# Social Network Analysis with Digital Behavioral Data

Lietz, Haiko; Schmitz, Andreas; Schaible, Johann

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

doi: 10.15464/easy.2021.005

# Social Network Analysis with Digital Behavioral Data

*Haiko Lietz, Andreas Schmitz & Johann Schaible*

*Our uses of digital technologies like social media platforms or email leave massive amounts of behavioral traces that are most interesting for social research. Other digital technologies like cell phones allow harnessing behavioral traces for research purposes. Such Digital Behavioral Data consists of genuinely relational records which can be thought of in terms of networks. However, this kind of data requires a shift of perspective from individuals to micro events (e.g., a post on social media) as units of observation and brings established techniques like Social Network Analysis to the center stage. We argue that, using this approach, obtaining individual attributes and attitudes as well as uncovering the micro-macro dynamics of behavior by mining patterns are potentially fruitful applications. We discuss methodological challenges and conclude that social theory is a constitutive pillar for the consolidation of Computational Social Science.*

*Keywords*: Digital Behavioral Data, Social Network Analysis, Computational Social Science, transactions, attributes and attitudes, patterns

Through digital media such as Facebook, we all became familiar with the idea of social networks. As "social networking," online practices have become part of our daily lives. But even beyond digital platforms, we encounter the concept of networking time and again: our job search, sports activities, the question with whom we (rather not) want to collaborate – more and more things are actually organized as networks. In fact, since its early beginnings in the 19th century, the social sciences have been interested in social relations and, subsequently, have increasingly dealt with networks in an explicit way. The idea is not only that social formations can be described in terms of nodes linked by edges and the structures that arise, but that they actually function as networks (White, 2008).

This theoretical perspective and the associated research methods have recently gained considerable importance. We are devoting an ever-increasing amount of our time to life in the digital world. These are lives in digital ecosystems where every action leaves a trace. Facebook logs who is friends with whom, Google logs who searched for what, Amazon logs who purchased what, all data collected in real time. But this is not all they do. Facebook knows which users are often friends of others together, Google knows which terms are often searched for together, Amazon knows which products are often purchased together. Platform operators mine the *patterns* that arise in the totality of things being selected together. Knowledge of patterns is the new gold in the digital economy. The platform operators sell it or use it to recommend new friends, search terms, or products. Since patterns take the form of networks, the network perspective has gained importance as associated meth-

ods turned out useful for processes of pattern mining (Nassehi, 2019). What is of particular interest for the social sciences is how to use this behavioral data that often arises as a byproduct from the operation of private businesses in a way that systematically relates to theoretical concepts.

> ***This 'telescope' of the social sciences is the simultaneous availability of massive amounts of behavioral data and the technological abilities to analyze it.***

Watts (2011, p. 266) compares the social sciences in the early 21st century to the beginning of modern astronomy in the early 17th century and proclaims that "we have finally found our telescope." This "telescope" of the social sciences is the simultaneous availability of massive amounts of behavioral data and the technological abilities to analyze it. Part of this technological innovation is that the "telescope" can process masses of microscopic events or transactions (e.g., posts on social media) for complete digital ecosystems, and it promises to enable a social science that had never been possible before. Computational Social Science (CSS), an emerging field at the intersection of social and computer science, is taking up the challenge (Lazer et al., 2020).

In this article, we offer a definition of Digital Behavioral Data, discuss its properties and insightful applications. We propose that it exhibits large potential in obtaining individual attributes and attitudes as well as in functioning as a macroscope that uncovers the micro-macro dynamics of behavior. We close with a discussion of five major restrictions and challenges related to using social science's "new telescope."

## Digital Behavioral Data in Context

We define Digital Behavioral Data (DBD) as the traces of behavior left by uses of or harnessed by digital technology. To start building our understanding of behavior, we characterize these traces as records of *transactions*, that is, observed phenomena that are not adequately defined as atomistic entities but rather as genuinely relational emanations. These objects have four dimensions (Figure 1): First, they comprise social relations among actors, usually a sender and one or several receivers. Second, they include communicative content and references. Third, entities have attributes (such as actors' attitudes) that can also be relationally produced. Fourth, transactions imply a temporal dimension, that is, as micro events, they are initiated at a point in time (Emirbayer, 1997). On Twitter, for example, transactions are called "tweets." They are social relations insofar as they are actions in a network of users and have meaning content (of up to 280 characters). Above that, user attributes can be obtained from profiles, and tweets are time-stamped.

