Data Quality of Digital Process Data: A Generalized Framework and Simulation/Post-Hoc Identification Strategy

Schmitz, Andreas; Riebling, Jan R.

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 491156185 / Funded by the German Research Foundation (DFG) - Project number 491156185

Empfohlene Zitierung / Suggested Citation:
Schmitz, A., & Riebling, J. R. (2022). Data Quality of Digital Process Data: A Generalized Framework and Simulation/ Post-Hoc Identification Strategy. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *74*(Supplement Issue 1), 407-430. https://doi.org/10.1007/s11577-022-00840-9

Köln Z Soziol (2022) (Suppl 1) 74:407–430
https://doi.org/10.1007/s11577-022-00840-9

Kölner Zeitschrift für Soziologie und Sozialpsychologie

ABHANDLUNGEN

Data Quality of Digital Process Data

A Generalized Framework and Simulation/Post-Hoc Identification Strategy

Andreas Schmitz · Jan R. Riebling

Received: 28 January 2022 / Accepted: 17 March 2022 / Published online: 20 May 2022

Abstract Digital process data are becoming increasingly important for social science research, but their quality has been gravely neglected so far. In this article, we adopt a process perspective and argue that data extracted from socio-technical systems are, in principle, subject to the same error-inducing mechanisms as traditional forms of social science data, namely biases that arise before their acquisition (observational design), during their acquisition (data generation), and after their acquisition (data processing). As the lack of access and insight into the actual processes of data production renders key traditional mechanisms of quality assurance largely impossible, it is essential to identify data quality problems in the data available—that is, to focus on the possibilities post-hoc quality assessment offers to us. We advance a post-hoc strategy of data quality assurance, integrating simulation and explorative identification techniques. As a use case, we illustrate this approach with the example of bot activity and the effects this phenomenon can have on digital process data. First, we employ agent-based modelling to simulate datasets containing these data problems. Subsequently, we demonstrate the possibilities and challenges of post-hoc control by mobilizing geometric data analysis, an exemplary technique for identifying data quality issues.

Online Appendix: https://kzfss.uni-koeln.de/sites/kzfss/pdf/Schmitz_Riebling.pdf

A. Schmitz
Department for Computational Social Sciences, GESIS – Leibniz-Institut für Sozialwissenschaften e.V.
Mannheim, Germany

Institut für Politische Wissenschaft und Soziologie, Abteilung Soziologie, Rheinische Friedrich-Wilhelms-Universität Bonn
Lennéstr. 27, 53113 Bonn, Germany
E-Mail: andreas.schmitz@uni-bonn.de

J. R. Riebling
Fakultät für Human- und Sozialwissenschaften, Bergische Universität Wuppertal
Gaußstr. 20, 42119 Wuppertal, Germany
E-Mail: riebling@uni-wuppertal.de

Keywords Digital process data · Data quality · Agent-based simulations · Relational data · Post-hoc identification · Mixed methods · Socio-technical systems

Datenqualität digitaler Prozessdaten

Ein generalisierter Orientierungsrahmen und eine Simulations-/Post-Hoc-Identifizierungsstrategie

Zusammenfassung Digitale Prozessdaten werden für die sozialwissenschaftliche Forschung immer wichtiger, doch ihre Qualität wurde in der Diskussion bisher stark vernachlässigt. In diesem Beitrag nehmen wir eine Prozessperspektive ein und argumentieren, dass Daten, die aus soziotechnischen Systemen extrahiert werden, im Prinzip denselben fehlerverursachenden Mechanismen unterliegen wie traditionelle Formen sozialwissenschaftlicher Daten, nämlich Verzerrungen, die vor ihrer Erfassung (Beobachtungsdesign), während ihrer Erfassung (Datengenerierung) und nach ihrer Erfassung (Datenverarbeitung) entstehen. Da der fehlende Zugang und Einblick in die eigentlichen Prozesse der Datenproduktion wichtige Mechanismen der traditionellen Qualitätssicherung weitgehend unmöglich machen, ist es unerlässlich, Datenqualitätsprobleme in den zur Verfügung stehenden Daten zu identifizieren – das heißt, sich auf die Möglichkeiten zu konzentrieren, die uns die post-hoc Qualitätsprüfung bietet. Wir entwickeln eine Post-hoc-Strategie der Datenqualitätssicherung, die Simulation und explorative Identifizierungstechniken integriert. Als Anwendungsfall illustrieren wir diesen Ansatz am Beispiel von Bot-Aktivitäten und den Auswirkungen, die dieses Phänomen auf digitale Prozessdaten haben kann. Dazu setzen wir zunächst eine agentenbasierte Modellierung ein, um Datensätze mit derartigen Datenproblemen zu simulieren. Anschließend demonstrieren wir die Möglichkeiten und Herausforderungen der Post-hoc-Kontrolle, indem wir die geometrische Datenanalyse einsetzen, eine exemplarische Technik zur Identifizierung von Datenqualitätsproblemen.

Schlüsselwörter Digitale Prozessdaten · Datenqualität · Agentenbasierte Simulationen · Relationale Daten · Post-hoc-Identifikation · Mixed-Methods · Sozio-technische Systeme

1 Introduction

Data generated on a massive scale and recorded automatically within digital contexts is increasingly forming the basis for contemporary social science research. One reason for this shift is that the object of study itself is transforming before our eyes: Social practices now take place to a great extent in digital spheres and social fields are extensively digitized. Depending on the viewpoint of the respective paradigm, digital environments allow people to perform everyday practices, to realize choices, to represent themselves, or to interact based on symbols, which makes data from digital contexts interesting for a variety of contemporary social scientists. Another reason for the enduring scientific popularity of these forms of data lies in the methodological advantages often attributed to them when compared with traditional data types, in

particular objectivity, unobtrusiveness, unbiasedness, reliability, and so on (see Boyd and Crawford 2012, p. 663). As a result, however, the actual quality of digital process data is a rather neglected topic in the current discourse.

Yet, in the context of data generated by modern socio-technical systems[1] (Dolata 2009; Riebling 2018), there are in fact quite serious problems regarding data quality. These problems include phenomena of bias, selectivity, erroneous aggregation, and recursive effects of the research instrument, in short, a whole series of abstract mechanisms that are anything but unfamiliar to social scientists. As with written surveys, digital data are equally capable of containing erroneous and biased information (see Japec et al. 2015, p. 854 f.; Sen et al. 2019; Baur et al. 2020; Diaz-Bone et al. 2020). Although in the context of written or online surveys, it is usually considered essential both to ensure data quality in the process of data generation and to correct potential quality problems, the systematic conceptualization of data quality for digital process data is still in its infancy. Even if numerous individual problems in data quality are known, or at least suspected, to exist in specific research settings, there is still a lack of a systematic understanding of the different quality-distorting mechanisms in the context of digital process data. This systematic deficiency occludes from our view the potentially considerable distortions of substantive findings and may, ultimately, threaten the legitimacy of research based on the analysis of digital process data. This is why the German Research Foundation (DFG 2020) recently emphasized the increasing importance of digital data and the need for quality assurance of these data.

In order to be able to deal with this major challenge of present and future research, social science requires two essential tools, which it in principle already has at its disposal: First, a systematic overview of abstract error mechanisms, and second, ways of inferring quality problems in existing data sets. For the first component, we delineate a systematic process perspective on error-inducing mechanisms of digital process data along three ideal-typical dimensions: *Observational design, data generation*, and *data processing*. For the latter component, we outline a mixed methods strategy of post-hoc quality control using a combination of simulation models and statistical identification techniques. Simulations are particularly promising, in that they allow researchers to (i) systematically control the data-generating process under different contextual conditions, (ii) model the various hypothetical error-generating mechanisms, and (iii) screen the resulting data using explorative techniques to determine what they can and cannot tell us about implemented quality impairments. As a practical use case, we will discuss a fundamental and far-reaching issue: The activities of non-human actors (bots). Bots account for the majority of global web traffic and bots of various types are active on most social media platforms, sometimes significantly affecting data structures, which thus poses one of the key challenges for modern quantitative social research and computational socioeconomics (see Gao

[1] Whereas the term "digital trace data" implies a rather objectivist view, the term "digital process data" emphasizes the institutional and technical preconditions and implications of data generation. The relational term "socio-technical systems" is intended to take into account the fact that usage practices and infrastructures impact on each other and jointly generate data that represent neither pure objective measurements nor mere technical artifacts (see also Häußling 2020).

et al. 2019, p. 94). For analyses based on the assumption that human actors are the entities to be observed, the presence of bots calls into question any supposedly unbiased "measurements"; equally, for approaches that conceive of bots as genuine elements of socio-technical systems, their identification is crucial too (see Venturini and Latour 2010, p. 8). Using agent-based simulations, we generate a series of artificial data sets including distortion effects from bots in systematic and controlled ways. Subsequently, we demonstrate the possibilities and challenges of post-hoc control by mobilizing geometric data analysis, an established exemplary technique for identifying issues with data quality. In the conclusion, we discuss how the generalized data quality framework can inform further research and what contribution our proposed combination of simulation and statistical screening can make as part of a much-needed multi-paradigmatic and mixed-method discourse on the quality of digital process data.

