Political Communication

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/upcp20

Computational Social Science and the Study of Political Communication

Yannis Theocharis & Andreas Jungherr

To cite this article: Yannis Theocharis & Andreas Jungherr (2021) Computational Social Science and the Study of Political Communication, Political Communication, 38:1-2, 1-22, DOI: 10.1080/10584609.2020.1833121

To link to this article: https://doi.org/10.1080/10584609.2020.1833121

Published online: 21 Oct 2020.

POLITICAL COMMUNICATION
2021, VOL. 38, NOS. 1–2, 1–22
https://doi.org/10.1080/10584609.2020.1833121

Routledge
Taylor & Francis Group

Computational Social Science and the Study of Political Communication

Yannis Theocharis [a] and Andreas Jungherr [b]

[a]Centre for Media, Communication and Information Research (Zemki), University of Bremen, Bremen, Germany; [b]Institute of Communication Science, Friedrich-Schiller-University Jena, Jena, Germany

ABSTRACT

The challenge of disentangling political communication processes and their effects has grown with the complexity of the new political information environment. But so have scientists' toolsets and capacities to better study and understand them. We map the challenges and opportunities of developing, synthesizing, and applying data collection and analysis techniques relying primarily on computational methods and tools to answer substantive theory-driven questions in the field of political communication. We foreground the theoretical, empirical, and institutional opportunities and challenges of Computational Communication Science (CCS) that are relevant to the political communication community. We also assess understandings of CCS and highlight challenges associated with data and resource requirements, as well as those connected with the theory and semantics of digital signals. With an eye to existing practices, we elaborate on the key role of infrastructures, academic institutions, ethics, and training in computational methods. Finally, we present the six full articles and two forum contributions of this special issue illustrating methodological innovation, as well as the theoretical, practical, and institutional relevance and challenges for realizing the potential of computational methods in political communication.

KEYWORDS

Computational Social Science; Computational Communication Science; Digital Trace Data; Political Communication; Big Data

Digital communication has opened up a host of new avenues for social and political interactions. These have radical effects on political information environments and the democratic attitudes and behaviors they shape (van Aelst et al., 2017). Not only are citizens able to produce content and have their voices heard in ways that were inconceivable two decades ago, but systemic changes within the broader media ecology, such as the expansion of choice in the media environment and the increasingly important role of social networking sites as sources of political information, are changing traditional political information production, distribution, and consumption dynamics (Jungherr et al., 2020b). Developments such as the mass supplanting of political information with entertainment, the diversification of media diets, the fragmentation of the media environment and the rising proliferation of misinformation, have not only impacted political communication processes, but have also added new challenges to the study of political communication. At the heart of these, there continue to lie questions political communication scholars have always asked: how can we reliably measure the reach of specific media outlets or political actors, identify the often-overlapping media and information

CONTACT Yannis Theocharis yannis.theocharis@uni-bremen.de Centre for Media, Communication and Information Research (Zemki), Linzer Str. 4, Bremen 28359, Germany

diets of people, and estimate the effects of information – especially in new, noisy and deeply confusing information environments?

But while the challenge of disentangling political communication processes and their effects has grown with the complexity of the political information environments, so have our toolsets and capacities as social scientists to better study and understand them. The sinking costs of computational power and broad access to data science toolkits, formerly confined to either highly specialized communities or to particular disciplines, have provided access to a wide variety of data containing political information, novel data collection practices, and a stream of new methods with which to make sense of them. These developments are often discussed under term Computational Communication Science (CCS) (van Atteveldt & Peng, 2018). We treat as computational *political* communication the research developing, synthesizing, and applying data collection and analysis techniques relying primarily on computational methods and tools, with the objective to answer substantive theory-driven questions in the field of political communication.

For those wishing to study political communication processes using entry-level computational methods and tools, CCS is very accessible (for examples see Habel & Theocharis, 2020). But critical questions emerge. Do the data we can access really help us answer the substantive questions we are interested in? Is the mere adjustment of R code from a tutorial enough to establish reliable findings? Can a colorful visualization in *Gephi* alone provide sufficient evidence that Twitter conversations around a hashtag are polarized? What theoretical puzzle does a purely descriptive overview of how people connect in online discussions help us answer, and what theory-building capacity do such insights have? How certain can we be that all those bots identified by *Botometer* are really non-humans? How do these methods square with developments toward open science, data sharing, and replicability?

There is little doubt that new types of data and methodological approaches allow a new vista into both existing and new communication processes. But while the fact that we can observe and describe a political communication process from a viewpoint that has so far not been attainable can sometimes be instructive in itself, does it also mean that we can necessarily say something new, meaningful or theoretically interesting about it? And does the fact that new and alternative ways of measuring our concepts are now at our disposal mean that these are not ridden with the same problems our previous approaches suffered from?

Much of computational communication research currently strives (and often achieves) to be not only conceptually clear and theory-driven, but to deploy sophisticated analytical methods that are rigorously validated and made transparent through openly accessible replication repositories. But the apparent 'magic' of computational tools and methods and the ease with which you can summarize certain types of information in seemingly insightful ways, might lure unsuspecting researchers into reading too much into their findings. In this, the apparent ease of the use of computational methods threatens achieving meaningful and valid insights. Even worse, despite the efforts of experts to make their tools available with immense documentation and support to everyone who has the will, resources, and capacity to invest time in building relevant skillsets, we feel that a rift runs along the lines well-known in other resource-intensive scientific fields: between those low and high in resources, those with access to high-quality data – including proprietary data from digital platforms – and those without, those with little and much institutional support, those in United States and elsewhere. In this, growing uses and demand for extensive computational methods and datasets thus risk exacerbating existing inequalities in the opportunities for contributions in the social sciences at large, and political communication research in particular. This is an issue many journals – including this one – and interest groups in the field increasingly commit to addressing.

Against this backdrop, the goal of this contribution, and of this Special Issue, is to foreground the theoretical, empirical, and institutional opportunities and challenges of CCS that are relevant to the political communication community. We believe that, despite its vast potential, as of yet CCS has only had marginal impact on core tenets in the field of political communication. New computational methods remain ill-connected with established approaches to social science research, and findings predominantly speak to isolated single-country cases. One reason for that is that computational methods are still pursued by a minority of political communication researchers (though this is changing rapidly, as the elevation of the ICA Computational Methods interest group – founded only in 2016 – into a division within a very short time, testifies). While one might argue that, until relatively recently, this was also the case with quantitative methods more broadly, the same need not be with computational methods. Technical knowledge, formerly mostly found in expensive methods textbooks, is now available in the form of both online guides/tutorials and code repositories with detailed instructions (Stan, for example, a language for Bayesian inference and optimization, comes free with open-source code and a 500-page manual) that can proliferate at a much faster rate and are accessible to a larger audience for free.

