pp. 8-15).

Gligorić, K., Anderson, A., & West, R. (2018, June). How constraints affect content: The case of Twitter's switch from 140 to 280 characters. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1).

Glymour, C. (2004). The automation of discovery. *Daedalus*, *133*(1), 69-77. https://doi.org/10.1162/001152604772746710

Goertzel, T. (1994). Belief in conspiracy theories. *Political Psychology*, *15*(4), 731-742. https://doi.org/10.2307/3791630

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40, https://doi.org/10.1146/annurev-soc-071913-043145

Grossman, M., & Tahiri, H. (2015). Community perceptions of radicalisation and violent extremism: An Australian perspective. Journal of Policing, *Intelligence and Counter Terrorism*, *10*(1), 14-24.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, & Roger Tourangeau. (2011). *Survey Methodology*, vol. 561. John Wiley and Sons.

Guidry, J. P., Carlyle, K., Messner, M., & Jin, Y. (2015). On pins and needles: how vaccines are portrayed on Pinterest. *Vaccine*, *33*(39), 5051-5056.

Hale, T., Petherick, A., Phillips, T., & Webster, S. (2020). Variation in government responses to COVID-19. *Blavatnik school of government working paper*, 31, 2020-11.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research*, *53*(6), 820-841.

Harezlak, J., Ruppert, D., & Wand, M. P. (2018). *Semiparametric regression with R*. New York, NY: Springer.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447.

Hassan, G., Brouillette-Alarie, S., Alava, S., Frau-Meigs, D., Lavoie, L., Fetiu, A., ... & Sieckelinck, S. (2018). Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence. *International Journal of Developmental Science*, *12*(1-2), 71-88.

Heuer, H., Hoch, H., Breiter, A., Theocharis, Y. (2021). Auditing the biases enacted by YouTube for political topics in Germany. In *Mensch und Computer 2021* (pp. 456-468).

Horgan, J. (2008). From profiles to pathways and roots to routes: Perspectives from psychology on radicalization into terrorism. The ANNALS of the American Academy of *Political and Social Science*, *618*(1), 80-94.

Hornsey, M. J., & Fielding, K. S. (2017). Attitude roots and Jiu Jitsu persuasion: Understanding and overcoming the motivated rejection of science. *American Psychologist*, *72*(5), 459.

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health Psychology*, 37(4), 307.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 2.

Humphreys, A., & Wang, R. J. H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research*, *44*(6), 1274-1306.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts. Online at: `http://otexts.org/fpp/`

Jasko, K., LaFree, G., & Kruglanski, A. (2017). Quest for significance and violent extremism: The case of domestic radicalization. *Political Psychology*, *38*(5), 815-831.

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in Psychology*, 6, 727.

Jiang, J., Chen, E., Yan, S., Lerman, K., & Ferrara, E. (2020). Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies*, *2*(3), 200-211.

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., ... & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261-265.

Jolley, D., & Douglas, K. M. (2014). The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. *British Journal of Psychology*, *105*(1), 35-56.

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*(8), 459-469.

Jungherr, A. (2018). Normalizing digital trace data. In *Digital discussions* (pp. 9-35). Routledge.

Jürgens, P., & Jungherr, A. (2016). A tutorial for using Twitter data in the social sciences: Data collection, preparation, and analysis. Retrieved from https://dx.doi.org/10.2139/ssrn.2710146

Kata, A. (2010). A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine*, *28*(7), 1709-1716.

Katella, K. (2020). Our new covid-19 vocabulary—what does it all mean. Yale Medicine.

Kay, A. C., Whitson, J. A., Gaucher, D., Galinsky, A. D. (2009). Compensatory control: Achieving order through the mind, our institutions, and the heavens. *Current Directions in Psychological Science*, *18*(5), 264-268.

