Using Citation Bias to Guide Better Sampling of Scientific Literature

Yuanxi Fu, Jasmine Yuan, and Jodi Schneider

{fu5, jyuan25, jodi}@illinois.edu
University of Illinois at Urbana-Champaign, School of Information Sciences, 501 E Daniel St, Champaign, IL 61820 (USA)

Abstract

It is rarely possible to cite every relevant work on a topic. When controversy exists in a field, work holding the same opinion as the citing paper (i.e., homophily) is more likely to be cited. Thus, readers may inadvertently select a non-representative sample of articles to read. Here, we begin to develop a method that guides better sampling of scientific literature by designing and testing two new network metrics. The first metric, the ratio between real and expected citation counts, guides users to papers that were cited many fewer times than expected and may represent marginalized findings. The second metric, the relative evidence coupling strength, guides users to papers that may present a unique view of the field. We test our metrics on a known case of citation bias: a network of 73 papers about whether stress is a risk factor for depression. Our metrics select a cross-section of 21 papers. The intersection of the two metrics selects 3 papers that represent all 3 positions of this claim network. In future work we will test our metrics on more datasets, and we will partner with domain experts to verify whether our metrics do identify a diverse sample of research articles.

Introduction

Authors must make choices when citing related work: it is rarely possible to cite every relevant article on a topic, and sometimes even identifying all relevant work is difficult. However, citation bias arises when "data critical of or refuting claims are systematically uncited in favour of data supporting a claim" (Greenberg, 2009). Studies have found that cited articles are more likely to report statistically significant results (Urlings, Duyx, Swaen, Bouter, & Zeegers, 2019b), outlier studies (Schrag, Mueller, Oyoyo, Smith, & Kirsch, 2011), or conclusions supportive of a hypothesis (Duyx, Urlings, Swaen, Bouter, & Zeegers, 2017; Urlings, Duyx, Swaen, Bouter, & Zeegers, 2019a). When controversy exists in a field, preference was given to articles holding the same opinion as the citing paper (i.e., homophily) (Trinquart, Johns, & Galea, 2016). Citation bias benefits authors by bolstering grants and papers, making them more easily accepted. However, it creates a "filter bubble" (Pariser, 2011) for people who want to use scientific literature as impartial evidence. For example, busy decision-makers only have time to read a few articles. They may end up inadvertently choosing articles belonging to a group of papers that reach the same findings. This selection of the literature gives the impression that there are no controversies on the given topic and that the evidence is well supported by many other papers.

This paper addresses an important challenge facing many literature users: how to sample diverse research articles. In this paper we propose a new sampling method, which uses insights from citation bias studies to design new network structure metrics to guide better sampling of the literature. First, we need to identify papers that get far fewer citations than chance permits, in case they are marginalized due to contradicting the dominant view of a field. Second, we need to identify papers that do not belong to any homophily group(s), because they may hold unique views towards the scientific question. Furthermore, the method should be accessible and intuitive, so that it can easily be adopted by a wide range of users. This paper reports our work in progress on developing such a method. First, we explain our current method, along with the design of two network structure metrics. Then we apply our method to a citation network. And finally, we discuss our plans for future work.

Data

To answer a specific scientific question, decision makers gather a body of literature relevant to some topic. This set of papers (nodes) and the citation relationships among them (directed edges) form a "claim-specific network," following Greenberg (2011). A claim-specific network is a subgraph of the entire citation network, since we omit all papers that do not concern a given topic (Greenberg, 2011). Thus, the papers (nodes) in a claim-specific network all address a single research question, for example, whether reducing salt intake brings health benefit at the population level, or whether stress is a risk factor for depression. Such a network captures the communication of ideas and the establishment of belief regarding a specific scientific question (Greenberg, 2011).

The dataset used here comes from a paper that analyzed citation bias in favor of positive findings (de Vries, Roest, Franzen, Munafò, & Bastiaansen, 2016). That study constructed a claim-specific citation network between 73 primary studies, all concerning risks of developing depression after stressful life events, associated with a gene known as 5-HTTLPR. Based on the outcome each primary study reported, 24 studies were coded as "Positive", 38 studies as "Negative", and 11 studies as "Unclear." The "Positive" articles received more citations than the "Negative" articles (de Vries et al., 2016).

