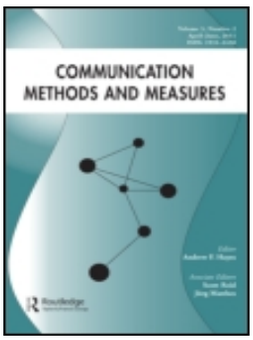

Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: http://www.tandfonline.com/loi/hcms20

# When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science

Wouter van Atteveldt & Tai-Quan Peng

**To cite this article:** Wouter van Atteveldt & Tai-Quan Peng (2018): When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science, Communication Methods and Measures, DOI: 10.1080/19312458.2018.1458084

**To link to this article:** https://doi.org/10.1080/19312458.2018.1458084

Published online: 20 Apr 2018.

COMMUNICATION METHODS AND MEASURES
https://doi.org/10.1080/19312458.2018.1458084

Routledge
Taylor & Francis Group



# When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science

Wouter van Atteveldt [a] and Tai-Quan Peng [b]

[a]Department of Communication Science, VU University Amsterdam, Amsterdam, the Netherlands; [b]Department of Communication, Michigan State University, East Lansing, MI, USA

**ABSTRACT**

The recent increase in digitally available data, tools, and processing power is fostering the use of computational methods to the study of communication. This special issue discusses the validity of using big data in communication science and showcases a number of new methods and applications in the fields of text and network analysis. Computational methods have the potential to greatly enhance the scientific study of communication because they allow us to move towards collaborative large-N studies of actual behavior in its social context. This requires us to develop new skills and infrastructure and meet the challenges of open, valid, reliable, and ethical "big data" research. By bringing together a number of leading scholars in one issue, we contribute to the increasing development and adaptation of computational methods in communication science.

## The role of computational methods in communication science

"We are on the cusp of a new era in computational social science" (Wallach, 2016). As evidenced by the many reviews, special issues, and position papers, a growing community of scholars is using computational methods for analyzing social behavior (e.g., Alvarez, 2016; Boyd & Crawford, 2012; Huberman, 2012; Lazer et al., 2009; Parks, 2014; Shah, Cappella, & Neuman, 2015; Trilling, 2017). The recent acceleration in the promise and use of computational methods for communication science is primarily fueled by the confluence of at least three developments:

1. A deluge of digitally available data, ranging from social media messages and other "digital traces" to web archives and newly digitized newspaper and other historical archives (e.g., Weber, 2018);
2. Improved tools to analyze this data, including network analysis methods (e.g., Lungeanu, Carter, DeChurch, & Contractor, 2018; Barabási, 2016) and automatic text analysis methods such as supervised text classification (Boumans & Trilling, 2016; Collingwood & Wilkerson, 2012; Odijk, Burscher, Vliegenthart, & de Rijke, 2013), topic modelling (Maier et al., 2018; Blei, Ng, & Jordan, 2003; Jacobi, Van Atteveldt, & Welbers, 2016; Roberts et al., 2014), word embeddings (e.g. Rudkovsky et al., 2018), and syntactic methods (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017); and
3. The emergence of powerful and cheap processing power, and easy to use computing infrastructure for processing these data, including scientific and commercial cloud computing, sharing platforms such as Github and Dataverse, and crowd coding platforms such as Amazon MTurk and Crowdflower.

Many of these new data sets contain communication artifacts such as tweets, posts, emails, and reviews; and many of these new methods are aimed at analyzing the structure and dynamics of human communication. As such, these developments are especially relevant for communication science and computational methods have been used to analyze a variety of communicative

**CONTACT** Wouter van Atteveldt wouter@vanatteveldt.com

phenomena (e.g., Colleoni, Rozza, & Arvidsson, 2014; Grimmer, 2016; Jungherr, 2014; Tucker et al., 2016) . Moreover, communication theories, such as agenda-setting, two-step flow of information, selective exposure, and interpersonal persuasion have been widely cited as major theoretical backbones in many computational studies (e.g., Russell Neuman, Guggenheim, Mo Jang, & Bae, 2014; Tan, Niculae, Danescu-Niculescu-Mizil, & Lee, 2016; Wu, Hofman, Mason, & Watts, 2011; Xu et al., 2013; Yang & Leskovec, 2010). As will be argued in this volume, these three developments have the potential to give an unprecedented boost to progress in communication science, provided we can overcome the technical, social, and ethical challenges presented by these developments.

