Important Topics in Causal Analysis: Summary of the CAWS 2021 Round Table Discussion

Erich Kummerfeld, PhD ERICHK@UMN.EDU
University of Minnesota, Institute for Health Informatics

Tom Woolf, PhD TWOOLF@JHU.EDU
Johns Hopkins University, Applied & Computational Mathematics

Will Glad, MS WTGLAD1@GMAIL.COM
Johns Hopkins University, Applied & Computational Mathematics

Michèle Sebag, PhD MICHELE.SEBAG@LRI.FR
TAU, CNRS – INRIA – LISN, Univ. Paris-Saclay

Sisi Ma, PhD SISIMA@UMN.EDU
University of Minnesota, Institute for Health Informatics, Dept of Medicine

Editor: Sisi Ma, Erich Kummerfeld

CAWS 2021 had 61 registered participants representing a wide range of researchers and data science professionals. These researchers have a variety of levels of familiarity with causal discovery and related disciplines, their interest and expertise span the spectrum of applying causal discovery methods for domain specific real-world problems to theoretical aspects of causal discovery. 12% of the participants are undergraduate students, 27% are graduate students, 9% are postdocs, 45% are researchers in academia, 15% are researchers in the industry. The participants are from 29 institutions in North America, Asia, and Europe.

At CAWS 2021, a round table discussion was held to identify the most important challenges facing causal analysis. The participants came up with 5 challenges, and discussed their importance and the current state of the field's progress on those issues. Overall, these issues were considered the most important topics in causal analysis at the time of the discussion, and the group agreed that if these problems could be solved, then causal analysis would be much better positioned to make enormous contributions to scientific discovery across many domains. In this paper, we briefly summarize these challenges, as a call for action to the entire causal analysis community.

The specific challenges are:

1. Methodology: Analyzing data with heterogeneous data types and distributions
2. Methodology: Incorporating non-causal relationships into causal models
3. Methodology: Collecting data sets with gold standards
4. Communication: Communicating causal analysis results to those outside the field
5. Education: Lowering the barrier to entry for people interested in causal analysis methods

Analyzing data with heterogeneous data types and distributions. Everyone in the discussion agreed that real world data sets almost universally contain a mixture of variables with different distributions. Examples include a mixture of continuous and categorical variables, a mixture of Gaussian and non-Gaussian variables, linear and nonlinear relationships, and so on. It was pointed out that some progress has been made on this issue, namely that there are methods in TETRAD for doing causal discovery on data with both categorical and Gaussian variables. However, the reliability of those methods on finite sample real data is uncertain at this time, and it is only a limited case among the many types of variable mixtures that exist.

Incorporating non-causal relationships into causal models. Both conceptually and in real world data sets, there are many constructs that are associated for non-causal reasons. For example, a depression score is often a sum of responses to several more specific depression items or questions. Since it is a sum, it is associated with the specific items, but it is wrong to say that the items cause the sum or vice versa. This is a clear violation of the most basic assumption of any currently existing causal analysis method, namely that the data was generated by a causal process. Similar situations, where variables are defined in relation to each other with equations, are ubiquitous across fields such as medicine and economics, and are commonly produced by data science processes such as feature construction and dimensionality reduction. In all these cases, causal analysts are currently forced to exclude some variables from the analysis. Ideally, it should be possible to model these variables all together, especially when it is already known a priori exactly how they are related to each other, and this information could be fed to an algorithm.

Collecting data sets with gold standards. At present, the vast majority of causal analysis methods do not use supervised learning, and can not use methods like holdout samples to produce an unbiased estimate of learning performance or accuracy. The primary reason for this is that there are scant data sets where the causal mechanism, the learning target of causal analysis methods, is known. All those participating in the discussion agreed that the lack of data sets with known data-generating causal mechanisms or models, a.k.a. "gold standards", is hindering the development and useful application of causal analysis methods. More resources should be devoted to collecting or producing data sets with gold standards.

Communicating causal analysis results to those outside the field. It was broadly agreed that it is difficult to communicate the results of causal analysis to readers, reviewers, and editors who are not familiar with these methods. Many scientists still believe that the kinds of causal analysis we do are not possible, and many readers can have knee-jerk reactions as soon as causality is discussed (or even mentioned) in the context of certain kinds of data (e.g. observational data, cross-sectional data). With a growing number of publications using causal analysis methods, it is clear that this hurdle is possible to overcome. Doing so is difficult, however, as there are no general guidelines or standard language to help causal analysis researchers communicate their findings. This is perhaps also related to the next issue, i.e. there is a lack of entry level educational material for causal discovery.

Lowering the barrier to entry for people interested in causal analysis methods. More people are becoming interested in causal analysis, however there is a severe lack of pedagogical material to aid prospective students who wish to learn about it. There is

no standardized educational material that the field agrees on. There is at least one text book intended to begin filling this gap, however it is not very complete and is targeted more at data scientists than at an applications-oriented audience. Students wishing to be self-taught, at present must rely on some primary source material that is obtuse and not intended for educational purposes, and/or a variety of recorded lectures from workshops of varying quality, content, and age. As such, the only reliable way to learn causal analysis at this time is one-on-one interaction with someone who understands the material, and this is unreliable, restrictive, and inefficient.

There are certainly many other challenges facing the field of causal analysis as well, however these 5 were the challenges identified at this round table. This indicates a consensus in the field that overcoming these challenges is of particular importance.