Metrics

In this pilot study, we will use two metrics. The first metric is the **ratio between the real and expected citation count**, which measures whether a paper is sufficiently cited. Currently, we define the expected citation count as the number of citations a paper would receive if all papers in the network published *two years or more later* cited it. This ratio is designed to select papers that would be less prominent in a purely citation-based heuristic.

To compute the ratio between the real and expected citation count for the 5-HTTLPR claim-specific citation network, we constructed two networks. First, we reconstructed the real claim-specific citation network from the published dataset (de Vries & Munafò, 2016), which has 73 nodes and 488 edges. Second, we constructed a new, simulated network, in which the nodes are identical to those in the real network, while the edges are determined by the gap between the publication years of the two articles. To be more specific, if two papers are two or more years apart, we add a directional edge from the younger paper to the older paper. The resulting simulated network has 73 nodes and 1799 edges. Currently, we compute the expected citation count of a node as its in-degree in this simulated network.

The second metric resembles the relative bibliographic coupling strength (Sen & Gan, 1983; Shen, Zhu, Rousseau, Su, & Wang, 2019). The difference is that we limit to the items within the claim-specific network, rather than the whole bibliography: thus, in the name we replace "bibliographic" with "evidence". Specifically, we define the **"relative evidence coupling strength"** as

$$\mathrm{RECS}(X,Y) = \frac{|E_X \cap E_Y|}{|E_X \cup E_Y|}$$

where E_X and E_Y represent the set of references for X and for Y in the claim-specific network. In other words, this metric characterizes the percentage of overlapping items (i.e., $E_X \cap E_Y$) in the collective evidence set (i.e., $E_X \cup E_Y$) between the pair of nodes. This metric is inspired by an observation we made in a study of disagreement between systematic reviews, in which we found that systematic reviews that synthesized different evidence sets reached different conclusions (Hsiao, Fu, & Schneider, 2020).

Although our two metrics were inspired by studies of citation bias, their ultimate utility is to help literature users sample a diversity of research articles. Users do not need to know whether a claim-specific network is biased or to what extent to make use of them.

Our script to construct the networks and compute the network metrics can be found in GitHub (https://github.com/infoqualitylab/citation_bias_study/tree/master/ISSI_2021).

Results

Figure 1(a) shows the distribution of our first metric, the ratio between the real and expected citation counts. Citations are quite unequal among those 73 papers. One paper, Caspi 2003, received 97% of its expected citation counts, whereas 16 papers (22%) (colored in pink in Figure 2) received less than 10% of their expected citation counts. Since they were cited many fewer times than expected, those 16 papers need more attention from readers, in case they represent a marginalized view of the topic. Also, 18 papers were published in the last two years covered in the datasets (2011 and 2012), and under our current construction, their expected citation counts are zero, and therefore do not have a value for this metric. Those papers are colored white in Figure 2.

We computed the relative evidence coupling strength for each of the 2628 pairs. Figure 1(b) shows the distribution of this metric. Forty-nine pairs (1.9%) shared more than 90% of their collective evidence set and 818 pairs (31%) shared less than 10% of their collective evidence set. Since a single paper can be involved in 72 pairs, it is quite common for a paper to find another paper in the dataset with which its relative evidence coupling strength is small. What we need to pay attention to are papers that *always* have a small relative evidence coupling strength with other papers. Those are what we call "unique" papers, as they may represent unique views of the field (i.e., looking at a different set of evidence). Currently, we set an arbitrary threshold of 0.3 and found 8 papers that never participated in any pair with relative evidence coupling strength more than 0.3. Those papers are colored in pink in Figure 3 and they should be sampled too.