2 Conceptualizing Quality Issues in Digital Process Data

In recent years, more and more researchers have noticed that social sciences' established quality concepts, such as coverage error and non-response, can also be applied and translated to digital process data (see Diaz et al. 2016, p. 3). Japec et al. (2015, p. 851) relate the concept of total survey error (see Biemer 2010) to digital process data in order to conceptualize "big data total error" (BDTE). Sen et al. (2019) have developed a total error framework for digital trace data inspired by traditional error-generating processes. Thus, after a phase of focusing on the apparent novelties and unique features of digital process data, it is being increasingly recognized that the prevailing quality problems in this context are in fact similar to the problems familiar from (survey) research.

To derive a systematic overview of the possible phenomena, sources, and mechanisms of errors, we employ a process perspective. We understand data production as processes emerging from the genuine interplay of social and technological entities. The systematic conception of the data-generating process and its accompanying errors have been examined both in the context of a process-oriented theory of survey research (Bachleitner et al. 2010) and as a "statistical chain"—that is, as a relational interplay in which different entities, objects, practices, and situations jointly generate data (see Desrosières 2009; Diaz-Bone 2018; Diaz-Bone et al. 2020, p. 319). Problems with data quality (as well as adequate interpretability) arise from the inconsistency of conventions between the different links in the chain of data production. Interviewers, data managers, statisticians, and recipients will differ in their data-related knowledge, definitions, implicit assumptions, practical choices, and their conceptions about the (realist or constructivist) status of the data and its constructs. On this analytical basis, successful attempts have been made, for survey data, to trace the process of data production from start to finish, to theoretically grasp the mechanisms of distortion, and to thereby make them accessible for investigation and, eventually, correction.

Expanding on this analytical tradition, we establish a generalized, ideal-typical model of error mechanisms in digital process data along three analytically separated

(yet empirically interacting) dimensions: *Observational design, data generation*, and *data processing*. *Observational design* describes the processes that precede the actual collection of data: The design of the architecture, programming rules (including the specification of data structures and information flow), conventions of handling information and events, and thus ultimately the arrangement of the social environment with and within which users interact and produce data. *Data generation* addresses the actual process of data gathering and thus the interaction processes between the users themselves and with the technical infrastructure. *Data processing* refers to the techniques and conventions of data handling employed by social scientists, i.e., practices such as restructuring, editing, and statistically analyzing data.

2.1 Observational Design

Analytically, the first step of quantitative data collection is a priori construction, i.e., designing the architecture that will collect the data. In the various lines of social science's tradition, systematic, controlled, and reflective access to data has been emphasized as an essential prerequisite for ensuring the validity and robustness of findings. To this end, social scientists are, ideally, substantially involved in defining the research question, constructing the survey instrument, and conducting it in the field. This allows the survey process to be controlled, and quality-reducing influences to be anticipated, identified, and managed. In this context, it is essential to enable the accurate reconstruction of problems and their consequences through careful documentation of the survey, the survey process, and the subsequent data management. Therefore, data quality is essentially associated with process control (Lyberg and Biemer, 2008). The problems of division of labor in the process of data production and usage are well known from the organization of standardized social research (Desrosières 2009; see Diaz-Bone 2018). The division of labor in the production process of data can lead to the data being used at the end of the production chain (e.g., in secondary analyses) without knowledge of the underlying conventions (Desrosières 2001b; see Diaz-Bone 2016) of data production and data management. In traditional survey research, data problems and the conventions of their treatment can be considered as rather congruent, at least from the perspective of the survey institute and the evaluating researchers.

In the context of digital process data, however, a particularly profound division of labor between providers, computer scientists, and social scientists can be said to prevail. In contrast to survey research, the organization of the data infrastructure on the part of a producer of digital data is considerably less oriented toward scientific interests, instead following considerations of economy or efficiency (see Schmitz et al. 2009). The fact that the data may also be used for scientific purposes is a subordinate criterion and genuinely scientific quality criteria are—according to this logic—often irrelevant (if not interfering with own quality criteria such as being fit for use). As a result, the opacity or impenetrability of the statistical chain can be said to be particularly pronounced in situations of this kind and the researcher's insight to be particularly limited (see Diaz-Bone et al. 2020, p. 324).

This problem is already evident in the elementary aspect of the definition of the units: Desrosières shows that, early in the history of statistics, the assumption

of a statistical equivalence space with clearly definable and identifiable units was a necessary prerequisite for subsequent analyses. In order to study the differences between units, we must first assume their unity (Desrosières 2001a). In past decades, in the context of questionnaire research (which made use of the contributions of sampling theory) it was comparatively easy to straightforwardly treat sampled actors as a statistical unit. However, according to what rules is the unit constructed in the context of digital process data? Which events and processes are attributed to an actor, and according to which rules? The scientist should know exactly how entities are defined a priori, what is actually treated (that is, kept) as a valid entity, and the principles on which they are collected in the data-generating process. The problem of unknown design decisions also manifests in the ways in which different users are handled differently, for example, in the form of offers that can be distributed differently with reference to time, specific user groups, or access type, such as when access from one specific residential area (as opposed to another) or via a smartphone (as opposed to a PC) results in a different offer in terms of price (see Morstatter et al. 2013).

Another example for design principles one would need to know are pre-structured sets of choices and interactions that can compromise the validity of interactional and network analyses as well as of constructs based thereon: Recommender systems such as those used by Netflix or Amazon that direct the users' attention by latent system-immanent choices as part of their customer retention strategy, and other algorithms implemented on similar platforms, pre-structure interactions and thereby shape the observation of interactional processes. For example, Twitter uses network indicators that define a position in the timeline, which makes the application of network analyses problematic; they do not "measure" the actual communication structure, but also the communication as structurally triggered by Twitter itself. The effects of suggestions on matching-based platforms suffer from a comparable problem (see Malik and Pfeffer 2016). Principles of data aggregation represent another example of a priori design decisions that can distort network analyses. Unknown pre-defined aggregation rules make it difficult to reconstruct original data structures based on end-user data (for example, network relationships (see Howison et al. 2011, p. 781)). One might think here of Twitter's mostly opaque agglomeration of retweets: User A retweets a post by C via user B's timeline. In the final data set, however, it looks as if A has referred directly to B. This can be particularly problematic if conclusions are to be formulated regarding the extent of the polarization of Twitter users, for example. Beyond that, the temporal mode of data storage can distort the processual structure of data and its temporal granularities. For example, several researchers who analyze Reddit data rely on the pushshift archive, which does not have a consistent way of keeping entries up to date. Events that may in fact be one minute apart in real time, can be stored and misrepresented in the data as co-occurring within one second. Similarly, Facebook's CrowdTangle stores time-series information about the evolution of likes for a post simply on an hourly basis.

As socio-technical systems are usually not embedded in scientific fields, problems of this kind are further exacerbated, as these systems are subject to constant (economic, political, legal, etc.) adaptation pressures. For example, with changing business strategies, the architecture of a platform can be repeatedly altered. Such

reorganizations of design can lead to significant transformations of data structures over time, thereby inducing structural breaks in the data (Think of YouTube's recent reorganization of how one can react to a video: today, only thumbs up and no longer thumbs down can be given). Design changes in the choice and interaction options offered to users can lead to far-reaching problems in the use of data. For instance, if profiles on platforms are redesigned, and new categories (e.g., additional gender categories) are introduced, interaction practices change, and so does the data structure. Although this can be compared with survey research, where panel studies can include modified item blocks, more radical changes can be introduced in digital environments, such as a redesign of recommender systems, which becomes the more problematic (e.g., for construct reliability) the less potential data users are aware of such reorganizations.

