A Roadmap for Computational Communication Research

Wouter van Atteveldt, Drew Margolin, Cuihua Shen, Damian Trilling, René Weber

CCR 1 (1): 1–11
DOI: 10.5117/CCR2019.1.001.VANA

Abstract

Computational Communication Research (CCR) is a new open access journal dedicated to publishing high quality computational research in communication science. This editorial introduction describes the role that we envision for the journal. First, we explain what computational communication science is and why a new journal is needed for this subfield. Then, we elaborate on the type of research this journal seeks to publish, and stress the need for transparent and reproducible science. The relation between theoretical development and computational analysis is discussed, and we argue for the value of null-findings and risky research in additive science. Subsequently, the (experimental) two-phase review process is described. In this process, after the first double-blind review phase, an editor can signal that they intend to publish the article conditional on satisfactory revisions. This starts the second review phase, in which authors and reviewers are no longer required to be anonymous and the authors are encouraged to publish a preprint to their article which will be linked as working paper from the journal. Finally, we introduce the four articles that, together with this Introduction, form the inaugural issue.

Keywords: computational communication science, computational social science, open science, research transparency

An increasing part of our daily life is organized and experienced online, from connecting with friends and reading news to shopping, entertainment,

and even dating. Most of these online actions leave 'digital traces' that offer unprecedented opportunities for scholars to explore, theorize, and test hypotheses about the way humans think, behave, and interact (Lazer et al., 2009; Shah, Cappella, & Neuman, 2015). In addition, human artifacts and knowledge such as scholarly and non-scholarly articles, records of historical events, song lyrics, stories, etc., that provide rich information on the context of human behavior, are increasingly available in digital form. Most of these online 'digital traces' are communicative in nature. Therefore, communication science, perhaps more than any other social science, is in a promising position to leverage these rich data sources to form a better understanding of human communication and behavior (Hilbert et al., 2019).

Computational Communication Science (CCS) is the label applied to the emerging subfield that investigates the use of computational algorithms to gather and analyze big and often semi- or unstructured data sets to develop and test communication science theories (Van Atteveldt & Peng, 2018). In recent years, scholarly interest in this subfield increased dramatically, as evidenced, for instance, by the strong growth of the Computational Methods Division within the International Communication Association (ICA), the largest international representation of communication scholars. One testament of this interest is the new open access journal Computational Communication Research, in which this article is published, and the many recent and upcoming special issues on computational communication science and related topics (see e.g. Alvarez, 2016; Peng, Liang, & Zhu, 2019; Shah et al., 2015; Van Atteveldt & Peng, 2018).

Method and theory development are necessarily synergistic (Greenwald, 2012). New methods, from the telescope to DNA sequencing, have often been instrumental to scientific progress by changing our perception of reality and allowing new questions to be asked (Hilbert et al., 2019). New methodologies and analytical approaches can lead to new findings which in turn can be used to formulate or refine theories. At the same time, theories suggest research questions that inspire the development of new methodologies. Neither methodological nor theoretical development is superior in science (Weber, Fisher, Hopp, & Lonergan, 2017). With its unique set of strengths and weaknesses, CCS is in a position to complement the traditional methodological toolkit and enhance the paradigm of method-theory synergy in communication science. For instance, going from self-reports in lab settings to modeling actual behavior in its natural social setting can alleviate many of the external and ecological validity issues of experimental studies. Moving from small-*N* cross-sectional surveys or panels with long time intervals to large-*N* real-time measurements can help overcome the

internal validity problems of current observational studies. Finally, although large data sets do not guarantee high quality data, more data points can help overcome problems of low statistical power and allow the researcher to zoom in on specific subpopulations or test more complex models than is possible with traditional behavioral studies.

That said, there are a number of specific challenges that will need to be addressed in a vibrant and critical community of computational communication scientists if CCS is to fulfill its full potential (see also Van Atteveldt & Peng, 2018). First, the ownership of many of the required data sets by (social) media companies and other commercial entities threatens the accessibility of data and the reproducibility of studies. Second, "big" data sets are often a by-product of naturally occurring behavior, and may not be representative for the actual behavior of interest: expressed attitudes on, for instance, Twitter, review websites, or dating apps might be quite different from the attitudes in the general public. Third, computational methods are not immune from replicability problems. A high number of researcher degrees of freedom combined with a lack of currently established standards for many new methods can jeopardize the scholarly scrutiny which is essential in assuring additive science and replicability. Finally, CCS requires unique skill sets (e.g. programming, data handling) which may lead to a rethinking of our educational programs and the institutional incentives for developing and maintaining these skill sets.