Kearney, M. W. (2020). TweetBotOrNot: An R package for classifying Twitter accounts as bot or not. Accessed 01 October 2020. https://github.com/mkearney/tweetbotornot

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, *21*(4), 507-525. http://dx.doi.org/10.1037/met0000091

Kerodal, A. G., Freilich, J. D., & Chermak, S. M. (2016). Commitment to extremist ideology: Using factor analysis to move beyond binary measures of extremism. *Studies in Conflict & Terrorism*, *39*(7-8), 687-711.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big data & society*, *1*(1). https://doi.org/10.1177/2053951714528481

Klausen, J., Campion, S., Needle, N., Nguyen, G., & Libretti, R. (2016). Toward a behavioral model of "homegrown" radicalization trajectories. *Studies in Conflict & Terrorism*, *39*(1), 67-83.

Klein, K. J., & Kozlowski, S. W. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*(3), 211-236.

Klein, C., Clutton, P., & Dunn, A. G. (2019). Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PloS one*, *14*(11), https://doi.org/10.1371/journal.pone.0225098

Kou, Y., Gui, X., Chen, Y., & Pine, K. (2017). Conspiracy talk on social media: collective sensemaking during a public health crisis. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1-21.

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods, 21*(4), 493. http://dx.doi.org/10.1037/met0000105

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790. https://doi.org/10.1073/pnas.1320040111

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How

significance quest impacts violent extremism. *Political Psychology*, 35, 69-93.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

LaFree, G., Jensen, M. A., James, P. A., & Safer-Lichtenstein, A. (2018). Correlates of violent political extremism in the United States. *Criminology*, *56*(2), 233-268.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475. `http://dx.doi.org/10.1037/met0000081`

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lahey, M. (2016). Invisible actors: Web application programming interfaces, television, and social media. *Convergence*, *22*(4), 426-439.

Lara-Cabrera, R., González-Pardo, A., & Camacho, D. (2019). Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter. *Future Generation Computer Systems*, *93*, 971-978.

Larsen, R. J., & Kasimatis, M. (1990). Individual differences in entrainment of mood to the weekly calendar. *Journal of Personality and Social Psychology*, *58*(1), 164–171. `https://doi.org/10.1037/0022-3514.58.1.164`

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon landing—therefore,(climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, *24*(5), 622-633.

Lin, Y. R., Margolin, D., & Wen, X. (2017). Tracking and analyzing individual distress following terrorist attacks using social media streams. *Risk analysis*, *37*(8), 1580-1605. `https://doi.org/10.1111/risa.12829`

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*(4), 256-265.

Lösel, F., King, S., Bender, D., & Jugl, I. (2018). Protective factors against extremism and violent radicalization: A systematic review of research. *International Journal of Developmental Science*, *12*(1-2), 89-102.

Loza, W. (2011). The prevalence of the Middle-Eastern extreme ideologies among some Canadians. *Journal of Interpersonal Violence*, *26*(7), 1388-1400.

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2004). Unemployment alters the set point for life satisfaction. *Psychological science*, *15*(1), 8-13.

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports*, *8*(1), 1-9. `https://doi.org/10.1038/s41598-018-25953-0`

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593-614.

Maddock, J., Starbird, K., & Mason, R. M. (2015). Using historical Twitter data for research: Ethical challenges of tweet deletions. In *CSCW 2015 workshop on ethics for studying sociotechnical systems in a Big Data World*. ACM.

Malik, M. M. (2018). *Bias and beyond in digital trace data* (Doctoral dissertation, Carnegie Mellon University). Retrieved from `http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html`.

Mneimneh, Z., Pasek, J., Singh, L., Best, R., Bode, L., Bruch, E., ... & Wojcik, S. (2021). Data Acquisition, Sampling, and Data Preparation Considerations for Quantitative Social Science Research Using Social Media Data. Working Paper.

Marres, N. (2015). Why map issues? On controversy analysis as a digital method. *Science, Technology & Human Values*, 1-32. `http://doi.org/10.1177/0162243915574602`

McCauley, C., & Moskalenko, S. (2017). Understanding political radicalization: The two-pyramids model. *American Psychologist*, *72*(3), 205.

McGlashan, J., Johnstone, M., Creighton, D., de la Haye, K., & Allender, S. (2016). Quantifying a systems map: Network analysis of a childhood obesity causal loop diagram. *Plos One*, *11*(10).