Another important barrier is that much of CCS research appears to lack connections to relevant theories, deploys measures that can be questionable, its capacities to reveal novel aspects of political communication processes are often misunderstood, and it remains largely descriptive or, on the other end, it can sometimes showcase methodological rigor at the expense of well-defined theoretical mechanisms. These are all understandable symptoms of an interdisciplinary field that has not yet matured, and they can, as numerous textbooks demonstrate, also be encountered in other types of social science research to a greater or lesser extent (Kellstedt & Whitten, 2018). Yet, as this is not the first time social science researchers are confronted with many of these issues), we have the advantage of learning from these enduring controversies and shortening the curve of progress (the use of content analysis since the 1950s, and more than 60 years of public opinion/survey research have taught us a lot about theory-testing and valid measurement – see, for example, Barberá, 2020). Finally, and this is something not based on quantitative indicators but rather comes from our experience and discussions with colleagues in relevant interest groups, CCS's highly interdisciplinary nature makes it institutionally cumbersome, often creating imbalances on who can substantively contribute to this research, but also on how scholars with relevant interdisciplinary expertise can position themselves on the market.

With this special issue, we plan to map the potential of CCS for the political communication community and demonstrate its broad appeal beyond that of highly technically skilled researchers, focusing on approaches and perspectives that not only demonstrate its methodological innovation but, most importantly, illustrate its theoretical, practical, and institutional relevance and the challenges in realizing its potential.

Defining Computational Communication Science

We position *Computational Political Communication* within the subfield of *Computational Ccommunication Science (CCS*), which itself is a variant of *Computational Social Science (CSS)* (Lazer et al., 2009). CSS is a developing interdisciplinary scientific subfield that is still lacking clear demarcations. We define *computational social science* as an interdisciplinary scientific field in which contributions develop and test theories or provide systematic descriptions of human, organizational, and institutional behavior through the use of computational methods and practices. On the most basic level, this can mean the use of standardized computational methods on well-structured datasets (e.g., applying an off-the-shelf dictionary to calculate how often specific words are used in hundreds of political speeches), or at more advanced levels the development or extensive modification of specific software solutions dedicated to solving analytically intensive problems (e.g., from developing dedicated software solutions for the automated collection and preparation of large unstructured datasets to writing code for performing simulations). Accordingly, CCS, and by extension *Computational Political Communication*, lie at the intersection of CSS and (political) communication, with a topical focus on theories and phenomena associated with communicative channels, objects, behavior, and effects.

The definition points to an important point of tension in precisely differentiating CSS from other fields in the social sciences. Nearly all contemporary work in the social sciences relies on computational methods. This includes the storage and processing of digital data – such as digital text, image, or audio files; computationally assisted data analysis – such as regression analyses and simulations; or data collection through digital sensors – such as eye tracking or internet of things enabled devices. In this work, computation is often a necessary precondition. For example, while it is possible to run multiple regressions with pen and paper, the success of this method in the social sciences depends on the digital representation of the underlying datasets and computational resources available to process the data. As in the most general reading of our definition the use of any computational method in data handling and analysis would qualify as computational social science, one could argue that nearly any form of contemporary social science would constitute computational social science. Obviously, this is not helpful in identifying constituting elements of the field and subsequent potentials and challenges.

It might be helpful to focus more on studies and research projects in which computational methods and practices are not used as plug-and-play solutions but instead demand for varying degrees of customization with regard to data collection, preparation, analysis, or presentation. Again, this is best thought of as a distinction in degree. On one end of the scale, we find projects that require some coding with regard to the sequential calling of preexisting or slightly modified functions or data management. On the other end of the scale we find research projects that demand the development of dedicated software solutions, for example, in automated and continuous data collection, preparation and structuring of large unstructured raw data, or the development of dedicated non-standardized analysis procedures. Projects at different ends of this scale share issues arising from their focus on social behavior, systems, or phenomena but they vary significantly with regard to their computational demands. Projects that use standardized computational methods might thus be basically indistinguishable from other areas in empirical social science research. On the other hand, projects at the other end of the scale are likely to face challenges indistinguishable from software development in computer science.

Often, CSS is discussed in the context of new datasets that have been made available through digital technology. This very famously includes data documenting user behavior in digital environments – so-called digital trace data (Freelon, 2014; Golder & Macy, 2014; Howison et al., 2011; Jungherr, 2015). Increasingly, however, other large datasets that are relevant for political communication research have become available digitally, such as large text corpora covering fields as diverse as newspaper coverage (Barberá et al., 2020), literature (Piper, 2018; Underwood, 2019), historical or contemporary parliamentary speeches (Rauh & Schwalbach, 2020), and images (Williams et al., 2020). All these datasets are legitimate objects of computational communication, and it is thus unnecessarily limiting to restrict one's definition of CSS to specific types of datasets.

Correspondingly, we find it unproductive to limit one's definition of CSS to one specific topical subfield. It is true that much early work in CSS focused on digital communication environments, but to us this is an artifact of early availability of datasets documenting user behavior on social media – especially Facebook and Twitter – and not a constitutive feature of CSS. Our understanding of CSS is thus not tied to a specific set of methods, datasets, or research interests. Instead, to us, the constituting element of CSS differentiating it from other approaches in the social sciences, and in political communication in particular, is the degree to which research projects demand the inclusion and development of computational methods over the course of a project. At the same time CSS is a specific subfield in computer science research in that it focuses on social systems and phenomena. Consequently, approaches and methods have to account for the specific conditions of this research area (Flyvbjerk, 2001). By focusing on these two constitutive characteristics of CSS – the examination of social systems, phenomena, and processes based on computational methods – we can identify and discuss the associated promises and challenges encountered in this field.

Promises, Promises

Accounts of CSS are usually accompanied by strong expectations with regard to the promises they hold for the study of societies and human behavior. This also holds for the study of political communication. Promises usually come in two forms: The first focuses on the increased coverage of social phenomena and human behavior through digital trace data and digital sensors, the second goes even further and expects a transformation of the nature of the social sciences.

On the most fundamental level, proponents of CSS agree in that the digital transformation has led to a massive increase in available data sources and types for social scientists. This is true for data that in principle had been available before but now are available at a significantly larger scale, such as newspaper corpora. Beyond this, we also have new data sources. For one, the interactions of users with online services create data traces. In principle, these digital trace data provide a comprehensive account of user behavior with, and mediated by, digital services, and can additionally provide environmental details that were previously impossible to acquire. For this reason, they are highly promising to political communication researchers as they allow an entry point for investigating processes and behaviors within what are probably the most vibrant political information environments of our time (Golder & Macy, 2014; Howison et al., 2011; Jungherr, 2015; Salganik, 2018).

In practice, however, most political communication researchers only have access to highly limited snap-shots of digital trace data and remain at the mercy of digital platforms regarding data access (Freelon, 2018). The upside of this is that this type of access is still better than not having even this narrow corridor into the data streams of platforms where much of today's exciting communication – and politics – happens. At the same time, however, it has made the realization of the full promise of digital trace data more elusive than originally hoped for, by adding strict barriers to the inferences that can be made about what these snapshots exactly represent and how, in the end, we can extract from massive data corpora something that is actually useful (Grimmer & Steward, 2013). Beyond this, we also encounter new data sources provided by digital sensors. This could be data emerging as a byproduct of another service, like satellite imagery (Weidmann & Schutte, 2017), or the output of sensors specifically designed by researchers (Pentland, 2008; Stopczynski et al., 2014). In principle, this data type is bound to increase with the availability and wide distribution of Internet of Things devices.