These considerations show that to be successful, CCS will have to emphasize research transparency, reproducibility, and collaboration (Klein et al., 2018; Nosek et al., 2015; Van Atteveldt, Strycharz, Trilling, & Welbers, 2019). Research transparency and reproducibility is needed to generate long-term trust in this new paradigm. Collaboration among a diverse set of stakeholders is needed to create synergies between methodological and theoretical progress, develop and maintain complex computational software, update criteria for hiring, tenure, and grant approvals, and provide researcher with access to proprietary data sets.

Why do we Need a New Journal?

Why do we need a new journal to tackle these challenges? While some may view computational research as simply a methodological extension to existing communication research techniques and topics, we believe it creates a broad and integrated set of opportunities and challenges for the field that include debates over epistemology, ethics and the role of publication

in the scientific process (Anderson, 2008; Kitchin, 2014; Lazer & Radford, 2017; Tufekci, 2015; Van Atteveldt & Peng, 2018). To address these opportunities and challenges an integrated, communal effort is needed to develop, debate, and demonstrate best practices–that is, to develop relevant paradigms–that guide future research (Margolin, 2018; Pfeffer, 1993).

Such work can continue, as it has over the past decade, in articles scattered among the top communication journals and computational social science conference proceedings. However, we believe there are important advantages to providing a specific outlet that addresses all facets of this conversation. First, many papers can contribute to important conversations within the computational community but, understandably, are not recognized as valuable by general interest or other, topic specific journals. Thus, the best judges of their contribution are editors and reviewers who share an interest and understanding of the relevant issues. Second, as much as computational communication studies provide unique opportunities, they also face unique challenges. As a consequence, the evaluation criteria applied to computational communication studies can differ significantly from those applied in other sub-fields (Margolin, 2018). Some traditional criteria may be not strict enough for computational work. For example, obtaining large samples with sometimes hundreds of thousands of observations is usually not a problem for computational studies, but renders classical hypothesis testing as problematic ("everything is significant"). Yet other criteria may be too restrictive, such as the still widespread tendency not to publish null findings. Reviewers selected mostly on substantive expertise may not appreciate these unique challenges in computational studies. This can lead both to methodologically flawed articles being accepted, and to good computational work being rejected because it is held to the standards of classical methodology.

The third motivation for the journal is to actively promote a consistent and coherent set of standards for addressing these unique challenges. The challenges of computational communication research apply across theoretical topics, methodological best practices, and ethical commitments. Inevitably, some of the ideal best practices will come into conflict. For example, accessibility and reproducibility can often conflict with ethical concerns. Here the journal can serve as both a forum to organize the conversation around these topics as well as a place to work towards and implement an emerging consensus. Finally, we recognize that the research topics of a computational communication research journal are intrinsically tied to a set of computational technologies that are rapidly developing. We thus believe it is important that a computational communication research

journal invites and welcomes innovations and discoveries that have the potential to push the envelope in state-of-the-art communication science, but also come with an elevated risk of failure. Scientific research is driven by a sound rationale and method, and should be inherently risky. We envision CCR to be on the leading edge of risky proposals to social scientific practice, with the hope that our collective successes (and failures) can inform the communication field more broadly.

What Kind of Research Does CCR Seek?

A journal needs to develop and articulate a clear picture of what it is looking for to guide the decisions of authors, reviewers, and editors. CCR welcomes research that contributes to our theoretical understanding of human communication. We define a theoretical contribution as one that is additive to prior work by altering the field's existing understanding of and expectations for communication phenomena. These contributions are best achieved by formulating hypotheses and research questions that are risky, that is, include claims that are not self-evident and in fact are likely to be wrong. Replications and studies that test the soundness and boundary conditions of existing theory also qualify as good strategies. Of course, a logical consequence of pursuing risky research is that computational scholars will see rejections or null-findings of their claims more often than their support. Given a well-argued claim, reliable and valid measures, as well as a sound analytical methodology, CCR is committed to value null-findings as a contribution that increases knowledge. If computational scholars honestly report what – against their expectation and best-practice efforts – has not worked, then other can learn, build on these efforts, and thereby contribute to additive science. This said, there are three primary ways in which articles can contribute:

1. By applying computational methods to new or existing theoretical questions. Importantly, CCR's emphasis on additive contributions means that research need not exclusively test hypotheses nor feel compelled to produce significant results. Nonetheless, whether deductive or inductive, analysis should be clearly linked to substantive theoretical questions and what is already known, or suspected to be known, with regard to them. Claims and conclusions should be explicit – naming boundary conditions and alternative explanations – and, of course, well supported by the data. Showing that a theory is at odds with data is a relevant finding, but only if alternative explanations can be reasonably

ruled out, and if accompanied by a clear argument indicating why the theory should have been applicable.