McGregor, I., Hayes, J., & Prentice, M. (2015). Motivation for aggressive religious radicalization: Goal regulation theory and a personality× threat× affordance hypothesis. *Frontiers in Psychology*, *6*(1325), 1-18.

McGilloway, A., Ghosh, P., & Bhui, K. (2015). A systematic review of pathways to and processes associated with radicalization and extremism amongst Muslims in Western societies. *International review of psychiatry*, *27*(1), 39-50.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mitts, T. (2019). From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West. *American Political Science Review*, *113*(1), 173-194.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269. `http:/doi.org//10.7326/0003-4819-151-4-200908180-00135`

Munn, L. (2019). Alt-right pipeline: Individual journeys to extremism online. *First Monday*, *24*(6). `http://doi.org/10.5210/fm.v24i6.10108`

Murphy, S. C. (2017). A hands-on guide to conducting psychological research on Twitter. *Social Psychological and Personality Science*, *8*(4), 396-412. `https://doi.org/10.1177/1948550617697178`

Neumann, P., & Kleinmann, S. (2013). How rigorous is radicalization research? *Democracy and Security*, *9*(4), 360-382.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, *45*(3), 211-236.

Nguyen, D., Smith, N. A., & Rose, C. (2011, June). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language*

technology for cultural heritage, social sciences, and humanities (pp. 115-123).

Nied, A. C., Stewart, L., Spiro, E., & Starbird, K. (2017, February). Alternative narratives of crisis events: Communities and social botnets engaged on social media. In Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 263-266).

Odag, Ö., Leiser, A., & Boehnke, K. (2019). Reviewing the role of the Internet in radicalization processes. *Journal for Deradicalization*, *21*, 261-300.

Oliver, J. E., & Wood, T. J. (2014). Conspiracy theories and the paranoid style (s) of mass opinion. *American Journal of Political Science*, *58*(4), 952-966.

Opp, K. D. (2019). *Analytical Criminology. Integrating Explanations of Crime and Deviant Behavior*. Routledge.

Opp, K. D., & Wippler, R. (Eds.) (1990). Empirischer Theorienvergleich. Erklärungen sozialen Verhaltens in Problemsituationen. [Empirical comparison of theories. Explanations of social behaviour in problem situations] Opladen: Westdeutscher Verlag.

Pape, R.A. (2009) Introduction: What is New About Research on Terrorism, *Security Studies*, *18*(4), 643-650. https://doi.org/10.1080/09636410903369100

Parekh, D., Amarasingam, A., Dawson, L., Ruths, D. (2018). Studying jihadists on social media: A critique of data collection methodologies. *Perspectives on Terrorism*, *12*(3), 5-23.

Pauwels, L. J., & Svensson, R. (2017). How robust is the moderating effect of extremist beliefs on the relationship between self-control and violent extremism?. *Crime & Delinquency*, *63*(8), 1000-1016.

Pawson, R. (2006). *Evidence-based policy: A realist perspective*. Sage.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Pearl, J., Glymour, C., & Jewell, N. P. (2016). Causal inference in statistics: A primer: Wiley.

Pearl, J., & MacKenzie, D. (2018). The book of why. New York: Basic Books.

Pelzer, R. (2018). Policing of terrorism using data from social media. *European Journal for Security Research*, *3*(2), 163-179. http://doi.org/10.1007/s41125-018-0029-9

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* 1532-1543.

Pretus, C., Hamid, N., Sheikh, H., Ginges, J., Tobeña, A., Davis, R., ... & Atran, S. (2018). Neural and behavioral correlates of sacred values and vulnerability to violent extremism. *Frontiers in Psychology*, 9, 2462. https://doi.org/10.3389/fpsyg.2018.02462

Prior, M. (2009). The immensely inflated news audience: Assessing bias in self-reported news exposure. *Public Opinion Quarterly*, *73*(1), 130-143.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media+ Society*, *7*(2), https://doi.org/20563051211019004

Pyrooz, D. C., LaFree, G., Decker, S. H., & James, P. A. (2017). Cut from the same cloth? A comparative study of domestic extremists and gang members in the United States. *Justice Quarterly*. https://doi.org/10.1080/07418825.2017.1311357

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/.

