Bibliometrics for Social Validation of High-Throughput Toxicology

Project Overview and Proposal

Daniel J. Hicks, Ph.D.

Background

The validation of novel scientific methods takes place at two levels. The first level, or *formal validation*, is the most familiar: methods are validated by showing that they are theoretically well-supported, their results can be replicated, and they agree with established methods (at least for cases where established and novel methods are expected to agree). The second level, or *social validation*, concerns the acceptance of novel methods by the relevant scientific community. A novel method may have high formal validation when performed by its developers, but fail to be generally accepted by the scientific community if it is difficult to use, requires equipment or materials that are expensive or otherwise difficult to obtain, is very slow, depends on assumptions that are not widely accepted or mathematical techniques that are not widely understood, and so on.[1]

The EPA's Chemical Safety for Sustainability program is developing an array of novel methods in high-throughput toxicology [HTT] to address the lack of toxicity and exposure data for tens of thousands of commercial chemicals. For some particular uses, HTT methods have been shown to have high formal validation (Browne et al. 2015). However, there has been no systematic study of the social validation of these methods. An analysis of the social validation of HTT could be useful whatever its findings: positive results (high social validation) could be useful when communicating the value of the research to audiences who are not familiar with HTT, while a careful analysis of negative results (low or mixed social validation) could help HTT researchers identify appropriate avenues for social validation.

[1] "Social validation" is my term for a concept that is widely studied in science and technology studies [STS] and philosophy of science. Key works on the role of social validation in scientific research include Fleck ([1935] 1979), Kuhn ([1962] 1996), Longino (1990).

Citation Network Analysis for Social Validation

This project proposes to measure the social validation of HTT methods developed by CSS researchers through a combination of bibliometrics, the analysis of citation network data, and qualitative literature review. In brief, social validation can be analyzed by examining the way in which a *core set* of research publications is embedded within an *extended network* of publications, representing the broader scientific community. To my knowledge, these kinds of computational approaches have not previously been used to examine the social validation of novel scientific methods.

Network construction

The core set for this analysis are articles in scientific research journals included in the CSS research products database through June 2015. The extended network is operationalized through a combination of *forward* and *backwards citation search*. In a forward citation search, given a paper A we attempt to find all papers B such that A is cited by B; we move forward in time along the true citation network. In a backward citation search, given a paper B we attempt to find all papers A such that A is cited by B; we move backward in time along the true citation network.

1. First, a forward citation search is used to find papers that cite the members of the core set. The papers found in this search are called *generation +1*.
2. Second, a backwards citation search is used to find papers cited by generation +1. These papers (that are not already in generation +1) are called *generation 0*. Note that the core set is a subset of generation 0.
3. Finally, a backwards citation search is used to find papers cited by generation 0. These papers (that are not already in generation 0 or generation +1) are called *generation -1*. The extended network consists of the union of generation +1, generation 0, and generation -1.

To conduct these searches, we use the online research database service Scopus. Scopus was chosen because (a) it is one of the four large-coverage research publication search databases (the others being Web of Knowledge / Web of Science, PubMed, and Google Scholar); (b) Scopus and PubMed both have well-documented APIs (systems for accessing the database using automated queries; see http://dev.elsevier.com), while I was unable to find documentation for the Web of Knowledge API;[2] and (c) Scopus contains more bibliographic information — that is, the contents of article bibliographies, which are needed to build a citation network — than PubMed.

[2] Google Scholar's Terms of Service prohibit automated queries.

The forward search in #1 is conducted manually; the backward searches in #2 and #3 are conducted using a Python script. Network analysis will also be conducted using Python scripts. These scripts will be written with replication in mind: the final versions will be made available on an open repository (a government repository, GitHub, Dryad, or some combination of these), with documentation enabling later users to easily understand the methods used in the analysis, replicate its results, and adapt it to similar projects (for example, for other ORD programs).