2. By developing, adapting, and/or validating methods. For this, the researcher needs to show that the method/tool is reliable and valid; that it is useful for understanding communication; and that it is better (by some measure) than existing tools that do that task. In most cases, tools or method papers should include quantitative validation on a gold-standard data set that was not used for development and that is representative of some use case relevant to communication research.
3. By creating or adapting datasets and making them accessible and searchable. Shared datasets are important because it makes it easier to compare and replicate research by offering a common point of reference. In publishing a description of a data set, it should be clear how it was gathered and preprocessed. Where possible, the raw data and cleaning procedure should be published alongside the final data set. Data should be as open and accessible as possible. For data that cannot be fully shared for legal or privacy reasons, as much as possible of the data should be shared openly (i.e. metadata, annotations, and/or anonymized versions), and where possible a procedure for acquiring the sensitive data should be given that is in principle accessible to all researchers.

CCR demands transparent and reproducible research. Computational analyses require many choices regarding design, preprocessing, and parameter tuning, and transparency are needed to allow scrutiny of these choices. As digital data and analysis code can be shared easily, computational research can be at the forefront of the open science philosophy (Munafò et al., 2017; Nosek et al., 2015). Most articles in CCR should be accompanied by an online appendix in a form that encourages reproducibility and reusability. For tool and software contributions, we expect software to be published open-source on GitHub or an equivalent service and in the repository that is normal for the programming language used, e.g. Pypi or CRAN. For articles presenting substantive and/or methodological analysis results and data contributions, we expect an online research compendium published on GitHub or an equivalent service. Such a compendium contains the data, code, and results, and makes it explicit how the code is used to derive the results from the raw data (Marwick, Boettiger, & Mullen, 2018; Van Atteveldt et al., 2019). By publishing this on GitHub rather than depositing it in a service such as DataVerse, the code can be a living document

rather than just a snapshot. Reproducibility and persistence is guaranteed by storing the final (and if applicable, raw) data on DataVerse in addition, and archiving the named release of the repository corresponding to the publication. An optional template for such a compendium, including code for automatically testing and generating containers, will be made available through the CCR website.

The CCR Review Process

Like most journals in our field, CCR will publish articles only after a rigorous peer-review process. However, in addition to employing a new substantive scope, open access publication, and openness for data and tool publications, CCR is also introducing a procedural innovation: a "two-phase review process" in the way articles are published.

In the first phase, a traditional double blind 'adversarial' review takes place, where the central task for the reviewer and editors is to judge whether a manuscript is (potentially) publishable: is it high-quality, novel (including direct replications), and relevant. The outcome of phase one is either rejection or an *intent to publish*: a conditional decision to accept the manuscript for publication dependent on satisfactory revisions. After this intent to publish decision, the author is encouraged to publish the manuscript via an open science archive like SocArXiv. The journal website will link to this manuscript as a 'working paper'. Any revisions in this phase are not required to be blinded. The reviewers also get the option to be publicly identified on the article if published.

The purpose of this two-phased approach is to better align the incentives of authors and reviewers so that work is published both more quickly and with higher quality. Specifically, the job of the first phase is to identify valuable, if not yet wholly optimized research. Blind review, and the somewhat adversarial nature of the process, are essential in this phase to distinguish high quality submissions. Once there is agreement on the overall value of the manuscript, however, the preprint process is designed to alleviate authors' anxiety (and potential hostility) regarding the status of their manuscript, as well as to encourage reviewers to focus on concrete, constructive changes rather than marshalling arguments to 'kill' the paper.

Additionally, we offer the option of pre-registering research. While it may not be equally applicable to all types of computational research, it can be a useful tool to help our goal of avoiding bias against null-findings. We therefore will also accept registered reports as submissions, in which a

introduction, theory, and methods are specified in advance, but data have not been collected and analyzed yet. In this case, the first phase of the review process is conducted on the basis of the preregistered report, meaning that the report will be sent out for review and an intent to publish the final article can be given on the basis of this review, independent of research outcomes but of course conditional on robust and transparent methodology in accordance with the preregistration. We encourage the use of preregistration services such as the Open Science Framework or aspredicted.org and/or the dissemination of the registered report as a preprint once intent to publish is given.