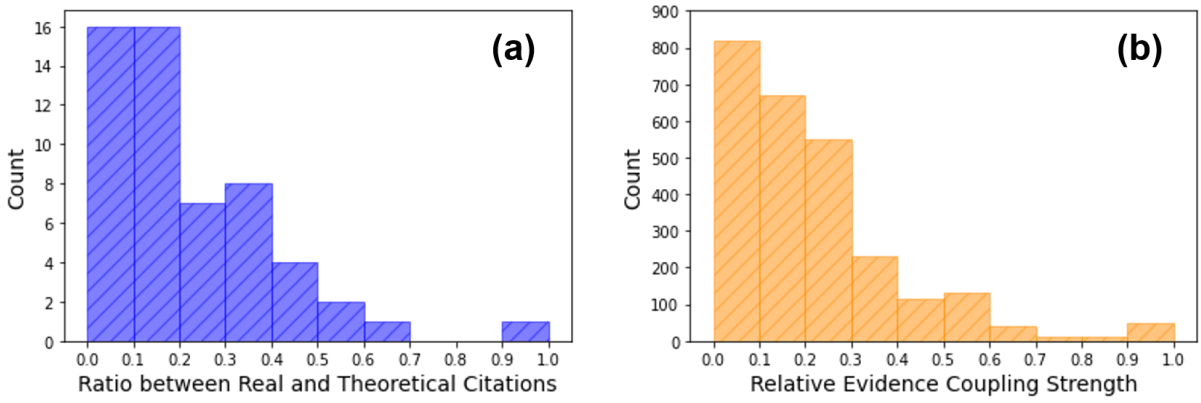

Figure 1. The distribution of the ratio between (a) the real and expected citation counts and (b) the relative evidence coupling strength for the 5-HTTLPR network

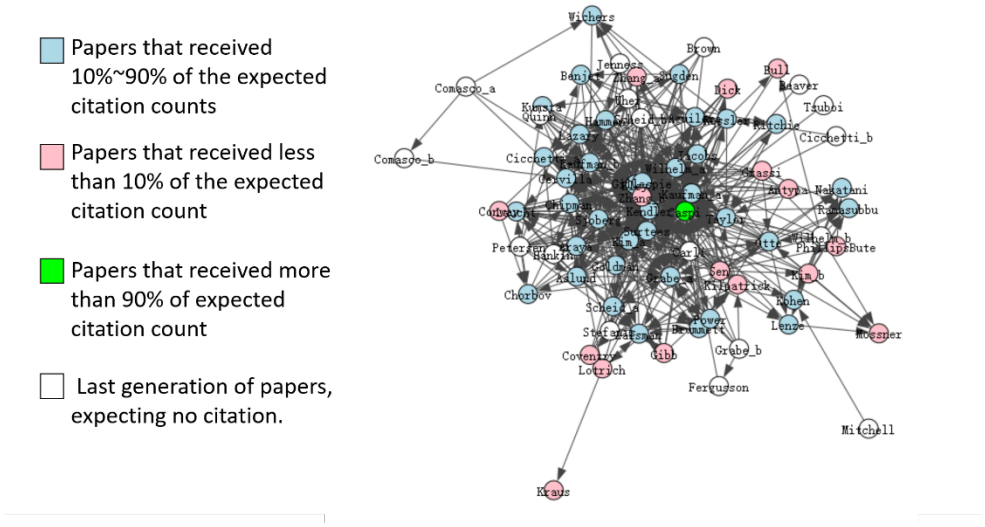

Figure 2. The ratio between real and expected citation count applied to the HTTLPR-5 network