Like "big data," the concept of computational methods makes intuitive sense but is hard to define. Sheer size is not a necessary criterion to define big data (Monroe, Pan, Roberts, Sen, & Sinclair, 2015), and the fact that a method is executed on a computer does not make it a "computational method"—communication scholars have used computers to help in their studies for over half a century (e.g., Nie, Bent, & Hull, 1970; Stone, Dunphy, Smith, & Ogilvie, 1966). Adapting the criteria given by Shah et al. (2015), we can give an ideal-typical definition by stating that computational communication science studies generally involve: (1) large and complex data sets; (2) consisting of digital traces and other "naturally occurring" data; (3) requiring algorithmic solutions to analyze; and (4) allowing the study of human communication by applying and testing communication theory.

Of course, computational methods do not replace the existing methodological approaches, but rather complement it. Computational methods are an expansion and enhancement of the existing methodological toolbox, while traditional methods can also contribute to the development, calibration, and validation of computational methods. Moreover, the distinction between "classical" and "computational" methods is often one of degree rather than of kind, and the boundaries between approaches are fuzzy: When does an on-line experiment turn into a computational analysis, and how do Facebook status updates really differ from self-reports? Nevertheless, the term computational methods is useful to make us realize that new datasets and processing techniques offer us possibilities beyond just scaling up our previous work; and to alert us to the potential challenges, pitfalls, and required expertise in using these methods.

This Special Issue originates from the first set of panels organized by the Computational Methods Interest Group of the International Communication Association. The purpose of the issue, and of this introduction in particular, is to provide an overview of the exciting work that is being done to study communication using computational methods, and highlight the advantages and challenges of this methodology. To start, in the next section we will review the potential benefits offered by computational methods. This is followed by a discussion of key limitations and challenges facing scholars using computational methods. We end with an overview of the articles published in this volume and pointers for the future direction of our field.

## Opportunities offered by computational methods

If computational methods simply allowed us to use our existing methodologies at greater speed, scale, or ease of use it would be welcomed, but it would hardly be revolutionary. Instead, we argue that computational methods allow us to analyze social behavior and communication in ways that were not possible before and have the potential to radically change our discipline at least in four ways.

### *From self report to real behavior*

Digital traces of online social behavior can function as a new behavioral lab available for communication researchers. These data allow us to measure actual behavior in an unobtrusive way rather than self-reported attitudes or intentions (e.g., Araujo, Wonneberger, Neijens, & de Vreese, 2017;

Dovidio, 1992). This can help overcome social desirability problems, and more importantly it does not rely on people's imperfect estimate of their own desires and intentions.

Digital traces on social media are bundled with important features that empower communication researchers to examine human communication phenomena with new perspectives. Specifically, the structural perspective has become more and more prominent in computational research with various social and interactive relations documented on social media platforms. The rapid advancement of network modeling techniques makes it feasible to study the intricate interplay between individuals and the local or global social structures they are embedded in (cf. Barabási, 2016; Diesner, Frantz, & Carley, 2005; Westlake & Bouchard, 2016). With voluminous time-stamped data on social media, it is methodologically viable to unravel the dynamics underlying human communication and disentangle the interdependent relationships between multiple communication processes (Monge & Contractor, 2003).

This can also help overcome the problems of linking content data to survey data. This a mainstay of media effects research but is problematic because of bias in media self-reports (Kobayashi & Boase, 2012; Scharkow & Bachl, 2017) and because news consumers nowadays often cherry-pick articles from multiple sites, rather than relying on a single source of news (Costera Meijer & Groot Kormelink, 2015). Using special-purpose mobile apps and browser extensions, we can now trace news consumption in real-time and combine it with survey data to get a more sophisticated measurement of news consumption and effects (Bodo et al., 2017; Guess, Nyhan, & Reifler, 2018; Kobayashi, Boase, Suzuki, & Suzuki, 2015).

### *From lab experiments to studies of the actual social environment*

A second advantage is that we can observe the reaction of persons to stimuli in their actual environment rather than in an artificial lab setting. In their daily lives, people are exposed to a multitude of stimuli simultaneously, and their reactions are also conditioned by how a stimulus fits into the overall perception and daily routine of people. Moreover, we are mostly interested in social behavior, and how people act strongly depends on their (perception of) actions and attitudes in their social network (Barabási, 2016).

The emergence of social media substantially facilitates the design and implementation of experiment research. First, crowdsourcing platforms on social media lowers the obstacles in research subject recruitment. Traditionally, many communication researchers counted on student subjects of small or modest size in experiment studies. Nowadays, it is no longer difficult to recruit thousands or even millions of diverse or specialized subjects from crowdsourcing platforms (e.g., Amazon Mechanical Turk and Volunteer Science) to participate in an experiment at low or zero costs.