Yet, when compared with (institutionally) collected survey data, sociologists are much less likely to be involved in the conceptualization, design, and planning phases of digital process data. Private companies such as social media platforms provide such data without the detailed documentation that would correspond to the information necessary for social scientists. In fact, because a private company is unlikely to have any interest in making these principles public, this can be seen as the default situation in large parts of current research. Thus, today, the assurance of process quality, traditionally so important in empirical research, is systematically hampered by a fundamental division of labor between private-sector providers, computer scientists, and sociologists (see Diaz-Bone et al. 2020, p. 324). Consequently, in the context of socio-technical systems, the procedures and conventions underlying data generation and data structures are not transparent to the scientist for the most part. A lack of understanding of the underlying conventions concerning the definition of data frames and the predefined data recording rules can lead to severe restrictions in the extent to which the data can be meaningfully used. If researchers are not informed about design-specific conditions, *a substantial lack of knowledge about the underlying observational design will engender the risk of misjudging the phenomena observed* (or to misconstrue opaquely constructed data as neutral, valid "measurement").

2.2 Data Generation

After the construction phase and the design of an observational instrument, analytically, the phase of actual data generation follows, which is traditionally referred to as the "field phase".[2] Here, apart from technically flawed observation and recording processes in the narrower sense, several biasing phenomena can occur.

Again, the first issue—representativeness—is well known from traditional social sciences (see Baur et al. 2020). As a consequence of a social media company's design principles, the population (of actors and events) under analysis in a digital context can represent a distorted image of the overall population, which is reminiscent of the coverage error and sampling error discussed in survey research (see

[2] It goes without saying that—depending on the paradigmatic orientation—there are several recursive steps.

Sen et al. 2019, p. 5). Further aspects such as the question of which client is used can impact representativeness, as some clients make it possible to avoid tracking; consequently, specific events of specific users remain unobserved, or are not connected with previous information (comparable with the differences between different modes of survey delivery such as landline vs. mobile). Human users may also generate distorted patterns whenever they do not actually constitute a single unit in the resulting data. In abstract terms, this reminds us of the situation where a third party influences a respondent during an interview, so that the resulting interview contains different patterns of response. This can be particularly problematic in the digital context, where different users can share an account (such as in streaming services) or one individual can create and operate a large number of user profiles (such as in dating services).

When it comes to the processes taking place in a socio-technical system, it is possible that entities that may not be defined as elements of the population actually become part of the data (similar to distorted selection frames in survey research). For example, bots can be designed to defraud or manipulate, or they can be used by the operators of the platform itself. Here, the problem does not end in the mere number of such artificial actors. Bot-generated events (such as contact patterns) represent even more problematic distortions for substantive analyses. For example, Schmitz et al. (2012) show that bots on a German dating platform make up less than 3% of all entities, but produce up to 33% of all first contacts and—by simultaneously exhibiting a high level of attractiveness and yet little selectivity—thereby establish a contact pattern in the data that diverges from human mating patterns, thus biasing statistical analyses of contact practices. Likewise, ideologically motivated third parties may try to exert external influence on digital interactions and communication contexts via bots or trolls (see Bratu 2017; Bulut and Yörük 2017; Starbird 2019).

To further complicate matters, digital process data are characterized by the fact that they are generated in a context that can have a decisive influence on the actors' practices—for example, by defining and sorting their interaction options. In the social science literature, this is referred to as the obtrusiveness of collection methods (Webb et al. 1966). From survey research, it is well-known that the influence of the data-generating context is particularly problematic when different mode effects occur, such as in a survey that is realized by a computer-assisted questionnaire on the one hand and a written one on the other (Shin et al. 2012). Mode effects can distort digital process data, too, such as when different users are treated differently by an algorithm or by moderators in commercial algorithmic systems. As a consequence, a platform's users and their practices can be subjected to different representations of the socio-technical environment (e.g., owing to client strategies), resulting in their actions and interactions being affected by different processes or modes. For example, YouTube recommendations are provided depending on a user's history, as the work of Faddoul et al. (2020) shows for exposure to conspiracy theories.

Conversely, in the context of digital interaction and communication, the observed units of inquiry can react substantially to the socio-technical system, a circumstance that social science has been referring to as reactivity. In addition to and in conjunction with the designers' influences, the users' reactions are of particular importance when it comes to potential issues with data quality. Similar to the way different respondents

react differently to the same survey instrument, thereby revealing different "response styles" (van Vaerenbergh and Thomas 2013), different usage styles can influence the content of data derived from digital interaction contexts (see Olteanu et al. 2019). Take the example of those Twitter users who are particularly opinionated about a topic and post long threads that span more than ten items. For the researcher, for example, ten separate tweets on one topic may result in that topic being perceived as ten times more frequent than an antagonistic statement that may be expressed in just one tweet.

Reactivity may also manifest as strategic reaction towards the technology; for example, privacy concerns can induce self-censorship and practices of subtweeting, or mock-retweeting (see Tufekci 2014). A particularly severe problem can result from the users' (legitimate) control over the data that they created in the past. For different reasons, some users may delete posts or accounts, thus creating gaps in the data that are difficult to track and can lead to misinterpretation of content. In socio-technical systems, strategies emerge in yet another respect: For surveys, it is known that respondents sometimes exhibit strategic response behavior by orienting themselves to norms of social desirability, e.g., in the social situation of a face-to-face interview (see Blasius and Thiessen 2012). Such strategic behaviors are disproportionately more likely to occur in digital interaction contexts. In contrast to surveys, relational, digital process data are not created in a context in which actors are socially independent of each other—instead, the actors observed strongly influence each other. Rather, they find themselves in a genuine social situation vis-à-vis other actors, thus reciprocally inducing strategic practices (e.g., on dating platforms, see Zillmann et al. 2011). Social desirability can manifest in the form of strategic postings with socially desirable content, which may lead to clickbait strategies. These strategies are part of the digital attention economy, with users of social platforms adjusting their publicly posted metrics in order to be placed higher in lists (referred to colloquially as "playing the algorithm game"). Other users respond reflexively to these metrics that are presented to them. On Twitter, for example, likes, retweets, and displayed "ratios" become the basis for reflexive practices that cannot be understood in the resulting data without understanding their context.

On top of that, in socio-technical systems, the provider can react to the users' behavior, in a further recursion. For example, provider-side handling of unwanted content can involve the complete removal of specific users' actions and communication acts from the database, e.g., when providers identify bot activities. But there are more subtle techniques that providers apply, such as "shadow banning," where users are not banned per se, but their output is made invisible; as such, the opportunities for being perceived are severely limited (or "throttled"). If, however, the probability of a user's actions being seen by other users is artificially decreased, their actions are not adequately represented in the resulting data. Such practices of the "silent truncation of data" have been observed, for example, in cases such as Sourceforge and Wikipedia dumps (see Howison et al. 2011). Although the problem of missing data is a familiar phenomenon (here reminiscent of item non-response and unit-nonresponse from survey tradition), the problem is even more far-reaching for digital interaction data. Within the relational environments of socio-technical systems, data that have been deleted may have elicited reactions (e.g., responses) prior to the

deletion. Such relations, however, cannot be observed and the original occasion of a response remains unknown.

When it comes to data dissemination, missing and biased data—again, familiar from written surveys—can result from providers being interested in only making available certain specific subsamples (of entities, events and their relations) for strategic reasons. Just as survey institutes can provide users with selective samples, selective provision (temporally, spatially, socio-structurally, etc.) of data excerpts will often result in a bias, especially because data that contain relational processes is not suitable for random samples (see Morstatter et al. 2013; Driscoll and Walker 2014; González-Bailón et al. 2014). Using Facebook as an example, Allen et al. (2021) show that censoring URLs with fewer than 100 public shares results in a biased data set that will overestimate the share of fake news by a factor of 4.