Rauchfleisch, A., & Kaiser, J. (2020). The False positive problem of automatic bot detection in social science research. *PloS one*, *15*(10). https://doi.org/10.1371/journal.pone.0241045

Reganti, A. N., Maheshwari, T., Das, A., Chakraborthy, T., & Kumaraguru, P. (2017). Understanding Psycho-Sociological Vulnerability of ISIS Patronizers in Twitter. Paper presented at the Proceedings of the *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, Sydney, Australia.

Rieger, D., Frischlich, L., & Bente, G. (2017). Propaganda in an insecure, unstructured world: How psychological uncertainty and authoritarian attitudes shape the evaluation of right- wing extremist internet propaganda. *Journal of Deradicalization*, *10*, 203-229.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27-42. https://doi.org/10.1177/2515245917745629

Rowe, M., & Saif, H. (2016). Mining pro-ISIS radicalisation signals from social media users. In Tenth International *AAAI Conference on Web and Social Media*.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Sageman, M. (2014). The stagnation in terrorism research. *Terrorism and Political Violence*, *26*(4), 565-580.

Samory, M., & Mitra, T. (2018, June). Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1).

Scarcella, A., Page, R., & Furtado, V. (2016). Terrorism, radicalisation, extremism, authoritarianism and fundamentalism: A systematic review of the quality and psychometric properties of assessments. *PloS One*, *11*(12).

Schmid, A. P. (2013). Radicalisation, de-radicalisation, counterradicalisation: A conceptual discussion and literature review. *ICCT Research Paper*, *97*(1).

Schmid, A. P., & Forest, J. J. (2018). Research Desiderata: 150 Un-and Under-Researched Topics and Themes in the Field of (Counter-) Terrorism Studies–a New List. *Perspectives on Terrorism*, *12*(4), 68-76.

Schuurman, B. (2018). Research on terrorism, 2007–2016: A review of data, methods, and authorship. *Terrorism and Political Violence*, 1-16.

Seabrook, E. M., Kern, M. L., Fulcher, B. D., & Rickard, N. S. (2018). Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. *Journal of Medical Internet Research*, *20*(5), e9267.

Selivanov, D., & Wang, Q.: text2vec: Modern text mining framework for R. (2018) Retrieved from: `https://CRAN.R-project.org/package=text2vec`

Sen, I., Flöck, F., Weller, K., Weiß, B., Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, *85*(S1), 399-422.

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*, *3*(2), 279-317.

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, *8*(1), 70. `https://doi.org/10.1186/1471-2288-8-70`

Simon, B., Reichert, F., & Grabow, O. (2013). When dual identity becomes a liability: Identity and political radicalism among migrants. *Psychological Science*, *24*(3), 251- 257.

Simpson, G. L. (2018). Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution*, 6, 149. `https://doi.org/10.3389/fevo.2018.00149`

Slootman, M., & Tillie, J. (2006). Processes of radicalisation. Why some Amsterdam Muslims become radicals. Amsterdam: *Institute for Migrations and Ethnic Studies*, University of Amsterdam, Amsterdam, 1-129.

Šrol, J., Mikušková, E.B., & Čavojová, V. (2021). When we are worried, what are we thinking? Anxiety, lack of control, and conspiracy beliefs amidst the COVID-19 pandemic. *Applied Cognitive Psychology*, *35*(3), 720-729. `https://doi.org/10.1002/acp.3798`

Starbird, K. (2017, May). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1).

Stempel, C., Hargrove, T., & Stempel III, G. H. (2007). Media use, social structure, and belief in 9/11 conspiracy theories. *Journalism & Mass Communication Quarterly*, *84*(2), 353-372.

Stern, J. (2016). Radicalization to extremism and mobilization to violence: What have we learned and what can we do about it? *The ANNALS of the American Academy of Political and Social Science*, *668*(1), 102-117.

Stoltz, D. S., & Taylor, M. A. (2019). Concept Mover's Distance: measuring concept engagement via word embeddings in texts. *Journal of Computational Social Science*, *2*(2), 293-313.