However, the implementation of experimental design on social media is not an easy task. Social media companies will be very selective on their collaborators and on research topics. As Wallach (2016) observed, the fear of losing reputation will probably cause social media companies to be more reluctant to share data in the future after the fallout of studies such as the Facebook mood manipulation study (Kramer, Guillory, & Hancock, 2014). The coordination of experiments on social media can also be extremely time-consuming. In a recently published study, it took the authors about five years to coordinate an experiment on social media (King, Schneer, & White, 2017). Furthermore, how to adequately address ethical concerns involved in online experiments has become a pressing ethical issue in scientific community.

### *From small-N to large-N*

Simply increasing the scale of measurement can also enable us to study more subtle relations or effects in smaller subpopulations than possible with the sample sizes normally available in communication research (Monroe et al., 2015). For example, the Facebook "voting study" showed that a stimulus message to vote also affects close friends of the people who received the message, but this

effect was so small that it was only significant because of the half a million subjects—but given the small margins in (American) elections even such a small effect can be decisive (Bond et al., 2012). Similarly, by measuring messages and behavior in real time rather than in daily or weekly (or yearly) surveys, much more fine-grained time series can be constructed, alleviating the problems of simultaneous correlation and making a stronger case for finding causal mechanisms.

In order to leverage the more complex models afforded by larger data sets we need to change the way we build and test our models. As argued by Hindman (2015), it is useful to consider techniques developed in machine learning research for model selection and model shrinkage such as penalized (lasso) regression and cross-validation which are aimed at out-of-sample prediction rather than within-sample explanation. Such techniques estimate more parsimonious models and hence alleviate the problems of overfitting that can occur with very large data sets. Additionally, methods such as Exponential Random Graph Modeling (ERGM; An, 2016) or Relational Event Modeling (Pilny, Schecter, Poole, & Contractor, 2016) can dynamically model network and group dynamics.

### *From solitary to collaborative research*

Currently, most empirical scholars in communication science gather, clean, and analyze their own data, either individually or as a group. Moreover, in many cases the tools and scripts used for data processing are also developed locally. Digital data and computational tools make it easier to share and reuse these resources. The increased scale and complexity also make it almost necessary to do so: it is very hard for any individual researcher to possess the skills and resources needed do do all the steps of computational research him or herself (Hesse, Moser, & Riley, 2015). An increased focus on sharing data and tools will also force us to be more rigorous in defining operationalizations and documenting the data and analysis process, furthering transparency and reproducibility of research.

A second way in which computational methods can change the way we do research is by fostering the interdisciplinary collaboration needed to deal with larger data sets and more complex computational techniques (Wallach, 2016). For example, measurements and analysis methods from neuroscience are being increasingly used in analyzing communication processes (Falk, Cascio, & Coronel, 2015; Weber, Mangus, & Huskey, 2015). Agent-based modeling uses computer models and empirical data to facilitate theory building (Palazzolo, Serb, She, Su, & Contractor, 2006). By offering a chance to zoom in from the macro level down to the individual data points, digital methods can also bring quantitative and qualitative research closer together, allowing qualitative research to improve our understanding of data and build theory, while keeping the link to large-scale quantitative research to test the resulting hypotheses (O'Brien, 2016).

## Challenges and pitfalls in computational methods

As argued above, computational methods offer a wide range of possibilities for communication researchers to explore new research questions and re-examine classical theories from new perspectives. By observing actual behavior in the social environment, and if possible of a whole network of connected people, we get a better measurement of how people actually react, rather than of how they (report or intent to) react in the artificial isolation of the lab setting; and the scale at which this is possible allows more complex or subtle causal relations to be tested and discovered. Large-scale exploratory research can help formulate theories and identify interesting cases or subsets for further study, while at the same time smaller and qualitative studies can help make sense of the results of big data research (cf. O'Brien, 2016). Similarly, "big data" confirmatory research can help test whether causal relations found in experimental studies actually hold in the "wild", i.e., on large populations and in real social settings (Monroe et al., 2015).

Using these new methods and data sets, however, creates a new set of challenges and pitfalls, some of which will be reviewed below. Most of these challenges do not have a single answer, and

require continued discussion about the advantages, disadvantages, and best practices in computational communication science. To contribute to this discussion, we pose a number of key questions about computational methods below, and give pointers to the relevant literature to understand the problems and possible solutions.

### How do we keep research datasets accessible?

Is it ironic or naive to ask about getting research datasets in the age of big data? Our answer is no. Although the volume, variety, velocity, and veracity of big data has been repeatedly bragged in both news reports and scholarly writings, it is a hard truth that many of the "big data" sets are proprietary ones which are highly demanding to access for most communication researchers (Boyd & Crawford, 2012; Lazer et al., 2009; Wallach, 2016). The privileged access to big data by a small group of researchers will make researchers with the access "enjoy an unfair amount of attention at the expense of equally talented researchers without these connections" (Huberman, 2012, p. 308). Moreover, studies conducted by researchers connected to these actors are generally based only on a single platform (e.g., Twitter or Facebook), which makes it challenging to develop a panoramic understanding of users behavior on social media as a holistic ecosystem and increases generalizability problems (Wallach, 2016). More importantly, such privileged access to big data will thwart the reproducibility of computational research which serves as the minimum standard by which scientific claims are judged (Peng, 2011).