In sum, the problem identified in the context of the definition of observational design is perpetuated. More often than not, the actual data generation process—and the potential issues of data quality entailed—remain unknown to the social scientist (see Baur et al. 2020, p. 220; Diaz-Bone et al. 2020, p. 324). But beyond that, the advantages of digital process data environments, namely the fact that actors proactively access them and perform their practices there (in contrast to surveys), and that recording is highly standardized and automated, go hand in hand with systematic challenges for data quality. Cultural norms and the architecture of platforms can influence user behavior, and vice versa, which may result in severe "measurement" errors (see Malik and Pfeffer 2016), a problem that has also been labelled as "platform affordances error" (see Sen et al. 2019, p. 8). Consequently, the objectivity sometimes attributed to digital process data when compared with survey data is called into question by the *complex reactivity relations and the diverse entities involved* that underlie socio-technical systems (and that are—again—hardly known to the end-user).

2.3 Data Processing

From an analytical perspective, quantitative data can be understood to be further processed after they have been collected. Although the production of digital data is characterized by a specific back and forth between data generation and data processing, it is still true that all kinds of (intermediate) data sets are handled, transformed, and processed in various ways, which can lead to errors that were not originally present. For survey research, we already know that the providers' (here: Survey institutes) decisions regarding data processing can be unsatisfactory for the end users. For example, an institute's technocratic quality conventions, which lead to specific respondents being sorted out of the final data, can render specific ideological positions invisible (Barth and Schmitz 2018). In the survey tradition, such mistakes subsequent to data collection have been referred to as "processing errors" (see Groves and Lyberg 2010). Analogously, providers of digital data can also compromise the quality of their data through faulty processing, such as biasing transformations of signals and entities (see Sen et al. 2019, p. 7 ff.). Japec et al. (2015, p. 854 f.) discuss the different steps of (digital) data processing, in which errors can be generated via "creating or enhancing meta-data," "record matching," "variable coding," "editing,"

"data munging (or scrubbing)," or "data integration" (linking records across disparate systems). Ideally, a researcher (whose particular methodological perspective and research interest will determine the *specific conception of data quality*) should be informed about the quality conventions according to which a company processes data internally. For example, the criteria of data cleansing may vary and so does, accordingly, which entities and events are considered to be irregular (e.g., bots) and are removed from the system and thus from our observation. Errors can also be caused by providers when aggregating, selecting, and reducing the data to be forwarded (Hellerstein (2008, p. 2) refers to this as "distillation error"), whereas the end user assumes the data were of high quality and unbiased. Of course, the end users themselves can also introduce numerous errors in any kind of data through faulty data management, e.g., through the incorrect identification of valid (or invalid) cases or incorrect recoding. In the context of digital process data, there is a more systematic problem that arises from the considerable distance between the links of the statistical chain described above, i.e., between provider and end user (see Diaz-Bone et al. 2020, p. 324). End users do not know the rules according to which data structures have been created during and processed after collection. Yet not knowing provider-based conventions of data processing (e.g., classifications, aggregations, transformations, etc.) can mislead the end user into wrongly assuming that the data structure thus obtained represents an adequate image of the original, underlying data structures and processes (see Venturini and Latour 2010, p. 8; Japec et al. 2015, p. 853). For example, entities can be "concealed in logs" (Van der Aalst 2016, p. 148).

Another more general phenomenon is the example of divergent definitions of units. Working from the assumption of a classical, individual-centered epistemology, social scientists sometimes still request or arrange the data in an actor-centric manner, i.e., as a two-dimensional flat file with human actors as the sole type of entity. However, this may be problematic when the original data structure is non-linear, e.g., relational, and contains multiple entities and interrelations, such as communicating parties and their communication acts nested in multiple and reciprocal hierarchies (see Sen et al. 2019, p. 6; Diaz-Bone et al. 2020, p. 334)[3].

Thus, in light of the problems that will occur during this process of data production, any statistical analysis based on data from digital sources can face severe difficulties (see Olteanu et al. 2019, p. 16 f.; Japec et al. 2015, p. 854 f.). Biases may appear in many different situations, for example, in univariate distributions, descriptive parameters, estimates of regression parameters, or network parameters. In sum, when misconstruing (and thus *misconstructing*) actually underlying complex data relations, inappropriate data management and data analysis conventions can lead to situations where *statistical models and substantive conclusions that draw on digital process data can be severely impaired.*

[3] Temporal relationality, a genuine trait of socio-technical systems, exacerbates this problem to a great extent.

3 Simulation and Post-Hoc Identification

To summarize up to this point: The quality of digital process data can be distorted in different, often unknown ways. Yet what we do know is that digital process data—much vaunted by some for its methodological virtues—are in fact haunted by issues that remind us of classic problems of data generation and data processing, such as selectivity, bias, validity, reliability, objectivity, etc. As outlined in the preceding section, a core difference between survey data and digital process data is that, for the latter, there is scarcely any comprehensive documentation for scientific end users, and scientists have little knowledge of, let alone control over, construction and collection processes. Under such conditions, however, both a priori control and process control of data quality become virtually impossible tasks and the adequacy of statistical analyses based on them must be fundamentally called into question.

For this reason, it is essential to identify data quality problems in the data available—that is, to focus on the possibilities that post-hoc quality assessment offers us. In fact, in addition to the goal of high-quality survey methodology and systematic process control, social scientists have traditionally conducted post-hoc analyses to develop hypotheses about possibly biasing phenomena. Employing descriptive statistics, and exploratory and visualization techniques such as cluster analyses (Bredl et al. 2012), classification models (Biemer 2010), or scaling approaches (Blasius and Thiessen 2015), scientists have been able to accumulate a great deal of insight by examining countless empirical data sets from different contexts. The practical knowledge gained in dealing with empirical contexts and the general phenomena of error mechanisms has proven instructive in asking the right questions of the data and using the statistical models in meaningful ways. Still, the problem of post-hoc control is that the actual error-generating mechanisms cannot be verified, but only assumed to be plausible. Whether, for example, social desirability actually guides actions in a questionnaire is ultimately an educated guess. This is why researchers have also used qualitative (cognitive) interviews with respondents to determine their perceptions of a survey.

Yet, in order to consolidate and systematize the insights gained from empirical data sets, a further methodological tool has proven to be of value in the context of survey data: Simulation techniques have been employed in order to systematically supplement the body of empirical experience. Some authors have worked with simulations that enable a more systematic understanding of data-generating (and error-generating) processes, and of what problems (and of what magnitude) are to be expected under specifiable conditions (Dijkstra et al. 1995; West 2013; McCarthy et al. 2017). Overall, in the tradition of survey research, the abductive interplay of human experience, empirical data, simulated mechanisms, and theory building has enabled social scientists to collectively achieve considerable advances in making quality problems more amenable to regulation.

In contrast to the survey tradition, however, there are not yet as many generally available empirical data sets for digital process data, and there is still a lack of practical experience, systematic comparability between contexts, and theoretical knowledge. Taking our argumentation so far into consideration, one would need to know how different kinds of biasing mechanisms are manifested in the highly

aggregated and selective digital data that are provided to end users from the field of social sciences. In order to achieve such systematic knowledge, we propose the implementation of simulation techniques to model the process of data production, including possible biases, and to thereby yield data sets where the error-generating mechanisms are known. Simulation techniques are particularly promising for the objective at hand, as they allow the otherwise merely hypothesized mechanisms of quality impairment to be generated and examined in a controlled manner, which is impossible for any kind of empirically collected data (unless we are dealing with experimental approaches). However, it is important to note that simulation models are not aimed at being complex and realistic. In fact, the majority of phenomena and effects that are known to be present must be consciously kept out of the model, with a few others added parsimoniously and incrementally. It is precisely this controlled approach that underlies the potential of simulations to systematically enrich the stock of empirical experience with a series of artificial datasets.