In our definition of DBD, we differentiate between two types of traces encountered in practice: found and designed DBD. Traces left by uses of digital technology correspond to *found* DBD (Howison et al., 2011). This subsumes data as logs from digital platforms where

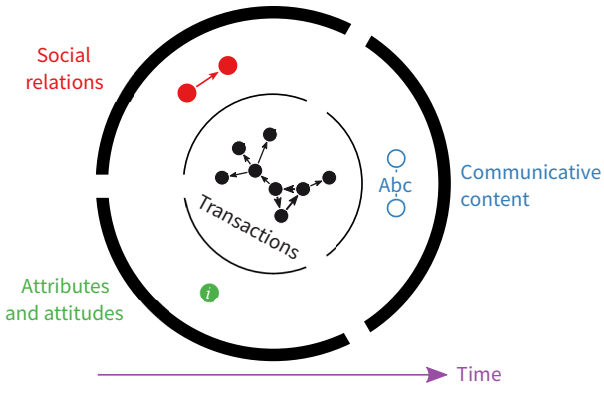


*Figure 1* Four Dimensions of Transactions as Micro Events

the trace is both the action and its technically facilitated observation. The ideal-typical example is the usage of social media and messaging services like Facebook, WhatsApp, or YouTube. Practices like posting, commenting, tagging, or liking are accounts of "digital life" (Lazer & Radford, 2017). Online purchasing, web browsing, or web searching are further examples. These particular traces consist of data on at least social relations and communicative content. Emailing, phone calling, and short messaging also leave traces by uses of the underlying technical infrastructures. But they are not forms of digital life because the traces are only records of actions, not the actions themselves (Lazer & Radford, 2017). In the case of emailing, social relations and communicative content are logged on mail servers. Call Detail Records that telephone providers nowadays sell are devoid of content.

Traces harnessed by digital technology correspond to *designed* DBD, that is, the data is not a byproduct of digital platform operations but is explicitly produced for a research purpose. Traces of face-to-face interactions are an example. Such traces can be recorded by wearable sensors that measure whether or not two carriers of a sensor are facing each other in close proximity. Such proxies of micro behavior are pure records of time-stamped social relations (Schaible et al., 2022). Other digital traces only have the transactional properties of time-stamped attributes, so-called metadata to micro events. Think of continuously logged GPS positions of a cell phone user (a physical attribute), sleep phases and heart rates recorded by fitness trackers (biological attributes), which other cell phone is proximate as measured by a Bluetooth sensor (a social attribute), or the noise level of a cell phone user's location (an environmental attribute). It is up to the scientist to map these traces to research concepts (e.g., sensor proximity to collaboration).

## Properties of Digital Behavioral Data

DBD is often referred to as "big data" defined by the three V properties of large volume (it exceeds the capacity of conventional hardware), large variety (it comes in many forms other than the square data frame), and/or large velocity (it is longitudinal and variable). In fact, DBD is quite diverse, and each data source requires its own processing routine. However, in practice, it can sometimes be handled on a laptop or desktop machine. A more fundamental aspect of DBD is its relationality. Figure 1 displays the social relation. But it can be any entities that relate to each other and that can be related to each other in a number of ways (e.g., transactions referring to previous transactions, words used together, or attributes occurring together).

> ***In principle, those with access and skills can analyze all transactions made on a digital platform.***

The meaning of DBD becomes clearer as we put it into context by comparing it to survey data (Table 1). In the tradition of survey methodology, which began a century ago, the unit of analysis about which information is to be obtained is society at large or one of its subpopulations. Yet, for methodological reasons, the individual became the paradigmatic unit of observation. Questioned in an interview, an individual self-reports her or his attributes and attitudes

*Table 1* Social Data in Comparison

| | Survey data | Digital Behavioral Data |
|---|---|---|
| Unit of observation | Individual | Transaction |
| Structure | Cross-sectional | Longitudinal |
| Scope | Representative | Exhaustive |
| Source of bias | Cognition | Feedback |

about a certain topic. Since full samples can hardly be realized, but representative samples may yield appropriate results, survey methodology was developed in close connection with sampling techniques. Methodic restrictions become even more virulent when researchers aim for longitudinal observations.

DBD, in contrast, is exhaustive in scope. In principle, those with access and skills can analyze all transactions made on a digital platform. This would constitute a full sample, with the units of observation being not individuals but transactions. Most often, DBD is severely skewed, with few users contributing a disproportional number of observations. But this is a property of the system under observation, not a problem. As a constant data stream (the velocity argument of big data), DBD poses no limits to analyzing these collective dynamics (Diaz et al., 2016).