In combination, this increase in available data sources allows for an increasing coverage and surfacing of social phenomena and human behavior. It also enables the examination of well-known phenomena at much higher temporal, behavioral, and procedural resolution especially when combined with other methods of social science inquiry. This might also allow for a more systems-level view of societies and human behavior (Golder & Macy, 2014; Lazer et al., 2009; Salganik, 2018). As an example, while the phenomenon of misinformation has in the past been studied using surveys and survey experiments (e.g., Kuklinski et al., 2000), organizing – and eventually matching – digital traces with individual-level data can now provide a far greater depth into mechanisms of exposure to dis- or misinformation (Grinberg et al., 2019; Guess et al., 2019) providing insights that would not be possible if survey data were used alone. Here researchers are not only using more and different data than before, but to achieve their goals they engage in high degree of customization when it comes to both the data generation and the analytical process.

More ambitiously, the availability of vast datasets documenting human behavior in interaction with digital services – or covered by digital sensors – has also led to the expectation that social sciences might transcend their status of a soft science into an actual scientific discipline with models allowing for the confident prediction of the future. In this view, more data do not only mean an increase of the coverage of social processes or human behavior but actually would allow for a "measurement revolution" (Watts, 2011) in the social sciences, allowing them to overcome their current state of after-the-fact explanation and develop into a science with true predictive power (Hofman et al., 2017). This hope rests on a view of society as being shaped by underlying context-independent laws that have mostly remained invisible to scientists due to the lack of opportunities to acquire data that can now be accessed (González-Bailón, 2017).

While we increasingly see studies that illustrate the first promise of CSS based on the expansive coverage of social phenomena, the second promise of a transformation of social science into a more strictly predictive science remains unfulfilled, especially when one looks at political communication research. While one might treat this as an indicator that we simply need even more data, we feel it is more plausible that the nature of the social sciences is the examination of context-dependent phenomena (Elster, 2015; Flyvbjerk, 2001; Gerring, 2012), and as such prediction in the social sciences is more an instrument of theory-testing and not an instrument of planning and design, as for example, in engineering or physics.

Overall, while these promises have been well articulated and prominently advanced, the challenges of realizing them remains predominantly buried in the discussion section of empirical papers. But looking back into what is now more than a decade of research allows us to identify at least three problem areas:

- CSS research continues to remain weak in connecting its research designs and findings to established theories, concepts, mechanisms, and discussions in the social sciences (Jungherr & Theocharis, 2017);
- While problems with data – especially when it comes to social media data – have been the subject of considerable debate in CSS (Japec et al., 2015; Sen et al., 2019; Stier et al., 2019), data generating processes and their effects on the composition, coverage, and interpretative meaning of signals in available datasets (Jungherr, 2019) are often treated as issues of – at best – secondary importance;
- CSS as an interdisciplinary research field struggles with establishing practices that connect it more strongly within the established social sciences, develop standards of transparency in data collection, preparation, harmonization and analysis, and surface and problematize conflicts of interest between researchers, industry, and the media (Jungherr et al., 2020a). Steps that have been taken to address especially the last of those issues (King & Persily, 2019) have been met with skepticism (Bruns, 2019).

For CSS, and therefore also computational political communication, to flourish and transcend its current niche existence among computational enthusiasts, among social scientists and socially curious computer scientists, these challenges have to be addressed.

Challenges

Data and Resource Requirements

Challenges in computational communication are not unlike those of CSS. The feature most often associated with CSS is probably the extraordinary size of datasets (Salganik, 2018). This development has given rise to the term “big data” in order to discuss associated research potentials (Lazer & Radford, 2017; Schroeder, 2016, 2019), though as of recently the term has lost some of its popularity given a growing awareness of its conceptual ambiguity.

On a very fundamental level, increasing sizes of datasets bring practical issues in their storage and processing. While it is true that processing power of computers increases, this does not make up for the increasing demands put on them through new types of datasets. This is already true for textual data, and all the more true for datasets with high-resolution images or videos which are becoming of increasing utility to political communication scholars, especially as research interest switches to platforms such as Instagram, YouTube, and Tik Tok. Projects collecting and using respective datasets increasingly face non-trivial data preparation and processing tasks that go well beyond the scope of typical projects in the social sciences.

Beyond these fundamental issues in the handling of large datasets with diverse types of data, further issues emerge from the mere fact that projects in political communication often use social media data collected through public interfaces provided by social media platforms. Until recently, this data access through so-called Application Programming Interfaces (APIs) was easy and provided researchers with comparatively rich data. But while some platforms gradually take steps to facilitate access and use of their data by academic researchers, others have come to restrict public data access through APIs thereby limiting the opportunities for collecting datasets of high quality. While some researchers have advocated for partnerships with social media companies as a possible way of mitigating risks related to current constraints in terms of reliability and reproducibility, and preserving user privacy by gaining access to research-grade data (Puschmann, 2019), others have advocated the development of dedicated data collection solutions independent from the access provided by platforms (Freelon, 2018). These examples illustrate that these data do not only hold potentials but also come with significant challenges. This makes it an area that increasingly demands for interdisciplinary teams of computer and social scientists to handle these demands (King, 2011).

Large datasets also bring considerable privacy concerns, data ownership and subsequent unresolved issues of research transparency and replicability. Also, in practice, it is likely that most users of online services remain unaware that their public contributions and interactions -as well as associated metadata - might be visible to others and subject to research projects in which these data are often used to infer their preferences, traits, or characteristics. This is even more true for the data from digitally enabled sensors or Internet of Things devices – all data sources of growing interest. Additionally, datasets with many data points per individual raise significant challenges to guarantee that it is not possible to identify the identity of individuals. This makes it difficult for companies to provide researchers with access to data (King & Persily, 2019; Levi & Rajala, 2020) and also raises issues after the completion of projects. While there is an increasing awareness in the sciences for transparent research practices and access provision to data underlying research projects (Christensen et al., 2019), privacy concerns and the often-proprietary nature of the underlying data make it much harder to establish similar standards. This often leads to papers remaining opaque when compared to developing transparency standards in the social sciences (Jungherr et al., 2020a).

Links: Theory and the Semantics of Digital Signals

A more hidden challenge in CSS, and therefore by extension to computational communication, lies in linking data and results to phenomena of interest and to relevant theories (Jungherr et al., 2020a). Social science is a field of highly contextual dependent findings, making grand-theorizing untenable (Flyvbjerk, 2001). Still, linking studies to existing theoretical mechanisms that have been shown to be at work when it comes to certain processes, phenomena or behaviors, allows researchers to build plausible research puzzles and establish a link to previous knowledge. This enables an assessment of what part of our previous understanding of reality is supported or contradicted by novel findings. Only this active linking between established theoretical ideas and novel research environments, phenomena, and methods will allow for the emergence of a cumulative body of evidence instead of a palimpsest of ill-connected and isolated findings (Schroeder, 2019).