This two-phase process and use of registered reports is experimental by design and should be seen as a first step in moving towards a more interactive and less adversarial review system. It is not clear how well it will work. Nonetheless it is one of the commitments of CCR to try new ideas that might improve the convoluted, and generally under-examined, publishing process.

Introduction to the first issue

The articles in this first issue present a snapshot of all aspects of computational communication research. *Hopp, Schaffer, Fisher, and Weber* present the Interface for Communication Research (iCoRe), a user-friendly web interface to access, explore, and analyze the Global Database of Events, Language, and Tone (GDELT). This interface makes it easier to work with GDELT to answer substantive communication questions, as well as enhancing the transparency and replicability of such work by providing a standardized query interface. The authors demonstrate in three theory-driven case studies the usefulness of iCoRe.

Pak uses Structural Topic Models (Roberts et al., 2014) to show how the Twitter feed of newspapers differ from their online content. This study shows how state-of-the-art analysis techniques can be used to study journalistic choices and how they differ for different audiences and contexts.

Haim and Nienierza present an open source browser plug-in that they use to observe both the content and context of the consumption of (public) Facebook posts. They also present a proof-of-concept study that, although highlighting the technical and social difficulties of recruiting participants for digital tracking studies, does show how the interaction with posts can be recorded, including scrolling, liking, and clicking links within a post.

Kim, Yang, Kim, Hemenway, Ungar, and Cappella used state-of-the-art recommender system techniques to create personalized health communication messages in a longitudinal study. Their results show that personalized messages have an improved effect compared to either showing the overall most preferred message or a random message.

Taken together, these four articles represent substantive computational scholarship in journalism health communication, and framing research. In addition, these articles contribute to making data and computational tools more accessible to communication scholars. We are confident that this is just the beginning of a stream of great research articles, and we look forward to your contributions and reviews.

Author Note

We would like to thank all reviewers, submitters, and editorial board members for contributing to the journal and for their feedback on this introduction. We would also like to thank Amsterdam University Press and especially our founding gold sponsors (Vrije Universiteit Amsterdam, The Network Institute, the University of Amsterdam / ASCoR) and founding silver sponsors (The Hebrew University of Jerusalem, The Center for Information Technology and Society at UC Santa Barbara, and the Computational Communication Science Lab of the University of Vienna), for making this journal possible.

References

Alvarez, R. M. (2016). *Computational social science*. Cambridge University Press.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*(2), 99–108. https://doi.org/10.1177/1745691611434210

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., ... Zhu, J. J. H. (2019). Computational communication science: A methodological catalyzer for a maturing discipline. *Accepted for publication in International Journal of Communication*.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. https://doi.org/10.1177/2053951714528481

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., ... Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*. https://doi.org/10.1525/collabra.158

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... Alstyne, M. V. (2009). Computational social science. *Science* (*New York, NY*), *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*, 19–39.

Margolin, D. (2018). *The computational contribution: A symbiotic approach to integrating computational research into the communication field*. Presented at the annual conference of the International Communication Association (ICA). Prague, Czech Republic.

Marwick, B., Boettiger, C., & Mullen, L. (2018). Packaging data analytical work reproducibly using r (and friends). *The American Statistician*, *72*(1), 80-88. https://doi.org/10.1080/00031305.2017.1375986

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Peng, T.-Q., Liang, H., & Zhu, J. J. (2019). Introducing computational social science for asia-pacific communication research. *Asian Journal of Communication*, *29*, 205–216. https://doi.org/10.1080/01292986.2019.1602911

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of management review*, *18*(4), 599–620.

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.

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. https://doi.org/10.1177/0002716215572084

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colorado Technology Law Journal*, *13*(2), 203 – 218.

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

Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Towards open computational communication science: A practical roadmap for reusable data and code. *Accepted for publication in International Journal of Communication*.

Weber, R., Fisher, J. T., Hopp, F. R., & Lonergan, C. (2017). Taking messages into the magnet: Method-theory synergy in communication neuroscience. *Communication Monographs*, *85*(1), 81–102. https://doi.org/10.1080/03637751.2017.1395059

About the authors

Wouter van Atteveldt: Vrije Universiteit Amsterdam
Correspondance address: wouter@vanatteveldt.com

Drew Margolin: Cornell University

Cuihua Shen: UC Davis

Damian Trilling: University of Amsterdam

René Weber: UC Santa Barbara