Samples of big data on social media are made accessible to the public either in its original form (data collected via Twitter public API) or in aggregate format (e.g., data from Google Trends). Moreover, as explicated by Matthew Weber (2018), external parties also create accessible archives of web data. However, the sampling, aggregation, and other transformation imposed on the released data is a black box, which poses great challenges for communication researchers to evaluate the quality and representativeness of the data and then assess the external validity of their findings derived from such data.

We need to make sure our data is open and transparent, and to make sure that research is not reserved to the privilleged few who have the network or resources to acquire data sets. To do this, is is vital that we stimulate sharing and publishing data sets. Where possible these should be fully open and published on platforms such as dataverse, where needed for privacy or copyright reasons the data should be securely stored but accessible under clear conditions. A corpus management tool like AmCAT (Van Atteveldt, 2008) can help alleviate copyright restrictions by allowing data to be queried and analysed even if the full text of the data set cannot be published. Additionally, we should work with funding agencies and data providers such as newspaper publishers and social media platforms to make standardized data sets available for all researchers.

### Is "big" data always good data?

With the increasing ease-of-use of computational algorithms, communication researchers themselves can retrieve megabytes or terabytes of digitized data from social media or other online sources. Both the nature of the new types of digital trace and social media data and the relative convenience in collecting such data are more appealing to communication researchers, in comparison to the time-consuming field work, relatively small sample, decreasing response rate, and heavily criticized biases in survey research. Do communication researchers need to bother with small-sample survey data when it is easy and cheap to get "big data" from social media? Big data, however, is not a panacea for all methodological problems in empirical research, and it has its obvious limitations despite its widely touted advantages.

First, big data is "found" while survey data is "made" (Taylor, 2013). Most of the big data are secondary and intended for other primary uses most of which have little relevance to academic research. On the other side, most of the survey data are "made" by researchers who design and

implement their studies and questionnaires with specific research purposes in mind. The big data is "found" and then tailored or curated by researchers to address their own theoretical or practical concerns. The gap between the primary purpose intended for big data and the secondary purpose found for big data will pose threat to the validity of design, measurement, and analysis in computational communication research.

Secondly, that data is "big" does not mean that it is representative for a certain population (Boyd & Crawford, 2012; Hargittai, 2015; Wallach, 2016). As Hargittai (2015) has shown based on representative survey data, people do not randomly select into social media platforms, and very limited information is available for communication researchers to assess the (un)representativeness of big data retrieved from social media. As Hazel Kwon and her colleagues point out in this volume, "specialized" actors on social media, such as issue experts, professionals, and institutional users, are over-represented while the ordinary publics are under-represented in computational research, which leads to a sampling bias to be carefully handled in empirical studies.

This also means that p-values are less meaningful as a measure of validity. There is a lively debate about the use and abuse of p-values and null-hypothesis significance testing (NHST; see, e.g., Leek & Peng, 2015; Vermeulen et al., 2015; Vidgen & Yasseri, 2016; Wasserstein & Lazar, 2016) and the leading political methodology journal *Political Analysis* decided to stop reporting p-values altogether (Gill, 2018). Especially for very large data sets, where representativeness and selection and measurement biases are a much greater threat to validity than small sample sizes, p values are not a very meaningful indicator of effect (Hofmann, 2015).

In general, we should recognize that size of data is neither a sign of validity nor of invalidity of the conclusions. Especially for big data studies, we should focus more on substantive effect size and validity than mere statistical significance, for example by showing confidence intervals and using simulations or bootstrapping to show the estimated real effects of the found relations.

### *Are computational measurement methods valid and reliable?*

The unobtrusiveness of social media data makes them less vulnerable to traditional measurement bias, such as instrument bias, interviewer bias, and social desirability bias. However, this does not imply that they are free of measurement errors.

Measurement errors can be introduced when text mining techniques are employed to identify semantic features in user-generated content, whether using dictionaries, machine learning, or unsupervised techniques (Boumans & Trilling, 2016; Yang, Adomavicius, Burtch, & Ren, 2018) and when social and communication networks are constructed from user-initiated behavior. For example, Soroka, Young, and Balmas (2015) found that different sentiment dictionaries capture different underlying phenomena and highlight "the importance of tailoring lexicons to domains to improve construct validity" (p. 108). In the same volume, González-Bailón and Paltoglou (2015) also observe the lack of correlation between sentiment dictionaries, and similarly argue for the need for domain adaptation of dictionaries. Similar to techniques like factor analysis, unsupervised methods such as topic modelling require the researcher to interpret and validate the resulting topics, and although quantitative measures of topic coherence exist these do not always correlate with human judgments of topic quality (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009).

