Open Collaborative Hyperpapers: A Call to Action

Alberto Dainotti
CAIDA, UC San Diego, USA
alberto@caida.org

Ralph Holz
University of Sydney, Australia
ralph.holz@sydney.edu.au

Mirja Kühlewind
ETH Zürich, Switzerland
mirja.kuehlewind@tik.ee.ethz.ch

Andra Lutu
Telefónica Research, Spain
andra.lutu@telefonica.com

Joel Sommers
Colgate University, USA
jsommers@colgate.edu

Brian Trammell
ETH Zürich, Switzerland
brian@trammell.ch

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

Drawing on discussions at various venues, we envision a publishing ecosystem for Internet science, supporting publications that are self-contained, interactive, multi-level, open, and collaborative. These publications, which we dub *hyperpapers*, not only address issues with reproducibility and verifiability of research in Internet science and measurement, but have the potential to increase the impact of our work and change how collaborations work in the field. This note announces initial experiments with Internet measurement hyperpapers with the help of common, tested technologies in data science and software development, and is a call to action to others to come build out this vision with us.

CCS CONCEPTS

• **General and reference → Computing standards, RFCs and guidelines**;

KEYWORDS

Publishing, Science, Reproducibility

1 MOTIVATION

Scientific papers were born as a means to share novel scientific knowledge. However, over time publications have also become the main metric for career advancement. This shift has influenced the whole publishing process, from the generation of ideas, data and results to how they are shared. If we step back and look at the currently established process for scientific paper authoring and publishing, including conventions and formats, it is clear there is room for optimization for the good of science and education (e.g., have we struck the right balance between confidentiality and openness? Are there opportunities from recent technologies and collaborative practices that we can leverage?)

Discussion at various venues, including the CAIDA AIMS workshop in March 2018 [1] and the seminar "Encouraging Reproducibility in Scientific Research of the Internet" at Schloss Dagstuhl in October 2018 [2], identified issues with the publishing ecosystem through which Internet measurement studies are disseminated that have an impact on reproducibility and verifiability of Internet science.

Problems we are facing today. The current publication ecosystem discourages incremental work by keeping the trade-off between required effort and consequent gain heavily unbalanced. On the one hand, the scarcity of data, tools and documentation to make research studies reproducible creates artificial barriers for building upon peers' work. On the other hand, we often put disproportionate emphasis on novelty, which can lead to rejection of valid work that *merely* extends previous results. The result is a vicious circle in which there is little incentive for researchers to release artefacts for others to work with and build upon. Our research community has recently boosted efforts to break this pattern. Some venues in our field are starting to experiment with artefacts such as code and data being submitted together with papers. While this is far better than the common status quo of boutique code running on secret data, these efforts are more suited to address archival requirements than those of repeatability and verifiability.

Even when artefacts are provided, problems remain. Tooling support is lacking to help reviewers to efficiently verify these artefacts as part of the paper review process, for example through repetition of experiments or analysis. The difficulty this causes makes it difficult to make partial verification of results during review the norm.

A culture of secrecy, both due to conditions of access on proprietary data sources as well as to the tradition of establishing academic priority, adds further barriers to scientific inquiry. Dead-ends and negative results stay secret within the groups that find them. Researchers new to the field, and

those without connections to established cliques of collaborators, can find it hard to get started in impactful Internet measurement.

It is not the goal of this note to advocate for the immediate death of the current publication process. Indeed, this process performs a valuable gatekeeper function, both by providing incentives for authors to submit papers for review and for reviewers to review them, as well as for bringing work from smaller communities (in our case, Internet measurement) to larger audiences. The spirit of our proposal is instead to suggest ways in which we, as a community, can address the issues above while retaining positive aspects of the current publication process.

2 A VISION FOR THE FUTURE

We propose to leverage recent developments in platforms and tools for data science and scientific collaboration to build an experimental publishing ecosystem for Internet measurements based on *hyperpapers*. The technical details of this experiment are left to future work, but the outlines of the properties of these papers already seem clear:

Hyperpapers are interactive and self-contained. Ideally, a full hyperpaper contains all the data from which results, plots, and conclusions in the paper are drawn, as well as source code implementing the analytic tasks distilling those results from the raw source data. A hyperpaper is interactive, allowing both changes to the raw source data and to the analysis code to be reflected in the analytic products (tables, charts, etc.) in the paper, supporting easy exploratory analysis of a dataset and/or measurement analysis methodology as well as incremental changes to published work. The hyperpaper includes one or more rendered versions, both for compatibility with existing publishing channels (e.g., PDF) and to minimize the requirements for the readers who will not want to use the paper's interactive features or whose devices do not allow it. Accessibility is thus maintained at least at the current level, although we emphasize that this is another area where improvement is needed.

The raw source data may be included with the hyperpaper either by value or by reference, and the analysis code can be configured to run either on a local machine or on some specified remote infrastructure. Hyperpapers can also require credentials to enable their interactivity, in order to support arrangements where raw and/or intermediate data is subject to access control.

Hyperpapers are multilevel. The initial view a reader will have of a full hyperpaper includes the typical prose of a paper (abstract, introduction, explanation of the questions answered, description of the methodology, results, graphs, and so on). Analysis products, such as charts and tables, can be expanded to show how they were derived.