The power of theory-driven research in computational communication can be illustrated by a recent study that also illustrates the aspect of customization with regards to the research design and analysis that we discussed earlier. Much of the current debate about the downsides of social media use is that it supposedly creates "echo chambers", and as a result people are not sufficiently exposed to information contradicting their prior beliefs. Scholars concerned about the adverse effects of echo chambers suggest that communication with people "from the other side" can enhance exposure to diverse viewpoints thereby reducing polarization. In a recent study, Bail et al. (2018) assess this hypothesis, but also go further and theorize the existence of a rival mechanism according to which such interactions have a backfire effect. To address this puzzle, they deployed an ingenious experimental research design that combined survey research, bot technology, and Twitter data. Their results revealed significant partisan differences in backfire effects, opening up new avenues for further exploring mechanisms behind the created backfire effect among Republicans in particular, and which their study was not able to reveal. Despite its limitations (discussed extensively by the authors), this theory-driven study is instructive in its deployment and craftiness of computational methods to better understand specific communication processes.

Experimental studies, such as this one, are a highly promising arena in computational methods-powered political communication research in which treatments can be rolled out in realistic environments for participants of unprecedented counts (Bail et al., 2018; Leeper, 2020; Salganik & Watts, 2009; Siegel & Badaan, 2020). But while this allows for high experimental control and the identification of small effect sizes, once the number of available observations increases, researchers also have to adjust the criteria by which they interpret results (Japec et al., 2015).

One example for the necessity of this adjustment is offered by Bond et al. (2012). The authors present a highly innovative experimental study in which they ran an experiment with 61 million Facebook users, a selection of whom they showed information if their friends had indicated that they had voted in a US-election. The authors identified very small effects of a specific variation of this information treatment. While they are careful in mentioning this as a limitation in the text of their paper, in the abstract and conclusion they speak prominently about the success of influencing people through information on Facebook. This and not the more careful reading accounting for effect sizes has come to dominate references of this paper in both scientific and public discourse. Here, the cautioning by the authors themselves goes out of the window and the study is predominantly cited as evidence for the tremendous manipulative power of Facebook in political communication and elections. In popular reception the high number of participants might thus have worked as a cue of the importance of the findings, when in fact the large participant count waters down the relevance of the reported statistical significance. This prominent study illustrates the necessity for researchers to adjust their reporting practices to the new conditions of big datasets.

Beyond theory, there is another link currently predominantly neglected in CSS, the one between signals found in data and the phenomena of interest to a study, the semantics of digital signals (Jungherr, 2019). Consider the following example that should be familiar to political communication scholars. Anyone doing research on social media knows that any given data point is a symbolic representation. It could represent something direct and unambiguous, such as for example, when a user clicking "like" below a post on Facebook about a friend's solidarity statement with the Black Lives Matter movement. But the data point could also represent something more indirect. For example, the "like" on Facebook could be an expression of support under a post voiced by another user, it could represent an agreement with a factual or interpretative statement, it could be an expression of sympathy to another user, or an act of social capital maintenance totally devoid of any direct connection to the content of the liked post. While one could argue that survey researchers are often faced with similar issues, linking signals with behaviors in social media environments is a much riskier endeavor. This is not only because of the diverse architectures of the platforms and the multiple social cues afforded by the specific activities offered by each platform's embedded functions. It is also because of the hard-to-measure error associated with the meaning assigned to specific acts. The semantics of signal and represented object might vary not only over time but also between types of content (e.g., text, images, or video) and services, especially given what we know about variation in platform affordances (Jungherr & Jürgens, 2013). This makes it crucially important for researchers to explicitly state their interpretation of the relationships between signals and objects of representation in order to foreground their underlying assumptions. Currently, the phenomena directly represented by digital traces are often conveniently ignored. Instead, scholars tend to project their interests onto signals found in digital trace data without worrying too much about establishing the link between their chosen signal and the phenomenon of interest (Jungherr et al., 2017).

Labels: Can We Infer People's Traits Based on Digital Traces?

The practice of labeling in CSS is tightly connected to the issues arising from establishing the semantic link between signals in data and the phenomena they are supposed to represent. One prominent feature in CSS is the attribution of labels to individuals or digital avatars based on behavior manifested in their digital traces. Examples abound. Activity on social media has been used to label users according to their political ideology (Barberá, 2015), psychological traits (Azucar et al., 2018), mental health (Chancellor & Choudhury, 2020), or the authenticity of their account (Rauchfleisch & Kaiser, 2020). Labels are powerful tools in CSS as they allow large-scale automated assignment of interventions based on perceived traits or preferences of users. In this, they resemble scoring procedures in other fields (Citron & Pasquale, 2014). Unsurprisingly, this raises a host of concerns and demands on researchers providing labeling propositions and solutions.

For academic researchers it is completely legitimate to identify correlations between signals in digital traces and other metrics documenting traits, preferences, or expected behavior by individuals. It is something else entirely if these labeling solutions become the basis of business models or policy interventions. For this, as the *Cambridge Analytica* case forcefully demonstrates, much greater scrutiny and public oversight is needed. For one, the discussion about labeling tends to focus on its surprising ease. Seemingly successful cases, once established in the public imagination, turn out to be hard to dislodge even if academic discussion moves beyond early enthusiasm to a more critical stance. Take the discussion about the supposed prevalence of (semi-)automated accounts – so-called bots –

in public discourse (Schneier, 2020). Here public and political imagination seems obsessed with visions of public debates in online spaces being overrun by manipulative and inauthentic accounts pushing narratives challenging to the political status quo. While methods of labeling social media accounts as bots abound (Varol et al., 2017), findings have been mixed (Keller et al., 2020). Increasingly, early enthusiasm is replaced by skepticism. Careful examination shows that methods labeling accounts as bots have not proved to be reliable, with mislabeling of actual authentic and legitimate accounts as bots (false-positive) and strong temporal decay in precision (out of sample prediction) (Rauchfleisch & Kaiser, 2020).

Labeling undoubtedly matters (Pasquale, 2015) and automatically labeling social media users as bots/-supporters of a political candidate in particular brings risks. While these risks remain manageable when labeling efforts stay in the confines of academic papers, they grow exponentially once unvetted and unsupervised labeling solutions are deployed widely on online platforms and become the basis for the roll-out of automated interventions. In the current heightened political climate and opaque governance practices of platforms, the scholarly community should be doubly careful with regard to how labels are assigned, how they are audited, and as the basis of which interventions they serve. Silencing automated accounts might be legitimate in specific circumstances. Silencing users who for some reason or other have been labeled as “bots” decidedly less so.

As it stands, CSS has not yet reflected this ethical responsibility in its practices sufficiently. Overall, the case of bot detection illustrates the need for CSS to demand much more vigilant reliability, validity, and robustness checks of proposed labeling procedures as overly-enthusiastic prototypes might develop hard to control societal effects that can become difficult to curtail once the collective imagination takes possession of them.