However, it should be noted that classical methods of manual content analysis are also no guarantee of valid or reliable data. Referring to the "myth of the trained coder", Rene Weber and co-authors in this volume show that using trained manual coders to extract subjective features such as moral claims can lead to overestimation of reliability and argue that untrained (crowd) coders can actually be better at capturing intuitive judgments.

The errors can introduce systematic biases in subsequent multivariate analysis and threaten the validity of statistical inference. This means that we need to emphasize the validity of measurements of social media and other digital data. (see e.g. De Choudhury et al., 2010; Newman, 2017; Paul & Dredze, 2011; Yang et al., 2018). The validity of a method or tool is dependent on the context in

which it is used, so even if a researcher uses an existing off-the-shelf tool with published validity results it is vital to show how well it performs in a specific domain and on a specific task. Additionally, a culture of sharing and reusing tools and methods and publishing the source code and validation sets of tools helps foster continuous and collaborative improvements to the measurement tools we all use.

### What is responsible and ethical conduct in computational communication research?

What responsible and ethical conduct should be adopted is another challenge in computational communication research. With the publication of several controversial large-scale experiments on social media (e.g., Bond et al., 2012; Kramer et al., 2014), the scientific community and the general public have expressed growing concern on ethical (mis)conduct in computational social science.

Such concerns can exist in different steps of computational communication research (Boyd & Crawford, 2012; Kosinski, Matz, Gosling, Popov, & Stillwell, 2015). For instance, in field experiments on social media, how can researchers get informed consent from the subjects? When users of a social media platform accept the terms of service of the platform, can researchers assume that the users have given an explicit or implicit consent to participate in any types of experiments conducted on the platform? As most of the datasets in computational communication research are directly produced by individuals, to what extent should the data be anonymized and sanitized for the sake of privacy protection, how can we achieve a balance between protecting individual privacy and advocating reproducible research? How do we deal with the findings that our digital traces can reveal a lot of very personal traits (Kosinski, Stillwell, & Graepel, 2013) and that there are techniques to de-anonymize supposedly "anonymized" data (Narayanan & Shmatikov, 2008)?

There are no unambiguous answers to all these questions, but we do not have the luxury of ignoring these problems and (further) losing the trust of the general public. This calls for a collective effort from the whole community to set up a responsible conduct of research in computational communication research.

### How do we get the needed skills and infrastructure?

Reaping the benefits of computational methods requires that as a scientific community we need to invest in skills, infrastructure, and institutions (Jagadish et al., 2014; Wallach, 2016). As we expect more and more research to include (some) computational aspects, it is increasingly important that as practitioners we are skilled at dealing with data and computational tools. As pointed out by Shah et al. (2015), many digital traces and other "big" data are textual rather than the numerical data most communication scholars are trained for and used to, and will require us to "hone [our] skills in natural language processing" (p. 21).

Of course, we can and should collaborate as much as possible with researchers in fields like computer scientists, computational linguists, and artificial intelligence. However, collaboration requires research that is innovative and challenging to both sides, and in many cases what we need is a good programmer to help us gather, clean, analyze, and visualize data rather than a computer scientist to invent a new algorithm. While the bigger (or richer) groups can afford to hire such programmers, many (young) scholars do not have these resources and, moreover, it is very difficult to select, supervise, and motivate a programmer without some computational skills.

Thus, we expect that doing research in communication science will increasingly demand at least some level of computational literacy. Not everyone can (or should) become a programmer, but modern computer languages, libraries, and toolkits have made it easier than ever to achieve useful results, and with a relatively limited investment in computing skills one can quickly become more productive at data driven research and better at communicating with programmers or computer scientists. We think it is vital that we make computational methods more prominent in our teaching to make sure the new generation of communication scientists and practitioners are stimulated and

facilitated to learn computational skills such as data analytics, text processing, or web scraping, as applicable.

Second, it is important to invest in research infrastructure and to move to a culture of sharing and reusing tools and data. One person or group cannot hope to master all skills and tools needed to gather, clean, process, and analyze data using innovative computational methods. Bigger teams have more skills and resources, of course, but there is no need for all steps of the process to be taken within the same team or project, and in many cases the value of data or tools are much greater than the context within which they were originally developed.