Simulation models make it possible to recreate the different ways in which the *observational design* underlying socio-technical systems influence the generation of data. Techniques such as agent-based simulation models (see Jun and Sethi 2008; Macal and North 2009) can simulate both the defined forms of permissible interactions between users and the principles of hierarchizing displayed information, as well as another essential structural feature of socio-technical systems: Mutual reactivity, and thus the relationships between interacting entities. These aspects can be implemented by choosing specific topologies in agent-based simulations, and thus the rules that govern, enable, and restrict interaction and communication. In contrast to a post-facto study of the quality of data, simulation studies enable the systematic exploration of elements of the observational design that would not be observable otherwise. This is crucial in situations where the scientific community is scarcely involved in the construction of the data-generating architecture. Likewise, in the actual *data generation* process, simulation models can take into account the myriad ways in which interaction in socio-technical systems can actually occur; this represents a powerful way to control and to partition out the concrete effects in which social scientists are interested. In doing so, error mechanisms can be added, such as information gaps that result from the users' or producers' clean-up activities. These aspects of the data-generating process can be implemented by specifying the strategies and rules of interactions the actors follow. Consequently, a variety of distortion mechanisms can be simulated not only by modelling different technically possible but also differing socially common ways of interacting. Finally, the simulation approach is useful as it enables us to study distorting mechanisms that may arise from applying specific *data processing* conventions, such as by transforming the simulated data to different data structures or selecting only very specific elements of the overall process.

In sum, this approach allows us to learn more about biasing mechanisms, and whether we can detect these built-in problems, for example, when we use statistical techniques of post-hoc identification.

4 A Case Study: Simulating and Identifying Bot Behavior

In order to illustrate the general strategy outlined above, we use the example of bot activity. The distortion effect we try to capture is the influence, in different scenarios, of bots on interaction patterns and on the resulting overall data structures. The population of our simulation model consists of two different agents, users and bots, and we assume a series of situations in which actors perceive, contact, and respond to each other in different scenarios, i.e., in different socio-technical systems that produce and (partly) provide data on the users' profile characteristics and actions. The simulations contain examples of the three aspects of the general principle outlined above: Observational design (the a priori specification of possible interactions based on profile selection), the process of data generation (as a result of the interactions of bots and users), and data processing (where the originally relational data will be transformed to and analyzed as a cross-sectional flat file).

Socio-technical systems differ according to their prevailing interaction structures and norms; the role of bots is, of course, different in each context, depending on the respective type of empirical environment (dating platform, platform for the exchange of political views, etc.). Therefore, we simulate different scenarios, each with a different parameterization of a crucial factor: What are the consequences of bot presence in contexts that differ only in terms of the prevailing interaction structure? That is, bot activity will remain constant over the series of scenarios, but the interaction style specific to the respective socio-technical system will be systematically modified. We model the differences between the scenarios as varying homophily thresholds. In terms of empirical examples, one may compare online dating platforms—where the homophily principle is more relevant—with consumer/business-to-consumer e-commerce websites such as eBay, where homophily is less prevalent (see Huber and Malhotra 2016; Šćepanović et al. 2017); one may think here, for example, about harvester or spam-bots. Furthermore, it has been shown that homophily is sometimes even used as a strategy for so-called astroturfing bots in order to influence discussions on certain social media and to reinforce specific opinions (Lazer et al. 2018). Although homophily cannot be considered a general principle of all interaction, it can serve as an established starting point for an exemplary, parsimonious model.

The agents are defined as being endowed with some invariant attributes. In reality, these attributes could take on many different forms, such as profile information, publicly displayed affiliations, or social status within an online community. We assign the values 0 or 1 to these invariant attributes for nine variables. For bots, we assume that they have been programmed in such a way that they are widely considered to be attractive or interesting (e.g., by publishing appealing content) in the eyes of a target audience. To instrumentalize this effect in the construction of our bots, we set one part of their profile information to 1, whereas the rest is filled randomly from a binomial distribution. As a result, an average bot is very similar to many human actors (in fact, even more similar than a randomly selected human).

Further, we assume that interactions are driven by some form of commitment in order to be sustained. As what is involved in interaction processes is not a mere deterministic process that runs given the constellation of two actors' attributes and

the interaction-process does not only involve manifest (dis)similarities between two actors, but rather the joint production of shared conceptions (such as complementary role conceptions or complementary hierarchy relations). For a real-world example, we might consider an online media discussion where people congregate around a shared interest or topic. As the interaction continues, it can develop in such a way that participants are no longer in sufficient agreement and thus terminate the interaction. The continuation of the interaction is modeled as being dependent on a given similarity as well on the participants' willingness to adapt to the ongoing changes caused by the interaction, that is, by additional similarity regarding attributes that can change during an interaction. Constellations in which no shared meanings are established will cease over the course of the interaction and the resulting processes of reciprocal classification; this pattern of communication cessation should be particularly typical for bots owing to their severely restricted communicative competencies. This is why our bots are simulated in such a way that they cannot adapt to the communication offers of human users.

The actual simulation process is set up as follows: First, the agents are randomly drawn from a population of all agents. If the newly drawn agent (ego) is not currently interacting with another agent, a new draw is made from the entire population (alter). If ego and alter have not interacted in the past, the Jaccard distance between all fixed attributes is calculated.[4] If the Jaccard distance falls below or is equal to a threshold value, both alter and ego are said to be interacting. The size of the threshold between 0 and 1 represents the degree to which interactions are structured by homophily, with values closer to 0 being indicative of interactions that have a higher requirement in terms of similarity (homophily) in order to be present. Following this step, a new agent is drawn from the pool, and we check again whether the agent is part of an interaction. If not, the process above is repeated until an agent with a raised interaction flag is drawn or the pool of available agents is empty. In the latter case, the next round begins by returning all agents to the pool. If the agent that has been drawn is marked as interacting, the interaction proceeds: In this case, the agent whose turn it is (ego) selects an element of its binary profile vector in a specified range and flips that bit, i.e., turns a 0 into a 1 and vice versa. After this change of a specific bit, the interaction partner (alter) can reciprocate by setting its corresponding bit to the same position as ego's. However, this reciprocity can only be exercised by non-bots. Therefore, if alter is not a bot, alter will reciprocate with a probability of p_r, determined by a global parameter.[5] After this change, the profile vectors are compared again, and the interaction is only continued if the Jaccard distance still falls below the threshold a.

Furthermore, during every interaction, the bot might terminate the interaction (depending on the base chance p_t times the duration of the ongoing interaction); this simulates a bot's tendency to follow a specific pattern, that is, trying to achieve a desired outcome, such as getting (exclusively human) users to click on a link.

[4] This measure is based on the overlap between two binary vectors, ranging between 0 (no overlap) and 1 (complete overlap) (Jaccard 1912).

[5] This parameter was invariably set to 0.5, meaning that there was a constant 50% chance for alter to reciprocate.

Finally, we include the users' competency to unmask bots dependent on the number of experiences gained with their prior bot interactions. The rule states that a user can employ a test for every interaction they have had with a bot: The test gives the user a chance of 25% of stopping the interaction. These simulations were run for a certain number ($n = 5000$) of interaction steps.

Subsequently, we selected three datasets for an exemplary statistical post-hoc analysis. In this way, we not only simulate data-generating and error-generating mechanisms, but rather the very situation in which sociologists usually find themselves—namely, having to work with highly aggregated, selective data and no definitive information on how these data are generated and which quality problems they might entail. In the specific case, we are interested in whether bot presence as generated in the simulated data sets is easily noticeable and to which extent cases are falsely classified as being problematic.

In each case, the resulting relational data were aggregated to a data format still common in the social sciences: A two-dimensional data extract, which is composed of the profile variables mentioned above, and augmented by selected continuous variables expressing the number of unique partners, the overall sum of interactions in which this agent participated, and the average interaction duration by contact partner. To select the sub-set of illustrative datasets, we apply the homogeneity criterion (Rosenberg and Hirschberg 2007) to both profile and interactional variables. This criterion serves to select scenarios that differ in their underlying multivariate data structure (and thus in the difficulty of their statistical classifiability). In doing so, we yield three datasets that differ with respect to their similarity thresholds ($a = 0.3$, 0.65, and 0.8, with 0 being the maximum possible and 1 the minimum possible similarity)[6]:

- Scenario 1: Strong homophily selection ($a = 0.3$)
- Scenario 2: Moderate homophily selection ($a = 0.65$)
- Scenario 3: Low homophily selection ($a = 0.8$)

For identifying bots in the artificially generated datasets, we mobilize geometric data analysis as an exploratory tool, which has proven useful in the context of error identification in survey data (see Blasius and Thiessen 2012). Explorative methods of this kind are appropriate when it can be assumed that error-inducing mechanisms are involved and that these errors might manifest in the resulting data, but when it is not known a priori which errors are actually present and in which ways they will be reflected in the data. The solutions yielded by geometric data analyses represent multivariate associations in the data, in the form of two kinds of interrelated spaces: Spaces of characteristics, which enable researchers to interpret the meaning of the variables analyzed, and spaces of individuals, which can indicate and visualize multivariate outliers.