The Shock of the New: Theory-driven Work as Stabilizer

Digital technology has expanded the reach of everyone and changed the composition and processes in social systems. While some political communication processes and phenomena can be parsed surprisingly well with existing scientific theories – see for example, the rich theoretical literature on digital media and polarization – others are inherently new or at least sufficiently different with regard to dynamics, reach, or effects as to demand new conceptualization and potentially even new theories (Neuman, 2016; Schroeder, 2018). Examples include the discussions on disinformation (Lazer et al., 2018) or the nature and effects of uncivil discourse in online communication spaces (Munger, 2017; Theocharis et al., 2020, 2016). In a fluid, high-choice, and complex political information environment, political communication researchers are therefore especially well-placed to employ theory-driven designs, as discussed above, to allow for the linking of findings to established discourses, but also to drive theorizing in order for the field to account for the changes to human behavior, institutions, and social structures we are living through.

At the time of writing, scholars are actively working on the impact of the COVID-19 pandemic. An event of momentous significance, it has elevated scientific and public communication to new heights. It is also triggering concerns about a number of issues, including how people change their information diets in response to the crisis, how racism and xenophobia in online conversations is impacting the health of online debates, how much misinformation is out there and what its consequences for public health might be. In times of social isolation, media consumption of all forms is bound to increase and the pandemic will, no doubt, offer itself for exciting research in the years to come, not least because never before has a singular and life-threatening event dominated media and communication for so long. That computational methods are going to play an important role in research associated with the event is in no doubt.

But, momentous though the COVID-19 pandemic may be, our existing understanding of political communication processes during major events (especially if it only pertains to public behavior on Twitter) does not exactly present us with momentous theoretical problems. We already have, for example, expectations and solid knowledge about conversation dynamics and information diffusion on Twitter in particular. We are also well informed as to why social media could prove pivotal for enabling groups and individuals to organize solidarity and collective action in their neighborhood, why certain types of misinformation will potentially polarize certain age groups but not others, and why social media affordances might contribute to certain individuals garnering support using falsehoods instead of facing complete demolition in the polls. Our aim here is not to pre-judge, or even give indication of what is, good and bad research. Nor is it to say do this research and not the other, or use this set of theories and tools and not the others. Rather our goal is to point out possible pitfalls in certain approaches that have become increasingly common, and explain the characteristics of research that we believe could, by now, be better represented and which can help the field address important questions in a more interesting manner.

The flip-side of theorizing, especially in new and unfamiliar environments or contexts, is the need for extensive and diligent descriptive work in order to map new phenomena widely and systematically (Swedberg, 2014). An inherently welcome maturation process in the social sciences over the last decades has given rise in some areas to neglect and active discouragement of descriptive work. While this might be justified for areas in which others are doing the descriptive heavy lifting – such as journalists, lawyers, or historians – in CSS it would be a mistake to adopt this attitude. The task of mapping the effects of the digital transformation in various fields and across cultural, temporal, or national contexts is far from trivial and crucial to public understanding and the further development of the field.

Practices: The Key Role of Infrastructures, Institutions, Ethics, and Training

CSS, an interdisciplinary field at the borders of various social sciences, computer science, and even some natural sciences, is having a significant impact on research practices in the social sciences and political communication in particular. While some communication is happening at the borders between researchers coming from these fields, in practice everyone brings the practices and standards from their original field to new endeavors in CSS. It thus comes as little surprise that CSS is not dominated by a coherent theoretical tradition, specific methods, or datasets. Instead we find a myriad of approaches and standards at work. This makes it difficult to find a coherent language and to develop a framework under which empirical findings from different traditions following different standards to contribute to a cumulative account of research (Schroeder, 2019).

While the dominant overviews of CSS clearly reflect the interdisciplinary nature of the field, in practice this is difficult to realize in research teams (Gilardi et al., 2020). While there are a number of high-profile CSS groups – predominantly in the USA – that are able to assemble interdisciplinary teams, in most academic contexts, field-specific hiring practices make this difficult. In practice, we either find loose interdisciplinary assemblages of research groups that within themselves remain more or less homogenous, or research groups situated in one field that try to pick up necessary skills from different fields on the fly. Of these three options, the first – establishment of dedicated interdisciplinary research groups – appears the most promising one regarding the development of CSS as a coherent field and allowing work on tough challenges. At the same time, employment in such a team might be risky for PhD students and post-docs as at present it is unclear if comparable teams spring up in sufficient size as to provide further employment opportunities and if more traditional job searches recognize their experience in interdisciplinary teams as valuable. The second option – loose interdisciplinary alliances between in themselves homogenous research groups – are somewhat risky with regard to the fragility of these efforts but still contribute to an interdisciplinary dialogue and potential standardization of CSS as a field. It also brings somewhat smaller risks for PhD students or post-docs as they might be working on interdisciplinary projects, but they maintain their affiliation with a clearly identifiable unit in their respective field. The third option – a homogenous group picking up skills from other areas on a need-to-know basis – is probably the most fragile option that at the same time contributes little if nothing to a standardization of CSS as an interdisciplinary field.

We stress here that true interdisciplinary research and teaching is not only difficult to attain practically, but requires the institutional open-mindedness and resources which, rhetoric aside, few institutions are willing or able to provide. It is no secret that interdisciplinary research requires not only bold decisions with an eye to the future, but also the funding of possibly high-risk, experimental initiatives – and these are both aspects that some academic systems privilege (and are able to financially sustain) much more than others. It is unsurprising that much of the development in computational communication begins its ambitious trajectory from the USA, where public but also private funding are responsible for the establishment and evolution of a series of excellent labs and centers that produce cutting edge work. This trajectory rarely continues to Europe and elsewhere, where not only funding for such initiatives is harder to come by, but where the few existing centers or labs have been mostly established in a handful of highly prestigious and heavily-funded institutions. As an example, of the “39 women doing amazing research in computational social science”, according to a *Sage Ocean* piece on diversity published October 2018,[1] 28 are based in the USA, and from the rest six are based at Oxford, Cambridge, the London School of Economics, and the newly found Alan Turing Institute – UK’s prestigious national institute for data science and artificial intelligence.

Troublingly, consistently the most publicly visible work in CSS is based on proprietary data that researchers gained access to through privileged partnerships with digital platforms. This is disconcerting for the future of this field for various reasons. For one, the need for proprietary data to do research reinforces existing power-imbalances. While researchers at Berkeley, Stanford, or MIT can rely on a strong alumni network to gain access and trust within companies providing digital platforms (Minsky, 2016), researchers from Europe and elsewhere cannot rely on this sort of access and therefore are consistently worse off than their well-resourced colleagues. This contrasts strongly to the cultural background and identity of the vast majority of international platform users. The underrepresentation of researchers from, e.g. Asia and India inevitably leads to a bias in research attention on the uses of platforms in Western democracies, especially the USA. This is of increasing importance as the USA continues to chart a very specific and contiguous course, and as a result CSS risks speaking predominantly to very specific momentary concerns of this particular country.