The unit of observation being transactions entails a rich set of opportunities for the unit of analysis. Information from the transactions (e.g., actors, words, even the event itself) can be used as network nodes, edges, or both. Units of analysis then depend on the level of analysis and can be anything from individual nodes and, in the case of actors, their micro behavior (e.g., their positional dynamics) to the whole network and its macro behavior (White, 2008). In the introduction, we have referred to macro behavior as patterns. Patterns can be anything from the existence of groups of actors to a structured discourse using words.

An important property of DBD is related to feedback, a mechanism by which individual behavior is influenced by past individual and collective behavior. It is central to how social structures come into being (Keuschnigg et al., 2018). A particular feature of found DBD (digital life) is that it is also subject to feedback loops put into effect by the platforms. It takes the form of patterns arising from, and later influencing, micro behavior (e.g., via recommendations or filtered event streams). Feedback is a real source of bias in found DBD, and it contrasts cognition as the main source of bias in survey data (e.g., desirability bias, memory loss).

The ability to study real-life settings, reconstruct manifest and symbolic relations, uncover the logics of behavior at micro and macro levels, and do so both with high temporal resolution and at scale are the main reasons why DBD fuels the emerging field of CSS. Next, we will discuss two types of applications based on relational analysis. We refer to other sources for avenues like machine learning, social simulations, or experiments (Lazer & Radford, 2017; Lazer et al., 2020).

## Applications of Digital Behavioral Data

### Scenario 1: Obtaining Individual Attributes and Attitudes

Survey methodology aims at identifying the attributes and attitudes of individuals. Many of these can also be inferred from DBD with a numeralizable error. Kosinski et al. (2013) surveyed detailed demographic profiles and performed several psychometric tests of almost 60,000 Facebook users, and attempted to predict the resulting variables only from their liking behavior using standard social science methods. Attributes like age and gender, political and religious views, and sexual orientation could all be predicted with at least 75% accuracy. Socio-psychological traits could be predicted much less accurately.

Multiple subfields of CSS are concerned with developing methods for mining opinions, recognizing emotions, identifying reasons, and detect sarcasm, irony, rumors, or stances from the textual part of DBD. Such inferences derive from methods of Natural Language Processing, that is, automated approaches to the procurement, management, and analysis of communication. These approaches became particularly famous in recent times as DBD makes the production and processing of meaning empirically accessible (Bail, 2014).

While attributes and attitudes have to be inferred from found DBD, attributes can also be harnessed directly by researchers employing digital technology. The Copenhagen Networks Study is a beacon in demonstrating the power of a mix of found DBD, designed DBD, and survey data as well as mixed methods. In 2013, researchers handed out 1,000 cell phones to students at the Technical University of Denmark and recorded their physical location (via the cell phone's GPS sensor), who was proximate to whom (via the Bluetooth sensor), who called whom, and who texted whom. In addition, their transactions on Facebook were collected, and their demographic and psychological traits were surveyed. Four types of social relations, taken from a publicly available portion of this dataset, are depicted in Figure 2.

Studying only the relational layer of physical proximity, researchers could uncover the micro behavioral roots of group formation (Sekara et al., 2016). In another project, they attempted to predict the students' Big Five personality traits (i.e., openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism) from their recorded behavior. Besides variables derived from their phone calling and text messaging behavior and the number of Facebook friendships, they also designed predictors measuring their mobility behavior (via GPS traces) and their geo-social embeddedness (via Bluetooth proximity). Out of the five traits, only extraversion could be predicted better than by chance, again indicating limits to predicting personality from behavior (Mønsted et al., 2018).

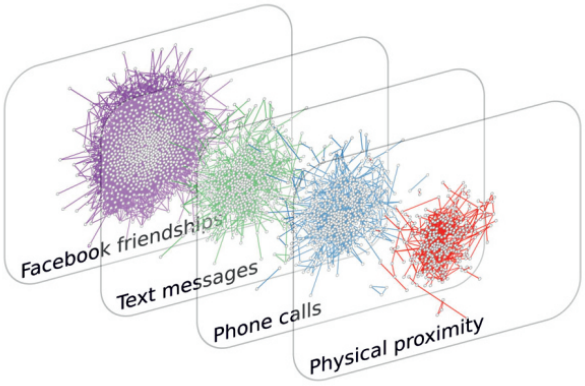


*Figure 2* Social Relations in the Copenhagen Networks Study

## Scenario 2: Uncovering the Micro-Macro Dynamics of Behavior

The second type of application that rests on DBD involves mining macro behavioral patterns. In its early, data-driven days, CSS had extensively devoted resources to exploring those. While it is worthwhile doing this, merely describing the macro level is not an explanation of its genesis. The promise of DBD is that it can function as a macroscope that allows for uncovering micro-macro dynamics of behavior. Macro behavioral patterns have importance as causes of micro behavior, aggregates of micro behavior, or part of a mechanism that integrates causes and aggregates. The feedback dynamics described above represent such a mechanism (Keuschnigg et al., 2018). Whether the micro level has primacy over the macro level or vice versa is one of the oldest questions in the social sciences, and DBD offers fresh answers. The essential step in addressing the problem is taking advantage of the longitudinal property of DBD.