The idea is that a researcher does not know the nature and extent of the problem, but is generally aware that bots may differ both in terms of their (e.g., more homogeneous) profile characteristics and in terms of their (e.g., more extreme) in-

[6] Figures D1–D4 in the online appendix illustrate the rationale for the selection of the three exemplary scenarios.

teraction patterns. As the data contain variables with different scales, we employ multiple factor analysis (Pagès and Bécue-Bertaut 2006), which integrates factorial solutions for categorical and continuous variables into a common solution. Online appendix 1 contains the space of characteristics for all analyses, separated by continuous (interaction) and categorical (profile) variables. The spaces of individuals, i.e., the dispersions, are shown below. For easier visibility, bots (blue) and human users (gray) have been color-coded differently and the dispersion of both categories is indicated by concentration ellipses.

Analyzing dataset 1, which exhibits the greatest extent of homophily-based selection, yields a space of characteristics where three of the binary profile variables (1, 2, 3), a high number of contact partners, and low average interaction length describe the lower, right area of the graph (see Figs. A1 and A2 in the online appendix). Based on existing knowledge, a researcher might surmise a pattern matching the characteristics and interaction modes of bots, as this specific pattern (many contacts, short interactions) is known from bots' activities on, for instance, dating platforms (see Schmitz et al. 2012). In fact, Fig. 1, which shows the space of individuals and thus the dispersion of all cases in the plane, allows us to identify outliers in the fourth quadrant (bottom right).

Yet most bots are still located within the distribution of human users (gray circle) and would not have been unmasked without further ado. Of the 30 bots actually present, the visual inspection of the outliers would identify only six to eight cases, while not suggesting false-positive assignments. This circumstance is the result of two effects: Bots are more likely to engage in an interaction (i.e., have high numbers of interactions) owing to their advantageous profile characteristics, which make them comparatively more similar to more contact partners than the average human user, resulting in a high number of different contact partners. However, the

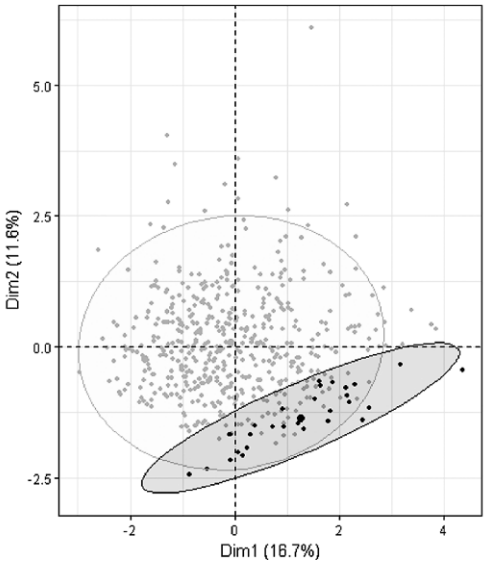

Fig. 1 Space of Individuals Dataset 1 (MFA)

intense demand for homophily makes continuing the interaction much harder, as bots cannot adapt to human users over the course of the symbolic interaction process, so that the distance between bots and human users increases during the interaction and the resulting average interaction length is very low. Nevertheless, this effect is not so extreme that bots could be immediately distinguished from humans in a highly distinctive form based on an initial exploratory analysis.

For dataset 2 (moderate homophily-based selection), the overall number of interactions and the average count of interaction events by contact partner are strongly positively correlated with each other, but strongly negatively correlated with the number of unique contact patterns (see Fig. B1 in the online appendix). This indicates, overall, a clearly ordered, uniform interaction structure in which agents only interact with a few contacts, although over a longer series of events. Assuming that bots do not establish numerous lasting interactions, and will have unsuccessfully tried to approach many unique contact partners, one might expect them to be located at the left side of the space. As in scenario 1, this region is also described by some of the profile properties, although no longer with the same discriminatory precision (see Fig. B2 in the online appendix); thus, as expected, the profile characteristics have less importance for the initiation and continuation of interactions. However, the space of individuals does not suggest any readily identifiable outliers (see Fig. 2). Although the bots are located with disproportionate frequency in the expected second and third quadrants (top and bottom left), they do not take extreme positions in the cloud of individuals.

In this scenario, by design, homophily is of less significance when it comes to continuing or terminating an interaction. The moderate pressure of homophily-based selection provides human users with sufficient potential interaction partners to which they can adapt and with whom they can form lasting, exclusive relationships, thereby

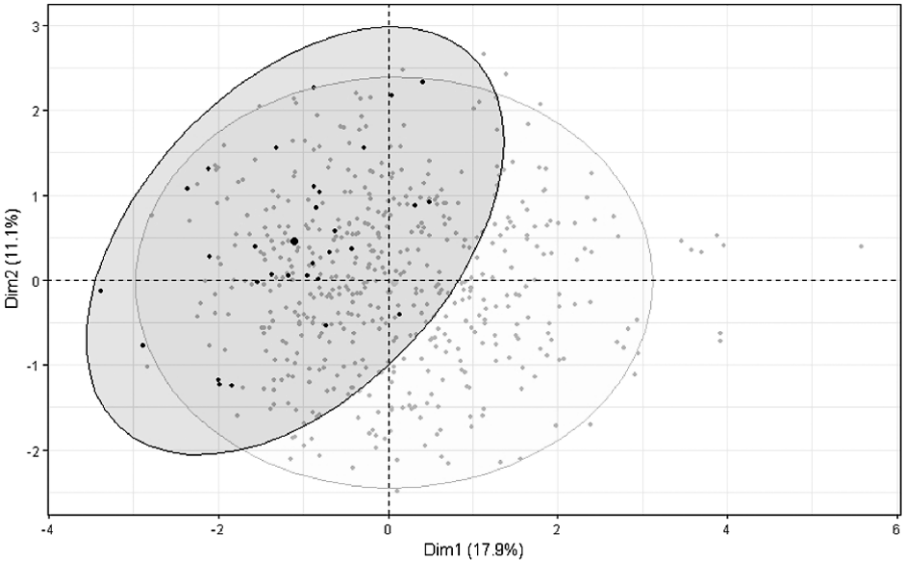

Fig. 2 Space of Individuals Dataset 2 (MFA)

obfuscating the specific interaction patterns that could potentially reveal the bots as outliers.

For dataset 3 (weak homophily-based choices), another pattern can be observed. Again, several profile indicators describe the right-hand part of the plane (1, 2, 3, and 4), but now, the overall sum of interaction events and the average duration of an interaction have the highest values along the first diagonal, whereas the number of unique interaction partners describes the second diagonal (see Fig. C1 and C2 in the online appendix). Thus, cases on the right side of the space possess a high number of interactions and long average interactions, as well as more interaction partners. One might assume that, in a context where the similarity and commitment of the interacting actors are not relevant, bots should be relatively successful in establishing numerous lasting interactions.