Even more worrying, studies based on proprietary data cannot be externally replicated. This is deeply problematic. For one, CSS is an emerging field where standards might shift over time. So even the most well-intentioned and most carefully designed study might need revisiting a few years after publication after a shift in standards or increased sensibility toward potential biases in data collection and analysis. Without a transparent replication regime this is not feasible. The reliance on proprietary data might actually endanger methodological progress in the field and the ongoing conversation of solutions on that front is to be greatly encouraged. At the same time opaqueness with regard to the underlying data and its selection process makes it essentially a faith-based decision to trust the findings or not. As companies providing access to said data are self-interested entities one might be forgiven to think this a weak criterion. More generally, in its reliance on access provided by companies who themselves, their governance-processes, and their business models are subject to researchers' findings, CSS is a field deeply mired in conflicts of interest between researchers, companies, and governments. As of now, the field has neglected to account for these conflicts and develop standards to make them transparent or avoid them (Jungherr et al., 2020a). This is a fundamental challenge for the subsequent maturation of CSS.

The interdisciplinary nature of the field also raises challenges for the review process. A researcher coming from a communication science background will review a paper written by a computer scientist based on the standards and practices of her field . She is therefore likely to find that paper falling short of these standards while a reviewer coming from computer science might have found it ready for publication. Our own experience from the review process in this special issue, in which reviewers came not from computer science but almost exclusively from the fields of political science and communication, already demonstrates that it is unlikely for CSS to develop a coherent and uniform core of theories, methods, and practices soon. Our sense is that it is necessary for editors and reviewers to reflect on these challenges and accordingly review papers with a somewhat broader mind than if they would be reviewing for an article at the core of their own field.

Finally, this challenge also arises with regard to education in CSS. How can one avoid to re-train social scientists into mediocre coders and computer scientists in mediocre social scientists? What is the right balance in providing a common core of CSS as to allow practitioners to use a common language and have a shared understanding of underlying challenges but also allow them to diverge in order to develop necessary specializations in theory, research design, and method? These are questions to which the field has no coherent answer yet but that are of fundamental importance in its maturation process.

Asking Better Questions: The Potential of CSS in Political Communication

The availability of new types of data and the development of computational tools and methods to make sense of them allow a multitude of political communication processes to be investigated from perspectives that were previously impossible. More detailed and more accurate information about people's news consumption and media diets can be acquired by matching individual-level data with data harvested via web-browser or social media trackers. This invites new insights as to how exposure to different types of content might be affecting different types of behavior, such as political participation and attitudes such as media trust, and can help inspect a number of classic media effects theories such as framing, priming or agenda-setting in new and more detailed light (Jungherr et al., 2019). This can allow scholars to better understanding changes in the traditional gatekeeping role of legacy media and professional journalists in an increasingly richer and more competitive media ecology.

Combining different types of data provides also a far more refined insight on how political information and content impacts individuals differently (Popa et al., 2020; Scharkow et al., 2020; Wells & Thorson, 2017), possibly exacerbating already existing inequalities in political information of high quality. Designs employing digital trace and individual level data are, similarly, able to answer a number of new questions related to political communication during electoral campaigns. These questions range from how people deploy communication strategies and language strategically to mobilize others, to how people are being impacted by it as they watch political debates and rallies.

Research into the communication strategies of social movements is also enriched by being able to look into patterns of diffusion of information across networks (Mercea & Bastos, 2016). New ways of sharing humorous political content originating from talk shows or social media memes can now be captured more precisely using digital trace data, and the impact of political humor can be better understood. Importantly, due to the more easily traceable textual and visual character of political discussion online, political communication scholars can today, assisted by a host of (automated) text analysis tools and methods, study in great detail political disagreement and the effects this might have not only on polarization, but also on uncivil behavior. Manifestations of intolerance in human communication, such as racism, misogynism, homophobia, all extremely difficult to measure using surveys, can now be observed on social media and analyzed using what are by now common mining methods and through a multitude of sophisticated text and network analysis methods in order to understand their effects (Benoit, 2020).

Building on this rich portfolio, our goal in this special issue was to attract contributions that help illustrate the kinds of problems computational methods help political communication scholars solve, and what theories they help us advance. We were also interested in contributions illustrating applications of computational methods for addressing major questions pertaining to a number of different topics. We were, finally, interested in garnering insights from scholars working in the field of computational methods and who could discuss their experiences pertaining to the interdisciplinary challenges in the field. We received a large number of high-quality submissions and are proud to present six full-scale research articles and two forum contributions. We are especially glad, that the collection of articles presented here manages to avoid some of the cultural and resource-driven biases in CSS reported above.

In their contribution to this Special Issue, *Lu and Pan* theorize that the expansion of social media use among the Chinese public has made it pertinent for the government to expand its propaganda proliferation strategies in ways that are different in kind to classic propaganda dissemination. Their study, which is one of the first demonstrating the benefits of combining ethnographic and computational methods, reveals the central role of metrics among Chinese propagandists and offers a number of novel insights on the different types of content used for achieving reach.

One of the most exciting advances in political communication research is the gradual entry of images as objects of analysis. The potentials – but also the caveats and extra caution when it comes to validation – of image analysis are powerfully demonstrated by two studies in this Special Issue. In the first one, *Haim and Jungblut* look into candidate imagery during the 2019 European Parliamentary Election using a comparative dataset with candidates across all 28 European members states. They provide a first, large-scale descriptive analysis and exploration of variation of visual communication of candidates across different platforms, demonstrate the value of descriptive analysis in CSS that we discussed earlier, and illustrate a number of different aspects of visual communication pertaining to non-verbal behavior. While the study's methodological approach involves third-party tools that are not only mired by a number of concerns (van Atteveldt & Peng, 2018) but also diverge from our understanding of CSS as discussed in this contribution, it nevertheless demonstrates how off-the-shelf providers for visual analysis can be applied in the study of visual communication and makes a strong case for using such tools with strong validation.

Boussalis and Coan's study asks to what extent nonverbal signals from candidates during televised debates might influence how voters form their level of support. Drawing on literatures on morphological features of candidates and the effects of facial signals of politicians, they theorize that specific emotional displays/facial expressions of candidates in televised debates might influence voter support. Combining frame-level facial display data of political candidates in the US with second-by-second continuous response measures of viewer reactions to debate participants, they find that facial signals of emotions by participants in televised debates may influence how viewers evaluate candidate performance.

Yarchi, Baden, and Kligler-Vilenchik address the important question of political polarization on social media platforms. The authors examine patterns in political talk on three online platforms on a political controversy in Israel. They find strong differences between the patterns identified on the three platforms with Twitter showing the strongest evidence of polarization across the three measures, interactions on WhatsApp growing depolarized over time, and Facebook showing the weakest evidence of polarization. This article provides a stark warning against drawing conclusions of digital media driving political polarization based on single-platform studies, especially if based on Twitter.