Finally, we need to make sure that our institutional arrangements stimulate scholars to invest in and share computational skills, methods, and data. Thus, it is important that researchers and institutions give credit to development and sharing of tools and data (cf. Crosas, King, Honaker, & Sweeney, 2015). As long as the basic "coins" in academia remain publications and citations, we need to make sure that data and tools can be published and cited, and that these publications and citations are valued. Especially important here is to also stimulate the maintenance and documentation of tools and data. A new tool can be published in a journal article, but it can be difficult to get credit for a new version or better documentation of a tool, especially for contributions to tools originally developed by another scholar. This stimulates a fragmentation of the ecosystem, even though it can be more useful for the community to have fewer but well-documented and well-maintained tools than to have many tools or data sets that lack documentation or maintenance. We also need to understand and appreciate the relative effort that goes into tools development and interdisciplinary research, where there is often a long lag before research becomes fully productive.

## Contents of this volume

The articles in this special issue covers a wide range of computational methods relevant to communication research, including text mining, network modeling, and the validity and provenance of big data sets.

On automatic text analysis, Daniel Maier and his colleagues review the topic modeling literature and discuss the advantages, challenges, and best practices in applying LDA in communication research. Their discussion of reliability and validity in topic modeling is of particular relevance to communication research. Rene Weber and co-authors describe how crowd coding can be an indispensable tool for extracting moral claims and other subjective information from text. They argue that the more intuitive judgments of (multiple) untrained crowd coders might actually yield a better measurement than expert coders, who can be influenced by group biases that cloud the intercoder reliability measures. Elena Rudkovsky and her colleagues use word embeddings to improve automatic sentiment analysis on Austrian newspapers. Finally, Damian Trilling and Jeroen Jonkman propose and discuss four criteria that a framework for automated content analysis should fulfill, including scalability, free and open source, adaptability, and accessibility via multiple interfaces.

On network analysis, Alina Lungeanu and her colleagues move beyond a dyadic person-to-person framework to construct social networks and adopt a hypergraph approach to account for the nesting of individuals in groups and the patterns of interlocks among groups. They discuss the conceptual innovation of the hypergraph approach and demonstrate how to address the methodological challenges in such a hypergraph approach by applying it to examine the assembly of scientific collaboration teams.

The last two authors discuss the challenges in obtaining valid "big" data sets. Matthew Weber gives an overview of the possibilities and challenges in using web archives for communication research. Web archives are an important and frequently used source of information on historical trends in Internet use, and by showing how issues of noise and representativeness can be addressed Weber paves the way for a more robust use of these resources in our field. Hazel Kwon and her

colleagues address sampling biases in digital research and demonstrate how supervised machine learning can reduce sampling bias induced from "proxy-population mismatch" on Twitter.

Together, these authors describe and apply a range of different computational methods and data sets, with a focus on critically assessing the utility and validity of computational methods for studying communication. They show how computational methods can be used to accelerate the pace of discovery in communication science, provided that we can tackle the challenges of conducting valid, open, and ethical research. In particular, we think that it is vital that as a community we move forward on at least three fronts: (1) build the infrastructure, skills, and institutional incentives required to use and maintain computational methods and tools; (2) work toward open, transparent, and collaborative research, with sharing and reusing datasets and tools the norm rather than the exception; and (3) continue developing, validating, and critically discussing computational methods in the context of substantive communication science questions. By producing this special issue, we hope to have contributed toward reaching these goals and to the further development and dissemination of computational methods.

## ORCID

Wouter van Atteveldt http://orcid.org/0000-0003-1237-538X
Tai-Quan Peng http://orcid.org/0000-0002-2588-7491

## References

Alvarez, R. M. (Ed.). (2016). *Computational social science: Discovery and prediction*. Cambridge, UK: Cambridge University Press.

An, W. (2016). Fitting ERGMs on big networks. *Social Science Research*, *59*, 107–119. doi:10.1016/j.ssresearch.2016.04.019

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, *11* (3), 173–190. doi:10.1080/19312458.2017.1317337

Barabási, A. (2016). *Network science*. Cambridge, UK: Cambridge University Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bodo, B., Helberger, N., Irion, K., Zuiderveen Borgesius, F., Moller, J., van de Velde, B., … de Vreese, C. (2017). Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents. *Yale JL & Technical*, *19*, 133.

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

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. doi:10.1080/21670811.2015.1096598

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). *Reading tea leaves: How humans interpret topic models*. Advances in neural information processing systems (pp. 288–296). New York, NY: Curran Associates.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*(2), 317–332. doi:10.1111/jcom.2014.64.issue-2

Collingwood, L., & Wilkerson, J. (2012). Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, *9*(3), 298–318. doi:10.1080/19331681.2012.669191

Costera Meijer, I., & Groot Kormelink, T. (2015). Checking, sharing, clicking and linking: Changing patterns of news use between 2004 and 2014. *Digital Journalism*, *3*(5), 664–679. doi:10.1080/21670811.2014.937149

Crosas, M., King, G., Honaker, J., & Sweeney, L. (2015). Automating open science for big data. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 260–273. doi:10.1177/0002716215570847

De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? *Icwsm*, *10*, 34–41.