Swami, V., Coles, R., Stieger, S., Pietschnig, J., Furnham, A., Rehim, S., & Voracek, M. (2011). Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. *British Journal of Psychology*, *102*(3), 443-463.

Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition*, *133*(3), 572-585.

Sweeney, M. M., & Perliger, A. (2018). Explaining the spontaneous nature of far-right violence in the United States. *Perspectives on Terrorism*, *12*(6), 52-71.

Taylor, W. D., Johnson, G., Ault, M. K., Griffith, J. A., Rozzell, B., Connelly, S., ... & Ness, A. M. (2015). Ideological group persuasion: A within-person study of how violence, interactivity, and credibility features influence online persuasion. *Computers in Human Behavior*, *51*, 448-460.

Tranfield, D., Denyer, D., & Palminder, S. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*, 207-222.

Tufekci, Z. (2014, May). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media*.

Vahratian, A., Siega-Riz, A. M., Savitz, D. A., & Zhang, J. (2005). Maternal prepregnancy overweight and obesity and the risk of caesarean delivery in nulliparous women. *Annals of Epidemiology*, *15*(7), 467-474. `https://doi.org/10.1016/j.annepidem.2005.02.005`

Vanderweele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 3, 211-219. `https://doi.org/10.1007/s10654-019-00494-6`

Van De Wijngaert, L., Bouwman, H., & Contractor, N. (2014). A network approach toward literature review. *Quality & Quantity*, *48*(2), 623-643.

van Mulukom, V., Pummerer, L., Alper, S., Cavojova, V., Farias, J. E. M., Kay, C. S., ... & Zezelj, I. (2020). Antecedents and consequences of COVID-19 conspiracy theories: a rapid review of the evidence. Preprint at: `https://psyarxiv.com/u8yah/`

van Prooijen, J. W., & Van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on Psychological Science*, *13*(6), 770-788. `https://doi.org/10.1177/1745691618774270`

van Prooijen, J. W., Krouwel, A. P., & Pollet, T. V. (2015). Political extremism predicts belief in conspiracy theories. *Social Psychological and Personality Science*, *6*(5), 570-578.

Vergani, M., Iqbal, M., Ilbahar, E., & Barton, G. (2018). The three Ps of radicalization: Push, pull and personal. A systematic scoping review of the scientific evidence about radicalization into violent extremism. *Studies in Conflict & Terrorism*, 1-32.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, 48, 865-885.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8): Cambridge University Press.

Weller, K., & Kinder-Kurlanda, K. E. (2016, May). A manifesto for data sharing in social media research. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 166-172). http://dx.doi.org/10.1145/2908131.2908172

Wojcik, S., & Hughes, A. (2019). Sizing up Twitter users. Pew Research Center, 24. Retrieved from: https://www.pewinternet.org/2019/04/24/sizing-up-twitter-users/

Wolfowicz, M., Litmanovitz, Y., Weisburd, D., & Hasisi, B. (2019). A field-wide systematic review and meta-analysis of putative risk and protective factors for radicalization outcomes. *Journal of Quantitative Criminology*, *36*(3), 407-447.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social Psychological and Personality Science*, *3*(6), 767-773.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Zeileis, A., & Kleiber, C. (2005). Validating multiple structural change models–A case study. *Journal of Applied Econometrics*, *20*(5), 685-690.

Zhirkov, K., Verkuyten, M., & Weesie, J. (2014). Perceptions of world politics and support for terrorism among Muslims: Evidence from Muslim countries and Western Europe. *Conflict Management and Peace Science*, *31*(5), 481-501.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe und die wörtlich oder dem Inhalt nach aus fremden Arbeiten entnommenen Stellen als solche kenntlich gemacht sind. Ferner versichere ich, dass ich die gleiche Arbeit noch nicht für eine andere wissenschaftliche Prüfung eingereicht und mit der gleichen Abhandlung weder bereits einen Doktorgrad erworben noch einen Doktorgrad zu erwerben versucht habe.

Ort, Datum	Unterschrift