The contributions by *Dun, Soroka and Wlezien*, and *Nicholls and Cullpepper* are better understood as speaking to the applications section of this special issue. The study by *Dun, Soroka and Wlezien* is situated within classic political communication approaches involving content analysis of media coverage, in this case of US defense spending. Relying on a large and longitudinal corpus of roughly 2 million articles between 1980 and 2018, they apply what they label as the *dictionary-plus-supervised-learning approach*. The results introduce new considerations as to whether the introduction of machine-learning brings sufficient benefits, but the authors propose that the two approaches need not compete and offer approaches for combining them.

One of the most prominent approaches to the study of political text is the analysis of frames. Here, the contribution by *Nicholls and Culpepper* offers interesting new perspectives. The automated discovery of frames is a thorny issue in the analysis of text. Often, researchers choose one approach without necessarily justifying their choice or providing comparisons to other methods. Nicholls and Culpepper illustrate the drawbacks of this practice by testing the performance of three different procedures in the automated identification of frames. They show that the quality of the approaches varies with regard to the nature of the corpus and different conceptual aspects of frames, and offer a powerful reminder that computational methods are not plug-and-play devices that can be deployed without adjustments.

We were happy to have two forum contributions addressing different challenging aspects of interdisciplinary research. *Windsor's* forum account is important in highlighting the complexities of setting up an interdisciplinary lab and developing a common language with scholars from computer science, and illustrates this in practice through a discussion of how different disciplines interpret and operationalize "cohesion„. *Van Atteveldt, Althaus, and Wessler* discuss the many issues emerging in collaborative endeavors that necessitate data sharing. As data sharing in CSS is often governed by copyright law and terms of service contracts, sustainable and ethical solutions that foster comparative work are very challenging, and their experience with short-term approaches is a useful guide for anyone beginning such endeavors with an interest to mitigate such problems.

Being able to open so many new avenues into political communication research, computational tools and methods have, quite clearly, a very broad appeal to political communication scholars. Yet, despite its broad appeal CSS is not only still limited, but it is also characterized by serious inequalities which only seem to grow over time. Why? We believe there are two reasons. For one, despite deceptively easy entry points CSS has a steep learning curve in acquiring competencies in computational methods that go beyond the use of out-of-the-box solutions . The second reason is the lack of comprehensive training in computational methods in the social sciences, which is itself partially an outcome of not enough people with this cross-disciplinary expertise being hired.

We hope that this Special Issue takes a first step towards demonstrating not only the challenges but also the broad and multifaceted application – and thereby appeal – of computational methods for the political communication community. Computational methods allow for approaching a multitude of existing problems and research puzzles that are particular to this era that is so much shaped by digital media from different angles and diverse ways. As we have shown, computational *political* communication, CCS, and CSS may differ based on the breadth of topics discussed, but they do not vary based on the underlying challenges of creating an interdisciplinary field at the borders of the social sciences, computer science, and the natural sciences. While the temptation to create subfields of subfields is strong, the field should carefully contemplate this development. There are still too few people working at this intersection at this point to begin with. Splitting these few further into silos risks slowing the development of interdisciplinary standards in favor of the emergence of various subfield-specific practices in the use of computational methods. While this development might increase the speed in which computational methods are accepted in specific social science subfields, a byproduct of it could be that the work on the hard questions of establishing interdisciplinary practices between social, computer, and natural scientists is pushed to the sidelines. We suspect that this could have consequences for the development of fresh methods and approaches to problems, and a possible strengthening of ready-made computational solutions to problems in the social sciences.

We see this Special Issue as a conversation-starter on why computational methods are of broader appeal to communication scholars and not the limited domain of highly technically skilled researchers. We believe that this can only be achieved by not only recognizing the necessity of interdisciplinary work but by keeping into perspective what is new and what is not, and that the pitfalls and boundaries of computational communication research relying on computational methods are not unlike those faced by the broader field of CSS.

Note

1. https://ocean.sagepub.com/blog/2018/9/28/39-women-doing-amazing-research-in-computational-social-science

Acknowledgments

The authors would like to thank Valeska Gerstung, Zoltan Fazekas, Sebastian Adrian Popa, Oliver Posegga, Spyros Kosmidis, Ralph Schroeder, Cristian Vaccari, Stefanie Walter, Chris Wlezien and the Editor for reading draft versions of this manuscript and providing immensely helpful comments. Special thanks go to the Editor, Claes de Vreese, for his guidance and support throughout the entire process of editing the special issue "Computational Political Communication: Theory, Applications and Interdisciplinary Challenges".

Disclosure Statement

No potential conflict of interest was reported by the authors.

Notes on contributor

Yannis Theocharis is Professor of Media and Communication with focus on Innovative Methods at the Centre for Media, Communication and Information Research (ZeMKI), University of Bremen Andreas Jungherr is Professor of Communication Science with special focus on Digital Transformation and Publics at the Friedrich-Schiller-University Jena.

ORCID

Yannis Theocharis http://orcid.org/0000-0001-7209-9669
Andreas Jungherr http://orcid.org/0000-0003-2598-2453

References

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*(1), 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2020). Automated text classification of news articles: A practical guide. *Political Analysis*, 1–24. https://doi.org/10.1017/pan.2020.8

Benoit, K. (2020). Text as data: An overview. In L. Cuirini & R. Franzese, (Eds.), *The sage handbook of research methods in political science and international relations* (pp. 461–497). Sage. https://doi.org/10.4135/9781526486387.n29

Berinsky, A.J. (2020). New Directions in Public Opinion Research (3rd Edition). Routledge

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

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

Chancellor, S., & Choudhury, M. D. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *Npj Digital Medicine*, *3*(43), 1–11. https://doi.org/10.1038/s41746-020-0233-7

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. University of California Press.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process fo automated predictions. *Washington Law Review*, *89*(1), 1–33. https://www.law.uw.edu/wlr/print-edition/print-edition/vol-89/1/the-scored-society-due-process-for-automated-predictions

Elster, J. (2015). *Explaining social behavior: More nuts and bolts for the social sciences* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107763111

Flyvbjerk, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge University Press.

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media*, *58*(1), 59–75. https://doi.org/10.1080/08838151.2013.875018

Freelon, D. (2018). Computational research in the post-api age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Gerring, J. (2012). *Social science methodology: A unified framework* (2nd ed. ed.). Cambridge University Press.

Gilardi, F., Baumgartner, L., Dermont, C., Donnay, K., Gessler, T., Kubli, M., … Müller, S. (2020). *Building research infrastructures to study digital technology and politics: Lessons from Switzerland*. (Working Paper). https://www.fabriziogilardi.org/resources/papers/Research_Infrastructures_Digital_Technology_and_Politics.pdf

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*(1), 129–152. https://doi.org/10.1146/annurev-soc-071913-043145

González-Bailón, S. (2017). *Decoding the social world: Data science and the unintended consequences of communication*. The MIT Press.