For example, there is much research interest in homophily, the frequently observed pattern that social relations preferentially link identities that are similar in some attribute. Kossinets and Watts (2009) studied 30,000 persons at a US university and the 7 million emails they exchanged over 270 days. They found that the closer two persons are in the communication network, the more similar they are regarding their sociodemographics. But what are the origins of this homophily? Does the pattern emerge from micro behavior because like associates with like? Or does the pattern constrain micro behavior such that like has no other choice but to associate with like? The authors' answer is: both. After studying the network dynamics, the authors conclude that micro behavior and structures of macro behavior are co-constitutive. Network proximity and attribute-similarity converge

as distant, but similar persons are drawn together, facilitated by shared activities like classes.

Another approach to modeling the mutual dependence of micro and macro levels is by way of generative models. Bayesian stochastic blockmodels of social networks are a very vivid example. The idea of stochastic blockmodels is that nodes in a network are grouped together if they have the same pattern of connections to nodes in other groups. A simple example would be nodes in the periphery being connected to nodes in the core but not to other peripheral nodes. The Bayesian algorithm learns a blockmodel from a network under the assumption that the blockmodel has generated the network. Peixoto (2015) studied a network of face-to-face interactions among high school students recorded with a high temporal resolution by social sensors. He showed that a blockmodel where students remain in the same groups over time best describes the dynamic network. This implies that the student network keeps on reproducing its macro behavioral state via a feedback mechanism.

These examples exploit the social dimension of transactions. But the analysis of natural language–the content dimension of transactions–is a straightforward way of studying the co-constitution of micro and macro behavior (Bail, 2014). As in the previous subsection, this is the stomping ground of natural language processing and machine learning. Topic models are generative models for automated text analysis. They are capable of uncovering latent macro behavioral patterns from which transactions are assumed to be produced. They are relational methods because topics are sets of words that are used together. For example, Stier et al. (2018) took an initial set of topics from survey data and found that politicians use different topics on social media. The digital patterns of discourse are diagnostic of micro behavioral practices: Facebook is heavily used for election campaigning, while Twitter serves as a channel for engaging in policy debates.

## Challenges of Using the "New Telescope"

Social science's "new telescope" is a metaphor for the availability of massive amounts of DBD and the tools to analyze it. DBD is genuinely relational. Established techniques such as network and text analysis take advantage of this relationality and belong to the main pillars of CSS. As this field grows in relevance, social scientists will have to familiarize themselves with the potential, the restrictions, and the challenges of DBD's methodological innovations. In our view, there are at least five core challenges for contemporary and future research.

(1) ***Data management.*** Explanations in CSS derive their power from the volume, variety, and velocity of DBD. Yet, this very complexity implies the need to learn how to manage such data (e.g., to link, aggregate, and analyze highly relational information). Today's social scientists must master not only statistical methods (as in past decades) but also advanced techniques for handling DBD.

(2) ***Data quality.*** A fundamental restriction of DBD is that the reconstruction of the users' subjectively intended meaning is not possible based on observed behavior. More fundamentally, some actors may not even be humans but bots. Thus, the advantages of DBD are accompanied by severe restrictions regarding the validity of constructs. In addition, various measurement and representation errors are possible in the research lifecycle. It is therefore recommended that these be identified, explicitly pointed out, and measures taken to alleviate them (e.g., complementing DBD with survey data; Sen et al., 2021; Schmitz & Riebling, forthcoming).

(3) ***Reproducibility.*** More recently, empirical social science has become increasingly

concerned with the need for studies to be reproducible. But this is precisely what is often difficult in the context of DBD. There is a trade-off inherent to DBD between its rich information on social relations and communicative content and its limited open availability. While several initiatives are underway to open the "closed shops" of private platform providers, researchers can already share their computer code to at least increase reproducibility by allowing others to re-run the analysis. For example, the GESIS Notebooks service at notebooks.gesis.org allows for executing computer code in a browser window without having to install a programming language (cloud computing).