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication networks from the enron email corpus "it's always about the people. enron is no different". *Computational & Mathematical Organization Theory*, *11*(3), 201–228. doi:10.1007/s10588-005-5377-0

Dovidio, J. F. (1992). New technologies for the direct and indirect assesment of attitudes. In J. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 204–237). New York, NY: Russell Sage Foundation.

Falk, E. B., Cascio, C. N., & Coronel, J. C. (2015). Neural prediction of communication-relevant outcomes. *Communication Methods and Measures*, *9*(1–2), 30–54. doi:10.1080/19312458.2014.999750

Gill, J. (2018). Comments from the new editor. *Political Analysis*, *26*(1), 1–2. doi:10.1017/pan.2017.41

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95–107. doi:10.1177/0002716215569192

Grimmer, J. (2016). Measuring representational style in the house: The tea party, obama, and legislators' changing expressed priorities. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (p. 307). Cambridge, UK: Cambridge University Press.

Guess, A., Nyhan, B., & Reifler, J. (2018). *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. Presidential campaign*. Retrieved from https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf

Hargittai, E. (2015). Is bigger always better? potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 63–76. doi:10.1177/0002716215570866

Hesse, B. W., Moser, R. P., & Riley, W. T. (2015). From big data to knowledge in the social sciences. *The Annals of the American Academy of Political and Social Science*, *659*(1), 16–32. doi:10.1177/0002716215570007

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62. doi:10.1177/0002716215570279

Hofmann, M. A. (2015). *Searching for effects in big data: Why p-values are not advised and what to use instead*. Proceedings of the 2015 winter simulation conference (pp. 725–736). New York, NY: IEEE Press.

Huberman, B. A. (2012). Sociology of science: Big data deserve a bigger audience. *Nature*, *482*(7385), 308. doi:10.1038/482308d

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. doi:10.1080/21670811.2015.1093271

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94. doi:10.1145/2622628

Jungherr, A. (2014). The logic of political coverage on twitter: Temporal dynamics and content. *Journal of Communication*, *64*(2), 239–259. doi:10.1111/jcom.2014.64.issue-2

King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. *Science*, *358*(6364), 776–780. doi:10.1126/science.aao1100

Kobayashi, T., & Boase, J. (2012). No such effect? The implications of measurement error in self-report measures of mobile communication use. *Communication Methods and Measures*, *6*(2), 126–143. doi:10.1080/19312458.2012.679243

Kobayashi, T., Boase, J., Suzuki, T., & Suzuki, T. (2015). Emerging from the cocoon? revisiting the tele-cocooning hypothesis in the smartphone era. *Journal of Computer-Mediated Communication*, *20*(3), 330–345. doi:10.1111/jcc4.2015.20.issue-3

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543. doi:10.1037/a0039210

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. doi:10.1073/pnas.1218772110

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

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Alstyne, M. V. (2009). Life in the network: The coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721. doi:10.1126/science.1167742

Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature News*, *520*(7549), 612. doi:10.1038/520612a

Lungeanu, A., Carter, D. R., DeChurch, L., & Contractor, N. (2018). How Team Interlock Ecosystems Shape the Assembly of Scientific Teams: A Hypergraph Approach. *Communication Methods and Measures*, *12*(2–3). doi:10.1080/19312458.2018.1430756

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2–3). doi:10.1080/19312458.2018.1430754