However, the paper can also be expanded in other ways. A section of prose may be linked to an alternate view, information for an alternate audience, related content, or allow to drill down on some interesting result. For example, the paper could contain an executive summary describing the utility of its insights for the general public, network operators, regulators, and so on; a methodology could be expanded to tell an expanded narrative of roads not taken, or unsuccessfully taken, and why; or an introduction could be expanded with introductory information on a measured protocol not necessary for practitioners in the field but useful for students just learning it; and so on. These sorts of expansions would replace the creation of multiple papers on a subject to multiple venues, e.g. [3] and [4].

Taken to its logical conclusion, multilevel hyperpapers allow a full explanation and report on a research project's activity and the evolution of the chain of hypotheses, addressing the dearth of negative results, while maintaining narrative coherence and conciseness in the paper's main "trunk" While this opens up new questions, e.g., for page counts and the review process, this could be addressed by considering all extensions beyond this trunk as appendices for review purposes.

Hyperpapers are open and collaborative. The interactive, self-contained, and multilevel nature of hyperpapers enables —indeed, may require— entirely new ways of working together as researchers. Starting incremental work on a paper, or beginning to verify and reproduce it, becomes a simple matter of forking it and given permission to access the data. To the extent that the hyperpaper infrastructure is integrated with a collaboration platform, papers can become open and living projects, with researchers performing associated work, interdisciplinary co-authors writing expansions for specific audiences beyond the usual networking conference audience, and even reviewers becoming part of a long-running, data-centered conversation about a particular hypothesis of how the Internet works.

3 HOW DO WE GET THERE?

As in all things scientific, by standing on the shoulders of giants. While this vision may seem like science fiction, we submit that enough of the underlying infrastructure behind this vision exists that rudimentary experimental hyperpapers for Internet measurements can be written today.

The perennial problem of setting up environments for data analysis without needing to replicate a full toolchain with dependencies from scratch is partially solved today by virtualization and containerization tools such as Vagrant and Docker. Problems of scale are addressed by the easy (if sometimes costly) widespread availability of cloud infrastructure from multiple providers. Integration of data analytics with

authoring environment interleaving text and interactive visualizations is supported by data analysis notebooks such as JupyterLab and Apache Zeppelin. GitHub has emerged as the de facto standard for integrating version control of digital artefacts with a collaboration environment, and its model of working is suited to open collaborative papers as we envision them, which have a fair amount in common with the long-running open-source projects GitHub was originally built to support. Of course, all of these technologies are supported by the web platform, decades of continuous investment in which has brought us to a world where almost all of the work of research can be done in a standard browser.

We have identified two main gaps in technical infrastructure necessary for a full initial realization of this vision:

- First, while some research studies can be done with data or models that can easily be stored in an ad-hoc format within the hyperpaper itself, large-scale Internet measurement studies need access to large data sets mediated through some interface. This exists for certain data sources (e.g., the RIPE Atlas and the BGPStream APIs), but a full realization would require the creation and standardization of interfaces for retrieval of data and metadata for each broad type of measurement activity.
- Second, the distribution of rendered versions of papers is currently possible for scientific notebook environments, but these render to a webpage that is not necessarily optimized for accessibility. Tooling to render a view of hyperpaper as a PDF according to the visual style for a given venue, for example, is necessary to support the full multi-rendering functionality of the vision above. We consider this a simple matter of engineering, though.

Addressing these gaps will take time; in the meantime, the authors are currently working to build initial hyperpaper versions of some of their previous and current work, one of which [5] has already been built using some of the technologies we identify as appropriate for hyperpapers. We welcome community collaboration as we develop it into an architecture document for an initial realization of a hyperpaper platform; see our GitHub organization at https://github.com/hyperpaper.

ACKNOWLEDGMENTS

This editorial is largely based on discussions at CAIDA's AIMS 2018 workshop and Dagstuhl Seminar 18412 on *Encouraging Reproducibility in Scientific Research of the Internet*; we thank the participants for the discussions leading to this work.

REFERENCES

[1] Alberto Dainotti. A Conversation on Hyperpapers & Open Co-Authoring, Mar 2018. http://www.caida.org/publications/presentations/2018/hyperpapers_open_coauthoring_aims/ —Accessed 31 December 2018.

[2] Dagstuhl Seminar 18412. Encouraging Reproducibility in Scientific Research of the Internet, Oct 2018. https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=18412 —Accessed 31 December 2018.

[3] Amogh Dhamdhere, David D. Clark, Alexander Gamero-Garrido, Matthew Luckie, Ricky K. P. Mok, Gautam Akiwate, Kabir Gogia, Vaibhav Bajpai, Alex C. Snoeren, and kc claffy. Inferring Persistent Interdomain Congestion. In *ACM SIGCOMM*, Aug 2018.

[4] David D. Clark, Amogh Dhamdere, k.c. claffy, Matthew Luckie, and Alexander Gamero-Garrido. Detecting Internet Congestion at Interconnection Points: An Empirical Analysis. In *TPRC46: Research Conference on Communications, Information and Internet Policy*, Sep 2018.

[5] Brian Trammell. On the Suitability of RTT Measurements for Geolocation, Aug 2017. https://github.com/britram/trilateration/blob/master/paper.ipynb —Accessed 6 December 2018.