Accordingly, the space of individuals shows some clearly identifiable outliers on the right-hand side (see Fig. 3) and 16 of the 30 bots would be identified (with two or three false-positive classifications) via visual inspection.[7]

As it turns out, albeit only for the two extreme specifications (strong and weak homophily-based contexts), we stumbled across bot-specific phenomena simply by performing a basic explorative analysis. For scenario 1, it can be assumed that strong homophily leads to a situation where bots have been sorted out by human users during the interaction process, owing to the bots' inability to adapt, whereas weak homophily in scenario 3 promotes situations where human users did not exclude

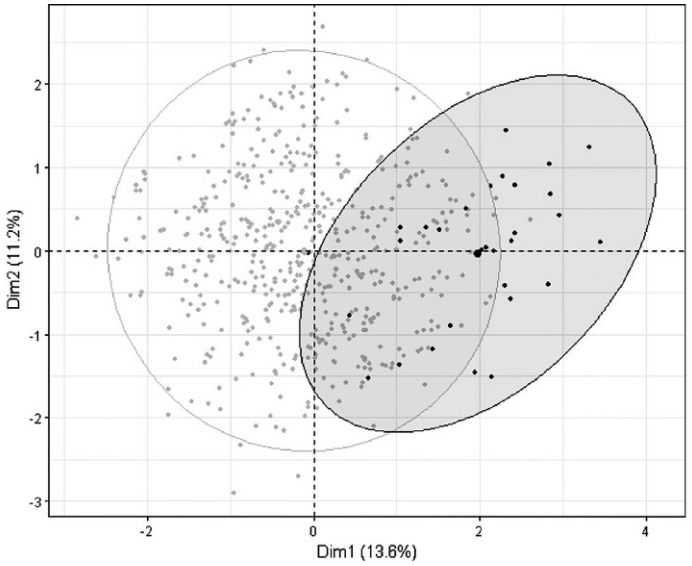

Fig. 3 Space of Individuals Dataset 3 (MFA)

[7] As homophily only played a weak role here, the pattern could also be the result of secondary effects: both the fact that bots can move on to new users after they are revealed as bots and their tendency to terminate unsuccessful interactions enables bots to establish an increased number of new interactions as well as more long-term interactions, whereas, for humans, these effects have no impact (as their random dispersion pattern suggests).

bots, but instead kept on interacting with them. Or, in short: Both constellations created discernible outlier patterns. In the case of strong homophily, bots were conspicuous in that they were rarely involved in (lasting) interactions, whereas in the case of weak homophily, bot presence became evident because they were very strongly involved in the interaction processes.

This nonlinear way in which homophily impacts the identifiability of bots demonstrates the need for further systematic simulations: Substantially more data sets need to be generated (a) over a wider parameter space, (b) with more varying parameters involved (such as heterophilous strategies), and (c) with multiple repetitions to control random variations in the individual scenarios. Only in this way can we specify the exact conditions in which the implemented errors can be understood and identified; without such a systematic, controlled comparison it is always possible that random effects will be mistaken for systematic effects.[8]

5 Discussion

While working with digital process data is becoming increasingly relevant to the practice of social scientists from diverse paradigmatic backgrounds, the topic of data quality is still in its infancy. To address this major challenge of modern empirical research, this paper has drawn on the body of social science knowledge with respect to the empirical phenomena, theoretical conceptualizations, and methodological controllability of quality distortion. Therefore, we analytically distinguished three generalized aspects that empirically interact in producing digital process data: Observational design, data generation, and data processing. These three ideal-typical dimensions describe mechanisms that may call into question the quality, validity, and reliability of digital process data on entities, events, and their relations. Whereas issues of quality of digital data can indeed be compared with traditional data types, a crucial meta-quality criterion of such data is *transparency*, i.e., the degree of inspectability and traceability into the underlying processes of data production.

Given the fact that transparency cannot be assumed, as sociologists usually have very little if any insight into the conventions and processes that underlie the production of digital process data, we discussed the promising role of combining simulation and post-hoc identification techniques. Simulation techniques represent a way to respond to the lack of control and insight and employing such approaches contributes to the body of empirical knowledge by adding artificial data, gaining experience about the effects of error mechanisms in the resulting data, and learning whether the traditional identification techniques at our disposal are helpful in drawing our attention to suspicious phenomena. We illustrated this approach using the example of the identification of bots, a phenomenon that can be said to genuinely belong to socio-technical environments and that cannot be understood as a mere "external" nuisance. Yet to the extent that we are unaware of their activities, we always run the risk of misinterpreting observed actions, interactions, and communications.

[8] Likewise, this method allows us to systematically evaluate the consequences of the other two implied effects (unmasking and termination).

In order to grasp such problems—as well as other distorting phenomena within observational design, data generation, or data processing—in a more systematic fashion, future research can build on our mixed-methods strategy by systematically generating data sets across a multitude of simulations and thereby accounting for random variations. Applying a "pipeline strategy" over a series of simulated data, varying the conditions and parameters, and running the simulations several times will enable social scientists to systematically evaluate the threshold values at which suspicious patterns come to light when using identification methods (geometric data analysis, but also finite mixture models, or clustering methods).

Yet, such automated strategies must be realized in conjunction with the interpretative competencies of the researcher. It is essential to have both an adequate understanding of the respective empirical phenomena and a theoretical understanding of possible distortion mechanisms, and, as in traditional survey research, insights must be acquired through practical engagement with empirical data sets. In doing so, knowledge gained from simulated datasets may sensitize empirical researchers who screen their real-life datasets for comparable patterns using similar identification techniques. For this purpose, entire large-N data sets cannot and need not be examined manually: The iterative, abductive examination of samples of conspicuous cases, as well as qualitative or ethnographic investigations, can be most useful in providing further clues regarding suspicious patterns and in specifying explorative identification models. In the context of survey research, researchers have already used qualitative (cognitive) interviews with respondents on their perceptions of the survey as well as ethnographic observations of the classification practices subsequently conducted by the researchers. In similar ways, future research will employ mixed-methods approaches to quality issues of digital process data. A great deal can be learned about observational design, data generation, and data processing procedures through expert interviews with programmers (who operate in comparable contexts of practice and can inform us about common conventions) and ethnographic observations of programming activities. Complementarily, qualitative interviews with platform users can reveal their perspectives, practices (such as their ways of identifying and interacting with bots), and effects on socio-technical systems. Such accounts are essential to increase our contextual knowledge, and they can be useful for further developing simulations (e.g., for implementing more realistic strategies).

Ultimately, such mixed-methods approaches to data quality in the digital realm can help us to understand that phenomena that are interpreted as mere distortions of otherwise accurate observational data are, in fact, constitutive and generative elements of socio-technical systems.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Allen, Jennifer, Markus Mobius, David M. Rothschild and Duncan J. Watts. 2021. *Research note: Examining potential bias in large-scale censored data*. Harvard Kennedy School Misinformation Review.

Bachleitner, Reinhard, Martin Weichbold and Wolfgang Aschauer. 2010. *Die Befragung im Kontext von Raum, Zeit und Befindlichkeit: Beiträge zu einer prozessorientierten Theorie der Umfrageforschung*. Wiesbaden: Springer VS.

Barth, Alice, and Andreas Schmitz. 2018. Response quality and ideological dispositions: an integrative approach using geometric and classifying techniques. *Quality & Quantity* 52(1):175–194.

Baur, Nina, Peter Graeff, Lilli Braunisch and Malte Schweia. 2020. The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age. *Historical Social Research/Historische Sozialforschung* 45:209–243.

Biemer, Paul P. 2010. *Latent class analysis of survey error*. Hoboken, NJ: John Wiley & Sons.

Blasius, Jörg, and Victor Thiessen. 2012. *Assessing the quality of survey data*. London: Sage.

Blasius, Jörg, and Victor Thiessen. 2015. Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research* 52:479–493.

Boyd, Danah, and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5):662–679.

Bratu, Sofia. 2017. The inexorable shift towards an increasingly hostile cyberspace environment: The adverse social impact of online trolling behavior. *Contemporary Readings in Law and Social Justice* 9:88–94.

Bredl, Sebastian, Peter Winker and Kerstin Kötschau. 2012. A statistical approach to detect interviewer falsification of survey data. *Survey Methodology* 38:1–10.

Bulut, Ergin, and Erdem Yörük. 2017. Digital populism: Trolls and political polarization of Twitter in Turkey. *International Journal of Communication* 11:4093–4117.

Desrosières, Alain. 2001a. Entre réalisme métrologique et conventions d'équivalence: les ambiguïtés de la sociologie quantitative. *Genèses* 43(2):112–127.

Desrosières, Alain. 2001b. How Real are Statistics? In *Social Research*, 339–355.

Desrosières, Alain. 2009. How to be real and conventional: A discussion of the quality criteria of official statistics. *Minerva* 47:307–322.

Deutsche Forschungsgemeinschaft (DFG). 2020. *Digitaler Wandel in den Wissenschaften. 28. Oktober 2020. Impulspapier*.

Diaz, Fernando, Michael Gamon, Jake M. Hofman, Emre Kıcıman and David Rothschild. 2016. Online and social media data as an imperfect continuous panel survey. *PLoS ONE* 11(1):e0145406.

Diaz-Bone, Rainer. 2016. Convention theory, classification and quantification. *Historical Social Research/Historische Sozialforschung* 41:48–71.

