An architecture for automated estimation of the reliability, reproducibility, and robustness of behavioral and social science research

Technical point of contact:
Jordan W. Suchow[1]
University of California, Berkeley
+1 914 582 2646
suchow@berkeley.edu

In response to DARPA-SN-17-57.

[1] Written in consultation with M Pacer and Tom Griffiths.

Technical Description

Automated assessment of the reliability, reproducibility, and robustness of behavioral and social science research is unlikely to have a small, tight, closed-form solution because the causal process that culminates in reports of research findings is complex, with many hundreds of contributing factors, any of which may have flaws affecting the reliability of the final product. These flaws include subtle bugs in statistical software packages that perturb otherwise-sound calculations, errors introduced during copyediting that invalidate mathematical proofs, cultural momentum that drives subfields down dead-ends, fuzzy definitions that befuddle, third-party adversaries that seek to undermine the validity of a group's or nation's research, missing data, insufficient power, poorly trained assistants, statistical methods misaligned with the underlying problem, selection biases in the recruitment of human participants, overblown conclusions, fabricated data, closed-source tooling, unavailable descriptions of methods, file-drawer biases in what gets reported, misalignment of experimental measures to the underlying constructs of interest, demand characteristics, uncontrolled eye movements, mistranslation of survey materials, failure to obtain consent, time-of-day effects, confounded research designs, jargon that obscures, ill-specified stopping rules that lead to arbitrary outcomes, and more. There are many ways that behavioral and social science research can go wrong.[2]

In this way, the problem of assessing the reliability, reproducibility, and robustness of research is similar in structure to the problem of determining whether a document contains clean, well-edited prose. Thousands of linguistic blunders are possible, including redundancy, jargon, illogic, clichés, misspelling, inconsistency, misuse of symbols, malapropisms, oxymorons, hedging, apologizing, pretension, and more. Though there is no small, tight, closed-form solution to the problem of assessing prose, there is a solution. Consider, for example, *Garner's Modern English Usage*,[3] an authoritative usage guide with 11,000 entries covering a broad range of advice that can help writers produce clear and idiomatic prose. Or consider the *Federal Plain Language Guidelines*,[4] a guide created by employees of the U.S. federal government to promote writing that is clear, concise, and well-organized. Professional conferences such as the annual meeting of the *American Copy Editors Society* are dedicated to sharing knowledge about editing prose. And within the academy, organizations such as the *American Psychological Association* publish manuals whose guidance on style has been adopted as a standard. In particular, there have been numerous attempts to create automated analyses that gauge the quality of writing by

---

[2] To be fair, there are also many ways it can go right. However, at the risk of being overly pessimistic in framing, any feature that bodes well for a paper's reliability, reproducibility, and robustness can be recast into the flaw of lacking that feature. Did a researcher "publish the raw data" or not "fail to publish the raw data"? Both.

[3] Garner, B. A. (2016). Garner's Modern English Usage. Oxford University Press.

[4] http://www.plainlanguage.gov/howto/guidelines/FederalPLGuidelines/FederalPLGuidelines.pdf

implementing usage guides in software, and several of these have become successful products used by many writers. These tools have succeeded by recognizing the structure of the problem and architecting a technological solution that leverages that structure to make many small steps towards the ultimate goal.[5]

Beyond the flaws detected within a research product, professional researchers often have a finely tuned sense of the prior probability of new research claims within their domain of expertise — they can provide information relevant to the reliability, reproducibility, and robustness of research even before having consumed the contents of the research product under question. Aggregation schemes such as prediction markets[6] and the Bayesian truth serum[7,8] can uncover the prior probability of a research claim in the minds of experts.

To automatically assess the reliability, reproducibility, and robustness of behavioral and social science research, then, we recommend an architecture that combines (1) prior beliefs derived from a model of aggregated expert opinion with (2) evidence in the form of flaws detected in the final research product, to produce a posterior measure of the reliability, reproducibility, and robustness of the research. This process begins by cataloguing all known flaws affecting the reliability, reproducibility, and robustness of behavioral and social science research. Then, for each of these catalogued flaws, an automated procedure can be developed to detect it automatically. Each detector takes as input a PDF of a published paper and returns a list of all detected instances of that flaw.[9] Alongside this battery of flaw detectors is a measure of prior probability obtained from a model of aggregate expert opinion. Together, they constitute an automated measure of the reliability of the paper, and they can be combined, summarized, and reduced in various ways to provide a small set of actionable metrics relevant to the particular application domain.

Framing the problem in this way has the benefit of decomposing one hard problem into four subproblems of lesser difficulty (gathering the corpus, developing the flaw detectors, modeling

---

[5] Pacer, M., & Suchow, J.W. (2016). Pacer, M. D. & Suchow, J. W. (2016). Linting science prose and the science of prose linting. *Proceedings of the 15th Python in Science Conference*.