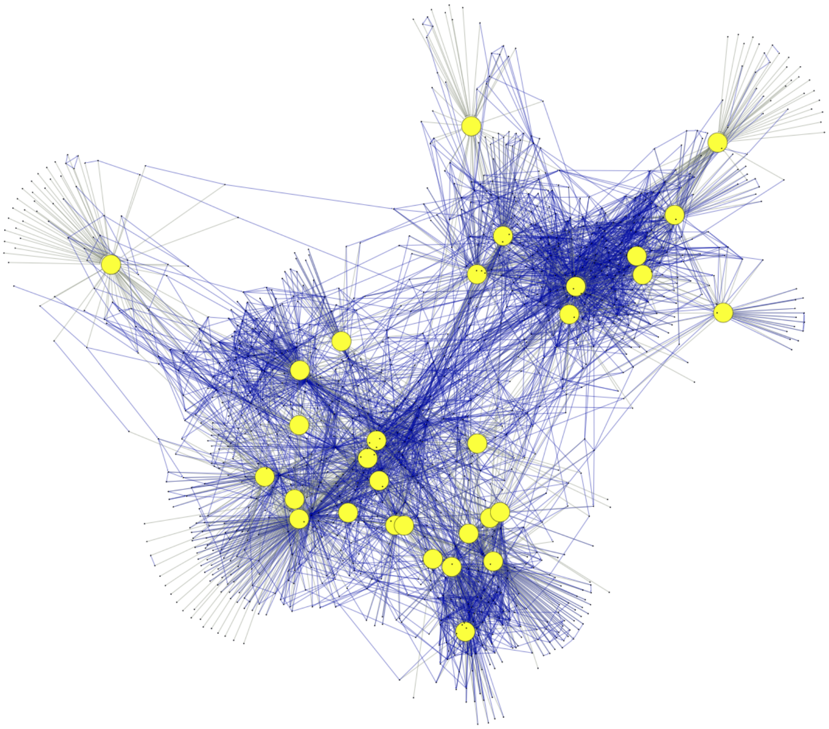

Figure 1: A preliminary citation network, based on a subset of the full core set. Large yellow nodes are in the core set; small blue nodes are in the extended network. Visualization in Gephi.

Quantitative Analysis

Within the extended network, the social validation of the core set can be measured in several ways. First, various *centrality measures* can be used to measure the impact or importance of individual papers, in more sophisticated ways than a conventional citation count. High centrality indicates more impact or importance, which suggests greater social validation. However, centrality measures apply to individual nodes; our interest here is in the core set as a collection of nodes.

Modularity measures can be used to measure the degree to which the core set is embedded in the extended network. One common definition of modularity compares the network to a random graph with the same number of nodes and same degree distribution (Newman 2006). This yields the modularity formula

$$Q = \frac{1}{2m} \sum_{v,w} \left(A_{vw} - \frac{k_v k_w}{2m} \right) \frac{s_v s_w + 1}{2}$$

where Q is the modularity statistic, m is the total number of edges in the network, v and w iterate over all nodes, A is the adjacency matrix ($A_{vw} = 1$ if there is an edge from v to w; 0 otherwise), k_v is the degree of node v, and s_v indicates the group membership of node v ($s_v = 1$ if v is in the core set, -1 otherwise).

The observed value of Q will be interpreted using several empirical comparisons:

- On the same extended network, we generate a large number (~1,000) random "core sets" by choosing random sets of nodes. Calculating the Q value for these random "core sets" gives us an empirical approximation for the sampling distribution of Q given the null hypothesis that the actual core set was produced by chance; and so this gives us a p-value for the actual core set. Because of the way the extended network was constructed, this null hypothesis is *a priori* implausible; but the p-value can instead be interpreted as the "distance" of the actual core set from one produced by chance (a smaller p indicates a less random core set).
- On the same extended network, we apply a community detection algorithm, such as the "Louvain algorithm" (Blondel et al. 2008). These algorithms attempt to find partitions of the network that maximize Q. Since the Louvain algorithm is non-deterministic and runs quickly, a large number of runs may be used to generate a sampling distribution and calculate a p-value, as with the random "core sets" comparison above. Here the p-value can be interpreted as the extent to which the core set cuts across different subcommunities within the extended network (a smaller p indicates more cross-cutting).
- Sample citation networks are available from the Stanford Network Analysis Project (https://snap.stanford.edu/index.html). The two methods described above can be applied to these networks, and the resulting sampling distributions can be compared to those of the extended network. This comparison will allow us to assess concerns that the construction of the extended network might somehow favor the core set.

Qualitative Analysis

The uses of the Q statistic described above can show that the core set is heavily discussed within the broader scientific community; but these statistics do not indicate the *valence* of this discussion. For example, controversial methods may be highly cited by critics, but these methods should be considered to have low or controversial social validity. To assess this valence, a random sample of generation +1 articles can be reviewed manually. Valence in general can be examined; in addition, the discussion of specific concerns surrounding HTT methods can be examined. Details of this phase of the project will be developed later.

(Catalini, Lacetera, and Oettl 2015) describes a text-mining approach that might be useful here.

Known Limitations

1. Every research database covers a limited (albeit large) set of the scientific literature and contains errors. The citation data from a research database should be understood as a large convenience sample, with the corresponding limitations. If desired, background research can be conducted to attempt to identify key limitations of the chosen database.

References

Blondel, Vincent, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment*, P10008. doi:10.1088/1742-5468/2008/10/P10008.

Browne, Patience, Richard Judson, Warren Casey, Nicole C. Kleinstreuer, and Russell Thomas. 2015. “Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model.” *Environmental Science and Technology* 49 (14): 8804–14. doi:10.1021/acs.est.5b02641.

Catalini, Christian, Nicola Lacetera, and Alexander Oettl. 2015. “The Incidence and Role of Negative Citations in Science.” *Proceedings of the National Academy of Sciences*, October, 201502280. doi:10.1073/pnas.1502280112.

Fleck, Ludwik. (1935) 1979. *The Genesis and Development of a Scientific Fact*. Edited by T.J. Trenn and R.K. Merton. Chicago: University of Chicago Press.

Kuhn, Thomas. (1962) 1996. *The Structure of Scientific Revolutions*. third edition. Chicago: University of Chicago Press.

Longino, Helen. 1990. *Science as Social Knowledge*. Princeton: Princeton University Press.

Newman, M.E.J. 2006. "Modularity and Community Structure in Networks." *Proceedings of the National Academy of Sciences* 103 (23): 8577–82. doi:10.1073/pnas.0601602103.