Diaz-Bone, Rainer. 2018. *Die „Economie des conventions". Grundlagen und Entwicklungen der neuen französischen Wirtschaftssoziologie*. Wiesbaden: Springer VS.

Diaz-Bone, Rainer, Kenneth Horvath and Valeska Cappel. 2020. Social research in times of big data. The challenges of new data worlds and the need for a sociology of social research. *Historical Social Research/Historische Sozialforschung* 45:314–341.

Dijkstra, Wil, Stasja Draisma and Johannes van Der Zouwen. 1995. Simulating response behavior in sociological survey interviews. *Journal of Mathematical Sociology* 20:127–144.

Dolata, Ulrich. 2009. Technological innovations and sectoral change: Transformative capacity, adaptability, patterns of change: An analytical framework. *Research Policy* 38:1066–1076.

Driscoll, Kevin, and Shawn Walker. 2014. Working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication* 8:1745–1764.

Faddoul, Marc, Guillaume Chaslot and Hany Farid. 2020. A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. arXiv preprint. arXiv:2003.03318.

Gao, Jian, Zhang, Yi-Cheng and Tao Zhou. 2019. Computational socioeconomics. *Physics Reports* 817:1–104.

González-Bailón, Sandra, Nina Wang, Alejandro Rivero, Jorge Borge-Holthoefer and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Social Networks* 38:16–27.

Groves, Robert M., and Lars Lyberg. 2010. Total survey error: Past, present, and future. *Public Opinion Quarterly* 74(5):849–879.

Häußling, Roger. 2020. Daten als Schnittstellen zwischen algorithmischen und sozialen Prozessen. Konzeptuelle Überlegungen zu einer Relationalen Techniksoziologie der Datafizierung in der digitalen Sphäre. In *Soziologie des Digitalen-Digitale Soziologie?* Eds. Sabine Maasen and Jan-Hendrik Passoth, 134–150. Baden-Baden: Nomos.

Hellerstein, Joseph M. 2008. *Quantitative data cleaning for large databases*. United Nations Economic Commission for Europe (UNECE).

Howison, James, Andrea Wiggins and Kevin Crowston. 2011. Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems* 12:767–797.

Huber, Gregory A., and Neil Malhotra. 2016. Political Homophily in Social Relationships: Evidence from Online Dating Behavior. *The Journal of Politics* 79(1):269–283.

Jaccard, Paul. 1912. The Distribution of the Flora in the Alpine Zone. *New Phytologist* 11(2):37–50.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil and Abe Usher. 2015. Big data in survey research: AAPOR task force report. *Public Opinion Quarterly* 79:839–880.

Jun, Tackseung, and Rajiv Sethi. 2008. Erratum: Neighborhood structure and the evolution of cooperation. *Journal of Evolutionary Economics* 18(1):103. Original in: 2007. *Journal of Evolutionary Economics* 17:623–646.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Lyberg, Lars E., and Paul P. Biemer. 2018. Quality assurance and quality control in surveys. In *International handbook of survey methodology*, 421–441.

Macal, Charles M., and Michael J. North. 2009. Agent-based modeling and simulation. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 86–98.

Malik, Momin M., and Jürgen Pfeffer. 2016. Identifying platform effects in social media data. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 241–249.

McCarthy, Jaki, James Wagner and Herschel Lisette Sanders. 2017. The impact of targeted data sourcing on nonresponse bias in an establishment survey: A simulation study of adaptive survey design. *Journal of Official Statistics* 33:857–871.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu and Kathleen M. Carley. 2013. Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In *Proceedings of the Seventh International AAAI Conference on Web and Social Media*. arXiv:1306.5204v1.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2:13.

Pagès, Jérôme, and Monica Bécue-Bertaut. 2006. Multiple factor analysis for contingency tables. In *Multiple Correspondence Analysis and Related Methods*, eds. Michael Greenacre and Jörg Blasius, 433–453. Boca Raton: Chapman & Hall.

Riebling, Jan R. 2018. The Medium Data Problem in Social Science. In *Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications*, Neue Schriften zur Online-Forschung of the German Society for Online Research (DGOF), eds. Cathleen M. Stuetzer, Martin Welker und Marc Egger, 77–103. Köln: Herbert von Halem.

Rosenberg, Andrew, and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 410–420. Prague, Czech Republic: Association for Computational Linguistics.

Šćepanović, Sanja, Igor Mishkovski, Bruno Gonçalves, Trung Hieu Nguyen and Pan Hui. 2017. Semantic homophily in online communication: Evidence from Twitter. *Online Social Networks and Media* 2:1–18.

Schmitz, Andreas, Jan Skopek, Florian Schulz, Doreen Klein and Hans-Peter Blossfeld. 2009. Indicating mate preferences by mixing survey and process-generated data. The case of attitudes and behaviour in online mate search. *Historical Social Research* 34(1):77–93.

Schmitz, Andreas, Olga Yanenko and Marcel Hebing. 2012. Identifying artificial actors in E-dating: A probabilistic segmentation based on interactional pattern analysis. In *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, eds. Wolfgang Gaul, Andreas Geyer-Schulz, Lars Schmidt-Thieme and Jonas Kunze, 319–327. Berlin, Heidelberg: Springer.

Sen, Indira, Fabian Floeck, Katrin Weller, Bernd Weiss and Claudia Wagner. 2019. A total error framework for digital traces of humans. arXiv preprint. arXiv:1907.08228.

Shin, Eunjung, Timothy P. Johnson and Kumar Rao. 2012. Survey mode effects on data quality: Comparison of web and mail modes in a US national panel survey. *Social Science Computer Review* 30:212–228.

Starbird, Kate. 2019. Disinformation's spread: bots, trolls and all of us. *Nature* 571:449–450.

Tufekci, Zeynep. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*. arXiv:1403.7400.

Van der Aalst, Wil. 2016. Getting the data. In *Process Mining*, 125–162. Berlin, Heidelberg: Springer.

Van Vaerenbergh, Yves, and Troy D. Thomas. 2013. Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research* 25:195–217.

Venturini, Tommaso, and Bruno Latour. 2010. The social fabric: Digital traces and quali-quantitative methods. In *Proceedings of future en seine*, 87–101.

Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz and Lee Sechrest. 1966. *Unobtrusive measures: nonreactive research in the social sciences*. Chicago: Rand McNally.

West, Brady T. 2013. The effects of error in paradata on weighting class adjustments: A simulation study. In *Improving surveys with paradata: Analytic uses of process information*, ed. Frauke Kreuter, 361–388. Somerset: Wiley and Sons.

Zillmann, Doreen, Andreas Schmitz and Hans-Peter Blossfeld. 2011. Lügner haben kurze Beine: Zum Zusammenhang unwahrer Selbstdarstellung und partnerschaftlicher Chancen im Online-Dating. *Zeitschrift für Familienforschung* 23(3):291–318.

Andreas Schmitz 1977, PD Dr., University Bonn, Department for Sociology, and Gesis, Cologne, Department Computational Social Science. Fields of research: relational data, digital process data, geometric data analysis, social structure analysis of digital interaction and communication contexts, sociology of fear. Publications: Social Network Analysis with Digital Behavioral Data. easy_social_sciences 66, 2021 (with H. Lietz and J. Schaible); Investigations of Social Space. Cham 2019 (ed. with J. Blasius, F. Lebaron and B. Le Roux); Objektivierung der Kritik, Kritik der Objektivierung. In: Themed issue, "Methoden & Gesellschaftskritik" in Psychologie & Gesellschaftskritik 44, 2020 (with A. Barth).

Jan R. Riebling 1980, Dr., research assistant at the chair for general sociology at the University of Wuppertal. Fields of research: Computational social science, methods and methodology as well as the formal and mathematical modeling of dynamic systems. Publications: Relating social and symbolic relations in quantitative text analysis. A study of parliamentary discourse in the Weimar Republic. Poetics 78, 2020 (with J. Fuhse, O. Stuhler and J. L. Martin); Eine Frage des Marktes? Regionale Unterschiede von Heimentgelten stationärer Pflegeeinrichtungen. Berliner Journal für Soziologie 27, 2017 (with R. H. Heiberger and B. Schwarzer); ZombieApocalypse: Modeling the Social Dynamics of Infection and Rejection. Methodological Innovations 9, 2016 (with A. Schmitz).