(4) ***Reflexivity.*** Polar attitudes towards CSS–either fundamental rejection or uncritical embrace–are insufficient: It is true that these data and methods must be approached, but in doing so, one must clarify the conditions under which they are produced as well as their analytical limitations. To adequately employ the "new telescope," scientists must better understand its underlying architectures and the modes of data generation, including the fundamental role of artificial intelligence and machine behavior in affecting and producing social phenomena (Wagner et al., 2021). Ultimately, working with DBD also has an ethical component.

(5) ***Theory.*** Provided that found DBD is a byproduct of digital platform operations, all research with it is inevitably data-driven to some extent. Nevertheless, only the use of theory guarantees that knowledge is produced in a meaningful and cumulative way. Given the potential of CSS to contribute to solving pressing issues on a global scale (e.g., sustainability), one way is to develop and apply solution-oriented middle-range theories (Watts, 2017). Yet, DBD represents a promising research field for the plurality of social science's paradigms: Beyond Social Network Analysis, the social sciences offer a wealth of established and elaborated perspectives on the social, such as practices, mechanisms, discourses, systems, fields, and functions. DBD represents a promising strategic research site for such different theories to be employed, developed, and adapted to digital life. These different approaches may prove to be useful in transcending the individual as unit of observation. They can become vivid communication paths between the social and computational sciences and, ultimately, a constitutive pillar for the consolidation of the field of CSS.

## References


Bail, C.A. (2014). The cultural environment: Measuring culture with Big Data. *Theory and Society,* 43(3–4), 465–482.

Diaz, F., Gamon, M., Hofman, J.M., Kıcıman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLoS ONE,* 11(1), e0145406.

Emirbayer. M. (1997). Manifesto for a relational sociology. *American Journal of Sociology,* 103(2), 281–317.

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), 767–797.

Keuschnigg, M., Lovsjö, N., & Hedström, P. (2018). Analytical Sociology and Computational Social Science. *Journal of Computational Social Science,* 1(1), 3–14.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences,* 110(15), 5802–5805.

Kossinets, G. & Watts, D.J. (2009). Origins of homophily in an evolving social network. *American Journal of Sociology,* 115(2), 405–450.

Lazer, D. & Radford, J. (2017). Data ex machina: Introduction to Big Data. *Annual Review of Sociology,* 43(1), 19–39.

Lazer, D., Pentland, A., Watts, D.J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M.J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational Social Science: Obstacles and opportunities. *Science,* 369(6507), 1060–1062.

Mønsted, B., Mollgaard, A., & Mathiesen, J. (2018). Phone-based metric as a predictor for basic per-

sonality traits. *Journal of Research in Personality*, 74, 16–22.

Nassehi, A. (2019). *Muster: Theorie der digitalen Gesellschaft*. München: C.H. Beck.

Peixoto, T.P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4), 042807.

Schaible, J., Oliveira, M., Zens, M., & Génois, M. (2022). Sensing close-range proximity for studying face-to-face interaction. In U. Engel, Quan-Haase, A., Liu, X., & Lyberg, L. (Ed.). *Handbook of Computational Social Science* (vol. 1, ch. 14). London: Routledge.

Schmitz, A. & Riebling, J. (forthcoming). Data quality of digital process data. A generalized framework and simulation/post-hoc-identification strategy. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*.

Sekara, V., Stopczynski, A., & Lehmann, S. (2016). Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences*, 113(36), 9977–9982.

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.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Political Communication*, 35(1), 50–74.

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, 595, 197–204.

Watts, D.J. (2011). *Everything Is Obvious*: *Once You Know the Answer*. New York, NY: Crown Business.

Watts, D.J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, 1, 0015.

White, H.C. (2008). *Identity and Control: How Social Formations Emerge*. Princeton, NJ: Princeton University Press.

**Haiko Lietz**

GESIS – Leibniz Institute for the Social Sciences

*E-mail* Haiko.Lietz@gesis.org

Haiko Lietz is a post-doc at GESIS, Cologne. His research interests lie in applying and developing relational theory and methodology by integrating sociology, complexity theory, and computational approaches.

**Andreas Schmitz**

GESIS – Leibniz Institute for the Social Sciences

*E-mail* Andreas.Schmitz@gesis.org

Andreas Schmitz is a researcher at GESIS, Cologne. His main research interests are relational social theory, relational methodology, the interplay between CSS and social theory, applied statistics, and generalized field theory.

**Johann Schaible**

EU|FH - Europäische Fachhochschule Rhein / Erft GmbH

*E-mail* j.schaible@eufh.de

Johann Schaible is a professor for applied computer science at the university of applied sciences EU|FH, Bruehl. His main research interest comprises Smart Cities with a focus on human mobility and in general on spatio-temporal data analysis.