[6] Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347.

[7] The *Bayesian truth serum* is an improvement to Galton's "one person, one vote" method for aggregating knowledge, in a way that extracts superior knowledge from experts in the crowd. See Prelec, D. (2004). A Bayesian truth serum for subjective data. S*cience*, *306*(5695), 462–466.

[8] Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532–535.

[9] This solution is essentially a linter, a form of static program analysis. A *linter* is a computer program that, like a spell checker, scans through a document and analyzes it, producing a list of syntactic or stylistic violation.

aggregate expert opinion, and creating actionable metrics), one of which, development of flaw detectors, can be broken down further into several hundred sub-subproblems, each corresponding to the development of an automated procedure for detecting one of the flaws. These four subproblems are not easy. However, considering the problem in this way makes it possible to leverage a steady stream of incremental improvements to algorithms, training data, aggregation methods, and automated flaw detectors to result in a revolutionary technological solution that cannot be achieved through a monolithic attempt to judge reliability, reproducibility, and robustness. Choosing the right architecture is the key.

Gathering the corpus. Ideally, the corpus is a large, representative sample of all published papers in the behavioral and social sciences. It can be obtained through publishers or by other means. A performer responsible for generating this corpus might also take on the auxiliary task of preprocessing the corpus to produce normalized intermediate representations that will be broadly useful for many of the automated flaw detectors. These may include, for example, running OCR algorithms over the PDFs to extract text; or using other algorithms to extract authorship information, journal names, copyright information, and other relevant metadata. Perhaps the corpus could be so large as to include a sizable number of retracted, corrected, or otherwise updated papers, providing labeled examples of particular flaws.

Developing the flaw detectors. The subproblem of developing automated detectors of flaws relevant to a paper's reliability, reproducibility, and robustness neatly decomposes into several hundred sub-subproblems, one for each possible feature. Specifically, each sub-subproblem is that of building an automated detector that takes as input a PDF of a published paper (perhaps making use of the provided intermediate representations) and outputs a list of detected instances of that flaw. Though these flaw detectors would have a common input and output format, they would likely vary widely in what they target, drawing on many different sources of information within and external to the particular paper. For example, one may detect inconsistency across papers in how a term is used. Another may detect that a paper was retweeted with negative sentiment by Twitter accounts belonging to researchers who study ethics in big data research. And another may detect that a cited paper was retracted. What is critical is that they all speak the same language: shared input and output formats. Flaws, like usage errors in writing, vary greatly in how difficult they are to detect automatically: the easiest are trivial, assignable as homework problems to an undergraduate students learning programming, whereas the hardest are likely to be AI hard.[10] Importantly, these flaw detectors can be discovered, implemented, and validated independently across different performers, keeping the complexity of the overall program low despite an ever-increasing library. Each flaw detector can be evaluated by an auditor on various metrics, including

[10] For a fuller discussion of this in the context of usage errors in writing, see Pacer & Suchow (2016), referenced in footnote 5.

false alarm rate. An existing example of one such detector is statcheck[11], which detects several flaws in the reporting of statistical tests. Another is Proselint,[12] which has modules that detect redundancy and misuse of psychological terms.[13]

Modeling aggregate expert opinion. The field of natural language processing (NLP) has developed sophisticated methods for automatic text summarization[14] that extract the key claims from a paper automatically. Ground-truth data regarding expert opinions on the prior probability of research claims in the papers in the corpus can be obtained from professionals in the relevant subfield using the Bayesian truth serum method, showing them the claims extracted from the target paper. What accounts for the sensibilities of these experts when evaluating the extracted claims? Various machine learning and NLP techniques could be used to develop a model of these expert opinions, learning the features of a research product that cause the aggregate opinion of experts to assign a low prior probability to a claim. For example, a model may assign low prior probability to claims containing the concepts "psi", "ESP", and "mind-reading" because it finds that, in aggregate, experts in the field assign low prior probability to claims with those words.

Defining the metrics for applications. The combined output of the flaw detectors is a list of flaws relevant to a paper's reliability, reproducibility, and robustness. But an enumeration of flaws is rarely the ultimate goal. The output of the flaw detectors can be aggregated into individual metrics that are relevant to the goals of the particular application. For example, when assessing a paper, a reader may want to understand the validity of the paper's assumptions. Knowing that a key assumption in a paper is based upon a retracted paper may dramatically undermine one's confidence in that paper. Equipping PDF viewers with tools that annotate a paper by noting weaknesses in cited results would help viewers evaluate a paper's claims. Papers that rely heavily on other weak work could be appropriately down-weighted.

---

[11] See http://statcheck.io/.

[12] See https://github.com/amperser/proselint.

[13] Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., & Waldman, I. D. (2015). Fifty psychological and psychiatric terms to avoid: a list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. *Frontiers in Psychology*, *6*.

[14] Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4, 192-195.