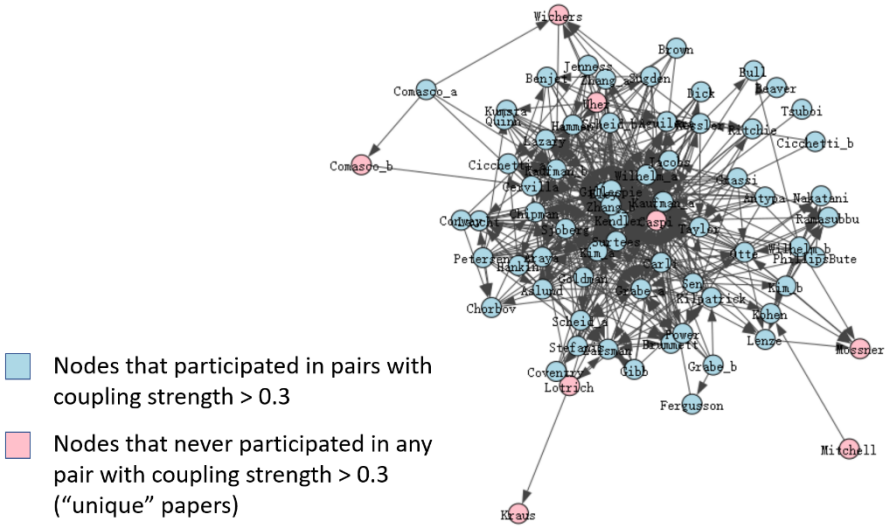

Figure 3. Relative evidence coupling strength applied to the HTTLPR-5 network

Our two metrics select different articles. As shown in Table 1, there are just 3 overlaps. Interestingly, each group selects papers with all 3 available outcomes (Positive, Negative, Unclear) – and the group selected by the intersection of the two metrics does this in just 3 papers, the minimum possible. Notably, there is only a 16.1% chance for a randomly selected set of three papers to cover all three outcomes. This finding suggests that these metrics are promising for further exploration, because the intersection included three papers with different outcomes, meeting our intention to select a diverse set of papers for readers to read.

Table 1. Papers selected by the two network metrics

Insufficiently cited papers	Dick 2007 Unclear Kilpatrick 2007 Unclear Bull 2008 Positive Zhang_a 2008 Negative PhillipsBute 2008 Negative Zhang_b 2009 Negative Kim_b 2009 Positive	Gibb 2009 Unclear Coventry 2009 Negative Grassi 2010 Negative Sen 2010 Positive Antypa 2010 Negative Conway 2010 Negative
Both	Mossner 2001 Positive Lotrich 2008 Unclear	Kraus 2007 Negative
"Unique" papers	Caspi 2003 Positive Comasco_b 2011 Unclear Wichers 2007 Negative	Uher 2011 Negative Mitchell 2011 Negative

Conclusions and Future Works

In conclusion, we proposed a method that may help users who need to select diverse papers to read to assist in their decision making. We designed two metrics inspired by previous studies of citation bias: First, the ratio between the real and expected citations identifies papers that received far fewer citations than expected, which may represent marginalized views. Second, the relative evidence coupling strength identifies papers with unique views towards a topic. We applied these metrics to a previously citation network of 73 research articles studying whether a gene increases the risk of depression after stressful events. Our sampling method resulted in partitioning the network into two groups: 21 papers that readers need to pay particular attention to, and 52 articles from which readers can choose a few to read.

Our current method has known limitations and needs evaluation and improvement in the future. First, the current expected citation count is in effect a maximum citation count (assuming a delay between publication and citation). Some of the later published papers were expected to cite as many as 55 papers, which is unrealistic and inflated the expected citation counts for other papers. Second, the relative evidence coupling strength also has low values among paper pairs whose publication years are far apart, and the method we are using now (i.e., selecting nodes that never participated in any pair with relative evidence coupling strength above a threshold value) over-selects early papers, which have a diminished chance to share reference items with papers published later.

For future work, our priority is to find domain experts to evaluate the two sets of papers and see whether the selected papers fit our expectations. Second, we need to evolve the metrics. We plan to create better algorithms to for computing more realistic expected citation counts. And we will also filter out papers selected purely due to their age. The visualization should be improved too, making it more interactive and informative. We envision an interface that allows users to zoom into one part of the network and click on a node to find information about a paper (e.g., title, DOI). Finally, the utility of our methods will need to be verified with larger-scale studies. We plan to use both published claim-specific networks (e.g., Duyx et al., 2019; Trinquart et al., 2016; Urlings et al., 2019a; Urlings et al., 2019b) as well as our own datasets constructed in collaboration with domain experts. We are currenting working with a kinesiology and community health expert and have constructed a claim-specific citation network with 439 articles relating to the effectiveness of exercise therapy for treating depression. This is larger than any claim-specific citation network that we are aware of. Those datasets will provide a

larger testbed for verifying whether our method is effective in helping readers obtain a diverse sample of research articles.