Monge, P. R., & Contractor, N. S. (2003). *Theories of communication networks*. New York, NY: Oxford University Press.
Monroe, B. L., Pan, J., Roberts, M. E., Sen, M., & Sinclair, B. (2015). No! formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, *48*(1), 71–74.
Narayanan, A., & Shmatikov, V. (2008). *Robust de-anonymization of large sparse datasets*. Security and privacy, 2008. sp 2008. ieee symposium on (pp. 111–125). New York, NY: IEEE Press.
Newman, M. (2017). *Measurement errors in network data*. arXiv preprint arXiv:1703.07376.
Nie, N. H., Bent, D. H., & Hull, C. H. (1970). *Spss: Statistical package for the social sciences*. New York, NY: McGraw-Hill.
O'Brien, D. T. (2016). Using small data to interpret big data: 311 reports as individual contributions to informal social control in urban neighborhoods. *Social Science Research*, *59*, 83–96. doi:10.1016/j.ssresearch.2016.04.009
Odijk, D., Burscher, B., Vliegenthart, R., & de Rijke, M. (2013). Automatic thematic content analysis: Finding frames in news. In Jatowt, A. et al. (eds.), *Social informatics*, SocInfo 2013. Lecture Notes in Computer Science, vol 8238. Cham: Springer.
Palazzolo, E. T., Serb, D. A., She, Y., Su, C., & Contractor, N. S. (2006). Coevolution of communication and knowledge networks in transactive memory systems: Using computational models for theoretical development. *Communication Theory*, *16*(2), 223–250. doi:10.1111/comt.2006.16.issue-2
Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication*, *64* (2), 355–360. doi:10.1111/jcom.2014.64.issue-2
Paul, M. J., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *Icwsm*, *20*, 265–272.
Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. doi:10.1126/science.1213847
Pilny, A., Schecter, A., Poole, M. S., & Contractor, N. (2016). An illustration of the relational event model to analyze group interaction processes. *Group Dynamics: Theory, Research, and Practice*, *20*(3), 181. doi:10.1037/gdn0000042
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. doi:10.1111/ajps.12103
Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, S., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12*(2–3). doi:10.1080/19312458.2018.1455817
Russell Neuman, W., Guggenheim, L., Mo Jang, S., & Bae, S. Y. (2014). The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, *64*(2), 193–214. doi:10.1111/jcom.2014.64.issue-2
Scharkow, M., & Bachl, M. (2017). How measurement error in content analysis and self-reported media use leads to minimal media effect findings in linkage analyses: A simulation study. *Political Communication*, *34*(3), 323–343. doi:10.1080/10584609.2016.1235640
Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. doi:10.1177/0002716215572084
Soroka, S., Young, L., & Balmas, M. (2015). Bad news or mad news? sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 108–121. doi:10.1177/0002716215569217
Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.
Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). *Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions*. Proceedings of the 25th international conference on world wide web (pp. 613–624). ACM Digital Library, New York, NY.
Taylor, S. J. (2013). *Real scientists make their own data*. Retrieved from https://seanjtaylor.com/post/41463778912/real-scientists-make-their-own-data
Trilling, D. (2017). Big data, analysis of. In *The J. Matthes, C. S. Davis, R. F. Potter (Eds.), international encyclopedia of communication research methods*. New York, NY: Wiley Online Library.
Tucker, J. A., Nagler, J., MacDuffee, M., Metzger, P. B., Penfold-Brown, D., & Bonneau, R. (2016). Big data, social media, and protest. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (p. 307). Cambridge, UK: Cambridge University Press.
Van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content* (dissertation). Charleston, SC: BookSurge.
Van Atteveldt, W., Sheafer, T., Shenhav, S., & Fogel-Dror, Y. (2017). Clause analysis: Using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 Gaza War. *Political Analysis*, *25*(2), 207–222. doi:10.1017/pan.2016.12
Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., van de Velde, B., & Oegema, D. (2015). Blinded by the light: How a focus on statistical "significance" may cause p-value misreporting and an excess of

p-values just below. 05 in communication science. *Communication Methods and Measures*, *9*(4), 253–279. doi:10.1080/19312458.2015.1096333

Vidgen, B., & Yasseri, T. (2016). P-values: Misunderstood and misused. *Frontiers in Physics*, *4*, 6. doi:10.3389/fphy.2016.00006

Wallach, H. (2016). Computational social science: Towards a collaborative future. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (p. 307). Cambridge, UK: Cambridge University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108

Weber, M. (2018). Methods and Approaches to Using Web Archives in Computational Communication Research. *Communication Methods and Measures*, *12*(2–3). doi:10.1080/19312458.2018.1447657

Weber, R., Mangus, J. M., & Huskey, R. (2015). Brain imaging in communication research: A practical guide to understanding and evaluating fmri studies. *Communication Methods and Measures*, *9*(1–2), 5–29. doi:10.1080/19312458.2014.999754

Westlake, B. G., & Bouchard, M. (2016). Liking and hyperlinking: Community detection in online child sexual exploitation networks. *Social Science Research*, *59*, 23–36. doi:10.1016/j.ssresearch.2016.04.010

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). *Who says what to whom on twitter*. Proceedings of the 20th international conference on world wide web (pp. 705–714). ACM Digital Library, New York, NY.

Xu, P., Wu, Y., Wei, E., Peng, T.-Q., Liu, S., Zhu, J. J., & Qu, H. (2013). Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2012–2021. doi:10.1109/TVCG.2013.221

Yang, J., & Leskovec, J. (2010). *Modeling information diffusion in implicit networks*. Data mining (icdm), 2010 IEEE 10th international conference on (pp. 599–608). New York, NY: IEEE Press.

Yang, M., Adomavicius, G., Burtch, G., & Ren, Y. (2018). Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, *29*, 4–24. doi:10.1287/isre.2017.0727