Grimmer, J., & Steward, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Nagler, J., & Tucker, J. A. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Habel, P., & Theocharis, Y. (2020). Citizens, elites, and social media methodological challenges and opportunities in the study of persuasion and mobilization. In E. Suhay, B. Grofman, &

A. H. Trechsel, (Eds.), *The oxford handbook of electoral persuasion* (pp. 1037–1058). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190860806.013.27

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

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. https://doi.org/10.17705/1jais.00282

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839–880. https://doi.org/10.1093/poq/nfv039

Jungherr, A. (2015). *Analyzing political communication with digital trace data: The role of twitter messages in social science research*. Springer. https://doi.org/10.1007/978-3-319-20319-5

Jungherr, A. (2019). Normalizing digital trace data. In N. J. Stroud & S. C. McGregor, (Eds.), *Digital discussions: How big data informs political communication* (pp. 9–35). Routledge. https://doi.org/10.4324/9781351209434-2

Jungherr, A., & Jürgens, P. (2013). Forecasting the pulse: How deviations from regular patterns in online data can identify offline phenomena. *Internet Research*, *23*(5), 589–607. https://doi.org/10.1108/IntR-06-2012-0115

Jungherr, A., Metaxas, P. T., & Posegga, O. (2020a). *The next 10 years of computational social science: Accounting for theory, transparency, and interests* (Working Paper).

Jungherr, A., Posegga, O., & An, J. (2019). Discursive power in contemporary media systems: A comparative framework. *The International Journal of Press/Politics*, *24*(4), 404–425. https://doi.org/10.1177/1940161219841543

Jungherr, A., Rivero, G., & Gayo-Avello, D. (2020b). *Retooling politics: How digital media are shaping democracy*. Cambridge University Press.

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*, *35*(3), 336–356. https://doi.org/10.1177/0894439316631043

Jungherr, A., & Theocharis, Y. (2017). The empiricist's challenge: Asking meaningful questions in political science in the age of big data. *Journal of Information Technology & Politics*, *14*(1), 97–109. https://doi.org/10.1080/19331681.2017.1312187

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political Communication*, *37*(2), 256–280. https://doi.org/10.1080/10584609.2019.1661888

Kellstedt, P. M., & Whitten, G. D. (2018). *The fundamentals of political science research* (3rd ed.). Cambridge University Press.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–721. https://doi.org/10.1126/science.1197872

King, G., & Persily, N. (2019). A new model for industry–academic partnerships. *PS, Political Science & Politics*. https://doi.org/10.1017/S1049096519001021

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, *62*(3), 790–816. https://doi.org/10.1111/0022-3816.00033

Lazer, D., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359* (6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M., & Christakis, N. (2009). Social science: Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, *43*(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Leeper, T. J. (2020). Raising the floor or closing the gap? How media choice and media content impact political knowledge. *Political Communication*, 1–22. https://doi.org/10.1080/10584609.2020.1753866

Levi, M., & Rajala, B. (2020). Alternatives to social science one. *PS, Political Science & Politics*. https://doi.org/10.1017/S1049096520000438

Mercea, D., & Bastos, M. T. (2016). Being a serial transnational activist. *Journal of Computer-Mediated Communication*, *21*(2), 140–155. https://doi.org/10.1111/jcc4.12150

Minsky, C. (2016). *Which colleges do Facebook, Google and other top employers recruit from?* The World University Rankings. Retrieved August 2, 2020, from https://www.timeshighereducation.com/student/news/which-colleges-do-facebook-google-and-other-top-employers-recruit

Munger, K. (2017). tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Neuman, W. R. (2016). *The digital difference: Media technology and the theory of communication effects*. Harvard University Press.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pentland, A. (2008). *Honest signals: How they shape our world*. The MIT Press.

Piper, A. (2018). *Enumerations: Data and literary study*. The University of Chicago Press.

Popa, S. A., Fazekas, Z., Braun, D., & Leidecker-Sandmann, -M.-M. (2020). Informing the public: How party communication builds opportunity structures. *Political Communication*, *37*(3), 329–349. https://doi.org/10.1080/10584609.2019.1666942

Puschmann, C. (2019). An end to the wild west of social media research: A response to axel bruns. *Information, Communication & Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Rauchfleisch, A., & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *Berkman Klein Center Research Publication*. https://doi.org/org/https://papers.ssrn.com/sol3/papers.cfm?Abstract_id=3565233

Rauh, C., & Schwalbach, J. (2020). *The parlspeech v2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies*. V1: Harvard Dataverse. https://doi.org/10.7910/DVN/L4OAKN

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press.

Salganik, M. J., & Watts, D. J. (2009). Web-based experiments for the study of collective social dynamics in cultural markets. *Topics in Cognitive Science*, *1*(3), 439–468. https://doi.org/10.1111/j.1756-8765.2009.01030.x

Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, *117*(6), 2761–2763. https://doi.org/10.1073/pnas.1918279117

Schneier, B. (2020). Bots are destroying political discourse as we know it. *The Atlantic*. https://www.theatlantic.com/technology/archive/2020/01/future-politics-bots-drowning-out-humans/604489/

Schroeder, R. (2016). Big data and communication research. In J. F. Nussbaum (Ed.), *Oxford research encyclopedia of communication*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.276

Schroeder, R. (2018). *Social theory after the internet: Media, technology and globalization*. UCL Press.

Schroeder, R. (2019). Big data and cumulation in the social sciences. *Information, Communication & Society*, *23*(11), 1593–1607. https://doi.org/10.1080/1369118X.2019.1594334

Sen, I., Floeck, F., Weller, K., Weiss, B., & Wagner, C. (2019). A total error framework for digital traces of humans. *arXiv*. https://arxiv.org/abs/1907.08228.

Siegel, A., & Badaan, V. (2020). #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, *114*(3), 837–855. https://doi.org/10.1017/S0003055420000283

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., & Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PLoS One*, *9* (4), e95978. https://doi.org/10.1371/journal.pone.0095978

Swedberg, R. (2014). *The art of social theory*. Princeton University Press.

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on twitter. *SAGE Open*, *10*(2), 1–15. https://doi.org/10.1177/2158244020919447

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–1031. https://doi.org/10.1111/jcom.12259

Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. The University of Chicago Press.

van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C. H. D., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *ICWSM 2017: Proceedings of the eleventh international aaai conference on web and social media* (pp. 280–289). Menlo Park, CA: Association for the Advancement of Artificial Intelligence (AAAI).

Watts, D. J. (2011). *Everything is obvious: How common sense fails us*. Random House.

Weidmann, N. B., & Schutte, S. (2017). Using night light emissions for the prediction of local wealth. *Journal of Peace Research*, *54*(2), 125–140. https://doi.org/10.1177/0022343316630359

Wells, C., & Thorson, K. (2017). Combining big data and survey techniques to model effects of political content flows in facebook. *Social Science Computer Review*, *35*(1), 33–52. https://doi.org/10.1177/0894439315609528

Williams, N. W., Casas, A., & Wilkerson, J. D. (2020). *Images as data for social science research*. Cambridge University Press. https://doi.org/10.1017/9781108860741