Acknowledgments

This research is supported by the Campus Research Board of the University of Illinois at Urbana-Champaign Grant RB21012. We acknowledge Dr. Ymkje Anna de Vries and Prof. Marcus Munafò for making their dataset available under the Non-Commercial Government Licence v2.0 and Zhonghe Wan for her assistance in statistical calculations.

References

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Selective citation in the literature on swimming in chlorinated water and childhood asthma: A network analysis. *Research Integrity and Peer Review*, *2*(1). http:/doi.org/10.1186/s41073-017-0041-z

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2019). Selective citation in the literature on the hygiene hypothesis: A citation analysis on the association between infections and rhinitis. *BMJ Open*, *9*(2), e026518. http:/doi.org/10.1186/s41073-017-0041-z

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ*, *339*, b2680. https://doi.org/10.1136/bmj.b2680

Greenberg, S. A. (2011). Understanding belief using citation networks: Citation networks. *Journal of Evaluation in Clinical Practice*, *17*(2), 389–393. http://doi.org/10.1111/j.1365-2753.2011.01646.x

Hsiao, T.-K., Fu, Y., & Schneider, J. (2020). Visualizing evidence-based disagreement over time: The landscape of a public health controversy 2002-2014. *Proceedings of the Association for Information Science and Technology* (Vol. 57, p. e315). https://doi.org/10.1002/pra2.315

Pariser, Eli. (2011). *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. New York: Penguin Press.

Schrag, M., Mueller, C., Oyoyo, U., Smith, M. A., & Kirsch, W. M. (2011). Iron, zinc and copper in the Alzheimer's disease brain: A quantitative meta-analysis. Some insight on the influence of citation bias on scientific opinion. *Progress in Neurobiology*, *94*(3), 296–306. https://doi.org/10.1016/j.pneurobio.2011.05.001

Sen, S., & Gan, S. K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, *30*, 78–82.

Shen, S., Zhu, D., Rousseau, R., Su, X., & Wang, D. (2019). A refined method for computing bibliographic coupling strengths. *Journal of Informetrics*, *13*(2), 605–615. https://doi.org/10.1016/j.joi.2019.01.012

Trinquart, L., Johns, D. M., & Galea, S. (2016). Why do we think we know what we know? A metaknowledge analysis of the salt controversy. *International Journal of Epidemiology*, *45*(1), 251–260. https://doi.org/10.1093/ije/dyv184

Urlings, M. J. E., Duyx, B., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2019a). Selective citation in scientific literature on the human health effects of bisphenol A. *Research Integrity and Peer Review*, *4*(1), 6. https://doi.org/10.1186/s41073-019-0065-7

Urlings, M. J. E., Duyx, B., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. A. (2019b). Citation bias in the literature on dietary trans fatty acids and serum cholesterol. *Journal of Clinical Epidemiology*, *106*, 88–97. https://doi.org/10.1016/j.jclinepi.2018.10.008

de Vries, Y. A., Roest, A. M., Franzen, M., Munafò, M. R., & Bastiaansen, J. A. (2016). Citation bias and selective focus on positive findings in the literature on the serotonin transporter gene (5-HTTLPR), life stress and depression. *Psychological Medicine*, *46*(14), 2971–2979. https://doi.org/10.1017/S0033291716000805

de Vries, Ymkje Anna, & Munafò, M. (2016). [Dataset] Citation bias and selective focus on positive findings in the literature on 5-HTTLPR, life stress, and depression. University of Bristol. Retrieved January 28, 2021, http://doi.org/10.5523/BRIS.Z7JCONXFBMDR1JJ3T0W4K1HWN